Universidade de Évora - Escola de Ciências e Tecnologia

Mestrado em Modelação Estatística e Análise de Dados

Dissertação

Modelação e Predição de Eventos Raros – um estudo comparativo

Lorena Ventura Santos

Orientador(es) | Paulo de Jesus Infante dos Santos

Anabela Afonso

Gonçalo João Jacinto

Évora 2025

Universidade de Évora - Escola de Ciências e Tecnologia

Mestrado em Modelação Estatística e Análise de Dados

Dissertação

Modelação e Predição de Eventos Raros – um estudo comparativo

Lorena Ventura Santos

Orientador(es) | Paulo de Jesus Infante dos Santos

Anabela Afonso

Gonçalo João Jacinto

Évora 2025

A dissertação foi objeto de apreciação e discussão pública pelo seguinte júri nomeado pelo Diretor da Escola de Ciências e Tecnologia:

Presidente	|	Dulce Gamito Pereira (Universidade de Évora)
Vogais	|	Lígia Henriques-Rodrigues (Universidade de Évora) (Arguente) Paulo de Jesus Infante dos Santos (Universidade de Évora)

Évora 2025

Agradecimentos

Encerrar este ciclo é mais do que concluir uma etapa académica - é por um ponto final num capítulo que me transformou. Esta dissertação é o resultado visível de um percurso invisível: de horas silenciosas, de dúvidas persistentes e de esforço constante.

O caminho até aqui esteve longe de ser linear. Foi um percurso de avanços e recuos, de entusiasmo e exaustão, de momentos de clareza e de outros em que tudo parecia escapar-me. Mas foi precisamente nesses intervalos – entre o tentar e o recomeçar - que aprendi mais sobre resiliência, curiosidade e sobre o valor de acreditar, mesmo quando tudo parece estar a desmoronar.

Nenhum caminho é percorrido sozinho, e este certamente não é exceção. Cada palavra, cada ideia, cada linha deste trabalho carrega um pouco das pessoas que, de uma forma ou de outra, caminharam comigo.

Por isso, é com profunda gratidão que dedico estas palavras àqueles que fizeram parte deste caminho e cuja presença, incentivo, carinho e confiança foram fulcrais.

Em primeiro lugar quero agradecer ao meu orientador, Professor Doutor Paulo Infante, por ser um verdadeiro pilar ao longo deste percurso. A sua ajuda, disponibilidade e dedicação foram incansáveis, estando presente em cada desafio, em cada dúvida, em cada momento que precisei de orientação ou confiança.

Agradeço também aos meus coorientadores, Professora Doutora Anabela Afonso e Professor Doutor Gonçalo Jacinto, pela presença permanente, pela disponibilidade e pelo rigor demonstrado em cada etapa deste trabalho. A vossa orientação atenta e o olhar crítico foram determinantes.

Aos meus pais, os verdadeiros investidores deste percurso, devo mais do que as palavras conseguem traduzir. Os quilómetros que nos separam nunca significaram distância de apoio – estiveram sempre presentes, mesmo quando o “estar perto” dependia apenas de uma chamada ou de uma palavra no momento certo. Este trabalho existe porque nunca me deixaram esquecer de onde venho, nem duvidar até onde posso chegar.

Ao João, à Anabela e ao Sr. Vítor, pela presença inestimável nestes últimos tempos, marcados por maior exigência. Agradeço-vos por me acompanharem e por me darem um ombro amigo.

Modelação e Predição de Eventos Raros: um estudo comparativo

Resumo

A modelação de eventos raros constitui um desafio central na ciência de dados aplicada à segurança rodoviária. Este estudo, centrado no distrito de Setúbal (2016–2023), analisou sinistros registados pela GNR, complementados com variáveis meteorológicas e infraestruturais. Testaram-se modelos estatísticos e de *machine learning* (Regressão Logística, Firth, Random Forest, XGBoost, C5.0 e Naive Bayes), avaliados por PR-AUC, ROC-AUC, F_1 e Brier *score*. Para mitigar o desequilíbrio extremo (≈2% casos graves), aplicaram-se técnicas de *oversampling* (ROSE e SMOTENC) apenas no treino, evitando *data leakage*, e definiu-se o ponto de corte pela maximização do F_2-*score*. O XGBoost e a Logística de Firth mostraram melhor compromisso entre sensibilidade e calibração, com AUC≈0,88. Conclui-se que a combinação de reamostragem adequada e calibração criteriosa melhora a previsão de sinistros graves, oferecendo suporte à definição de políticas de prevenção baseadas em evidência.

Palavras-chave: desequilíbrio de categorias; eventos raros; *machine learning*; reamostragem

Modelling and Prediction of Rare Events – a comparative study

Abstract (English)

Modelling rare events remains a central challenge in data science applied to road safety. This study focuses on severe road accidents in the district of Setúbal (2019–2023), using data from the National Republican Guard (GNR), complemented with meteorological and infrastructural information. Several statistical and machine learning models (Logistic Regression, Firth, Random Forest, XGBoost, C5.0 and Naive Bayes) were evaluated through PR-AUC, ROC-AUC, F_1 and Brier score metrics. To address the strong class imbalance (≈2% severe accidents), oversampling techniques (ROSE and SMOTENC) were applied only to the training set, avoiding data leakage, and thresholds were defined by maximising the F_2-score. The XGBoost and Firth logistic models achieved the best balance between sensitivity and calibration (AUC≈0,88). Results demonstrate that combining appropriate resampling with careful calibration enhances the prediction of severe road accidents, supporting evidence-based decision-making in road safety policies.

Keywords: class imbalance; rare events; Firth logistic regression; *machine learning*; resampling

Índice

Índice de Figuras

Índice de Tabelas

Glossário

ANSR - Autoridade Nacional de Segurança Rodoviária

AUC - Área sob a curva, medida global da capacidade discriminativa de um modelo

Brier score - Erro quadrático médio entre as probabilidades previstas e os resultados observados

IC95% - Intervalo de Confiança a 95%

COVID-19 - Doença causada pelo coronavírus SARS-CoV-2

df - Graus de liberdade

FN - Falsos Negativos

FP - Falsos Positivos

F_1-*score* - Média harmónica entre precisão e sensibilidade, atribuindo igual peso a ambas

FG - Feridos graves

FL - Feridos ligeiros

GAM – Modelos Aditivos Generalizados (Generalized Additive Models)

GEE - Equações de Estimação Generalizadas (Generalized Estimating Equations)

GIVF - Fator de inflação da variância generalizado (Generalized Variance Inflation Factor)

GLM - Modelos Lineares Generalizados (Generalized Linear Models)

GLMM - Modelos Lineares Generalizados Mistos (Generalized Linear Mixed Models)

G-mean - Média geométrica entre sensibilidade e especificidade

GNR - Guarda Nacional Republicana

IGR - Índice de gravidade (número de mortos por 100 acidentes com vítimas)

KDE - Estimativa por núcleo (Kernel Density Estimation)

M/FG - Mortes e/ou feridos graves

ML - *Machine Learning*

MPL - Modelo de Probabilidade Linear

MV - Máxima Verosimilhança

NB - Naive Bayes

OOF - *Out-of-Fold* (fora da amostra de treino em validação cruzada)

PR - Precisão (*Precision*)

PR-AUC - Área sob a curva Precisão–Sensibilidade

RF - Random Forest

ROC - Receiver Operating Characteristic

ROC-AUC - Área sob a curva ROC

ROSE – Random Over-Sampling Examples

ScFG - Sinistro com feridos graves

ScV - Sinistro com vítimas

ScVM - Sinistro com vítimas mortais

SMOTENC - Synthetic Minority Over-sampling Technique – Nominal Continuous

TN - Verdadeiros Negativos

TP - Verdadeiros Positivos

VIF - Fator de inflação da variância (Variance Inflation Factor)

XGBoost - Extreme Gradient Boosting

1. Introdução

A segurança rodoviária continua a ser uma preocupação central para os governos, autoridades de trânsito e sociedade em geral. Entre os vários tipos de sinistros que ocorrem nas estradas, os sinistros graves - que envolvem mortes e/ou feridos graves (M/FG) – são eventos raros, mas com consequências devastadoras para as vítimas, famílias e a comunidade em geral. Além do impacto emocional, esses eventos acarretam custos sociais, económicos e de saúde pública substanciais. Assim, a modelação e predição de sinistros graves, classificados como eventos raros, são essenciais para o desenvolvimento de estratégias eficazes de mitigação e prevenção.

A presente dissertação tem como principal objetivo comparar metodologias estatísticas e de *machine learning* na modelação e previsão de eventos raros, aplicando-as ao caso da sinistralidade rodoviária grave no distrito de Setúbal, de forma a identificar abordagens que maximizem o desempenho preditivo e a interpretabilidade dos modelos em contextos de forte desequilíbrio entre categorias.

O horizonte temporal do estudo abrange os anos de 2016 a 2023. Foram excluídas da análise as observações correspondentes ao período compreendido entre 11 de abril de 2020 e 30 de abril de 2021, correspondente à fase mais restritiva da pandemia de COVID-19, devido às alterações significativas nos padrões de tráfego e mobilidade observadas nesse intervalo.

Dado o caráter raro destes eventos, a escassez de dados e o desequilíbrio entre feridos leves (FL) e M/FG constituem desafios críticos para a modelação e predição. Para abordar estas limitações, este estudo irá aplicar uma combinação de técnicas de modelação clássicas e modernas, incluindo metodologias estatísticas e de *machine learning*. A eficácia dessas técnicas será avaliada com base na sua capacidade preditiva e no desempenho global na identificação de sinistros graves.

Para alcançar este objetivo geral, definem-se os seguintes objetivos específicos:

- Avaliar e comparar o desempenho de diferentes modelos preditivos, incluindo métodos estatísticos (Regressão Logística Clássica e de Firth) e algoritmos de *machine learning* (Random Forest, C5.0, XGBoost e Naive Bayes), na previsão de sinistros com mortos ou feridos graves.

- Analisar o impacto do desequilíbrio de categorias na qualidade da modelação, testando estratégias alternativas de mitigação — nomeadamente técnicas de reamostragem (ROSE, SMOTENC) e ponderação por pesos inversos — aplicadas de forma controlada à fase de treino.
- Explorar e otimizar a calibração e os limites de decisão dos modelos através de métricas adequadas a eventos raros (Precisão – Sensibilidade*, AUC*, *ROC-AUC*, F_2*-score*, *Brier score*, e parâmetros de calibração), assegurando uma avaliação robusta em validação cruzada e conjunto de teste independente.
- Identificar limitações metodológicas e potenciais vieses (como *data leakage* e sobreajuste) nas abordagens de modelação, discutindo estratégias para mitigação e propondo linhas futuras de investigação em modelação estatística aplicada a fenómenos raros de segurança rodoviária.

2. Enquadramento teórico

O sistema rodoviário é uma parte integral da vida moderna, influenciando diretamente o quotidiano dos cidadãos. O mesmo abrange uma ampla variedade de formas de deslocamento, desde meios não motorizados, como a caminhada e o ciclismo, até veículos motorizados, como carros particulares e transportes públicos. A mobilidade rodoviária é vital para as atividades pessoais e profissionais, conectando e movimentando a sociedade. No entanto, com essa interconectividade surgem também riscos, nomeadamente a possibilidade de sinistros rodoviários, que podem ter consequências avassaladoras (Valente, 2025).

A condução é uma atividade de elevada responsabilidade, que exige um conhecimento profundo das regras e dinâmicas do sistema rodoviário. O domínio destas competências é essencial para que os condutores possam desempenhar um papel ativo na segurança rodoviária, protegendo-se a si e aos outros cidadãos da via. A falta desse conhecimento ou a negligência na aplicação das normas podem resultar em sinistros com consequências graves. Neste sentido, a segurança rodoviária continua a ser uma preocupação e uma prioridade central para os governos, autoridades de trânsito e sociedade em geral. A elevada taxa de sinistros e as suas consequências ressaltam a necessidade urgente de implementar medidas eficazes para melhorar a segurança nas estradas (Tribunal de Contas Europeu, 2024).

2.1 Terminologia Acidente vs. Sinistro

A terminologia utilizada para descrever sinistros rodoviários tem sido amplamente debatida na literatura, especialmente quanto ao uso do termo “acidente”. Embora social e academicamente enraizado, esse termo é problemático por associar os eventos a imprevisibilidade e aleatoriedade (Perez, 2011), sugerindo ocorrências inevitáveis e desconsiderando fatores de prevenção.

Diante dessa limitação terminológica, organismos como a *National Highway Traffic Safety Administration* (NHTSA) e a Organização Mundial de Saúde (OMS) têm vindo a substituir o termo “acidente” para afastar a ideia de casualidade, enfatizando a influência de fatores humanos, mecânicos e ambientais, e destacando que os sinistros podem ser analisados e prevenidos por meio de medidas corretivas.

Encontrar uma alternativa adequada é complexo, mas essencial para transmitir a verdadeira natureza desses eventos. Nesse contexto, termos como "sinistro rodoviário", proposto por Pérez (2011) e Tabasso (2012), ganham relevância por destacarem a capacidade de investigação e correção, além de sensibilizarem a sociedade para políticas de segurança. A adoção dessa linguagem reflete a evolução tecnológica e científica, que permite compreender as causas e propor soluções, mesmo em casos parcialmente inevitáveis.

2.2 Sinistros Rodoviários

A sinistralidade rodoviária em Portugal mantém-se um desafio crítico, com flutuações significativas nos indicadores entre 2016 e 2023, conforme os dados do Relatório de Sinistralidade a 24 horas e Fiscalização Rodoviária de Maio de 2023 da Autoridade Nacional de Segurança Rodoviária (ANSR). Na Tabela 1 estão apresentados os dados comparativos dos sinistros rodoviários entre os diferentes anos. Os mesmos referem-se exclusivamente a Portugal continental, excluindo as regiões autónomas dos Açores e da Madeira.

Tabela 1 - Evolução da sinistralidade rodoviária no Continente.

Ano		ScV	ScVM + ScFG	ScVM	FL	FG	M	IGR
2016		32299	2201	416	39121	2102	445	1,38
2017		34416	2397	488	41787	2198	510	1,48
2018		34235	2337	468	41356	2141	508	1,48
2019		35704	2403	429	43202	2301	474	1,33
2020		26501	1975	372	30706	1829	390	1,47
2021		29217	2221	367	34217	2106	390	1,33
2022		32788	2352	428	38456	2243	462	1,41
2023		34974	2569	431	41058	2437	467	1,34

Nota: *ScV: Sinistros com vítimas, ScVM: Sinistros com vítimas mortais, ScFG: Sinistros com feridos graves, FL: Feridos ligeiros, FG – Feridos graves, M: Mortes, IGR: Índice de gravidade)*

Fonte - Relatório de Sinistralidade a 24 horas e Fiscalização Rodoviária de Maio de 2023

Entre 2016 e 2023, observa-se um aumento de 8,3% nos Sinistros com Vítimas (ScV), passando de 32299 para 35974 casos. Esse crescimento, no entanto, não foi linear: em 2020, houve uma queda abrupta para 26501 sinistros, provavelmente devido às restrições da COVID-19.

No que diz respeito às vítimas mortais (M), registou-se um crescimento de 4,9% no período analisado, passando de 445 mortes em 2016 para 467 em 2023. Os feridos graves (FG) também apresentaram uma tendência preocupante, com um aumento de 15,9% entre 2016 (2102 casos) e 2023 (2437 casos), sendo 2020 o ano com o menor registo (1829), reflexo direto da redução da mobilidade durante a pandemia.

Os feridos ligeiros, por sua vez, tiveram um crescimento moderado de 5%, subindo de 39121 para 41058 no mesmo intervalo. Apesar do aumento nos números absolutos de mortes e feridos, o índice de gravidade (IGR) apresentou uma redução de 2,9%, passando de 1,38 em 2016 para 1,34 em 2023. Essa diminuição sugere uma menor letalidade por sinistro. O ano de 2020 destacou-se como atípico, com quedas expressivas em todos os indicadores. No período pós-pandemia (2021-2023), observou-se uma retoma gradual dos valores. Em 2023, os sinistros com vítimas atingiram 34974 casos, valor próximo ao pico de 35704 registado em 2019.

2.3 Evento Raro

Um evento raro é definido como um fenómeno que ocorre com muita baixa frequência, independentemente da natureza da variável associada (categórica ou numérica). Exemplos comuns incluem desastres naturais, doenças raras, *crash* na bolsa, entre outros. A natureza rara desses eventos significa que, muitas vezes, os conjuntos de dados disponíveis apresentam um desequilíbrio muito acentuado entre eventos e não eventos. Tal desequilíbrio pode comprometer a performance de modelos preditivos tradicionais, levando à necessidade de desenvolver abordagens específicas para lidar com essa escassez de dados (King e Zeng, 2001a). A identificação de padrões e a previsão de eventos raros é essencial, dado o impacto económico, social e humano que estes eventos podem ter nas áreas em que ocorrem, apesar da sua baixa frequência.

2.4 Modelação e Predição de Eventos Raros

A literatura demonstra que modelos clássicos, nomeadamente a regressão logística estimada por máxima verosimilhança, tendem a apresentar *viés* na estimação das probabilidades quando aplicados a eventos raros, subestimando a probabilidade de ocorrência da categoria minoritária (King & Zeng, 2001). Este problema é frequentemente agravado pela utilização de métricas de avaliação globais, como a *accuracy*, que se revelam pouco informativas em contextos de forte desequilíbrio entre categorias.

No âmbito da modelação estatística, diversas abordagens foram propostas para mitigar estes problemas, destacando-se a regressão logística penalizada, em particular a correção de Firth, que permite reduzir o *viés* das estimativas e lidar com situações de separação completa ou quase completa dos dados.

Paralelamente, técnicas de *machine learning* têm vindo a ser aplicadas à predição de eventos raros, explorando a sua capacidade de capturar relações não lineares e interações complexas entre variáveis. Algoritmos baseados em árvores, como Random Forest e métodos de *boosting*, têm demonstrado bom desempenho discriminativo em contextos desequilibrados. No entanto, a literatura reconhece limitações associadas à interpretabilidade e à calibração das probabilidades previstas, aspetos críticos em aplicações de apoio à decisão.

Outro aspeto central identificado é a necessidade de estratégias adequadas para lidar com o desequilíbrio de categorias, bem como a adoção de métricas de avaliação que reflitam corretamente o desempenho na identificação do evento raro. Adicionalmente, tem sido sublinhada a importância da calibração das probabilidades previstas, de forma a garantir a utilidade prática dos modelos em contextos reais.

Em síntese, a literatura evidencia que a modelação de eventos raros requer uma abordagem integrada, que combine modelos estatísticos robustos, técnicas de *machine learning*, métricas de avaliação adequadas e procedimentos rigorosos de validação e calibração.

3. Metodologias

3.1 Seleção de Métricas para Eventos Raros

Diversos são os estudos que indicam que, em conjuntos de dados desequilibrados, métricas convencionais como *accuracy* pode induzir em erro. Modelos que tendem a favorecer a categoria maioritária, ou negativa, podem exibir uma elevada precisão global enquanto falham redondamente a deteção da categoria minoritária, ou positiva (He; Garcia, 2009). Por exemplo, num cenário em que o conjunto de dados é composto por 95% de observações negativas e apenas 5% positivas, um modelo que prevê todas as observações como negativas alcançaria 95% de *accuracy*, mas falharia completamente em identificar os casos positivos (Japkowicz, 2000).

Esta limitação exige a adoção de métricas que quantifiquem corretamente o desempenho da categoria minoritária. Métricas como sensibilidade, F_1*-score, G-mean* e a AUC são mais adequadas para avaliar o desempenho nesses cenários, pois operacionalizam o equilíbrio entre sensibilidade (capacidade de identificar corretamente a categoria minoritária) e o controlo da taxa de falsos positivos (Saito; Rehmsmeier, 2015).

Entre essas métricas, a sensibilidade mede a proporção de verdadeiros positivos identificados corretamente, calculada por:

$$Sensibilidade = \frac{VP}{VP + FN},$$

sendo VP o número de casos positivos corretamente classificados e FN o número de casos positivos classificados incorretamente.

A especificidade mede a proporção de verdadeiros negativos identificados corretamente, calculada por:

$$Especificidade = \frac{VN}{VN + FP},$$

sendo VN o número de casos negativos corretamente classificados e FP o número de casos negativos classificados incorretamente.

Uma vez que esta métrica apenas avalia o desemprenho da categoria maioritária, em conjuntos de dados desequilibrados, o modelo pode atingir uma especificidade muito alta simplesmente por classificar todos os casos como negativos.

A precisão mede a proporção de verdadeiros positivos entre todas as previsões positivas, calculada por:

$$\text{Precisão} = \frac{VP}{VP + FP}.$$

Quando a precisão e a sensibilidade são igualmente importantes, utiliza-se o *F1-score*, definido como a média harmónica entre ambas:

$$F_1\text{-}score = 2 \times \frac{Precisão \times Sensibilidade}{Precisão + Sensibilidade}.$$

Esta métrica atribui maior peso a valores baixos, penalizando desequilíbrios entre precisão e sensibilidade (Dobriban et al, 2014).

No presente estudo, o F_1-*score* é calculado para a categoria minoritária (sinistro grave), uma vez que é a de maior interesse analítico. Alternativamente, poderiam ser utilizadas versões agregadas, como o F_1-*score* macro, micro ou ponderado, conforme o objetivo da análise.

O F_1-*score* é um caso particular da medida F_β, quando $\beta = 1$, onde o parâmetro β indica a importância da sensibilidade sobre a precisão. A expressão geral de F_β-*score* é:

$$F_\beta\text{-}score = (\beta^2 + 1) \times \frac{Precisão \times Sensibilidade}{\beta^2 \times Precisão + Sensibilidade}.$$

Ou seja, enquanto F_1-*score* atribui igual peso à precisão e sensibilidade, quando $\beta = 2$ obtém-se a métrica F_2-*score* que considera que sensibilidade é duas vezes mais importante que a precisão. Esta medida deve ser usada quando se pretende que o modelo detete mais os verdadeiros positivos, sendo por isso mais adequada em situações de eventos rados.

Complementarmente, o *G-mean* avalia o equilíbrio entre a taxa de verdadeiros positivos e verdadeiros negativos:

$$G\text{-}mean = \sqrt{(Sensibilidade \times Especificidade)}.$$

Adicionalmente, em problemas com categorias desequilibradas, é comum recorrer à Área sob a Curva Precisão-Sensibilidade (PR-AUC), que mede o desempenho global no

modelo considerando a relação entre precisão e sensibilidade em diferentes limites de decisão, sendo mais adequada do que a AUC-ROC nestes cenários.

A curva *ROC* oferece uma visão geral do desempenho do modelo em diferentes limites de classificação (Kubat; Matwin, 1997), e a AUC mede a capacidade de o modelo classificar corretamente as observações:

$$AUC = \frac{Sensibilidade}{2} + \frac{VN}{2(VN + FP)},$$

sendo VN o número de casos negativos classificados corretamente.

3.2 Modelo Estatístico de Regressão Logística

A análise da regressão teve início com Francis Galton (1822−1911), que investigou a hereditariedade da altura. Em 1886, Galton introduziu o conceito de "regressão à média" ao estudar a hereditariedade de caraterísticas como a altura nos seres humanos (Galton, 1886). Nesse estudo, o autor observou que, embora pais excecionalmente altos tendessem a ter filhos também altos, estes não mantinham a extrema altura dos pais, mas direcionavam-se para valores mais próximos da média da população.

Karl Pearson (1857 – 1936) formalizou essa observação na década de 1890 ao desenvolver a "linha de melhor ajuste" entre variáveis, utilizando o método dos mínimos quadrados (Pearson, 1900). No entanto, os modelos de regressão linear não são adequados para todos os tipos de dados. Embora sejam úteis para prever variáveis contínuas, esses modelos enfrentam limitações quando a variável dependente é binária, ou seja, assume apenas dois valores (0 ou 1).

A inadequação da regressão linear clássica (originalmente desenvolvida para respostas contínuas) quando aplicada a variáveis dicotómicas pode ser ilustrada pelo "modelo de probabilidade linear" (MPL). Nesta abordagem, a variável resposta binária $Y \in \{0, 1\}$ é reinterpretada como a probabilidade $P\,(Y = 1 \mid X)$, modelada com uma função linear das covariáveis. Contudo, o MPL frequentemente viola os pressupostos estatísticos (como normalidade e homocedasticidade dos resíduos) e por não impor restrições ao intervalo de previsão produz previsões fora do intervalo [0, 1] (Aldrich & Nelson, 1984). Essa fragilidade reforça a necessidade de uma abordagem não linear, capaz de modelar relações complexas sem comprometer a interpretabilidade.

Para lidar com esse tipo de situações, Joseph Berkson contribuiu significativamente ao popularizar o termo "*logit*" e demonstrar a equivalência entre a função logística e a maximização da verosimilhança (Berkson, 1944). Paralelamente, David Cox (1924-2022) desenvolveu, na década de 1950, a regressão logística. O principal objetivo desta técnica é estimar a probabilidade de ocorrência de um evento binário com base nas variáveis explicativas, transformando a relação entre elas numa função logística (Cox, 1958). A principal vantagem da regressão logística é que ela transforma uma relação linear entre variáveis independentes e a probabilidade de um evento ocorrer numa função que mapeia a saída para um intervalo entre 0 e 1. A função logística, definida como $f(z) = \frac{e^z}{1+ e^z}$, garante que as probabilidades permanecem dentro do intervalo]0, 1[, resolvendo assim um dos entraves dos modelos lineares.

A regressão logística múltipla é uma extensão do modelo de regressão logística simples, em que duas ou mais variáveis explicativas, contínuas ou categóricas, são utilizadas para prever a probabilidade de ocorrência de um evento binário (Hosmer et al., 2013). A principal vantagem deste modelo é que ele permite a análise simultânea do impacto de diversas variáveis sobre o resultado binário. Isso é particularmente útil em situações em que múltiplos fatores podem influenciar a probabilidade de um evento.

A flexibilidade da regressão logística múltipla permite que se usem os métodos tradicionais de seleção de variáveis *forward*, *backward* e *setpwise*, bem como técnicas avançadas, como por exemplo o LASSO (*Least Absolute Shrinkage and Selection Operator*), que utiliza a regularização L_1para identificar preditores relevantes ao reduzir os coeficientes menos significativos para zero, evitando o *overfitting* e superando limitações de técnicas tradicionais como Regressão Logística Penalizada (LASSO/Elastic Net) (Tibshirani, 1996; Hastie et al., 2015).

Outra abordagem importante é a Regressão Logística de Firth, uma abordagem penalizada introduzida por Firth (1993) que corrige o *viés* das estimativas de máxima verosimilhança. Esta técnica é particularmente adequada em situações com eventos raros ou separação completa – condição e, que uma ou mais variáveis explicativas permitem distinguir perfeitamente as observações entre as duas categorias (por exemplo, quando todas as observações com uma determinada caraterística pertencem apenas a uma das categorias). Nesses casos, o modelo logístico tradicional pode não

convergir ou produzir coeficientes infinitos. A Regressão Logística de Firth aplica uma penalização baseada na *Jeffreys prior*, o que permite obter estimativas mais estáveis e sem *viés*, mesmo em amostras pequenas ou desequilibradas.

O modelo de Regressão Logística é amplamente aplicado em diversas áreas do conhecimento, como saúde, economia, ciências sociais, entre outras. No contexto da sinistralidade rodoviária, por exemplo, a regressão logística é frequentemente utilizada para identificar fatores de risco associados a sinistros graves, como sinistros com vítimas mortais ou feridos graves. Pode-se prever a probabilidade de um sinistro ter consequências graves com base em variáveis como velocidade, condições meteorológicas adversas, caraterísticas da via, entre outras.

Além disso, a interpretação dos coeficientes no contexto múltiplo pode fornecer indicações detalhadas sobre as relações entre as variáveis independentes e a variável dependente, além de permitir a estimação de probabilidades ajustadas, essenciais para uma análise mais precisa e eficaz.

Representação matemática

A probabilidade de um evento ocorrer (denotado como Y) é dada por:

$$\pi(x_1, x_2, \dots, x_n) = E(Y|X_1 = x_1, X_2 = x_2, \dots X_k = x_k) = \frac{e^{\beta_0+\beta_1 x_1+ \beta_2 x_2+\cdots+ \beta_k x_k}}{1 + e^{\beta_0+\beta_1 x_1+ \beta_2 x_2+\cdots+ \beta_k x_k}},$$

onde:

- $X_1, X_2, \dots, X_k$ são as variáveis explicativas que influenciam a probabilidade de ocorrência do evento;
- $\pi(x_1, x_2, \dots, x_k)$ representa a probabilidade condicional da variável resposta Y ser igual a 1 (i.e., o evento ocorrer), dado o conjunto de variáveis explicativas $X_1, X_2, \dots, X_k$;
- β_0 é o intercepto do modelo, ou seja, o valor da *log-odds* quando todas as variáveis explicativas são iguais a zero;
- $\beta_1, \beta_2, \dots, \beta_k$ são os coeficientes associados às variáveis $X_1, X_2, \dots, X_k$, e medem o impacto de cada variável na *log-odds* do evento ocorrer.

A curva logística apresenta uma forma de "S" (sigmoide), onde mudanças nas variáveis explicativas têm um maior impacto nas probabilidades próximas a 0,5 e um impacto menos acentuado próximo aos extremos 0 e 1.

A notar que, no caso das variáveis categóricas ordinais e nominais, com c categorias, estas variáveis são transformadas em $c-1$ variáveis *dummy*. Por exemplo, considerando a variável "período do sinistro" com três categorias:

1. "manhã" (06h00–12h00) – categoria de referência,
2. "tarde" (12h00–18h00),
3. "noite" (18h00–06h00),

será transformada em:

- X_1 que representa a categoria "tarde", assumindo o valor de 1 se o sinistro ocorreu nesse período e 0 caso contrário,
- X_2 que representa a categoria "noite", assumindo o valor de 1 se o sinistro ocorreu nesse período e 0 caso contrário.

Assim, no caso de um sinistro ocorrer no período da:

- manhã temos $X_1 = 0$ e $X_2 = 0$,
- tarde temos $X_1 = 1$ e $X_2 = 0$,
- noite temos $X_1 = 0$ e $X_2 = 1$.

O coeficiente X_1 mede o *log-odds* do sinistro ser à tarde vs. manhã e o coeficiente X_2 mede o *log-odds* do sinistro ser à noite vs. manhã.

Essa abordagem permite quantificar o efeito relativo de cada categoria na probabilidade do evento (Hardy, 1993).

Suposições do modelo

Para garantir a validade do modelo, devem ser atendidas algumas suposições:

- Independência das observações: as observações devem ser independentes umas das outras, o que significa que o resultado de uma observação não deve influenciar o resultado de outra. Esta suposição é avaliada principalmente pelo delineamento do estudo (por exemplo, amostragem aleatória, ausência de medições repetidas no mesmo indivíduo, etc). Quando o delineamento não assegura a independência (por exemplo, com dados longitudinais agrupados), a sua avaliação e modelação podem ser feitas através de modelos de efeitos

mistos (GLMM) ou de equações de estimação generalizadas (GEE), que são projetados para lidar com a estrutura de dependência dos dados.

- Ausência de multicolinearidade entre as variáveis explicativas: refere-se à inexistência de correlação elevada entre as variáveis independentes do modelo. Essa suposição é avaliada pelo fator de inflação de variância (VIF), que mede o quanto a correlação entre uma variável e as outras variáveis do modelo compromete a precisão da estimativa do seu coeficiente. Um VIF entre 5 e 10 indica multicolinearidade moderada, porém se for superior a 10 a multicolinearidade passa a ser grave (Singh, 2024);
- Linearidade na escala do *logit*: pressupõe-se que a relação entre as variáveis independentes e o *logit* (transformação logarítmica das *odds*) seja linear. Para verificar essa suposição, realiza-se a análise de resíduos. Se os pontos se distribuírem aleatoriamente em torno de zero, sem padrões curvos ou sistemáticos, a linearidade é válida. No caso de os padrões não serem aleatórios, como por exemplo, no formato de U, existe a violação do pressuposto.

Transformação Logit

O "*logit*" é uma transformação matemática que ajuda a linearizar a relação entre as variáveis explicativas e a probabilidade de o evento ocorrer. Por outras palavras, ele converte a equação da probabilidade π para uma forma mais simples, chamada *log-odds.*

A equação do *logit* é dado por:

$$logit\big(\pi(x_1, x_2, \dots, x_n)\big) = g(x_1, x_2, \dots, x_n) = ln\left(\frac{\pi(x_1,x_2,\dots,x_n)}{1-\pi(x_1,x_2,\dots,x_n)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

O processo de transformação é:

1. $\frac{\pi(x_1,x_2,\dots,x_n)}{1-\pi(x_1,x_2,\dots,x_n)}$: esta fração é designada de *odds* (chances) e calcula a razão entre a probabilidade de um evento ocorrer e a probabilidade de ele não ocorrer. A *log-odds* aplica o logaritmo natural a essa razão.
2. $ln\left(\frac{\pi(x_1,x_2,\dots,x_n)}{1-\pi(x_1,x_2,\dots,x_n)}\right)$: aplica o logaritmo natural às *odds*, obtendo-se as *log-odds*;

3. A equação resultante $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ mostra como a combinação das variáveis explicativas $X_1, X_2, \ldots X_k$ afeta a *log-odds* do evento ocorrer. Ou seja, ela transforma a relação não linear entre as variáveis explicativas e a probabilidade numa forma linear.

Portanto, em vez de se modelar diretamente a probabilidade de $\pi(x)$, modela-se a *log-odds* que pode assumir valores entre -∞ e +∞, garantindo uma relação linear com as variáveis explicativas.

Interpretação dos coeficientes

Os coeficientes estimados β_j (para $j = 1, 2, \ldots, k$) possuem interpretações específicas:

- Para variáveis X_j numéricas:
    - $\beta_j > 0$, indica que um aumento em X_j está associado a um aumento na probabilidade de o evento ocorrer $(Y = 1)$;
    - $\beta_j < 0$, indica que um aumento em X_j está associado a uma redução na probabilidade de o evento ocorrer;
- Para as variáveis X_j categóricas:
    - β_j indica o impacto de pertencer a uma categoria específica em relação à categoria de referência.

A regressão logística distingue-se pela interpretabilidade da exponencial dos coeficientes como *odds ratio*. Para variáveis $\mathrm{X_j}$:

- Numéricas: e^{β_j} indica que por cada aumento unitário em X_j, as *odd ratio* do evento ocorrer multiplicam por e^{β_j}. Por exemplo, se $\beta_j = 0{,}7$, então $e^{0{,}7} \approx 2{,}01$, o que indica que cada unidade adicional em X_j duplicam as *odds* do evento ocorrer, mantendo as outras variáveis constantes.
- Categóricas: e^{β_j} indica as chances de o evento ocorrer se pertencer à categoria *j* relativamente à categoria de referência da variável $\mathrm{X_j}$. Por exemplo, seja $X_j = 1$ se pertencer à categoria *j* e $\beta_j = 0{,}7$, então $e^{0{,}7} \approx 2{,}01$, o que indica que as *odds* de o evento ocorrer quando $X_j = 1$ são o dobro de quando $X_j = 0$, mantendo as outras variáveis constantes.

Estimação dos parâmetros

Os coeficientes $\beta_0, \beta_1, \dots, \beta_k$ precisam de ser estimados para que o modelo consiga fazer previsões com base nos dados. O método utilizado para obter essas estimativas é o da máxima verosimilhança, que maximiza a probabilidade de observar os dados amostrais. A função de verosimilhança, que mede essa probabilidade, é expressa por:

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i} \, (\, 1 \, - \, \pi_i \,)^{1 \, - \, y_i},$$

sendo $\pi_i = P(Y_i = 1)$.

A função de verosimilhança envolve um produto de várias probabilidades. Trabalhar diretamente com produtos pode ser matematicamente complicado, especialmente quando se está a otimizar a função para encontrar os coeficientes $\beta_0, \beta_1, \dots, \beta_k$. Por isso, de forma a facilitar os cálculos, é usual aplicar-se o logaritmo natural (ln) à função de verosimilhança, obtendo-se a log-verosimilhança:

$$L(\beta) = \sum_{i=1}^{n} [y_i \ln(\pi_i) + (\, 1 \, - \, y_i) \ln(\, 1 \, - \, \pi_i)],$$

que transformou a multiplicação das probabilidades numa soma, o que torna o processo matemático mais simples.

O estimador de máxima verosimilhança é o veotr de parâmetros que maximiza esta função, sendo definido como:

$$\hat{\beta} = arg \underset{\hat{\beta}}{max} \; \textstyle\sum_{i=1}^{n} [y_i \ln(\pi_i) + (\, 1 \, - \, y_i) \ln(\, 1 \, - \, \pi_i)].$$

Este estimador apresenta importantes propriedades assintóticas, nomeadamente:

- Consistência: o estimador $\hat{\beta}$ é consistente, isto é, converge em probabilidade para o verdadeiro vetor de parâmetros β quando o tamanho da amostra tende para infinito;
- Normalidade assintótica: à medida que o tamanho da amostra aumenta, a distribuição do estimador de máxima verosimilhança aproxima-se de uma distribuição normal multivariada;
- Eficiência assintótica: o estimador de máxima verosimilhança é assintoticamente eficiente, ou seja, atinge o limite inferior de variância de Cramér-Rao, apresentando a menor variância possível entre todos os estimadores consistentes assintoticamente normais.

Estas propriedades permitem realizar inferência estatística sobre os parâmetros do modelo, nomeadamente a construção de intervalos de confiança e a realização de testes de hipóteses, mesmo em amostras de grande dimensão.

Avaliação do Modelo

A qualidade do ajuste é avaliada por métricas como:

- Curva *ROC/AUC*

A Curva ROC é uma ferramenta gráfica útil para visualizar o equilíbrio entre a taxa de VP e a taxa de FP para diferentes pontos de corte. A AUC é um valor escalar único que varia entre 0 e 1. Este quantifica a capacidade discriminativa do modelo, indicando a probabilidade de o modelo classificar corretamente um caso positivo face a um negativo escolhido aleatoriamente. Um *AUC* superior a 0,7 reflete um bom desempenho do modelo.

- Teste de Hosmer-Lemeshow

O teste de Hosmer-Lemeshow avalia a calibração de modelos probabilísticos, verificando a concordância entre as probabilidades previstas e as frequências observadas. Este teste segue uma distribuição qui-quadrado, onde os valores não significativos indicam a adequação do modelo (ou bondade do ajustamento) às observações empíricas.

- R^2

As medidas de R^2 avaliam a melhoria explicativa de modelos estatísticos em relação a um modelo nulo. Dois exemplos amplamente utilizados em modelos logísticos são:

- McFadden: quantifica o ganho relativo na função *log-verosimilhança* ao incluir preditores;

- Nagelkerke: expande o McFadden ao reescalar o intervalo para [0, 1].

Ambos indicam a proporção da variância explicada encontrada em modelos lineares generalizados.

Além disso, é também realizada uma análise de resíduos com o objetivo de identificar observações mal ajustadas, violações de pressuposto e problemas estruturais.

Após a apresentação da fundamentação teórica, é necessário compreender como é que esses conceitos se traduzem em etapas práticas de construção e validação do modelo.

Assim, na sequência, descrevem-se os procedimentos adotados para a aplicação do modelo em contexto empírico, desde o ajustamento inicial até à seleção de variáveis e definição do modelo final.

1) <u>Ajustar o modelo nulo</u>

O ajuste do modelo nulo na regressão logística serve para estabelecer uma linha de base de comparação, avaliar a significância das variáveis independentes, calcular métricas de ajuste e interpretar a variabilidade explicada. Este modelo não inclui nenhuma variável preditora e a sua fórmula de calculo é dada por:

$$\log\left(\frac{p}{1-p}\right) = \beta_0,$$

onde p é a probabilidade de o evento de interesse acontecer.

2) <u>Seleção das variáveis independentes (análise univariada)</u>

O objetivo desta etapa é identificar as variáveis que têm uma relação estatisticamente significativa com a variável resposta, a um nível de significância previamente definido. Para cada variável independente realiza-se o teste da razão de verosimilhança de modo a comparar o modelo nulo com o modelo que inclui apenas essa variável.

Antes de aplicar o teste, os valores ausentes na variável são removidos da amostra, o modelo é reajustado com os dados disponíveis, e verifica-se se a inclusão da variável resulta numa diferença estatisticamente significativa em relação ao modelo nulo.

No que concerne a variáveis com elevada proporção de valores omissos, não existe um limite absoluto para a sua exclusão. Trata-se de um julgamento baseado no equilíbrio entre o valor informativo da variável e a quantidade de dados que se está disposto a perder. Em casos de *missing values* excessivos, as variáveis podem ser descartadas a priori, por inviabilizarem a manutenção de uma amostra robusta para a análise.

3) <u>Modelo múltiplo preliminar e exclusão de variáveis</u>

O processo de construção do modelo multivariado final segue uma abordagem de eliminação retroativa (*stepwise backward*), com base nos princípios de parcimónia e significância estatística.

O modelo inicial numa primeira fase inclui todas as variáveis identificadas como significativas na análise univariada e de seguida é realizada uma seleção iterativa de variáveis mediante a aplicação do teste da razão de verosimilhança. Adota-se um critério de significância mais ríspido, garantindo assim uma maior robustez.

4) <u>Agrupamento de categorias</u>

Para o agrupamento de categorias de uma variável categórica na regressão logística, o procedimento utilizado envolve duas etapas que são executadas de forma conjunta: a análise dos coeficientes estimados e a verificação da significância estatística. Essas etapas permitem identificar categorias com efeitos semelhantes sobre a variável dependente, facilitando o agrupamento.

- <u>Análise dos coeficientes estimados</u>

  Após a execução do modelo de regressão logística, são analisados os coeficientes atribuídos a cada categoria da variável categórica, exceto à categoria de referência. Esses coeficientes refletem o impacto de cada categoria na variável dependente em relação à categoria de referência. Categorias com coeficientes similares indicam efeitos parecidos, sugerindo a possibilidade de agrupamento

- <u>Verificação da significância estatística</u>

  Além de analisar a magnitude dos coeficientes obtidos na regressão logística, é fundamental avaliar a significância estatística de cada um deles. Esse procedimento é realizado através do teste de razão de verosimilhança, que permite verificar se a inclusão de uma variável ou categoria específica melhora de forma relevante ao ajustamento do modelo. Quando se observa que duas ou mais categorias apresentam coeficientes semelhantes e valor de *p-value* acima do nível de significância definido, isso indica que os seus efeitos sobre a variável dependente não têm uma diferença estatisticamente significativa. Ou seja, essas categorias têm comportamentos equivalentes no modelo. Nesses casos, é possível agrupar categorias semelhantes, seja com a categoria adjacente ou com outra cujo agrupamento seja justificável, o que simplifica o modelo sem perda de informação relevante.

5) Verificação da Linearidade

Para garantir que o modelo de regressão logística captura de forma adequada a relação entre as variáveis preditivas e a probabilidade de ocorrência do evento, é essencial verificar o pressuposto da linearidade. Esta análise avalia se a relação entre cada variável e o *logit* é efetivamente linear. Para o efeito pode aplicar-se o método GAM que ajusta um modelo de regressão aditiva generalizada (GAM) com a função "*logit*".

6) Incorporação de interações

No processo de modelação, a incorporação de interações entre variáveis presentes no modelo ajuda a compreender melhor como a combinação de diferentes fatores afeta a variável independente. O objetivo é determinar se a inclusão dessas interações melhora significativamente o ajuste do modelo. Para alcançar esse objetivo, ajusta-se uma série de modelos de regressão logística, cada um contendo diferentes interações, e através do teste de razão de verosimilhanças, avalia-se a significância da inclusão da interação relativamente ao modelo sem essa interação.

Além disso, é fundamental que as interações testadas não apenas apresentem significância estatística, mas também sejam coerentes com o contexto do problema em análise. Dessa forma, garante-se que as adições ao modelo sejam interpretáveis e úteis para a compreensão do fenómeno em estudo.

7) Verificação da qualidade do modelo

Após a inclusão das interações significativas no modelo múltiplo, o modelo é refinado para assegurar que o modelo final se ajusta adequadamente aos dados, apresente uma capacidade discriminativa sólida e tenha capacidade de fornecer previsões úteis e informativas sobre a gravidade dos sinistros. Essa avaliação é essencial para garantir que o modelo não apenas representa de forma precisa a relação entre as variáveis preditoras e a gravidade do sinistro. As principais atividades realizadas nesta etapa incluem:

a. Multicolinearidade
b. Bondade do Ajustamento
c. Capacidade discriminativa
d. Análise de Resíduos
e. Validação do modelo

Análise de multicolinearidade

Para garantir a robustez do modelo final e a confiabilidade dos coeficientes estimados, é preciso avaliar a presença de multicolinearidade entre as variáveis independentes. A multicolinearidade pode inflacionar as variâncias dos coeficientes, dificultando a interpretação precisa do modelo.

A avaliação da multicolinearidade é realizada pelo cálculo do VIF (*Variance Inflation Factor*). O VIF mede o quanto a variância de um coeficiente estimado é inflacionado devido à correlação com outras variáveis no modelo.

Em modelos que incluem interações ou variáveis com múltiplos graus de liberdade (df), utiliza-se o GVIF (*Generalized VIF*), uma extensão do VIF para esses casos específicos (Fox & Monette, 1992). Para facilitar a interpretação, o GVIF pode ser ajustado utilizando a fórmula $GVIF^{(1/(2df))}$, que normaliza o valor de GVIF em casos de variáveis com mais de um grau de liberdade. De acordo com Gujarati (2004), valores de VIF superiores a 10 sugerem uma forte colinearidade, o que pode afetar a precisão das estimativas dos coeficientes. No caso do GVIF, valores de $GVIF^{(1/(2df))}$ superiores a 2,5 podem indicar que as variáveis com múltiplos graus de liberdade estão a apresentar colinearidade significativa.

Bondade do Ajustamento

- R^2 de Nagelkerke

O R^2 de Nagelkerke é uma adaptação do R^2 tradicional para modelos de regressão logística. A sua função centra-se em quantificar a proporção da variância da variável dependente explicada pelo modelo. Esta medida calcula-se comparando a verosimilhança do modelo em estudo com a verosimilhança do modelo nulo. Os seus valores variam entre 0 e 1, onde 0 indica que o modelo não explica qualquer variabilidade da reposta e 1 corresponde a um modelo com poder explicativo máximo.

- Teste de Hosmer e Lemeshow

Para avaliar o ajuste do modelo, utiliza-se o teste de Hosmer-Lemeshow, que verifica a adequação das probabilidades previstas pelo modelo em relação às observações reais.

Capacidade discriminativa

- Curva ROC

A curva ROC (*Receiver Operating Characteristic*) oferece uma representação visual da relação entre a sensibilidade e a especificidade do modelo, à medida que se varia o limite de decisão.

O eixo Y da curva representa a sensibilidade, enquanto o eixo X representa 1 – especificidade. A linha representada corresponde à curva ROC do modelo, que nos dá uma ideia de quão bem o modelo consegue discriminar entre categorias.

As métricas dadas por este gráfico são:

→ Sensibilidade

→ Especificidade

→ Área sob a Curva (AUC)

Análise de Resíduos

- Resíduos e influência

Realiza-se uma análise detalhada dos resíduos para identificar *outliers* e pontos influentes que poderiam afetar as estimativas dos parâmetros. Usa-se métricas como a distância de Cook e os resíduos de *deviance* para detetar e analisar essas observações influentes.

Validação do modelo

Quando se constrói um modelo, especialmente com muitos dados, ele pode ajustar-se bem aos dados que já se tem, mas não funcionar tão bem com novos dados (dados desconhecidos). Isso é conhecido como *overfitting*. A validação é uma forma de verificar se o modelo funciona bem em novos conjuntos de dados e não apenas nos dados utilizados para o construir.

- *Bootstrap*

Neste sentido, para avaliar a robustez do modelo começa-se por realizar uma validação com a técnica *bootstrap*. O *bootstrap* é uma técnica onde se criam "observações" diferentes dos dados originais (com reposição) para simular como o modelo se comportaria com novos dados. Essa técnica permite calcular a estabilidade do modelo e verificar o seu desempenho.

O modelo é validado utilizando o procedimento de "*Backwards Step-down*", que escolhe aleatoriamente um número de observações para ajustar os modelos e comparar os vários ajustamentos.

- Calibração

A calibração avalia a precisão das probabilidades previstas por um modelo, verificando se estas efetivamente se aproximam das frequências observadas na realidade.

- Validação Cruzada

A avaliação do modelo é conduzida por meio da matriz de confusão, que permite observar as previsões feitas pelo modelo em comparação com os casos reais. Além disso, métricas como *accuracy*, sensibilidade, especificidade e F_1-*score* podem ser calculadas para fornecer uma visão abrangente do desempenho do modelo. Essas métricas ajudam a identificar a eficácia do modelo na deteção de casos da categoria que se quer prever, permitindo uma análise crítica das suas capacidades preditivas.

8) <u>Apresentação do modelo final</u>

Na etapa final da regressão logística ocorre a sistematização e comunicação dos resultados obtidos após todas as fases de ajuste, seleção de variáveis e validação do modelos, sendo assim apresentado o modelo final ajustado. Esta fase não se restringe à exposição dos coeficientes e medidas estatísticas, mas envolve a interpretação dos efeitos estimados e a avaliação da qualidade do ajustamento e do desempenho preditivo. O modelo final apresentado é, portanto, aquele que concilia a capacidade explicativa, a parcimónia e a robustez, assegurando que os resultados sejam consistentes e com potencial de generalização para novas observações.

Para concluir, a regressão logística destaca-se como um método estatístico essencial para modelar variáveis binárias, superando as limitações da regressão linear ao garantir probabilidades entre 0 e 1. Desenvolvida a partir de fundamentos históricos de Galton e Cox, a sua estrutura baseia-se na função *logit* e permite interpretar coeficientes como impactos nas *odds* do evento. Este modelo oferece flexibilidade para incorporar

múltiplas variáveis e técnicas avançadas, como regularização, embora exija atenção aos pressupostos como é o caso da linearidade no *logit* e na independência das observações.

3.2 Modelo Estatístico de Regressão Logística de Firth

Na regressão logística binária, modela-se a probabilidade de ocorrência do evento (por exemplo, *sinistro grave*) em função de um vetor de covariáveis $x_i \in \mathbb{R}^p$. Para $i = 1, \dots, n$, considere-se $Y_i \in \{0,1\}$ com

$$Y_i \sim \text{Bernoulli}(\pi_i), \pi_i = \mathbb{P}(Y_i = 1 \mid x_i).$$

O modelo logístico especifica

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \beta_0 + x_i^\top \beta,$$

logo

$$\pi_i(\beta) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}.$$

A estimação por máxima verosimilhança (MV) baseia-se na função de verosimilhança

$$L(\beta) = \prod_{i=1}^{n} \pi_i(\beta)^{y_i}[1 - \pi_i(\beta)]^{1-y_i},$$

ou, equivalentemente, no logaritmo da verosimilhança

$$\ell(\beta) = \sum_{i=1}^{n} \{y_i \log(\pi_i(\beta)) + (1 - y_i)\log(1 - \pi_i(\beta))\}.$$

Em problemas de eventos raros (classe positiva muito pouco frequente) e/ou com muitos preditores categóricos, podem ocorrer duas dificuldades clássicas:

1. *Viés* de pequena amostra (*small-sample bias*): os estimadores de máxima verosimilhança em modelos logísticos podem apresentar *viés* não negligenciável quando o número de eventos é pequeno relativamente ao número de parâmetros.

2. Separação (completa ou quase completa): existe separação quando uma combinação linear das covariáveis discrimina perfeitamente as classes (por exemplo, sempre que $x$assume certo padrão, $Y = 1$). Nesse caso, a verosimilhança aumenta sem limite e os estimadores de máxima verosimilhança divergem (alguns $\hat{\beta} \to \pm\infty$), levando a instabilidade numérica e probabilidades previstas extremas (≈ 0 ou ≈ 1).

A regressão logística penalizada de Firth foi proposta precisamente para reduzir o *viés* de pequena amostra e fornecer estimações finitas mesmo sob separação.

Ideia central: penalização de Jeffreys e correção de viés

O método de Firth pode ser visto como uma maximização de uma verosimilhança penalizada. Em vez de maximizar $\ell(\beta)$, maximiza-se:

$$\ell_F(\beta) = \ell(\beta) + \frac{1}{2}\log \mid I(\beta) \mid,$$

onde $I(\beta)$é a matriz de informação de Fisher (observada/esperada, dependendo da formulação; na prática usa-se a forma padrão em GLM).

Esta penalização corresponde ao uso do prior de Jeffreys (em interpretação Bayesiana) e, do ponto de vista frequentista, produz uma redução do viés de primeira ordem do estimador de máxima verosimilhança.

No caso do modelo logístico, definindo:

- Xcomo a matriz de desenho $n \times p$(incluindo intercepto),
- $W(\beta) = \text{diag}(w_i(\beta))$com $w_i(\beta) = \pi_i(\beta)(1 - \pi_i(\beta))$,

tem-se a informação de Fisher:

$$I(\beta) = X^\top W(\beta) X.$$

Equações de estimação: scores modificados

Na regressão logística padrão, o vetor score é:

$$U(\beta) = \frac{\partial \ell(\beta)}{\partial \beta} = X^\top (y - \pi),$$

onde $y = (y_1, \dots, y_n)^\top$e $\pi = (\pi_1, \dots, \pi_n)^\top$.

Em Firth, o termo penalizador altera o score para:

$$U_F(\beta) = U(\beta) + \frac{1}{2}\frac{\partial}{\partial\beta}\log\ |\ I(\beta)\ |.$$

Uma forma prática e muito usada desta correção é escrever o score modificado como:

$$U_F(\beta) = X^\top(y - \pi + a),$$

em que $a = (a_1, \ldots, a_n)^\top$é um vetor de ajuste dependente da alavancagem, com

$$a_i = \left(\frac{1}{2} - \pi_i\right) h_i,$$

onde $h_i$são os elementos diagonais da matriz "hat" do GLM:

$$H = W^{1/2}X\ (X^\top WX)^{-1}X^\top W^{1/2}, h_i = H_{ii}.$$

Assim, o método de Firth equivale a resolver:

$$X^\top(y - \pi + a) = 0,$$

o que pode ser interpretado como uma substituição do vetor-resposta efetivo (ou "pseudo-resposta") que corrige o *viés*.

Consequências práticas: o que distingue Firth da logística clássica

A regressão logística penalizada de Firth distingue-se da regressão logística clássica em aspetos críticos para eventos raros:

- Estimativas finitas sob separação: enquanto a máxima verosimilhança pode divergir, Firth produz $\hat{\beta}_F$finitos, estabilizando o ajuste.
- Redução do *viés* em amostras pequenas: especialmente relevante quando o número de eventos é reduzido (por exemplo, poucos sinistros graves) e existem muitos níveis/categorias.
- Probabilidades previstas menos extremas: ao contrariar a separação e a instabilidade, evita previsões degeneradas (0/1), o que tende a beneficiar a calibração.
- Inferência mais robusta em eventos raros: a estimação finita permite obter intervalos e testes mais interpretáveis quando a máxima verosimilhança falha.

Importa notar que Firth é uma penalização diferente de *Ridge/Lasso*: não visa seleção de variáveis, mas sim correção de *viés* e robustez sob separação. Em termos conceptuais, enquanto *RidgeLlasso* impõem uma penalização direta em $\| \beta \|$, Firth penaliza pela geometria da informação (via $\log | I(\beta) |$).

Testes e intervalos: razão de verosimilhança penalizada (preferível)

Em modelos com Firth, é comum reportar inferência baseada em razão de verosimilhança penalizada:

$$\Lambda = 2\left[\ell_F(\hat{\beta}_F) - \ell_F(\hat{\beta}_{F,0})\right],$$

onde $\hat{\beta}_{F,0}$é o estimador sob a hipótese nula (por exemplo, removendo uma covariável). Esta abordagem tende a ser mais estável do que testes Wald em cenários com eventos raros e/ou separação.

Enquadramento no presente trabalho

Dado o forte desequilíbrio entre categorias (sinistros graves como categoria minoritária), a logística penalizada de Firth foi incluída como alternativa ao GLM logístico clássico por duas razões:

(i) aumentar a estabilidade do ajuste e reduzir *viés* em presença de poucos eventos;

(ii) (ii) garantir estimações finitas e previsões probabilísticas úteis mesmo em cenários onde combinações de covariáveis possam induzir separação.

Assim, Firth constitui uma opção metodológica particularmente adequada quando se pretende manter um modelo interpretável, com coeficientes e odds ratios bem definidos, num contexto de eventos raros.

Em síntese, a logística de Firth é uma versão "regularizada" da regressão logística clássica, desenhada para funcionar melhor quando há poucos eventos, reduzindo viés e evitando coeficientes infinitos em situações de separação — um risco real em dados muito desequilibrados.

3.3 *Machine Learning*

Os modelos de *machine learning* usados neste trabalho enquadram-se nos modelos de aprendizagem supervisionada.

A maioria dos modelos considerados (Random Forest, C5.0 e XGBoost) baseiam-se em árvores de decisão, ou seja, estruturas em árvore que representam conjuntos de decisões.

Para garantir a validade estatística dos modelos e evitar o *overfitting*, os dados originais são divididos em dois subconjuntos distintos:

- Treino: estes dados são usados para a construção e ajuste do modelo.
- Teste: estes dados que não foram usados na fase de construção e ajuste do modelo são usados para avaliar a capacidade de generalização do modelo.

3.3.1 Naive Bayes

O Naive Bayes é um algoritmo de *machine learning* baseado no Teorema de Bayes, que permite calcular probabilidades condicionais "invertidas". Desenvolvido a partir do trabalho do matemático Thomas Bayes (século XVIII), o teorema revolucionou a inferência estatística ao propor como atualizar as probabilidades iniciais (probabilidade *a priori*) com base em evidências observadas, resultando numa probabilidade *a posteriori*. A fórmula central é dada por:

$$P(Y \mid X) = \frac{P(X|Y) \times P(Y)}{P(X)},$$

onde:

- $P(Y)$: probabilidade *a priori* – estimativa inicial da probabilidade da categoria Y
- $P(X|Y)$: verosimilhança – probabilidade de observar as caraterísticas X dado que a categoria é Y
- $P(X)$: evidência – probabilidade marginal das caraterísticas X, atuando como fator de normalização
- $P(Y|X)$: probabilidade a posteriori – probabilidade atualizada da categoria Y após considerar as evidências X

Para evitar que probabilidades condicionais sejam zero quando uma caraterística não aparece numa determinada categoria, Pierre-Simon Laplace introduziu conceito de "suavização" (*Laplace smoothing*). A técnica adiciona um valor α (geralmente $\alpha = 1$) às contagens de frequência:

$$P(X_i|Y) = \frac{Contagem\ (X_i\ em\ Y) + \alpha}{Total\ de\ contagens\ em\ Y + \alpha \times c}, i = 1, \ldots, c,$$

onde c é o número de valores únicos, ou categorias, que X_i pode assumir. Assim, mesmo que uma caraterística esteja ausente no treino, a sua probabilidade não será zero. O modelo passa a considerar uma probabilidade mínima, permitindo que outras caraterísticas influenciem a classificação.

Na prática, estamos a classificar sinistros rodoviários como "M/FG" ou "FL" com base em variáveis específicas. Supondo que as variáveis analisadas são "tipo de acidente", "Veículos Pesados" e "Hora" ($c = 3$), e que no treino temos:

- "M/FG"
  - "Atropelamento": 20 observações
  - "Veículo pesado": 25 observações
  - "20h-6h": 15 observações

  } total de observações = 60
- "FL"
  - "Atropelamento": 2 observações
  - "Veículo pesado": 1 observação.
  - "20h-6h": 0 observações

  } total de observações = 3

A probabilidade de "Veículo Pesado = Sim" em sinistros com "FL" sem a suavização de Laplace seria:

$$P(\text{Veículo Pesado} \mid \text{FL}) = \frac{0}{3} = 0.$$

Com a suavização de Laplace ($\alpha = 1$) ajusta-se as contagens adicionando $\alpha = 1$ a cada palavra e atualiza-se o denominador:

$$P(\text{Veículo Pesado} \mid \text{FL}) = \frac{0+1}{3+1 \times 3} = \frac{1}{6} \approx 0{,}1667.$$

Mesmo que "Veículo Pesado = Sim" nunca tenha sido registado em sinistros de "FL" durante o treino, a sua probabilidade agora não é nula. Isso permite que outros fatores (como "Atropelamento" ou "Hora") contribuam para a decisão final, evitando que o modelo falhe ou fique bloqueado por causa de uma variável ausente.

Este algoritmo adapta-se a diferentes tipos de dados através das seguintes variantes, cada uma projetada para lidar com caraterísticas específicas:

- Multinomial Naive Bayes

Foi projeto para trabalhar com dados discretos, como contagem de palavras em textos. O seu funcionamento baseia-se no cálculo de probabilidades a partir das frequências do evento.

- Gaussian Naive Bayes

Variante do algoritmo projetada para lidar com dados contínuos, como temperatura, valores biométricos (altura, peso...). O seu funcionamento assume que os dados seguem uma distribuição normal (Gaussiana), o que permite estimar probabilidades condicionais a partir da média (μ) e da variância (σ^2) de cada caraterística por categoria. Um aspeto crítico é a sua sensibilidade a *outliers*, já que a distribuição Gaussiana pressupõe simetria nos dados. Se um valor extremo estiver presente, a estimativa de probabilidade pode ser distorcida, afetando a precisão do modelo. Para mitigar isso, é recomendado aplicar técnicas de pré-processamento, como normalização ou remoção de *outliers*, antes do treino.

- Bernoulli Naive Bayes

É especializado em caraterísticas binárias, como presença ou ausência de palavras num documento. O seu funcionamento baseia-se na estimativa de probabilidades

condicionais para cada estado binário ($X_i = 1$ ou $X_i = 0$) dentro de uma categoria. Uma diferença crucial em relação com Multinomial Naive Bayes é que o Bernoulli ignora a frequência de ocorrência, focando apenas na existência ou não de uma caraterística.

A tabela que se segue, Tabela 2, apresenta uma comparação entre os três principais tipos de Naive Bayes – Multinomial, Gaussiana e Bernoulli. Esta comparação permite compreender qual a versão do Naive Bayes é mais adequada dependendo da natureza dos dados (discretos, contínuos ou binários) e do contexto da aplicação.

Tabela 2 – Comparação de algoritmos de Naive Bayes.

	Multinomial Naive Bayes	**Gaussian Naive Bayes**	**Bernoulli Naive Bayes**
Tipos de Dados	Dados discretos (contagens/ frequências)	Dados contínuos (valores numéricos)	Dados binários (0 ou 1)
Suposição	Distribuição multinominal (contagens)	Distribuição Gaussiana (normal)	Distribuição de Bernoulli (presença/ausência)
Aplicação	Classificação de texto, análise de sentimentos, categorização de documentos	Diagnóstico médico, reconhecimento de padrões em sensores, previsão de falhas	Deteção de fraudes, classificação binária de documentos, filtros de *spam* simplificados
Cálculo de $P(X_i\|Y)$	Frequências relativas com suavização de Laplace	Média (μ) e variância (σ^2) por categoria	Probabilidade de presença $P(X_i = 1\|Y)$ ou ausência $P(X_i = 0\|Y)$
Tratamentos de Zeros	Suavização de Laplace para evitar $P(X_i\|Y) = 0$	Não aplicável (usa distribuição contínua)	Suavização de Laplace opcional
Limitações	Ignora ordem de palavras em texto	Sensível a *outliers* e distribuições não Gaussianas	Ignora frequência

Sendo Y a variável resposta do tipo nominal (categorias a serem previstas) e $X = (X_1, \dots, X_k)$ o vetor das variáveis explicativas (caraterísticas), o algoritmo funciona da seguinte forma:

- Probabilidades *a priori* ($P(Y)$)

  O objetivo é calcular a probabilidade inicial $P(Y)$ de cada categoria de Y com base nos dados de treino, geralmente estimada pela frequência relativa. O cálculo é:

  $$P(Y = y) = \frac{Número\ de\ observações\ da\ categoria\ y\ no\ treino}{Total\ de\ observações\ no\ treino}.$$

  Este procedimento reflete a distribuição das categorias no treino. Se uma categoria é rara, a sua probabilidade a priori será baixa.

- Probabilidades condicionais

  Nesta etapa o objetivo é calcular a probabilidade de observar as caraterísticas X dado que a categoria é $Y = y$.

  Para variáveis contínuas, assume-se uma distribuição normal (Gaussiana) e estimam-se a média e a variância por categoria;

  Para variáveis discretas, utiliza-se as contagens de frequência, muitas vezes com a suavização de Laplace para evitar probabilidades iguais a zero.

- Classificação (Cálculo da Probabilidade *a posteriori* - ($P(Y|X)$)

  Na última fase o objetivo é determinar a categoria mais provável para uma nova observação com caraterísticas X, utilizando a fórmula do Teorema de Bayes:

  $$P(Y = y \mid X) = \frac{P(X \mid Y = y) \times P(Y = y)}{P(X)},$$

  onde a categoria com maior probabilidade é atribuída à observação.

Este algoritmo é simples e eficiente, ideal para problemas como a classificação da sinistralidade rodoviária com múltiplas variáveis preditivas. Ele é escalável e possui complexidade computacional linear, sendo adequado para grandes volumes de dados, requerendo poucos dados de treino, o que o torna útil em cenários com dados limitados. Além disso, é robusto a ruídos e *outliers* devido à suposição de independência.

Porém, a sua principal limitação é a suposição de independência entre as características, conhecida como "independência condicional", o que pode prejudicar a precisão quando as variáveis estão correlacionadas. Além disso, o Naive Bayes pode ter dificuldades para

fornecer boas estimativas de probabilidade quando há desequilíbrio de categorias ou quando uma categoria não aparece no treino (problema conhecido como "*zero-frequency*", mitigado pela suavização de Laplace). Outra limitação é a sensibilidade a preditores irrelevantes: se muitas variáveis não informativas foram incluídas, o desempenho pode degradar-se.

3.3.2 Random Forest

O *Random Forest*, é um algoritmo de *machine learning*, mais especificamente de *ensemble learning* (aprendizagem por conjunto), proposto por Leo Breiman e Adele Cutler em 2001 (Breiman, 2001). Para compreender este algoritmo primeiro é necessário entender o seu comportamento básico: árvores de decisão.

De acordo com o IBM (2024) uma árvore de decisão é um modelo não paramétrico de aprendizagem supervisionada, utilizado para classificação e regressão. A sua estrutura assemelha-se a uma estrutura em forma de árvore, que inicia com um nó raiz, seguindo-se os nós internos (ou nós de decisão) em que cada um dos nós representa um teste aplicado a uma variável explicativa, cada ramo representa o resultado desse teste e cada nó folha (ou nó terminal) contém o resultado, i.e., uma etiqueta de classe (para classificação) ou um valor contínuo para regressão. No entanto, apenas uma árvore de decisão é altamente sensível a pequenas variações nos dados de treino e é precisamente para superar essa limitação que o Random Forest foi criado.

O algoritmo *Random Forest* combina múltiplas árvores de decisão para produzir previsões mais precisas e robustas do que uma árvore isolada. A essência do *Random Forest* reside na diversificação: ao construir várias árvores com subconjuntos aleatórios dos dados e variáveis, o modelo reduz a variância e evita o *overfitting*.

O *Random Forest* opera em três etapas principais: *bootstrap aggregating (bagging)*, construção de árvores com seleção aleatória de *features* e agregação de resultados. Cada etapa é projetada para introduzir aleatoriedade e diversidade, garantindo que as árvores sejam independentes e complementares.

1) ***Booststrap Aggregating (Bagging)***

- Amostragem com reposição: cada árvore é treinada com um subconjunto dos dados de treino, gerado pela amostragem com reposição (técnica de *bootstrap*).

Isso significa que, para uma base de dados com n observações, cada subconjunto terá n amostras. No entanto, devido à reposição, algumas observações originais podem ser selecionadas várias vezes, enquanto outras não são selecionadas. Quando $n \to \infty$, a probabilidade de uma observação nunca ser escolhida é de aproximadamente 0,37. Deste modo, em média, para treinar cada árvore são usadas aproximadamente 63% das observações originais e as 37% restantes não são usadas nesse treino. A este conjunto de observações excluídas dá-se o nome de *out-of-bag* (Hastie et al., 2009).

- Versatilidade: a amostragem aleatória garante que cada árvore "veja" dados ligeiramente diferentes, reduzindo a correlação entre as árvores e melhorando a generalização do modelo.

2) Seleção Aleatória de Variáveis (*Feature Randomness*)

Em cada divisão de um nó da árvore, apenas um subconjunto de m variáveis (geralmente $m = \sqrt{k}$, sendo k o número total de variáveis para classificação) é considerado. Essa seleção aleatória evita que uma única variável dominante influencie todas as árvores, promovendo diversidade (Breiman, 2001).

- Critério de divisão: para cada subconjunto de variáveis, a árvore escolhe a melhor divisão usando critérios como *Gini impurity* (classificação) ou redução da variância (regressão).

3) Construção de Árvores e Agregação

Cada árvore é construída independentemente até à sua profundidade máxima, o que a torna propensa a *overfitting*. No entanto, a agregação de múltiplas árvores compensa esse *viés* e tenta fazer previsões usando os dados amostrados. Isso reduz o risco de *overfitting*, um problema em árvores de decisão únicas.

4) Previsão final:

A etapa final de previsão requer a agregação das árvores que é determinado pela natureza da variável resposta. Quando se está perante uma classificação, a variável resposta é categórica e por isso cada árvore do *ensemble* produz uma previsão de

categoria. A previsão final neste caso é obtida pela moda das previsões individuais, ou seja, é a categoria mais frequente entre todas as árvores.
Quando se está perante uma Regressão, a variável é contínua e cada árvore produz uma previsão numérica. A previsão final nestas situações é dada pela média aritmética das previsões.

Os hiperparâmetros do *Random Forest* são configurações que controlam o treino do modelo, influenciando a sua precisão, velocidade e capacidade de generalização. Entre os mais relevantes estão:

- **Número de árvores ($n_{estimators}$):** mais árvores aumentam a estabilidade, mas têm custos computacionais.
- **Número de Variáveis por Divisão (m):** controla a diversidade. Valores menores reduzem a correlação entre árvores.
- **Profundidade Máxima das Árvores:** limitar a profundidade previne *overfitting* individual, porém árvores muito rasas podem estar sujeitas a *underfitting*.

A escolha adequada destes parâmetros é essencial para equilibrar o desempenho e a complexidade, evitando *overfitting* ou *underfitting*.
Esta abordagem permite avaliar a relevância de cada variável para as previsões do modelo. A importância de cada variável é calculada de duas formas:

- **Gini Importance:** mede quantas vezes uma variável reduz a impureza (Gini) nas divisões, ponderada pelo número de amostras afetadas.
- ***Permutation Importance:*** avalia a queda na precisão do modelo quando os valores da variável são aleatoriamente permutados (Lundberg; Lee, 2017)

Relativamente às vantagens e limitações, destaca-se positivamente a robustez a dados ruidosos e *outliers*, pela capacidade de lidar com relações não lineares entre variáveis e pela avaliação interna de desempenho via amostras *out-of-bag*. Contudo, a sua principal limitação reside no custo computacional elevado para grandes bases de dados e na interpretabilidade reduzida, já que a "floresta" de árvores dificulta a explicação de previsões individuais.

Devido à sua versatilidade, o *Random Forest* é amplamente utilizado em áreas como medicina (diagnóstico de doenças), finanças (avaliação de risco de crédito), ecologia (modelação de *habitats*) e *marketing* (segmentação de clientes). Na sinistralidade rodoviária, esta técnica é aplicada para:

- Identificar combinações de fatores de risco (ex: geometria da via, condições ambientais, comportamento do condutor, etc).
- Priorizar intervenções preventivas (ex: classificação do troço por nível de perigo).
- Prever sinistros graves com base em padrões complexos (ex: deteção de relações não lineares).

A sua eficácia em grandes conjuntos de dados, aliada á capacidade de quantificar a relevância de variáveis, consolida-o como uma ferramenta analítica valiosa. Contudo, a complexidade computacional inerente à construção de diversas árvores e a menor interpretabilidade comparativamente a modelos individuais representam compromissos a considerar.

Em síntese, o algoritmo equilibra precisão preditiva e generalização, tornando-se indispensável para problemas de classificação e regressão, onde a estabilidade e adaptabilidade a cenários heterogéneos são prioritárias.

3.3.3 Algoritmo C5.0

O algoritmo C5.0, desenvolvido por Ross Quinlan na década de 1990, representa a evolução mais avançada dos algoritmos de árvore de decisão criados pelo autor. Quinlan, reconhecido como pioneiro da área do *machine learning*, estruturou uma linha cronológica de modelos, iniciada com o ID3 (Iterative Dichotomiser 3) em 1986, seguido pelo C4.5 em 1993 e culminando no C5.0 (Quinlan, 1993; Kuhn & Johnson, 2013).

No que concerne aos antecedentes, os primeiros algoritmos de Quinlan surgiram para resolver desafios centrais de aprendizagem supervisionada: criar modelos interpretáveis capazes de prever uma variável dependente com base em atributos descritivos. O ID3 introduziu conceitos inovadores, como o uso de entropia e o ganho de informação para selecionar divisões ótimas na árvore.

A entropia é uma medida de impureza ou desordem num conjunto de dados S. Matematicamente, calcula-se através da equação que se segue:

$$\text{Entropia } (Y) = \ -\sum_{i=1}^{c} p_i \log_2(p_i)\,, \qquad 0 \leq Entropia(Y) \leq log_2(c),$$

onde p_i é a proporção de observações da categoria i da variável Y, e c o número de categorias distintas de Y.

Por exemplo, supondo que 85% dos condutores não estiveram envolvidos em sinistros no último ano (categoria "Não") e 15% estiveram envolvidos (categoria "Sim"), ou seja:

- Categoria "Não": $p_{não} = 0{,}85$,
- Categoria "Sim": $p_{sim} = 0{,}15$,

o valor da entropia é:

$$Entropia\ (Y) = \ -\big(0{,}85 \times log_2(0{,}85) + 0{,}15 \times log_2(0{,}15)\big) \approx 0{,}609.$$

Neste caso ($S \approx 0{,}609$), a entropia está mais próxima do mínimo (0) do que do máximo (1), indicando que há uma alta homogeneidade nos dados. A maioria dos condutores partilha um comportamento semelhante (não se envolvem em sinistros), o que reduz a desordem na previsão de comportamentos futuros.

Se as categorias estiverem igualmente distribuídas (ex.: 50% "Sim", 50% "Não"), a entropia é máxima (1). O valor mínimo de entropia (0) é obtido quando todas as observações pertencem a uma única categoria.

O ganho de informação, por sua vez, quantifica quanto um atributo X reduz a entropia de Y após dividir os dados S com base na variável X (Quinlan, 1993). Por exemplo, ao utilizar o atributo "Idade do condutor" (condutores até 25 anos vs. condutores com mais de 25 anos) para dividir os registos de sinistros, calcula-se:

1) A entropia original do conjunto completo;
2) A entropia de cada subconjunto (condutores até 25 anos vs. condutores com mais de 25 anos);
3) A diferença entre a entropia original e a média ponderada das entropias dos subconjuntos.

$$Ganho\ (Y, X) = Entropia\ (Y) - \sum_{v \in Valores(X)} \frac{|Y_v|}{|Y|}\ Entropia\ (Y_v)$$

O atributo com maior valor é selecionado para a divisão, pois maximiza a homogeneidade dos subconjuntos.

Apesar da inovação, a abordagem do ID3 apresentava falhas críticas ligadas justamente a esses conceitos:

- *Viés* do ganho de informação: atributos com muitos valores únicos geravam ganhos artificialmente altos, mesmo sem relevância preditiva;
- Atributos contínuos: exigia discretização manual prévia, o que limitava a sua aplicação em dados numéricos;
- Ausência de poda (*pruning*): resultava em árvores excessivamente complexas, sem mecanismos de simplificação, reduzia a generalização. Aqui eram capturados ruídos em vez de padrões;
- *Missings*: ignorava *missings* e não suportava tarefas de regressão (Quinlan, 1996).

O C4.5 (1993) superou essas limitações com avanços significativos, nomeadamente:

- Discretização automática de atributos contínuos: o algoritmo identifica, de forma dinâmica, o ponto de corte ideal para variáveis numéricas, transformando-as em condições binárias durante a construção da árvore. Esse processo é realizado mediante a ordenação dos valores e avaliação de possíveis pontos de corte entre as diversas categorias distintas, selecionando aquele que maximiza o ganho de informação.

  Ganho da razão (*Gain Ratio*), ajusta o *viés* de atributos multivariados, penalizando aqueles com alta cardinalidade (grande quantidade de dados com mínima repetição).

$$Gain\ Ratio\ (X) = \frac{Ganho(Y, X)}{SplitInformation(X)},$$

  com

$$Split\ Information\ (X) = -\sum_{i=1}^{k} \frac{|S_i|}{|S|} \times\ log_2 \frac{|S_i|}{|S|},$$

  em que:

k: número de subconjuntos (ramos) gerados pela divisão do atributo X;

$|S_i|$: número de instâncias no i-ésimo subconjunto (ramo);

$|S|$: número total de instâncias no nó.

- Poda pós construção (*post-pruning*) para simplificação;
- Tratamento probabilístico valores omissos (*missings*), distribuindo as observações conforme a frequência observada;
- Geração de regras para maior interpretabilidade (Quinlan, 1993).

Apesar dos avanços, o C4.5 mostrou-se inadequado para determinadas caraterísticas:

- Ineficiência computacional: consumo excessivo de memória em grandes conjuntos de dados;
- Poda não otimizada: a poda *post-hoc* gerava desperdício de recursos ao simplificar a árvores apenas após a sua construção completa;
- Desequilíbrio de categorias: o desempenho era insatisfatório em categorias minoritárias;
- Falta de suporte nativo a técnicas de *ensemble* (combinação de múltiplos modelos para melhorar a performance): impossibilitava a implementação de abordagens como *boosting* (técnica que combina modelos sequencialmente, onde cada novo modelo corrige os erros do anterior) ou *bagging* (método que combina modelos independentes treinados em subconjuntos aleatórios dos dados), limitando a sua capacidade de reduzir a variância e melhorar a generalização (Kuhn & Johnson, 2013).

Foi nesse contexto que surgiu o C5.0, como uma tentativa de endereçar as limitações práticas do C4.5 e tornar o algoritmo mais eficiente, escalável e competitivo frente a novas abordagens emergentes em *machine learning*.

Neste sentido, o C5.0 é um algoritmo de classificação baseado em árvores de decisão. O seu principal objetivo é prever variáveis dependentes categóricas a partir de atributos preditivos, construindo uma estrutura hierárquica que divide os dados conforme a capacidade discriminativa das variáveis. A sua eficiência em lidar com dados heterogéneos (numéricos e categóricos) e a sua robustez contra o *overfitting* fazem com que o mesmo seja utilizado em diversas áreas como marketing, medicina, entre outras (Quinlan, 2014).

A construção deste modelo segue uma abordagem recursiva, com etapas que incluem seleção de atributos, divisão de dados e otimização pós-construção.

Na preparação dos dados, os mesmos devem ser estruturados em formato tabular, onde:

- Linhas representam observações;
- Colunas correspondem a atributos preditivos e à variável dependente;
- Os dados podem incluir tanto variáveis contínuas como categóricas bem como *missings*.

Para a construção da árvore, o processo é composto por três fases:

1) **Seleção de atributos:**

A seleção de atributos é baseada no ganho de informação, ajustado pelo *Gain Ratio* para mitigar o *viés* em atributos multivariados (Quinlan, 1993).

O atributo com maior *Gain Ratio* é escolhido, priorizando divisões que gerem subgrupos homogéneos.

O processo de seleção de divisões envolve a avaliação de todas as possíveis divisões dos dados para cada atributo, escolhendo a que proporciona o maior ganho de informação. Essa escolha sequencial dos melhores atributos resulta em uma árvore de decisão que hierarquiza as caraterísticas mais informativas, refletindo os padrões subjacentes no conjunto de dados.

2) **Divisão do conjunto de dados:**

O processo de divisão do conjunto de dados é fundamental para a construção da árvore e difere consoante o tipo de variável:

- Variáveis categóricas: A divisão é efetuada pelos valores únicos da variável. Cada ramo da árvore corresponde a um valor possível (ex: para a variável "Tipo de veículo", seriam criados ramos para "Veículo particular" ou "Veículo comercial");

- Variáveis numéricas: o processo de divisão baseia-se na identificação de um ponto de corte ótimo. O algoritmo avalia sequencialmente possíveis pontos de corte ao longo da distribuição dos valores, selecionando aquele que maximiza a homogeneidade (ou minimiza a impureza) dos subconjuntos. Esta divisão binária separa o conjunto de dados em dois subconjuntos, cada um direcionado para um ramo distinto da árvore, consoante a veracidade da condição de desigualdade estabelecida.

3) **<u>Critérios de finalização:</u>**

O processo repete-se recursivamente até que pelo menos um dos seguintes critérios seja satisfeito:

- Todos os exemplos num nó pertençam à mesma categoria;
- Não haja atributos para divisão;
- Limites predefinidos sejam atingidos.

Após a construção inicial da árvore, o C5.0 aplica técnicas de otimização para garantir equilíbrio entre precisão e generalização:

- Poda (*Pruning*): remove ramos redundantes para evitar *overfitting*. O critério baseia-se numa avaliação estatística de custo-complexidade. Um ramo é considerado redundante se a sua remoção não provocar um aumento significativo da taxa de erro de classificação, ou seja, se a sua contribuição para a redução da impureza (ex: entropia) for inferior a um determinado limite de ganho mínimo predefinido.
- Peneiramento (*Winnowing*): descarta atributos que contribuem pouco para a redução de entropia, aumentando a simplicidade e eficiência da árvore final.

O C5.0 é a culminação de uma trilha evolutiva iniciada por Ross Quinlan com o ID3 e o C4.5 que resolve limitações históricas e estabelece novos padrões em modelos de classificação. Ao integrar avanços como o *Gain Ratio* (corrige o *viés* de atributos multivariados), a discretização automática de variáveis contínuas e técnicas de poda otimizada, o C5.0 destaca-se pela eficiência computacional, interpretabilidade e robustez contra *overfitting*. A sua capacidade de hierarquizar atributos informativos, aliada a métodos como o peneiramento, produz árvores adaptáveis a dados heterogéneos, mantendo o equilíbrio entre precisão e generalização.

3.3.4 XGBoost

O XGBoost (*Extreme Gradient Boosting*) emergiu como um dos algoritmos mais influentes na história do *machine learning*, revolucionando a forma como problemas de classificação e regressão são abordados. Desenvolvido em 2014 por Tianqi Chen e Carlos

Guestrin, este algoritmo combina a robustez teórica do *Gradient Boosting* com otimizações computacionais inovadoras, tornando-o uma ferramenta indispensável em cenários que exigem precisão, eficiência e escalabilidade.

A criação do XGBoost foi motivada por lacunas no *Gradient Boosting* tradicional, proposto por Jerome H. Friedman em 2001. Embora o método de Friedman permitisse a construção iterativa de modelos preditivos precisos, três desafios persistiam:

- **Ineficiência computacional:** o treino sequencial de árvores (cada nova árvore corrige os erros da anterior) tornava o processo lento, especialmente em grandes volumes de dados.
- **Fragilidade a *overfitting*:** a falta de mecanismos de controle de complexidade levava os modelos a memorizar os dados de treino, prejudicando a generalização.
- **Dificuldade de implementação:** a ausência de otimizações restringe a escalabilidade (Friedman, 2001).

Para superar tais limitações, Chen e Guestrin (2016) introduziram três avanços fundamentais:

- **Paralelização e otimização computacional**

No *Gradient Boosting* tradicional, cada árvore é treinada sequencialmente, ou seja, uma árvore só começa a ser construída após a conclusão da anterior. O XGBoost substitui o treino sequencial por estratégias paralelas em múltiplos níveis:

- Paralelização de nível de árvore: enquanto as árvores são construídas sequencialmente, o cálculo das melhores divisões (*splits)* em cada nó é paralelizado. O algoritmo divide o conjunto de dados em partes menores (blocos – estruturas de dados compactas) e avalia divisões para diferentes caraterísticas simultaneamente utilizando diferentes núcleos do processador CPU (Chen & Guestrin, 2016). Isso permite que os cálculos das divisões sejam feitos mais rapidamente, acelerando o processo.

- Algoritmo aproximado para encontrar divisões: utiliza histogramas para agrupar os dados em categorias de intervalo. Isso simplifica os cálculos, reduzindo a complexidade de $O(n)$ para $O(logn)$, onde n é o número de amostras (Chen & Guestrin, 2016). Isso significa que, em vez de se analisar cada amostra individualmente – o que levaria um

tempo proporcional ao tamanho do conjunto de dados $\big(O(n)\big)$, o algoritmo consegue encontrar divisões de forma mais rápida, examinando apenas uma pequena parte dos dados a cada passo $\big(O(logn)\big)$.

- Suporte a ambientes distribuídos: o treino do modelo por ser dividido entre várias máquinas, em vez de ser executado em apenas um computador. Isso é feito através de um *cluster* (grupo de máquinas que trabalham juntas), permitindo que grandes quantidades de dados sejam processadas de forma mais rápida e eficiente, pois cada máquina executa uma parte do trabalho.

- **Técnicas avançadas de regularização (L1/L2)**

A regularização é uma técnica para evitar que o modelo se torne demasiado complexo e perca a capacidade de generalização.

O XGBoost incorpora termos de penalização na função de perda (*loss function)* para evitar *overfitting*:

- Regularização L1 (Lasso) adiciona uma penalização proporcional ao valor absoluto dos coeficientes do modelo. Isso força o algoritmo a eliminar variáveis irrelevantes.

- Regularização de L2 (Ridge) penaliza o quadrado dos coeficientes, suavizando o impacto de variáveis extremas, evitando que *outliers* dominem o modelo.

A equação da perda é dada por:

$$Loss = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \lambda \sum_{j=1}^{k} |b_j| + \alpha \sum_{j=1}^{k} b_j^2,$$

onde:

- L: é a função de perda ou custo,
- $L(y_i, \hat{y}_i)$: erro de predição,
- λ: penalidade L1(Lasso) para eliminar variáveis irrelevantes,
- α: penalidade L2 (Ridge) para suavizar coeficientes,
- k: número de variáveis do modelo,
- b_j: peso associado à *j-ésima* variável.

- **Sistema de gestão de memória e eficiência**

O sistema de armazenamento foi otimizado através de:

- Estrutura de dados em blocos: armazena os dados em blocos compactos, permitindo acesso rápido e reduzindo a sobrecarga de memória.

- Formato de coluna comprimido: comprime colunas de dados poupando espaço, reduzindo o espaço do disco.

- *Cache-Awareness*: algoritmos antecipam quais os dados que serão necessários, armazenando-os para acesso rápido.

Um exemplo prático é a organização de registos de sinistralidade rodoviária. Em vez de registos desorganizados, eles são agrupados por género (blocos) e colocados em ficheiros identificados, facilitando a procura.

O artigo de 2016 "XGBoost: A Scalable Tree Boosting System", detalha estas inovações, posicionando o XGBoost como uma ferramenta dominante. De forma resumida, a tabela abaixo, Tabela 3, compara o XGBoost com o *Gradient Boosting* tradicional, evidenciando as melhorias introduzidas pelo XGBoost em termos de desempenho, regularização e eficiência computacional.

Tabela 3 – Comparação dos algoritmos de Gradient Boosting.

Caraterística	Gradient Boosting Tradicional	XGBoost
Paralelização	Sequencial por árvore	Paraleliza o cálculo das divisões dos nós e distribui o treino por *clusters*
Regularização	Não suportada	Adiciona termos L1/L2 diretamente na função perda para controlar o *overfitting*
Memória	Armazenamento não otimizado	Estruturas de dados compactas
Algoritmo de *Splits*	Busca exata de divisões $O(n)$	Usa histogramas para aproximar as divisões, reduzindo tempo de $O(n)$ para $O(logn)$
Tratamento de valores omissos	Requer pré-processamento	Deteta automaticamente padrões para lidar com valores omissos

Em suma, o XGBoost surge não apenas como uma evolução técnica do *Gradient Boosting* tradicional, mas como uma resposta sistémica para desafios históricos: ineficiência computacional, *overfitting* e falta de escalabilidade. Ao integrar paralelização de nível de

árvore, algoritmos otimizados para cálculo de divisões e suporte a ambientes distribuídos, o modelo acelera o treino e permite a modelação de grandes volumes de dados. A regularização L1 e L2, por sua vez, introduz equilíbrio entre precisão e generalização, mitigando riscos de *overfitting*.

3.4 Técnicas de Reamostragem

No campo da análise de dados, um elevado desequilíbrio de categorias impacta significativamente a construção de modelos preditivos, pois os modelos tendem a ser mais sensíveis à categoria maioritária, subestimando as caraterísticas da categoria minoritária. Por outras palavras, o modelo é exposto a muitos mais exemplos da categoria maioritária do que da categoria minoritária, criando um *viés* nos algoritmos, que aprendem mais facilmente os padrões mais frequentes nos dados.

Diante desse desafio, diversas técnicas foram desenvolvidas para equilibrar a distribuição das categorias e melhorar o desempenho dos modelos preditivos. Entre as abordagens mais comuns, destacam-se:

- *Oversampling*

O *oversampling* é uma técnica que aumenta a quantidade de observações da categoria minoritária, frequentemente replicando as observações existentes ou criando observações sintéticas.

- *Undersampling*

O *undersampling* é uma técnica que envolve a redução do número de observações da categoria maioritária, eliminando algumas observações para equilibrar as categorias. Embora essa técnica possa ser eficaz para simplificar o problema, ela também pode resultar na perda de informação.

Ambas as técnicas visam criar um conjunto de dados equilibrado, permitindo que os algoritmos de *machine learning* identifiquem padrões presentes em todas as categorias de forma mais precisa. No entanto, a aplicação indiscriminada dessas técnicas pode levar a problemas como o *overfitting* e *underfitting*. O *overfitting* ocorre quando o modelo se ajusta em demasia aos dados de treino. Como resultado, o modelo apresenta um bom desempenho nos dados de treino, mas um desempenho fraco nos novos dados. Isso faz

com que o modelo não consiga gerar boas previsões. Já o *underfitting* acontece quando o modelo é muito simples para capturar as complexidades dos dados e quando não está bem ajustado aos dados, resultando num modelo com baixo desempenho tanto nos dados de treino quanto em novos dados.

Para superar essas limitações, têm sido propostas técnicas mais sofisticadas e eficientes, como o ROSE (*Random Over-Sampling Examples*) e o SMOTE (*Synthetic Minority Over-sampling Technique*). Essas técnicas geram novas observações sintéticas para a categoria minoritária, preservando ao mesmo tempo as caraterísticas intrínsecas dos dados originais.

3.4.1 ROSE (Random Over-Sampling Examples)

A técnica ROSE (*Random Over-Sampling Examples*), apresentada em 2014 pelos autores Nicola Lunardon, Giovanna Menardi e Nicola Torelli (Lunardon et al., 2014), foi proposta para mitigar o problema de categorias desequilibradas em conjuntos de dados de classificação, nas diversas aplicações de métodos de *machine learning*. Ao contrário dos métodos tradicionais que se limitam em replicar o número de observações da categoria minoritária ou a reduzir dados da categoria maioritária, o ROSE combina elementos de *bootstrap* com a estimativa de densidade *kernel* (KDE) para gerar novas observações sintéticas mais realistas. Esta abordagem considera tanto os dados contínuos quanto categóricos, reduz o risco de *overfitting* e melhora a capacidade de generalização dos modelos de *machine learning*, através:

1) **<u>Criação de dados sintéticos mais diversificados:</u>**

O processo tem início com a divisão da base de dados em conjuntos de treino e teste, podendo essa divisão ser estratificada ou temporal, conforme a natureza dos dados. O *oversampling* é então aplicado exclusivamente ao conjunto de treino, garantindo que o conjunto de teste permanece inalterado e representativo da distribuição original . o que evita problemas de *data leakage*.

No treino, procede-se à seleção aleatória de observações da categoria minoritária através de *bootstrap* - técnica de amostragem aleatória com reposição onde as observações podem ser selecionadas diversas vezes. De seguida, é calculada a KDE em cada ponto selecionado de forma a obter uma distribuição de probabilidade suavizada

à volta de cada observação original, permitindo que os novos pontos sejam gerados nas proximidades de forma que sejam diversificados e realistas. Por exemplo, existe um registo de um sinistro rodoviário ocorrido às 3h da manhã, o ROSE neste caso cria casos sintéticos com horários próximos, tal como 2h ou 4h da manhã (valores ligeiramente diferentes), seguindo a distribuição natural dos dados originais. O modelo que resulta deste procedimento aprende a reconhecer padrões mais amplos em vez de memorizar casos específicos, melhorando significativamente a capacidade de generalização.

2) **Balanceamento da distribuição das categorias:**

O número de novos casos sintéticos gerados depende do grau de equilíbrio pretendido nos dados, i.e., o quão próximo se pretende que esteja o número de observações nas duas categorias (minoritária e maioritária). Por exemplo, considerando um cenário onde existem 100 observações de vias sem sinistros graves e apenas 10 observações de vias com sinistros graves, o ROSE pode gerar 90 observações sintéticas de vias com sinistros graves, cada uma com pequenas variações em relação aos dados originais. Este valor não é fixo, depende do método de balanceamento escolhido assim como do objetivo do modelo. Este balanceamento faz com que a categoria minoritária tenha peso suficiente no processo de treino. O resultado é um conjunto de dados onde ambas as categorias contribuem de forma equilibrada para o modelo.

3) **Suavização da fronteira de decisão:**

Um dos aspetos inovadores do ROSE é a capacidade de lidar com a presença de zonas ambíguas – regiões do espaço de caraterísticas (*feature space*) onde as observações de diferentes categorias se sobrepõem, tornando a classificação incerta. Esta abordagem é particularmente relevante em várias situações do quotidiano, como a avaliação da sinistralidade rodoviária, uma vez que a distinção entre as categorias raramente é bem definida, o que pode fazer com que as fronteiras de decisão criadas sejam artificiais e rígidas.

Em resumo, o ROSE destaca-se como uma abordagem sofisticada e eficaz para lidar com conjuntos de dados desequilibrados, superando as limitações dos métodos tradicionais de *oversampling*. Ao integrar *bootstrap* com a KDE, esta técnica não só equilibra a

distribuição entre as categorias, como também gera observações sintéticas que refletem a complexidade dos dados. A grande diferença está na capacidade de modelar zonas ambíguas, onde as fronteiras entre as categorias são naturalmente difusas. O resultado é um modelo com maior capacidade preditiva, que aprende transições graduais em vez de divisões abruptas.

3.4.2 SMOTENC (Synthetic Minority Over-sampling Technique-Nominal Continuous)

A técnica SMOTE-NC foi desenvolvida em 2002 por Chawla, Bowyer, Hall e Kegelmeyer para superar uma limitação do SMOTE tradicional: a incapacidade de processar variáveis categóricas em conjuntos de dados (como tipo de veículo, ou estado da via). Enquanto o SMOTE tradicional é eficaz na geração de observações sintéticas para variáveis contínuas por meio de interpolação linear, ele falha ao lidar com variáveis discretas, podendo gerar novos valores de forma inapropriada, que resulta em dados inválidos (exemplo: "0,5" entre "chuva" e "nevoeiro"). O SMOTE-NC resolve essa lacuna com três adaptações metodológicas propostas.

Considerando que existem k_1 variáveis contínuas e k_2 variáveis nominais, com $k_1 + k_2 = k$, o algoritmo SMOTE-NC envolve os seguintes passos:

1) **<u>Cálculo da mediana dos desvios padrões</u>**

Para cada variável contínua X_j $(j = 1, \dots, k_1)$ na categoria minoritária, calcula-se o desvio padrão (s_j). De seguida, calcula-se a mediana de todos esses desvios, $\tilde{x}_s = mediana(s_1, \dots, s_{k_1})$, que será posteriormente usada como referência para penalizar a diferença nas variáveis nominais. A mediana é escolhida pela sua robustez a *outliers*. Diferentemente da média, a mediana não é distorcida por valores extremos.

2) **<u>Cálculo da distância euclidiana modificada e do vizinho mais próximo</u>**

Esta etapa visa quantificar a desigualdade entre as amostras da categoria minoritária, integrando variáveis contínuas e categóricas numa única métrica de distância adaptada. Deste modo, a distância euclidiana modificada entre a amostra X (referência) e amostra Z_m (vizinha) é dada por:

$$d(X, Z_m) = \sqrt{\sum_{j=1}^{k_1} (X_j - Z_{m_j})^2 + \sum_{j=1}^{k_2} I_{X_j \neq Z_{m_j}} \tilde{x}_s},$$

onde:

- X_j e Z_{m_j} são os valores da j-ésima variável contínua nas amostras X e Z_m.
- $I_{X_j \neq Z_{m_j}}$ é uma função binária que assume o valor:

  - 1 se a categoria da j-ésima variável nominal difere entre as amostras X e Z_m;

  - 0 se as categorias são idênticas nas duas amostras X e Z_m.
- $\tilde{x}_s$ é a mediana dos desvios padrão das variáveis contínuas, calculada previamente, garantindo que a penalização por diferenças categóricas seja proporcional à variabilidade natural dos dados numéricos. De notar que a penalização $\tilde{x}_s$ é incorporada no cálculo da distância euclidiana modificada tantas vezes quantas as variáveis nominais cujas categorias diferem entre X e Z_m. Além disso, a incorporação de $\tilde{x}_s$ garante que diferenças categóricas sejam ponderadas de forma equivalente a uma diferença de $\tilde{x}_s$ unidades nas variáveis contínuas.
- O termo $\sum_{j=1}^{k_1} (X_j - Z_{mj})^2$ corresponde à distância euclidiana clássica entre as variáveis contínuas, ponderando diferenças maiores quadraticamente.
- O termo $\sum_{j=1}^{k_2} I_{X_j \neq Z_{mj}} \tilde{x}_s$ adiciona uma penalização fixa ($\tilde{x}_s$) para cada variável categórica em que X e Z_m divergem. Essa penalização reflete a variabilidade médias das variáveis contínuas.

  A função indicadora (I) transforma diferenças categóricas em valores numéricos binários (0 ou 1), permitindo que sejam integradas à métrica de distância. Cada diferença categórica adiciona $\tilde{x}_s$ à distância total.

Após calcular as distâncias para todas as observações Z_m da categoria minoritária, os K vizinhos mais próximos são selecionados com base nas menores distâncias euclidianas modificadas, conforme proposto por Chawla et al. (2002). Esses vizinhos são utilizados na etapa seguinte para gerar observações sintéticas, preservando a coerência semântica das categorias.

A escolha do número de vizinhos (K) tem um impacto direto na qualidade das observações sintéticas geradas. Este parâmetro, definido *a priori* pelo utilizador, deve equilibrar dois riscos:

- Valores baixos de K (ex.: K = 1)

As observações sintéticas tornam-se quase réplicas da observação original, o que pode ser problemático se essa observação contiver ruídos ou *outliers*. Por exemplo, se Z_m for um erro de medição (como um registo incorreto de velocidade), a observação sintética reproduzirá esse erro, que resultará em dados artificias pouco diversificados e potencialmente enviesados.

- Valores altos de K (ex.: K > 15)

As observações geradas são mais genéricas, pois combinam informações de múltiplos vizinhos. O risco aqui é perder detalhes importantes da categoria minoritária. Por exemplo, num conjunto de sinistros graves, um K muito elevado pode misturar padrões distintos. Apesar das amostras serem mais diversificas, as mesmas vão ser menos específicas.

3) Geração da amostra sintética

Esta etapa visa criar observações sintéticas para a categoria minoritária, combinando informações da amostra de referência X e dos seus K vizinhos mais próximos. O processo é dividido em duas partes, conforme o tipo de variável:

- Geração de variáveis contínuas

  Para cada variável contínua X_j $(j = 1, \dots, k_1)$, a nova observação sintética é feita por meio de uma interpolação linear estocástica entre o valor da amostra de referência X_j e o valor do vizinho selecionado $Z_{knn,j}$:

$$X_{syn,j} = X_j + \gamma\left(Z_{knn,j} - X_j\right),$$

  onde γ é um número aleatório uniformemente distribuído no intervalo [0, 1].

  O objetivo é introduzir diversidade nas observações sintéticas, evitando sobreposição excessiva com as observações originais; e preservar a distribuição estatística das variáveis contínuas da categoria minoritária.

- Geração de variáveis nominais:

  Para cada variável nominal, $X_j (j = 1, \dots, k_2)$, o valor da nova amostra sintética é definido como a moda entre os K vizinhos mais próximos da amostra de referência X:

$$X_{syn,j} = mode(Z_{1j}, Z_{2j}, \dots, Z_{Kj}).$$

Em resumo, conforme detalhado na Tabela 4, o SMOTE-NC surge como uma evolução crucial no campo da reamostragem para dados desequilibrados, superando as limitações do SMOTE tradicional ao integrar estratégias adaptativas para conjuntos de dados mistos (contínuos e categóricos). Ao incorporar uma distância euclidiana modificada – que pondera diferenças categóricas com base na variabilidade das variáveis contínuas – e ao definir valores nominais sintéticos via moda dos vizinhos, a técnica preserva a coerência semântica dos dados, evitando a geração de categorias inválidas ou irrealistas.

A eficácia desta técnica é respaldada por aplicações recentes em domínios críticos, como saúde, finanças, onde a heterogeneidade de variáveis é comum.

Na Tabela 4, apresentam-se as principais características das técnicas SMOTE, SMOTE-NC e ROSE.

Tabela 4 – Comparação das técnicas de reamostragem para dados desequilibrados.

Critério	SMOTE	SMOTENC	ROSE
Tipos de Variáveis Suportadas	Apenas variáveis contínuas.	Variáveis contínuas e categóricas.	Variáveis contínuas e categóricas.
Geração de Amostras Sintéticas para Variáveis Categóricas	Ignora variáveis categóricas ou gera valores inválidos.	Utiliza a moda (valor mais frequente) dos vizinhos, preservando categorias válidas.	Observações sintéticas baseadas na distribuição original.
Métrica de Distância	Distância Euclidiana padrão (só variáveis contínuas).	Distância Euclidiana modificada incorporando penalizações para diferenças categóricas.	Aplica estimativas de densidade de *kernel*.
Preservação Semântica	Não preserva integridade de categorias.	Mantém a coerência semântica, evitando categorias intermediárias ou inválidas.	Preserva relações contextuais e combinações plausíveis.

Critério	SMOTE	SMOTENC	ROSE
Tratamento de Dados Mistos	Ineficaz em datasets com variáveis mistas.	Integra variáveis contínuas e categóricas de forma equilibrada.	Lida naturalmente com dados mistos, mantendo coerência
Robustez a Outliers	Sensível a Outliers em variáveis contínuas.	Usa a mediana dos desvios padrão, mais robusta a *outliers*.	Robusto através da estimativas de densidade.
Aplicações Típicas	Dados puramente numéricos.	Dados heterogéneos.	Dados heterogéneos.

4. Metodologia de Modelação Preditiva

4.1 Preparação dos Dados e Desequilíbrio

Na fase inicial desta dissertação, as análises exploratórias focaram-se no comportamento dos algoritmos de classificação quando confrontados com um acentuado desequilíbrio entre categorias. O objetivo primordial passava por compreender de que forma diferentes abordagens de reamostragem poderiam mitigar essa desproporção e, consequentemente, melhorar a capacidade dos modelos em identificar casos raros.

A variável resposta de interesse é a ocorrência de um sinistro com vítimas graves ou mortos, sendo a categoria negativa ocorrer um sinistro com feridos leves. Trata-se um problema de eventos raros, situação que tende a enviesar os classificadores para a categoria maioritária (FL) reduzindo a sensibilidade dos modelos (Chawla et al., 2002).

A fim de mitigar este desequilíbrio, aplicaram-se duas técnicas de reamostragem:

- ROSE: que gera observações sintéticas via *bootstrap* com suavização de *kernel*;
- SMOTENC: uma extensão SMOTE clássico, que lida com conjuntos de dados de natureza mista, criando observações sintéticas por interpolação e combinação das variáveis categóricas.

Ambas as técnicas foram testadas sob três estratégias representativas:

- *Oversampling* total, em que as categorias ficam aproximadamente equilibradas;
- *Oversampling* parcial, em que o desequilíbrio entre as categorias é atenuado, mas não eliminado;
- Combinação de *undersampling* da categoria maioritária com *oversampling* da categoria minoritária, resultando em categorias equilibradas.

Numa fase inicial do trabalho, a reamostragem foi aplicada antes da divisão dos dados em conjuntos de treino e teste. Esta prática, comum em estudos exploratórios, permitia trabalhar com um conjunto de dados equilibrado, proporcionando maior estabilidade durante o treino. Contudo, verificou-se posteriormente que esta estratégia poderia ser problemática à luz de desenvolvimentos metodológicos mais recentes (Demircioglu, 2024). Estes estudos demonstraram que aplicar a reamostragem antes da separação treino/teste pode introduzir enviesamentos significativos nos resultados devido ao fenómeno de *data leakage*, ou fuga de informação. Este problema ocorre quando o

modelo tem acesso, direta ou indiretamente, a informações do conjunto de teste durante o processo de treino, comprometendo a validade da avaliação final.

No caso específico do *oversampling*, o *leakage* surge porque as observações sintéticas são criadas tendo em conta todas as observações da base de dados antes da divisão treino/teste. Parte da estrutura estatística do conjunto de teste - incluindo distribuições, relações entre variáveis e fronteiras de decisão - acaba por ser parcialmente incorporada no treino. Mesmo que o modelo nunca "veja" explicitamente as observações de teste, ele é treinado sobre padrões artificiais que derivam desses mesmos dados. Como consequência, o desempenho medido pode parecer artificialmente superior ao verdadeiro, uma vez que o modelo é avaliado sobre informações cuja estrutura já conhece.

4.2 Divisão Temporal e Validação Cruzada

Reconhecendo este risco metodológico, e face aos novos desenvolvimentos metodológicos descobertos após uma fase avançada da dissertação, a estratégia de modelação foi integralmente revista, garantindo uma separação rigorosa entre treino e teste e eliminando qualquer potencial partilha de informação.

Para evitar enviesamentos temporais e simular um cenário de aplicação real, a divisão dos dados respeitou a cronologia: treino = 2016-2022 e teste = 2023. Assim, o modelo aprende no passado e é avaliado no futuro, evitando *look-ahead bias* (Hyndman & Athanasopoulos, 2021). Todos os pré-processamentos (transformação em variáveis *dummy*, normalizações) foram ajustados apenas no treino e posteriormente aplicados ao teste, prevenindo *data leakage* (Kuhn & Johnson, 2013). Dessa forma, a nova abordagem segue princípios consolidados de *machine learning* e predição temporal, assegurando validade estatística, consistência temporal e comparabilidade entre modelos. Para além de corrigir o *data leakage*, esta revisão metodológica procurou aproximar o processo de treino e validação às verdadeiras condições de previsão. Em contextos temporais como o da sinistralidade rodoviária, onde os padrões mudam com o tempo e novas condições surgem anualmente, é essencial que o modelo aprenda apenas com o passado e seja testado sobre o futuro.

Dessa forma, a nova sequência metodológica passou a incluir um conjunto estruturado

de etapas, concebidas para maximizar a imparcialidade e a robustez do processo de modelação:

- **Divisão temporal:** os dados de 2016-2022 foram usados para treino, e o ano 2023 foi reservado para teste independente, garantindo que o modelo é avaliado sobre um período totalmente não visto;
- **Reamostragem apenas no treino:** o desequilíbrio (≈ 2-3 % de sinistros com feridos graves ou mortos) foi corrigido dentro do treino, preservando a distribuição natural do teste;
- **Validação cruzada estratificada (5×2):** dentro do treino, cada *fold* manteve a proporção da categoria rara, assegurando estabilidade estatística na comparação entre modelos (Kuhn, 2008). A estratificação é particularmente recomendada em cenários de elevada desproporção entre categorias, garantindo que cada *fold* contém uma representação mínima da categoria positiva;
- **Reamostragem dentro dos *folds*:** o método ROSE foi aplicado em cada sub-treino da validação cruzada, permitindo que o conjunto de validação permanecesse intacto – uma prática essencial para evitar qualquer fuga de informação interna;
- ***Threshold* de decisão:** após a validação cruzada, o ponto de corte ótimo (máx. F_2-*score*) foi determinado a partir das predições *out-of-fold* (OOF), proporcionando uma calibração baseada em evidência empírica e não apenas heurística;
- **Avaliação final:** todas as métricas foram calculadas sobre o teste 2023, com intervalos de confiança (IC 95 %) obtidos por *bootstrap* estratificado e análise de calibração (interceto e declive), permitindo quantificar a incerteza associada às estimativas e avaliar o grau de sobreajuste.

A seguir, foram consideradas três estratégias distintas de tratamento do desequilíbrio, avaliadas de forma comparável.

4.2.1 Estratégia A - ROSE (fora da validação)

Nesta abordagem, o ROSE foi aplicado uma única vez ao conjunto de treino completo, antes da validação cruzada. Este procedimento permite criar um conjunto de treino equilibrado, combinando *oversampling* da categoria minoritária e *undersampling* da categoria maioritária. Serve como configuração de referência original, permitindo avaliar o risco de *data leakage*, já que as observações sintéticas podem incorporar padrões presentes em toda a base de treino.

4.2.2 Estratégia B - SMOTENC (fora da validação)

O SMOTE-NC foi utilizado como alternativa ao ROSE, também fora da validação cruzada, para bases de dados mistas (numéricas + categóricas). O algoritmo cria observações sintéticas da categoria minoritária interpolando variáveis contínuas e combinando variáveis categóricas por vizinhança. Esta implementação foi utilizada através do pacote UBL, permitindo comparar diretamente com o ROSE e avaliar o efeito de diferentes técnicas de reamostragem aplicadas de forma global ao conjunto de treino.

4.2.3 Estratégia C - ROSE (dentro de cada *fold*)

Para eliminar qualquer risco de *data leakage*, o ROSE foi aplicado apenas dentro de cada *fold* da validação cruzada, ou seja, sobre o subconjunto de treino interno de cada interação. Dada a raridade da categoria positiva, utilizou-se o método ROSE, que gera observações sintéticas por *smoothed bootstrap*, suavizando fronteiras de decisão e melhorando o ajuste em contextos desequilibrados (Lunardon, Menardi, & Torelli, 2014). A geração de amostras foi aplicada apenas dentro de cada *fold* da validação cruzada, evitando contaminação entre treino e validação. Este cuidado assegura que as métricas *out-of-fold* (OOF) são imparciais.

Esta abordagem assegura que o conjunto de validação permaneça intacto, permitindo uma avaliação mais confiável da generalização do modelo. Comparar este método com o ROSE global permite quantificar o impacto de uma segmentação temporal correta sobre métricas como AUC, F_1-*score* e sensibilidade.

4.2.4 Estratégia D - SMOTENC (dentro de cada *fold*)

Nesta abordagem, o SMOTENC foi aplicado apenas dentro de cada *fold* da validação cruzada, sobre o subconjunto de treino interno de cada interação. Este procedimento elimina o risco de *data leakage*, garantindo que as observações sintéticas sejam geradas exclusivamente a partir dos dados de treino de cada *fold*. O algoritmo interpolou variáveis numéricas e combinou variáveis categóricas por vizinhança, permitindo um equilíbrio local e realista.

Esta comparação permitiu não apenas configurar os efeitos do *leakage* sobre o desempenho, mas também quantificar o ganho obtido com a aplicação consistente da reamostragem dentro da validação cruzada, reforçando a credibilidade dos resultados. Por fim, esta secção preserva parte dos resultados obtidos na fase inicial, não como evidência de desempenho, mas como testemunho da evolução metodológica do trabalho. Esses resultados servem para ilustrar de que forma a reamostragem incorreta e a ausência de calibração do *threshold* podem afetar significativamente as estimativas de métricas como AUC, F_1-*score* e precisão, conduzindo a interpretações excessivamente otimistas do desempenho do modelo.

4.3 Modelos e Avaliação

Para avaliar o desempenho preditivo da deteção de sinistros graves, foram ajustados seis modelos de classificação supervisionada, representando métodos lineares, baseados em árvores de decisão e probabilísticos:

- **Regressão Logística (GLM):** modelo linear clássico que oferece elevada interpretabilidade e coeficientes que permitem compreender o efeito de cada variável nas previsões (Hosmer, Lemeshow, & Sturdivant, 2013);
- **Regressão Logística Penalizada de Firth:** abordagem desenvolvida para corrigir o *viés* que pode surgir em amostras pequenas ou com eventos raros, mostrando-se particularmente adequada à deteção da categoria minoritária (Firth, 1993; Heinze & Schemper, 2002);

- **Random Forest:** método baseado na combinação (*ensemble*) de múltiplas árvores de decisão, robusto a variáveis correlacionadas, sendo capaz de capturar relação não lineares complexas (Breiman, 2001);
- **Extreme Gradient Boosting (XGBoost):** algoritmo de *boosting* altamente eficiente, otimizado para grandes volumes de dados e com capacidade para modelos padrões complexos (Chen & Guestrin, 2016);
- **Naïve Bayes:** modelo probabilístico simples, frequentemente utilizado como *baseline* pela sua rapidez e facilidade de interpretação, servindo de referência para comparar o desempenho com métodos mais sofisticados (John & Langley, 1995);
- **C5.0 Decision Tree: versão avançadas das** árvores de decisão tradicionais, oferecendo interpretabialdade e métricas de importância de variáveis úteis para compreender o processo de decisão do modelo (Quinlan, 1993).

Todos os modelos foram avaliados considerando diferentes dimensões do desempenho, com foco especial na categoria minoritária:

- **AUC-ROC**: discriminação global entre as categorias;
- **AUC-PR**: avalia o desempenho em categorias raras, sensível a desequilíbrio, sendo mais adequada que a AUC-ROC em situações de eventos raros por medir diretamente o compromisso entre sensibilidade e precisão (Saito & Rehmsmeier, 2015);
- $\boldsymbol{F_1}$ ***e*** $\boldsymbol{F_2}$***-score***: compromisso entre precisão e sensibilidade, com F_2-*score* priorizando a sensibilidade para detetar sinistros graves mesmo que à custa de alguma perda de precisão (Davis & Goadrich, 2006);
- **Brier score**: calibração probabilística, medindo a proximidade entre probabilidades previstas e observadas (Brier, 1950);
- **Matriz de confusão**: interpretação operacional, permitindo analisar falsos positivos e negativos.

A evolução na estratégia de definição do *threshold* reflete um alinhamento metodológico mais rigoroso com os objetivos da investigação. Inicialmente, o *threshold* foi escolhido de forma empírica, procurando equilibrar a sensibilidade e especificidade.

Embora intuitiva, essa abordagem genérica não otimizava o modelo para a principal prioridade: detetar os sinistros graves.

Na versão final, o *threshold* passou a ser determinado automaticamente pela maximização do F_2-*score*, métrica que atribui maior peso à sensibilidade, valorizando a capacidade de o modelo identificar corretamente os casos positivos. Hand e Christen (2018), destacam que essa otimização foi realizada exclusivamente com os dados de treino, através da validação cruzada, e o valor obtido foi posteriormente mantido fixo para avaliar o desempenho no conjunto de teste (ano de 2023). Dessa forma, assegurou-se uma medição imparcial e realista do desempenho do modelo em dados completamente novos.

A escolha desta métrica reflete também uma decisão consciente sobre o custo relativo dos erros: num contexto de segurança rodoviária, um falso negativo (não identificar um sinistro grave) tem consequências potencialmente mais sérias do que um falso positivo (assinalar incorretamente um caso como grave). Assim, esta calibração permitiu privilegiar a deteção de sinistros graves, mesmo que isso implique aceitar um aumento controlado do número de falsos negativos.

Para garantir a robustez das métricas com contexto de categorias desequilibradas, foram calculados intervalos de confiança de 95% por *bootstrap* estratificado com 1000 repetições (Efron & Tibshirani, 1993), o que permite avaliar a estabilidade dos resultados e a sua variabilidade estatística. Além disso, foi analisada a calibração probabilística do modelo através do *intercept* e do declive da regressão de calibração, indicadores que permitem verificar se as probabilidades previstas estão bem ajustadas à realidade observada (Van Calster et al., 2019). Esta análise ajuda também a quantificar a incerteza das previsões e a detetar eventuais sinais de sobreajustamento, garantindo uma avaliação mais fiável do comportamento dos modelos em diferentes cenários.

4.4 Pesos das categorias (e diferenças face a SMOTENC/ROSE)

A presente abordagem visa eliminar totalmente a geração de observações sintéticas, compensando o desequilíbrio entre categorias através da ponderação das observações na função de perda. Em vez de “criar” novas observações artificiais, altera-se o custo

atribuído aos erros de classificação, penalizando de forma mais intensa os erros cometidos na categoria minoritária — neste caso, M/FG. Este princípio segue a lógica das abordagens de *cost-sensitive learning*, amplamente reconhecidas na literatura como alternativas robustas ao *oversampling* ou *undersampling* (He & Garcia, 2009; Fernández et al., 2018).

Definição dos Pesos

Os pesos são definidos de modo que cada categoria contribua igualmente para o risco esperado do modelo, garantindo equilíbrio sem inflacionar o tamanho efetivo da amostra.

Seja n_+o número de observações positivas (M/FG) e n_- o número de observações negativas (FL) no conjunto de treino. O peso atribuído a cada observação i é:

$$w_i = \begin{cases} 0.5/n_+, & \text{se } y_i = 1 \text{ (Mortes/Feridos Graves)} \\ 0.5/n_-, & \text{se } y_i = 0 \text{ (Feridos Leves)} \end{cases}$$

Desta forma, a soma total dos pesos por categoria é igual a 0,5, forçando ambas as categorias a contribuírem de forma simétrica para o risco empírico. Tal estrutura estabiliza o processo de estimação e evita a variância inflacionada típica do *oversampling* "*hard*" (King & Zeng, 2001; Branco, Torgo, & Ribeiro, 2016).

Onde aplicar os pesos (e "só dentro") ?

Os pesos são calculados exclusivamente com base no conjunto de treino de cada *fold* durante a validação cruzada (v = 5, r = 2), assegurando ausência total de *data leakage*. Durante a fase de treino, o modelo é ajustado com ponderação das observações. No conjunto de validação de cada *fold* e no teste final (2023), os pesos não são utilizados - os modelos são apenas aplicados (*scored*) sem qualquer reponderação. No ajuste final, baseado em todos os dados de treino (2016-2022), os pesos são novamente calculados sobre esse período e aplicados apenas ao ajuste; o conjunto de teste (2023) é avaliado de forma neutra, preservando a independência temporal.

Modelos e incorporação dos pesos

A integração dos pesos depende da estrutura de cada algoritmo:

- **Regressão Logística (GLM)** - *weights = w_tr* altera a verosimilhança ponderada, equivalendo a replicar frações de casos da categoria minoritária sem aumentar o tamanho aparente da amostra (King & Zeng, 2001).
- **Regressão Logística de Firth** - *weights = w_tr* combina a correção de *viés* para eventos raros com ponderação por categoria, mostrando robustez em cenários de separação quase-completa (Heinze & Schemper, 2002).
- **Random Forest (*ranger*)** - *case.weights = w_tr* altera o critério de divisão e o processo de *bagging*, permitindo que cada árvore reflita a importância relativa das categorias sem ajustar manualmente probabilidades (Wright & Ziegler, 2017).
- **XGBoost** - *weights = w_tr* é transmitido diretamente ao *booster*, permitindo uma forma mais granular do parâmetro global *scale_pos_weight*, adaptada à distribuição efetiva do treino (Chen & Guestrin, 2016).
- **C5.0** - aceita *weights = w_tr* de forma nativa, ajustando as estimativas de entropia em função das ponderações.
- **Naive Bayes** - não suporta pesos diretos na implementação do *caret*; neste caso, foram fixadas probabilidades a *priori* iguais (*prior* = c(0.5, 0.5)), garantindo neutralidade no desequilíbrio inicial.

Seleção do *threshold*

Para manter a comparabilidade entre cenários e assegurar uma avaliação imparcial, seguiu-se o mesmo protocolo de decisão já estabelecido:

- Validação OOF (com pesos):

  O *threshold* foi escolhido para maximizar o F_2-$score$, que privilegia a sensibilidade. Para evitar *thresholds* extremos, aplicaram-se restrições leves, como uma taxa mínima de precisão e um número mínimo de positivos previstos no conjunto OOF.

- Teste (2023):

  No conjunto de teste, o *threshold* foi definido usando o *percentile matching*, ou seja, pelo quantil da distribuição de *scores* que produz uma taxa prevista positiva próxima de 3% (análise principal) e 5% (análise de sensibilidade).

Caso a distribuição de *scores* seja quase uniforme, aplicam-se *fallbacks* hierárquicos:
(i) usar o quantil direto $1 - rate$ no teste;
(ii) se necessário, adotar o *threshold* $F_2\text{-}score$ obtido na validação OOF.

Métricas e intervalos de confiança

Para lidar com o desequilíbrio entre as categorias, foram consideradas duas abordagens complementares:

- a ponderação de categorias (Pesos), que ajusta a função de perda sem alterar a estrutura original dos dados originais;
- as técnicas de reamostragem sintética (ROSE e SMOTENC), que geram novas observações artificiais para reforçar a categoria minoritária.

As principais métricas utilizadas para avaliação foram:

- PR-AUC (mais informativa em contextos de categorias raras, Saito & Rehmsmeier, 2015);
- ROC-AUC, precisão, sensibilidade, $F_1\text{-}score$, *G-mean*, *accuracy* e *Brier score*.

Os intervalos de confiança a 95% são obtidos por *bootstrap* estratificado no teste (B = 1000), com correção automática da direção das probabilidades sempre que o ROC-AUC < 0,5, substituindo p por $1 - p$.

A Tabela 5 sintetiza as principais diferenças entre as duas abordagens utilizadas para lidar com o desequilíbrio das categorias. Esta comparação permite avaliar as vantagens e limitações de cada abordagem, auxiliando na escolha da estratégia mais apropriada para diferentes cenários e objetivos de análise.

Tabela 5 - Ponderação de categorias versus técnicas de reamostregm

Dimensão	PESOS	SMOTENC/ ROSE
Dados utilizados	Apenas dados reais; modifica a função de perda.	Cria observações artificiais (interpolações no SMOTE-NC; amostragem kernel em ROSE).

Dimensão	PESOS	SMOTENC/ ROSE
Risco de *leakage*	Nulo se calculado dentro de cada *fold*; não gera novas linhas.	Elevado se aplicado fora dos *folds* ou antes da separação temporal (corrigido nesta investigação).
Variância e calibração	Menor variância e melhor calibração, sobretudo em GLM/Firth.	Maior variância; pode distorcer a fronteira de decisão e exigir calibração adicional.
Modelos mais adequados	GLM, Firth, RF, XGB e C5.0 integram pesos nativamente.	Útil em modelos sensíveis ao balanço, mas suscetível a *overfitting* local.
Natureza da correção	Ajuste de custo (reponderação).	Reamostragem (alteração da distribuição empírica).

Em síntese, a ponderação de categorias constitui uma abordagem mais conservadora e estatisticamente coerente para lidar com desequilíbrios severos, mantendo a integridade amostral e a interpretabilidade dos coeficientes (He & Garcia, 2009; Branco et al., 2016). Embora técnicas sintéticas como SMOTENC e ROSE possam aumentar a sensibilidade, fazem-no frequentemente à custa da calibração e da precisão, sendo menos adequadas quando se pretende comunicação transparente de probabilidades ou quando a integridade temporal da amostra é crítica.

4.5 Calibração isotónica das probabilidades

O balanceamento por pesos 0,5/0,5 altera a função de perda e, com isso, o *baseline* das probabilidades previstas. Em modelos de árvores/*ensembles* - e mesmo em GLM sob forte desequilíbrio - é comum obter *scores* mal calibrados (sub- ou sobre-confiança). Por isso, após treinar cada modelo com pesos, calibrámos as probabilidades por regressão isotónica, um método não paramétrico que aprende uma transformação monótona das *scores* para aproximá-las a probabilidades bem calibradas (Zadrozny & Elkan, 2002; Niculescu-Mizil & Caruana, 2005). Ao contrário do *Platt scaling* (*logit*), a isotónica não impõe forma funcional, acomodando relações não lineares entre *score* e probabilidade (Platt, 1999; Kull, Silva Filho, & Flach, 2017).

Protocolo sem *data leakage*

Para evitar *leakage*, o calibrador é aprendido apenas com previsões OOF (*out-of-fold*) do período de treino:

- Geração OOF (v = 5, r = 2, com pesos):

  Em cada *fold*, ajusta-se o modelo no treino do *fold* com pesos e prevê-se a probabilidade no *validation* do mesmo *fold*. Agregando todos os *folds*, obtêm-se pares (p_i, y_i) sem contaminação (Zadrozny & Elkan, 2002).

- Correção de direção (robustificação):

  Se ROC-AUC $< 0{,}5$ nas OOF, inverte-se a direção dos *scores* ($p \leftarrow 1 - p$), garantindo monotonia entre *score* e probabilidade.

- Ajuste isotónico (PAV):

  Ajusta-se g: $[0,1] \rightarrow [0,1]$que minimiza $\sum_i (y_i - g(p_i))^2$, sob a restrição de monotonia não decrescente. O algoritmo *Pool-Adjacent-Violators* (PAV) produz uma função em degraus $\hat{g}$ (Zadrozny & Elkan, 2002).

  Se os *scores* OOF tiverem variância quase nula (modelo degenerado), define-se $\hat{g}(p) \equiv \hat{\pi}$(prevalência OOF); todas as saídas são truncadas a $[10^{-6}, 1 - 10^{-6}]$.

- Aplicação no teste (2023):

  As probabilidades no teste, p_{test}, são corrigidas na mesma direção e transformadas por $\hat{p}_{\text{cal}} = \hat{g}(p_{\text{test}})$. A calibragem é independente da escolha de *threshold* e anterior à análise de *trade-off* (PR/ROC, F_2, etc.).

Avaliação da calibração

- Brier score (Brier, 1950) é uma métrica que combina resolução e calibração (valores menores são melhores):

$$\text{Brier} \;=\; \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{p}_{\text{cal},i})^2.$$

- Interceto (*calibration-in-the-large*) e declive (*calibration slope*):

  Ajusta-se, no teste, a regressão onde $\alpha \approx 0$ e $\beta \approx 1$ indicam calibração ideal.

$$\text{logit}(y) \;=\; \alpha + \beta\, \text{logit}(\hat{p}_{\text{cal}}),$$

  Onde:

  - $\beta < 1$: *overconfidence* (probabilidades demasiado extremas);

- $\beta > 1$: *underconfidence* (probabilidades "encolhidas").

O *intercetp* também pode ser interpretado como correção de base-rate: $\alpha \approx \text{logit}(\bar{y}) - \text{logit}(\hat{p}_{\text{cal}})$.

- Gráficos de confiabilidade (opcional): curvas observada vs. prevista por *bins* de $\hat{p}_{\text{cal}}$ auxiliam inspeção visual da calibragem; foram utilizados essencialmente para verificação qualitativa, não para decisão.
- Incerteza: métricas no teste (incluindo Brier) têm IC95% por *bootstrap* estratificado, refletindo a variabilidade amostral sem supor normalidade assintótica.

Porquê isotónica (e não apenas *Platt sclaing*)?

A calibração isotónica não assume forma funcional entre *score* e probabilidade, sendo preferível quando o mapeamento é não-linear (situação comum com árvores/*boosting* e com reponderação por pesos).

O risco de sobreajuste da isotónica é mitigado por:

(i) treino OOF (não usa o teste);
(ii) restrição de monotonia;
(iii) truncagem em $[10^{-6}, 1 - 10^{-6}]$.

Platt (*logistic scaling*) é mais parcimoniosa e por vezes suficiente para modelos quase lineares (e.g., GLM); contudo, pode subajustar padrões sistemáticos de má calibragem quando a relação verdadeira não é logito-linear (Niculescu-Mizil & Caruana, 2005; Kull et al., 2017).

Modelos sem calibração isotónica (logísticos e Firth)

Nesta secção do estudo, optou-se deliberadamente por não aplicar calibração isotónica aos modelos logísticos ou de Firth, tanto na versão base como na versão com interações e pesos de categoria. Esta decisão fundamenta-se em razões metodológicas e conceptuais claras:

- **Modelos probabilísticos por construção.**

  Tanto a regressão logística como o modelo de Firth são modelos paramétricos probabilísticos, em que a ligação *logit* garante que a saída $\hat{p} = \text{logit}^{-1}(X\beta)$já corresponde a uma estimativa da probabilidade condicional $P(Y = 1 \mid X)$. Diferentemente de algoritmos não paramétricos (e.g., *Random Forest*, XGBoost), estes modelos produzem previsões naturalmente calibradas, salvo situações extremas de separação quase completa (King & Zeng, 2001).

- **Ausência de amostras sintéticas.**

  Ao contrário dos cenários com SMOTENC ou ROSE, em que a geração de observações artificiais altera a distribuição empírica das categorias e pode distorcer as probabilidades previstas, o presente *pipeline* com pesos de categorias mantém integralmente os dados reais. Os pesos ajustam apenas a função de perda (penalizando mais fortemente os erros na categoria minoritária), sem inflacionar o número efetivo de observações nem modificar a base de cálculo probabilística.

- **Correção de *viés* em eventos raros (modelo de Firth).**

  O estimador de Firth (penalização de Jeffreys) reduz o *viés* de máxima verosimilhança em amostras pequenas ou altamente desequilibradas, melhorando simultaneamente a estabilidade dos coeficientes e a calibração intrínseca das probabilidades (Heinze & Schemper, 2002; Puhr et al., 2017).

- **Invariância das métricas ao escalonamento monotónico.**

  As métricas utilizadas (PR-AUC, ROC-AUC, F_1*-score, G-mean*) dependem apenas da ordenação das probabilidades, sendo invariantes a transformações monotónicas, ou seja, uma calibração isotónica não alteraria os resultados substantivos, apenas a escala das probabilidades.

Assim, a exclusão da calibração isotónica garante maior comparabilidade entre os modelos logísticos e de Firth, concentrando a análise na contribuição das interações e dos pesos de categorias para o poder discriminativo e equilíbrio entre sensibilidade e precisão.

Interação com pesos de categoria

- A ponderação 0,5/0,5 altera a verosimilhança durante o treino (custo por categoria), o que pode deslocar as probabilidades previstas da *base rate* observada no teste.
- A calibragem isotónica reancora as probabilidades num mapeamento orientado por dados sem violar a ordenação (monotonia). Isso é crucial quando as decisões operacionais dependem de *thresholds* por taxa prevista positiva (≈3%/≈5%): a calibragem melhora o Brier e a coerência probabilística, mantendo o PR-AUC (baseado na ordenação) essencialmente inalterado.

Salvaguardas e *edge cases*

- *Scores* quase constantes: usar calibrador constante $\hat{g}(p) \equiv \hat{\pi}$ evita instabilidade; documenta-se o caso e considera-se retirar o modelo do *ensemble* operacional.
- Inversão de direção: verificação sistemática (AUC OOF) evita calibrar *scores* "ao contrário".
- Extrapolação: como $\hat{g}$ é função em degraus definida em $[0,1]$, não há extrapolação; usa-se *clipping* para extremos numéricos.

4.6 Interações em modelos lineares e de Firth

A introdução de termos de interação pretende capturar efeitos de moderação (isto é, quando o efeito de uma variável depende do nível de outra). Em teoria, isso pode melhorar a discriminação quando a relação $X \rightarrow Y$ é verdadeiramente não aditiva (Harrell, 2015; Hastie, Tibshirani, & Friedman, 2009). Contudo, em dados observacionais, raros e desequilibrados, existem várias razões pelas quais as interações podem não se traduzir em ganhos de predição em teste temporal:

***Viés*–variância e complexidade excessiva**

Cada interação aumenta a dimensão do espaço de parâmetros (via produtos, sobretudo com *dummies* para categorias), elevando a variância do estimador e o risco de sobreajustamento a padrões locais de 2016-2022 que não se replicam em 2023 (Babyak, 2004; Kuhn & Johnson, 2013; Hastie et al., 2009). Mesmo com Firth (que reduz o *viés* em

separação/quase separação), a variância preditiva pode crescer e anular ganhos aparentes de treino/OOF.

Sinais práticos:

- ganhos de métrica em OOF que desaparecem ou invertem no teste temporal;
- grande sensibilidade do resultado a pequenas alterações de definição das interações.

Esparsidade e separação local

Combinações raras (p. ex., certos níveis de *tipovia2* × *concelho2* × *HaVeicMoto*) geram células com baixas contagens. Em logística clássica, isso favorece quase-separação, coeficientes instáveis e previsões degeneradas. O estimador de Firth ajuda, mas pode "congelar" efeitos extremos em regiões pouco suportadas, penalizando a generalização (Heinze & Schemper, 2002; King & Zeng, 2001).

Sinais práticos:

- avisos de separação, coeficientes muito grandes, *scores* muito "achatadas" ou quase binárias em subgrupos.

Deriva temporal (*dataset shift*)

Interações capturam padrões contextuais (infraestrutura × composição do tráfego × condições), vulneráveis a mudanças entre anos: obras, *enforcement*, clima atípico, *mix* de frota, etc. O que é "verdade" em 2016.2022 pode mudar em 2023 - logo, as interações perdem valor preditivo fora de amostra (Quiñonero-Candela, Sugiyama, Schwaighofer, & Lawrence, 2009).

Sinais práticos:

- melhorias em CV aleatória que não se mantêm no *holdout* temporal.

Multicolinearidade e pseudo-replica de não linearidades

Interações entre contínuas/categóricas podem imitar não linearidades que seriam melhor modeladas por *splines* (*restricted cubic splines*, *tensor-product smooths*) em vez

de simples produtos. O resultado são coeficientes instáveis e ganhos ilusórios em treino, sem benefício em teste (Harrell, 2015; Wood, 2017).

Boa prática:

- preferir funções de suavização para contínuas (e, quando necessário, interações via *tensor-product splines*) antes de proliferar produtos de *dummies*.

Interação + desequilíbrio + *thresholding*

Em categorias raras, o desenho de interações pode alterar a distribuição das *scores* da minoria, tornando o *threshold* operacional mais sensível. Se o limite não for harmonizado (via *rate-matching* em teste), falsas perdas (precisão/sensibilidade=0) podem surgir por *threshold* demasiado conservador, não por falta de sinal (Kuhn & Johnson, 2013).

Boa prática:

- corrigir a direção das probabilidades no teste e usar percentil por *rate* (3% e 5%), como adotado.

Diagnósticos recomendados (e que justificam decisões)

- Suporte mínimo por célula de interação. Quantificar contagens por combinação; *lump* de níveis raros quando necessário;
- Estabilidade temporal. Comparar efeitos e métricas por subperíodos ou com CV "*rolling-origin*" (quando viável);
- Robustez a *threshold*. Fixar *rate* no teste (3%/5%) e verificar se alterações pequenas do quantil mudam drasticamente precisão/sensibilidade;
- Alternativas suavizadas. Testar *splines* para contínuas e, se necessário, interações suaves (*te()*) em GAMs (Wood, 2017), ou *ridge* (glmnet) para estabilizar coeficientes (Friedman, Hastie, & Tibshirani, 2010);
- Hierarquia de modelação. Cumprir o princípio hierárquico (não incluir interação sem os termos principais) e pré-especificar um conjunto pequeno e plausível de interações (Harrell, 2015).

Considerações sobre a aplicação das interações

No contexto deste estudo, resultados inferiores em modelos com interações não implicam que "interações não existam"; significam que, dado o *split* temporal e a amostra disponível, os custos associados à complexidade/variância podem superar os ganhos de redução de *viés*.

Após a aplicação de correção da direção das *scores* e *rate-matching* no conjunto de teste, as interações com métricas nulas foram eliminadas, e os modelos com interações demonstraram tendência à recuperação, particularmente no cenário com *rate* $\approx 5\%$. Apesar disso, os ganhos permanecem modestos com o facto de que a base aditiva mais os pesos já capta grande parte do sinal preditivo.

Para fins operacionais, recomenda-se priorizar modelos estáveis, como Firth ponderado, e incorporar apenas interações que:

- Possuam suporte estatístico suficiente;
- Persistam no tempo;
- Melhorem PR-AUC/sensibilidade sem degradar Brier de forma relevante.

4.7 Discussão crítica das escolhas metodológicas

As opções metodológicas adotadas refletem o esforço em equilibrar rigor estatístico, relevância prática e limitações inerentes ao problema dos eventos raros.

A escolha da PR-AUC como métrica principal justifica-se pelo forte desequilíbrio entre categorias. Em problemas de eventos raros, a ROC-AUC tende a subestimar o desempenho, pois atribui igual peso às categorias positiva e negativa, sendo pouco sensível ao número desproporcionadamente elevado de negativos. Já a PR-AUC concentra-se na relação entre sensibilidade e precisão, oferecendo uma avaliação mais informativa da capacidade do modelo em identificar corretamente os casos positivos (Saito & Rehmsmeier, 2015). Em contextos como a deteção de sinistros graves, a PR-AUC fornece uma métrica mais realista e discriminativa do que a ROC-AUC.

Adicionalmente, a utilização do F_2*-score* para a seleção do ponto de corte reforça essa prioridade metodológica. Enquanto métricas simétricas como o F_1*-score* tratam igualmente precisão e sensibilidade, o F_2*-score* dá maior peso à sensibilidade, refletindo a preocupação em minimizar falsos negativos mesmo à custa de um aumento nos falsos

positivos (Davis & Goadrich, 2006). Esta decisão traduz a prioridade prática em não falhar a deteção de sinistros graves, alinhando-se ao princípio da precaução em saúde pública e segurança rodoviária.

Do ponto de vista do pré-processamento e modelação, seguiu-se uma sequência estruturada que combina rigor estatístico e técnicas de *machine learning*. Os dados foram divididos temporalmente, com 2016-2022 para treino e 2023 para teste, garantindo uma avaliação adequada correta. Para lidar com o desequilíbrio extremo, aplicou-se reamostragem *intra-fold* via ROSE e SMOTENC, preservando a integridade dos *folds* de validação cruzada e evitando estimativas excessivamente otimistas (Lunardon, Menardi, & Torelli, 2014).

Em termos de modelos treinados, optou-se por uma abordagem híbrida, incorporando:

- métodos estatísticos tradicionais, como regressão logística clássica (GLM) e regressão penalizada de Firth;
- algoritmos de machine learning, incluindo Naive Bayes, Random Forest, C5.0 e XGBoost.

Esta diversidade permitiu comparar o desempenho de abordagens paramétricas e não paramétricas, fornecendo *insights* sobre robustez e interpretabilidade. O processo metodológico adotado encontra-se no Apêndice 3.

A avaliação de desempenho foi cuidadosamente delineada para eventos raros, utilizando métricas de sensibilidade, precisão, F_1/F_2-*scores*, PR-AUC, ROC-AUC e *Brier Score*. A definição do *threshold* de decisão priorizou a maximização do F_2-*score*, complementada por análise de sensibilidade considerando diferentes taxas previstas positivas (≈3% e 5%). Posteriormente, a calibração das probabilidades foi realizada via regressão isotónica, curvas de calibração e validação por *bootstrap*, reforçando a confiança na interpretação das predições.

Essa abordagem metodológica estruturada evidencia que cada etapa - divisão temporal, pré-processamento, reamostragem, treino, avaliação e calibração - foi cuidadosamente projetada para maximizar a robustez, reduzir *vieses* e produzir modelos confiáveis para a identificação de eventos raros.

4. Análise dos Dados

A base de dados analisada neste estudo reflete os registos de sinistros rodoviários ocorridos no distrito de Setúbal entre os anos de 2016 e 2023, fornecidos pela GNR de Setúbal. Estes dados foram complementados com informações adicionais provenientes de outras fontes relevantes como:

- Instituto Português do Mar e da Atmosfera (IPMA), que disponibilizou dados meteorológicos, tais como as condições climáticas no momento dos sinistros (chuva, nevoeiro, etc.).
- Infraestruturas de Portugal (IP), que contribuiu com informações sobre as caraterísticas físicas e operacionais das vias, incluindo o tipo de pavimento, sinalização, condições de manutenção, entre outros aspetos que podem afetar a segurança rodoviária.

Inicialmente, a base de dados continha 53649 observações, que englobam tanto "Feridos Leves" como "Mortes/Feridos Graves" e 1198 variáveis. Posteriormente, foi decidido excluir o período da pandemia, compreendido entre 11 de abril de 2020 até 20 de abril de 2021. Esse período foi marcado por medidas governamentais rigorosas de prevenção à COVID-19, como confinamentos obrigatórios, limitações de deslocações, restrições de horários, entre outras. Essas medidas tiveram um impacto significativo no volume de tráfego nas estradas, resultando numa redução substancial no número de veículos em circulação. Esta redução, por sua vez, influenciou diretamente a frequência e a natureza dos acidentes registados durante esse intervalo de tempo.

Também foram excluídos os concelhos que não pertenciam ao distrito de Setúbal, nomeadamente, Amadora, Lisboa, Loures, Sintra e Vila Franca de Xira.

Ao final da exclusão do período da pandemia e dos concelhos que não pertencem ao distrito de Setúbal, a base de dados foi consolidada em 47731 observações. Para este estudo, o objetivo principal é modelar e prever a gravidade de um sinistro, que foi tratada como a variável resposta. Esta é uma variável de natureza qualitativa nominal, com duas categorias: "Feridos Leves" e "Mortes/Feridos Graves". Neste sentido, as variáveis independentes analisadas, selecionadas com base na sua relevância para a previsão da gravidade do sinistro, encontram-se sintetizadas e descritas no Anexo 1. Esta decisão visa evitar distorções nos dados devido às alterações significativas no

comportamento do tráfego e nas condições rodoviárias. Isso assegura a que os resultados obtidos reflitam de maneira precisa e equitativa as verdadeiras tendências e os fatores associados à sinistralidade rodoviária no distrito de Setúbal.

5.1 Modelo Estatístico de Regressão Logística Binomial

A análise de dados por meio de um modelo de regressão logística foi a abordagem utilizada para compreender a relação entre a variável dependente binária e as variáveis independentes.

Seguiu-se a metodologia descrita por Hosmer-Lemeshow (Hosmer et al., 2013), para ajustar o modelo regressão logística.

5.1.1 Seleção das Variáveis Independentes (Análise Univariada)

Devido ao grande número de variáveis disponíveis (1198 no total), foi necessário priorizar e selecionar apenas aquelas consideradas mais relevantes para a análise da regressão. Numa primeira fase, variáveis com uma taxa de valores omissos muito elevada foram automaticamente excluídas, por representarem um risco para a robustez dos modelos, podendo introduzir enviesamento e reduzir o poder estatístico da análise. Após esta triagem, aplicou-se o teste da razão de verosimilhanças com um nível de significância de 5%, de modo a identificar as variáveis que têm uma relação estatisticamente significativa com a variável resposta. A identificação completa das variáveis significativas resultantes desta análise univariada encontram-se no Anexo 2.

5.1.2 Modelo Múltiplo Preliminar e Exclusão de Variáveis

Inicialmente, foi criado um modelo onde foram incluídas apenas as variáveis que se revelaram significativas na análise univariada. Este modelo inicial, serviu como ponto de partida para a seleção de variáveis que seriam mantidas no modelo final.

Utilizando um nível de significância de 1%, procedeu-se à exclusão progressiva das variáveis que se tornaram não significativas, com base no teste de razão de verosimilhanças. Priorizou-se a exclusão de variáveis com um elevado número de categorias, independentemente do valor de *p-value* associado, por uma questão de

parcimónia. A lista final das variáveis selecionadas, assim como a sua classificação, encontram-se apresentadas no Anexo 3.

5.1.3 Agrupamento de Categorias

Com o objetivo de reduzir a complexidade do modelo e assegurar a sua estabilidade estatística, procedeu-se ao agrupamento de categorias em algumas variáveis explicativas. Esta etapa tem como objetivo evitar problemas de sobreajuste associados a categorias com baixa frequência, aumentar a parcimónia do modelo e, simultaneamente, preservar a capacidade explicativa.

O processo de agrupamento baseou-se na significância estatística das categorias individuais, garantindo que as categorias com comportamentos semelhantes fossem consideradas em conjunto. A seguir, apresentam-se as variáveis sujeitas a este processo, bem como os respetivos agrupamentos definidos.

a. Variável Concelho ("concelho"): a categoria de referência foi definida como "ALCACER DO SAL". Foram realizados os seguintes agrupamentos:
   i. As categorias "ALCOCHETE", "GRANDOLA", "SEIXAL", "SINES" e "PALMELA" foram agrupadas sob a nova categoria "AGSSP".

As categorias "SANTIAGO DO CACEM" e "SETUBAL" foram agrupadas sob a nova categoria "SS".

b. Variável Tipo de Via ("tipovia"): a categoria de referência foi "A-Auto Estrada". Os agrupamentos foram:
   i. As categorias não significativas ("Arruamento", "EF – Estrada Florestal", "IP – Itinerário Principal", "Outra Via", "PNT – Ponte" e "VAR – Variante") foram agrupadas com a categoria de referência sob a nova categoria "AE/A/EF/IP/O/P/V".
   ii. As categorias significativas "EM – Estrada Nacional", "IC – Itinerário Complementar" e "ER – Estrada Regional" foram agrupadas sob a nova categoria "EN/IC/ER".
c. Variável Percentagem de condutores masculinos envolvidas no acidente (categorizada) ("PercCondMCat"): A categoria de referência foi "[0,25)". As categorias "[25,50)" e "[50,75)" foram consideradas não significativas e,

portanto, agrupadas com a categoria de referência sob a nova categoria "Perc25- 75".

d. Variável Hora do Acidente ("horaacid"): A categoria de referência foi "0". O agrupamento foi o seguinte:

   i. A categoria "6" foi agrupada sob a nova categoria "6h".

   ii. As categorias "8", "9", "10", "11", "12" e "13" foram agrupadas sob a nova categoria "8h-13h".

   iii. A categoria "7" foi agrupada às categorias "14", "15", "16", "17", "18", "19", "20", "21", "22", "23", "0", "1", "2", "3", "4" e "5", formando a nova categoria "14h-5h", dado que os coeficientes estimados revelaram-se próximos e o teste da razão de verosimilhança não evidenciou diferenças estatisticamente significativas entre estas categorias.

5.1.4 Verificação da Linearidade

Após a aplicação do método GAM (Modelo de Regressão Aditiva Generalizada), a análise da linearidade entre a variável índice de gravidade e o *logit*, revelou que o comportamento não era linear, conforme ilustrado na Figura 1.

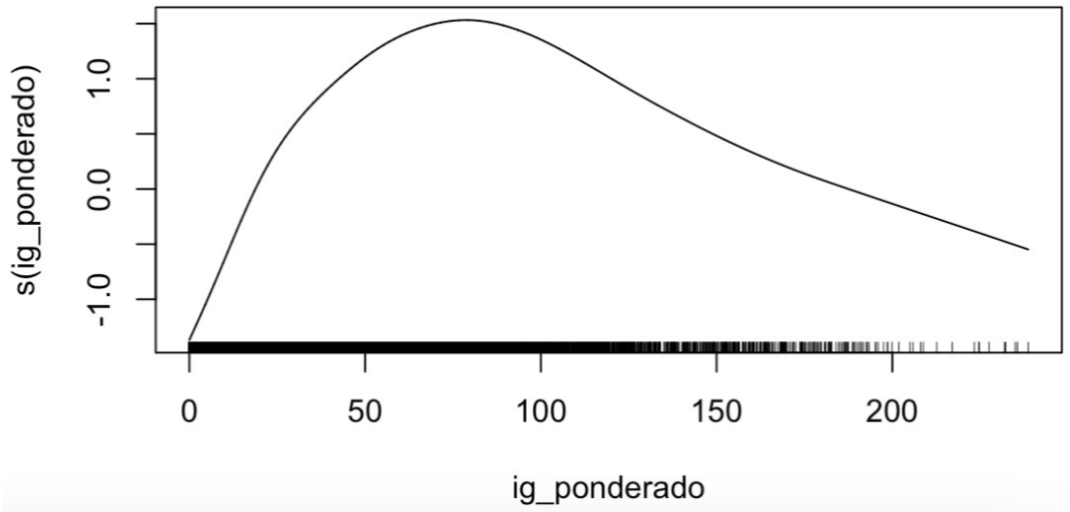

Figura 1 – Representação da função spline (s) resultante da aplicação de um GAM para verificação do pressuposto de linearidade para a variável índice de gravidade (ig_ponderado).

A comparação do AIC (Critério de Informação Akaike) entre os dois modelos indica que o GAM apresenta um AIC menor (7197,020) em comparação com o modelo final

(7539,934). Como o AIC penaliza modelos mais complexos, valores menores indicam um melhor equilíbrio entre ajustamento e parcimónia. Ambos os modelos possuem o mesmo número de graus de liberdade (19), o que sugere que a diferença no AIC não se deve à complexidade, mas sim à capacidade de o modelo explicar a variabilidade dos dados. Neste sentido, o GAM é preferível para inferência e previsão, de acordo com o princípio da parcimónia, uma vez que oferece um melhor ajustamento sem necessidade de aumentar a complexidade.

5.1.5 Incorporação de Interações

No processo de modelação, a incorporação de interações entre variáveis presentes no modelo ajuda a compreender melhor como a combinação de diferentes fatores afeta a variável independente. O objetivo é determinar se a inclusão dessas interações melhora significativamente o ajuste do modelo. Para alcançar esse objetivo, ajusta-se uma série de modelos de regressão logística, cada um contendo diferentes interações, e através do teste de razão de verosimilhanças, avalia-se a significância da inclusão da interação relativamente ao modelo sem essa interação.

Neste caso, adotou-se um nível de significância de 1% para avaliar a relevância estatística das interações, assegurando um maior rigor na seleção das interações e reduzindo o risco de incluir aquelas que não apresentem um impacto substancial sobre a variável dependente. Além disso, é fundamental que as interações testadas não apenas apresentem significância estatística, mas também sejam coerentes com o contexto do problema em análise. Dessa forma, garante-se que as adições ao modelo sejam interpretáveis e úteis para a compreensão do fenómeno em estudo.

5.1.6 Verificação da Qualidade do Modelo

Nesta etapa, procedeu-se à análise da qualidade do modelo, conforme se descreve abaixo.

Análise de multicolinearidade

De modo a garantir a robustez e fiabilidade das estimativas do modelo, avaliou-se a existência de multicolinearidade entre as variáveis preditivas. Os resultados para o

modelo final encontram-se apresentados na Tabela 6, onde se identificam as variáveis e as interações que apresentam uma colinearidade elevada.

Tabela 6 - Medidas de multicolinearidade e identificação de colinearidade elevada nas variáveis explicativas.

Variável	GVIF	df	$GVIF^{(1/(2df))}$	Colinearidade Elevada
concelho2	2551,67	3	3,70	X
tipoacid	373,17	2	4,40	X
tipolocal2	12,26	1	3,50	X
tipovia2	8,83	2	1,72	
horaacid1new	1,11	2	1,03	
fuga	1,01	1	1,00	
PercCondMCat2	1,08	1	1,04	
HaVeicPesado	1,32	1	1,15	
HaVeicLig	1,98	1	1,41	
HaVeicMoto	3,06	1	1,75	
HoraLaboral	1,15	1	1,07	
MedianaIdadeVeic	1,06	1	1,03	
ig_ponderado	54,60	1	7,39	X
concelho2*tipoacid	48261,74	6	2,46	
tipoacid*tipolocal2	31,73	2	2,37	
tipovia2*HaVeicMoto	5,11	2	1,50	
ig_ponderado*concelho2	237,36	3	2,49	
ig_podenrado*tipoacid	23,68	2	2,21	
ig_podenrado*tipovia2	8,28	2	1,70	

Embora os valores observados não atinjam os níveis críticos que indicam uma colinearidade severa (valores superiores a 10), a existência de valores elevados ainda aponta para uma possível correlação significativa entre algumas variáveis.

Bondade do Ajustamento

- R^2 de Nagelkerke

O modelo de regressão logística apresentou um Pseudo R^2 de Nagelkerke de 0,2607 o que indica que 26,07% da variabilidade da variável dependente foi explicada pelas variáveis independentes. Embora esse valor possa parecer baixo em comparação com os R^2 da regressão linear, na regressão logística é comum que o Pseudo R^2 tenha valores mais moderados, uma vez que o modelo lida com probabilidades e não com variáveis contínuas.

O valor de 0,2607 sugere que o modelo consegue capturar uma porção significativa da relação entre as variáveis, sendo capaz de distinguir as categorias da variável dependente de forma razoável. Em modelos logísticos, valores acima de 0,2 podem ser considerados aceitáveis, especialmente em contextos onde a variabilidade não explicada pode ser atribuída a fatores não incluídos no modelo.

- Teste de Hosmer e Lemeshow

O teste de Hosmer-Lemeshow forneceu um valor de *p-value* de 0,501, logo não há evidências estatísticas para rejeitar a hipótese nula de que o modelo se ajusta bem aos dados. Portanto, os resultados sugerem que o modelo de regressão logística apresenta um ajuste adequado aos dados.

Capacidade discriminativa

- Curva ROC

A avaliação da capacidade discriminativa do modelo foi realizada através da curva ROC, apresentada na Figura 2. Esta curva constitui uma das ferramentas mais utilizadas para aferir o desempenho de modelos de classificação, uma vez que sintetiza a relação entre verdadeiros positivos e falsos positivos. Ao representar graficamente este equilíbrio, a curva ROC permite avaliar em que medida o modelo consegue distinguir corretamente as categorias de interesse. Quanto mais a curva se afastar da diagonal aleatória, maior será a sua capacidade discriminativa.

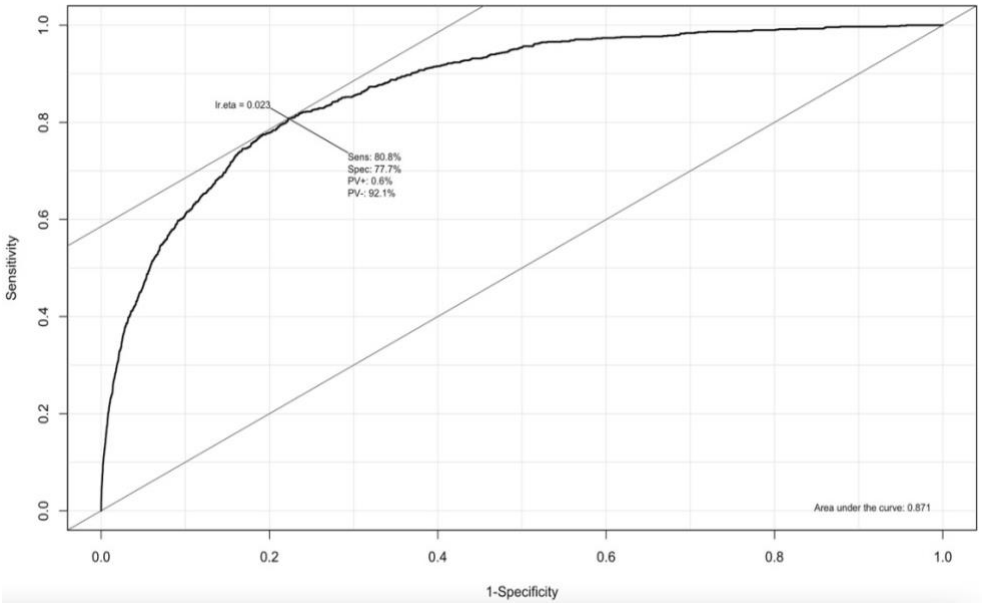

Figura 2 - Curva ROC do modelo de regressão logística final para 43312 observações.

Interpretando as métricas sabemos que:

→ Sensibilidade:

Um valor de sensibilidade de 80,8% indica que o modelo conseguiu identificar corretamente 80,8% dos eventos que ocorreram.

→ Especificidade:

Um valor de 77,7% indica que o modelo foi capaz de reconhecer corretamente 77,7% das situações onde o evento não ocorreu.

→ Área sob a Curva (AUC):

Com um AUC de 0,871, o modelo mostra uma boa capacidade discriminativa, uma vez que valores próximos a 1 refletem um desempenho muito bom. Isso significa que há uma grande probabilidade de o modelo classificar corretamente um caso positivo como positivo e um caso negativo como negativo.

→ Intervalo de confiança para AUC:

O IC de 95% para AUC varia de 0,8599 a 0,8813, indicando que há 95% de confiança de que o valor "real" do AUC está dentro desse intervalo.

Validação do modelo

- *Bootstrap*

Para avaliar a estabilidade e precisão das estimativas do modelo, foram geradas 5000 e, posteriormente, 10000 observações *bootstrap*.

Após a validação do modelo através do procedimento de "*Backwards Step-down*", as variáveis que mantiveram significância estatística e relevância prática foram:

- tipo de sinistro
- horário do sinistro
- presença de veículos pesados
- presença de motociclos
- mediana da idade dos veículos

- Calibração

Partindo para a análise da calibração, as Figuras 3 e 4 apresentam uma curva de calibração que compara a probabilidade prevista pelo modelo com a probabilidade observada no conjunto de dados. O eixo horizontal representa as probabilidades previstas pelo modelo, enquanto o eixo vertical mostra as probabilidades observadas, ou seja, a proporção real de ocorrências de "Mortes/Feridos Graves".

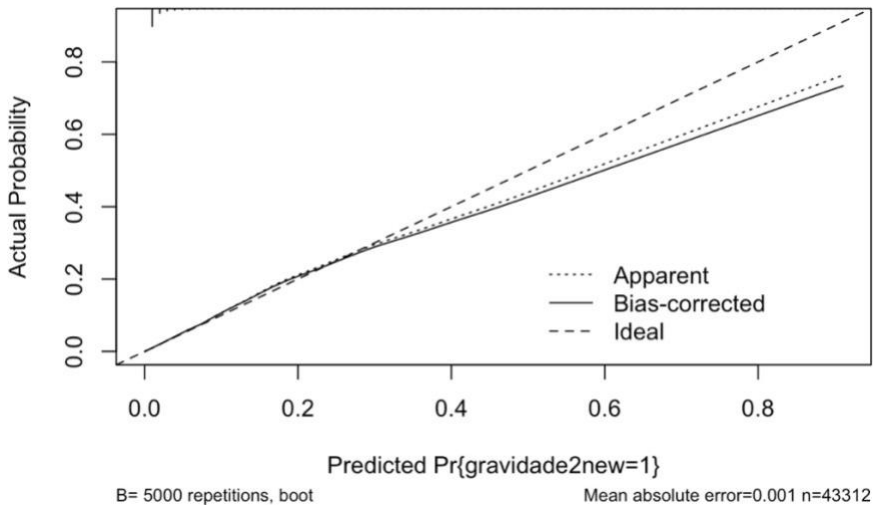

Figura 3 - Calibração para 5000 repetições de bootstrap

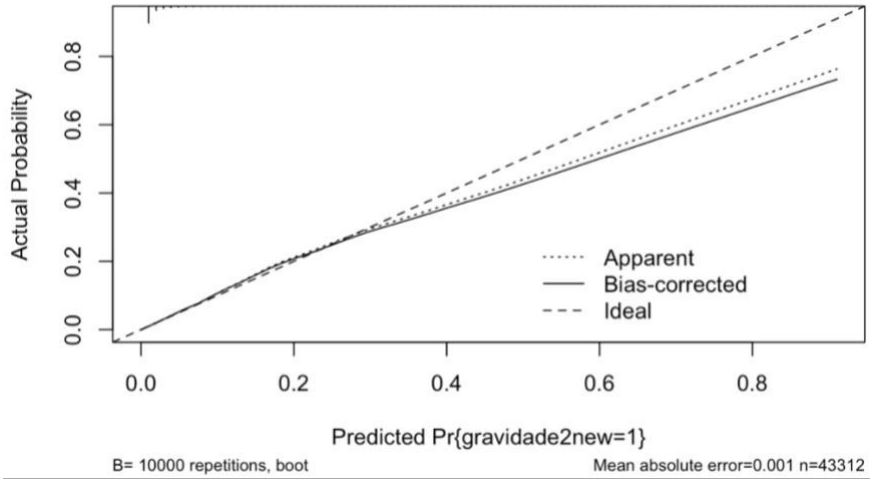

Figura 4 - Calibração para 10000 repetições de bootstrap

Para ambos os casos, os gráficos indicam que, para valores de probabilidade previstos abaixo de 0,3, o modelo tem um desempenho relativamente bom, com a linha aparente e a linha corrigida pelo otimismo bastante próximas da linha ideal. Isso sugere, para esses casos, o modelo está bem calibrado e as suas previsões refletem de forma adequada a realidade observada.

Entretanto, para valores de probabilidade mais altos (acima de 0,3), tanto a linha aparente quanto a corrigida ficam abaixo da linha ideal, indicando uma subestimação das probabilidades reais. Ou seja, o modelo tende a prever probabilidades menores do que as efetivamente observadas.

Neste caso, o modelo apresenta uma boa calibração para previsões de probabilidade baixa, mas demonstra uma leve tendência de subestimação para probabilidades mais elevadas, mesmo após a correção pelo otimismo. O erro absoluto médio de 0,001 e a utilização de 5000 ou 10000 repetições de *bootstrap* indicam que o ajuste é estável e bem fundamentado, considerando-se também o tamanho da amostra (n=43312).

- Validação Cruzada

Os dados foram divididos em conjunto de treino e teste, onde 70% das observações pertencem ao treino e 30% das observações pertencem ao teste. Na Tabela 7, encontram-se os valores referentes a cada subconjunto, particularmente a cada categoria.

Tabela 7 - Divisão dos dados do modelo de regressão logística (43312 observações) em dois subconjuntos: treino e teste e respetivo número de observações por categoria em cada subconjunto.

	Regressão Logística – 42000 Observações	
	Treino	**Teste**
Feridos Leves	29613	12704
Mortes/Feridos Graves	706	289
	30319	12993

5.1.7 Apresentação do modelo final

Na Tabela 8, apresentam-se as variáveis que integram o modelo final, as suas respetivas categorias e a classificação atribuída a cada uma delas.

Tabela 8 - Designação e classificação das variáveis independentes do modelo final de regressão logística para 43312 observações.

Variável	Categorias	Classificação
Concelho	Alcácer do Sal Alcochete, Grândola, Seixal, Sines e Palmela Almada, Barreiro, Moita, Montijo e Sesimbra Santiago do Cacém e Setúbal	Qualitativa Nominal
Tipo de acidente	Atropelamento Colisão Despiste	Qualitativa Nominal
Localização do acidente	Dentro das localidades Fora das localidades	Qualitativa Nominal
Tipo de via	Autoestrada, Arruamento, Estrada Florestal, Itinerário Principal, Outra Via, Ponte e Variante Estrada Municipal Estrada Nacional, Itinerário Complementar e Estrada Regional	Qualitativa Nominal
Hora com minutos a zero do acidente	14 – 17h 6h 8h – 13h	Qualitativa Nominal
Acidente com fuga	Não Sim	Qualitativa Nominal

Variável	Categorias	Classificação
% de condutores masculinos envolvidos no acidente	Perc [25 – 75] [75 – 100]	Qualitativa Nominal
Existência de veículos pesados	Não Sim	Qualitativa Nominal
Existência de veículos ligeiros	Não Sim	Qualitativa Nominal
Existência de veículos ciclomotores e motociclos	Não Sim	Qualitativa Nominal
Acidente ocorreu no horário Laboral	Não Sim	Qualitativa Nominal
Mediana da idade da matrícula dos veículos	Não Sim	Quantitativa Numérica
Índice de gravidade	Não Sim	Quantitativa Numérica

Nota: o modelo representado na Tabela 8 foi o modelo aplicado em todas as abordagens desenvolvidas no estudo.

Na Tabela 9 é apresentado o modelo final ajustado, no qual se encontram as variáveis incluídas com as categorias correspondentes e os respetivos coeficientes estimados.

Tabela 9 - Modelo logístico múltiplo para a existência de Mortes/Feridos graves nos sinistros com vítimas (p-value do teste de Wald).

Variável	Categorias	Coeficiente	Std. Error	P-value
Concelho	**ALCÁCER DO SAL**			
	AGSSP	0,1353	0,5723	<0,001
	ABMMSS	0,5420	0,5509	0,8060
	SS	-0,7007	0,5556	-0,3292
Tipo de Acidente	**Atropelamento**			
	Colisão	-1,4680	0,6070	0,2224
	Despiste	-0,4714	0,5754	<0,001
Localização do Acidente	**Dentro das localidades**			
	Fora das localidades	0,1940	0,2406	0,4086
Tipo de Via	**AE/A/EF/IP/O/P/V**			
	Estrada Municipal	0,1892	0,2489	0,4201
	EN/IC/ER	0,9815	0,1220	0,4455
Hora com minutos a zero do acidente	**14h – 5h**			
	6h	0,6952	0,1889	<0,001
	8h – 13h	-0,3099	0,0818	<0,001
Acidente com fuga	**Não**			
	Sim	-1,4477	0,2673	<0,001
% de condutores masculinos envolvidos no acidente	**Perc 25 – 75**			
	[75, 100]	0,3048	0,0804	<0,001
Existência de veículos pesados	**Não**			
	Sim	1,0485	0,1314	<0,001
Existência de veículos ligeiros	**Não**			
	Sim	0,6088	0,1304	<0,001
Existência de veículos ciclomotores e motociclos	**Não**			
	Sim	2,6520	0,1228	<0,001
Acidente ocorreu no horário laboral	**Não**			
	Sim	-0,4670	0,0738	<0,001
Mediana da idade da matrícula dos veículos		0,0359	0,0049	<0,001
Índice de gravidade		0,0824	0,0075	<0,001
	Concelho2AGSSP* tipoacidColisão	-0,0503	0,5745	0,9302

Variável	Categorias	Coeficiente	Std. Error	P-value
	Concelho2ABMMS* tipoacidColisão	-0,9020	0,5814	0,1208
	Concelho2SS* tipoacidColisão	0,3938	0,6317	0,5330
	Concelho2AGSSP* tipoacidDespiste	-0,3217	0,5691	0,5719
	Concelho2ABMMS* tipoacidDespiste	-1,0773	0,5792	0,0629
	Concelho2SS* tipoacidDespiste	0,0280	0,6282	0,9644
	tipoacidColisão* tipolocal2 Fora das Localidades	0,5704	0,2564	<0,001
	tipoacidDespiste*tipolocal2 Fora das Localidades	0,3353	0,2705	0,2151

As principais métricas de avaliação calculadas da matriz de confusão, encontram-se na Tabela 10 com a respetiva interpretação. Estes resultados correspondem ao modelo final.

Tabela 10 - Resultados das métricas da matriz de confusão do modelo final de regressão logística aplicado a 43312 observações para identificação de Mortes/Feridos Graves

Métrica	Resultado	Observação
Ponte de Corte	0,021	Valor que separa as observações e duas categorias.
Accuracy	0,7657	O modelo classifica corretamente 76,57% das observações.
IC (95%)	(0,7583; 0,773)	Intervalo de Confiança de 95% para a *accuracy*.
Kappa	0,0907	O modelo sugere um desempenho muito baixo.
Mcnemar's Test P-Value	<0,001	Reflete uma diferença não significativa entre as taxas de erro de classificação nas duas categorias.

Métrica	Resultado	Observação
Sensibilidade	0,7716	O modelo identificou corretamente, aproximadamente, 77,16% dos casos Mortes/Feridos Graves.
Especificidade	0,7656	O modelo identificou corretamente, aproximadamente, 76,56% dos casos de Feridos Leves.
Valor Preditivo Positivo	0,0697	Das observações classificadas como positivas pelo modelo, 6,97% são verdadeiras positivas.
Valor Preditivo Negativo	0,9933	Das observações classificadas como negativas pelo modelo, 99,33% são verdadeiras negativas.
F1-score	0,8647	Bom desempenho do modelo.
AUC	0,8538	O modelo tem uma boa capacidade de discriminação.
Precisão	0,0697	Aproximadamente 6,97% observações classificadas como positivas são mesmo positivas.

Em suma, o modelo apresenta um bom desempenho geral, mas mostra limitações na precisão das previsões positivas, conforme evidenciado pelos valores preditivos e pelo Kappa. O elevado valor da AUC e do *F1-score* sugere que o modelo possui uma capacidade relevante de discriminação entre as categorias.

5.2 Resultados com correção temporal e reamostragem *intra-fold*

5.2.1 Resultados Preliminares e Impacto do *Oversampling* Pré-divisão

Durante a fase inicial do trabalho foram realizadas experiências exploratórias com o objetivo de testar diferentes técnicas de reamostragem para lidar com o forte desequilíbrio entre casos de sinistros fatais e não fatais. Testou-se a influência de diferentes graus de desequilíbrio e de técnicas de reamostragem (ROSE e SMOTENC) sobre o desempenho preditivo dos modelos.

Nessas versões preliminares, o *oversampling* foi aplicado antes da divisão dos dados em treino (70%) e teste (30%), procedimento que, apesar de comum em estudos iniciais, induz *data leakage*, contaminando a amostra de teste com observações sintéticas geradas a partir de todo o conjunto de dados.

Apesar de esta prática ser comum em abordagens exploratórias, resulta numa sobrestimação das métricas preditivas, dado que os modelos acabam por "ver" padrões parciais da amostra de teste durante o treino.

Além disso, nessa versão inicial o *threshold* para a classificação binária foi definido sem otimização explícita do F_2-*score*, métrica que, na versão final, foi usada para calibrar o compromisso entre precisão e sensibilidade.

Esta secção preserva parte desses resultados, não como evidência de performance, mas como testemunho da evolução metodológica do trabalho e como demonstração do impacto que a reamostragem incorreta pode ter nas estimativas de AUC, F_1-*score* e precisão.

5.2.2 Resultados Preliminares com ROSE (Pré-Divisão)

As Tabelas A12–A16 da versão anterior da dissertação (disponíveis no Apêndice A) apresentavam os resultados obtidos após a aplicação de *undersampling* e *oversampling* combinados, antes da divisão aleatória 70/30.

A Tabela 11 compara os resultados obtidos com a abordagem inicial — em que o *oversampling* (no caso, o método ROSE) era aplicado a todo o conjunto de dados antes da divisão treino/teste — com os resultados corrigidos, obtidos após aplicar a reamostragem apenas no conjunto de treino e ajustar o ponto de corte pelo F_2-*score*.

Tabela 11 - Impacto do oversampling pré-divisão (exemplo com ROSE)

Modelo	PR-AUC (antes)	PR-AUC (depois)	Δ	ROC-AUC (antes)	ROC-AUC (depois)	Δ	F1 (antes)	F1 (depois)	Δ
GLM	0,41	0,16	-0,25	0,93	0,87	-0,06	0,49	0,08	-0,41
RF	0,44	0,15	-0,29	0,94	0,86	-0,08	0,45	0,13	-0,32
XGB	0,46	0,22	-0,24	0,95	0,88	-0,07	0,49	0,11	-0,38
NB	0,39	0,17	-0,22	0,91	0,87	-0,04	0,45	0,08	-0,37
C5.0	0,40	0,16	-0,24	0,92	0,85	-0,07	0,48	0,14	-0,34

Ambos os conjuntos foram extraídos da mesma base de dados original de sinistros rodoviários do distrito de Setúbal (2016-2023), contendo o mesmo número de observações e variáveis preditoras. A diferença entre "antes" e "depois" não reflete, portanto, qualquer alteração nas fontes de dados, mas exclusivamente a correção metodológica associada à eliminação do *data leakage* e à adoção de um critério de decisão mais apropriado para eventos raros.

Os resultados "antes" exibiam métricas substancialmente mais elevadas, em particular no PR-AUC e no F_2-*score* sugerindo um desempenho artificialmente otimista. Após a correção, observou-se uma redução generalizada das métricas (-0,25 a -0,30 pontos no PR-AUC; -0,30 a -0,40 no F_1-*score*), refletindo uma avaliação mais realista da capacidade de generalização dos modelos.

Apesar desta diminuição, a hierarquia relativa entre modelos manteve-se, com o XGBoost e o Random Forest a exibirem desempenho consistentemente superior ao GLM e ao Naive Bayes, o que confirma a estabilidade estrutural das relações modeladas — apenas as magnitudes das métricas estavam inflacionadas no cenário anterior.

5.2.3 Resultados Preliminares com SMOTENC (Pré-Divisão)

De forma análoga, a aplicação do SMOTENC antes da separação treino/teste gerou métricas igualmente elevadas. As Tabelas B8–B17 (ver Apêndice B) mostravam, em geral, ganhos aparentes de desempenho, com ROC-AUC entre 0,92 a 0,96 e F_1-*score* médios próximos de 0,45 a 0,50.

A Tabela 12 contém o resumo dos resultados representativos obtidos nesta etapa.

Tabela 12 - Resumo dos resultados representativos

Modelo	PR_AUC	ROC_AUC	Precisão	Sensibilidade	*F1-score*	*Accuracy*
GLM	0,40	0,93	0,34	0,96	0,48	0,91
RF	0,43	0,94	0,36	0,96	0,46	0,92
XGB	0,45	0,95	0,38	0,97	0,48	0,93
NB	0,37	0,90	0,31	0,96	0,46	0,90
C5.0	0,38	0,91	0,33	0,94	0,47	0,91

Fonte: resultados exploratórios, fase inicial (Apêncide B)

Tal como no caso anterior, o desempenho mais elevado resulta do contacto indevido entre observações artificiais (geradas por SMOTENC) e observações reais no conjunto de teste.

5.2.4 Discussão Crítica

O conjunto das análises exploratórias oferece valor científico ao demonstrar empiricamente como erros de desenho experimental podem alterar profundamente a perceção de desempenho.

A passagem de AUCs próximas de 0,95 para valores realistas em torno de 0,87-0,88 confirma que a separação temporal e a reamostragem restrita ao treino são condições indispensáveis para avaliação honesta em contextos de eventos raros.

Do ponto de vista metodológico, esta análise comparativa é particularmente relevante:

- evidencia o impacto negativo do *data leakage* na avaliação de modelos de classificação com desequilíbrio extremo;
- demonstra a importância de otimizar o ponto de corte em função do objetivo analítico (neste caso, maximizar a sensibilidade sem degradar em excesso a precisão);
- e reforça a necessidade de uma validação rigorosa e estratificada, garantindo que as métricas refletem o desempenho em dados verdadeiramente não observados.

Em síntese:

- Os resultados pré-divisão não devem ser interpretados como estimativas válidas, mas sim como caso de estudo sobre o impacto do *data leakage*.
- A consistência da hierarquia de desempenho entre modelos reforça a robustez estrutural das conclusões qualitativas.
- As tabelas completas foram preservadas no Apêndice A, garantindo transparência e reprodutibilidade, mas a discussão quantitativa principal deve basear-se exclusivamente nos resultados da secção seguinte.

5.2.5 ROSE fora da validação

Os resultados obtidos com o ROSE aplicado fora da validação estão apresentados na Tabela 13 e na Tabela 14, mostrando métricas de desempenho, matrizes de confusão e indicadores de calibração.

A coluna "Prioridade" indica o critério adotado na avaliação e seleção dos modelos. No contexto de eventos raros, as métricas clássicas de classificação, como *accuracy* ou mesmo o F_1-*score*, tendem a ser pouco informativas, uma vez que o desequilíbrio extremo entre categorias pode mascarar o verdadeiro desempenho do modelo na deteção da categoria minoritária. Assim, a análise deu prioridade às métricas mais sensíveis a este tipo de problema: a área sob a curva Precisão-Sensibilidade (PR-AUC) e a área sob a curva ROC (ROC-AUC), que avaliam, respetivamente, a capacidade de distinguir corretamente os casos graves e de manter baixo o número de falsos positivos. Além disso, o limite de decisão em cada modelo foi ajustado com base no F_2-*score*, uma métrica que atribui maior peso à sensibilidade relativamente à precisão. Esta escolha reflete o objetivo fundamental do estudo, maximizar a identificação de sinistros graves, ainda que à custa de um maior número de falsos alarmes, o que é coerente com uma perspetiva de prevenção e segurança rodoviária.

Assim, a designação "PR-AUC, ROC-AUC; limiar por F_2" sintetiza a estratégia global de avaliação: os modelos foram comparados principalmente pela sua discriminação (PR-AUC e ROC-AUC), sendo o ponto de corte ajustado de modo a otimizar o F_2-*score*.

As colunas "CAL-INTERCEPT" e "CAL-SLOPE" representam os parâmetros clássicos de calibração dos modelos preditivos, avaliando até que ponto as probabilidades estimadas correspondem à frequência real dos eventos observados.

O *calibration intercept* (interceto de calibração) mede o desvio médio entre as probabilidades previstas e as observadas. Um valor próximo de 0 indica ausência de *viés* sistemático; valores negativos sugerem sobrestimação do risco (probabilidades previstas demasiado elevadas), enquanto valores positivos indicam subestimação.

Já o *calibration slope* (declive de calibração) avalia a dispersão das probabilidades previstas. O valor ideal é 1, correspondendo a uma calibração perfeita: valores inferiores a 1 refletem excesso de confiança do modelo (probabilidades extremas demasiado amplificadas), enquanto valores superiores a 1 indicam um modelo demasiado conservador, com previsões comprimidas em torno da média.

A inclusão destas métricas é essencial em contextos de eventos raros, onde a calibração probabilística tem impacto direto na utilidade prática do modelo permitindo, por exemplo, distinguir se uma probabilidade prevista de 10% corresponde efetivamente a um risco real próximo desse valor, aspeto crucial para aplicações em segurança rodoviária e decisão operacional.

Tabela 13 - Métricas no teste (ponto e IC95%) - ROSE fora da validação

Modelo	Prioridade	PR-AUC	ROC-AUC	Precisão	Sensibilidade	*F1-score*	*G-mean*	*Acurracy*	*Brier*	*CAL-INTERCEPT*	*CAL-SLOPE*
GLM	PR-AUC, ROC-AUC; limiar por F2	0.160 [0.123, 0.210]	0.875 [0.851, 0.896]	0.044 [0.037, 0.051]	0.964 [0.934, 0.988]	0.084 [0.072, 0.096]	0.670 [0.656, 0.682]	0.478 [0.466, 0.490]	0.180 [0.174, 0.186]	-4.069 [-4.248, -3.912]	0.887 [0.801, 0.997]
RF	PR-AUC, ROC-AUC; limiar por F2	0.149 [0.113, 0.196]	0.859 [0.834, 0.881]	0.069 [0.059, 0.079]	0.885 [0.831, 0.927]	0.128 [0.111, 0.146]	0.784 [0.759, 0.803]	0.700 [0.689, 0.710]	0.122 [0.116, 0.128]	-3.819 [-4.023, -3.623]	0.548 [0.477, 0.617]
XGB	PR-AUC, ROC-AUC; limiar por F2	0.221 [0.168, 0.293]	0.880 [0.854, 0.907]	0.059 [0.050, 0.069]	0.927 [0.889, 0.967]	0.110 [0.094, 0.128]	0.759 [0.742, 0.777]	0.629 [0.617, 0.641]	0.137 [0.132, 0.143]	-3.829 [-4.022, -3.640]	0.819 [0.733, 0.931]
NB	PR-AUC, ROC-AUC; limiar por F2	0.170 [0.125, 0.226]	0.868 [0.843, 0.890]	0.039 [0.033, 0.045]	0.988 [0.969, 1.000]	0.075 [0.063, 0.086]	0.609 [0.598, 0.619]	0.390 [0.379, 0.402]	0.077 [0.073, 0.082]	-3.106 [-3.323, -2.884]	0.643 [0.582, 0.708]
C5.0	PR-AUC, ROC-AUC; limiar por F2	0.158 [0.118, 0.207]	0.849 [0.820, 0.878]	0.075 [0.063, 0.089]	0.770 [0.706, 0.836]	0.137 [0.116, 0.159]	0.764 [0.732, 0.796]	0.759 [0.749, 0.769]	0.107 [0.102, 0.113]	-3.484 [-3.712, -3.287]	0.440 [0.378, 0.528]

Tabela 14 - Matrizes de confusão e métricas derivadas (Teste 2023) - ROSE fora da validação

Modelo	*Threshold*	TP	FN	FP	TN	Precisão	Sensibilidade	Especificidade	*Accuracy*	*F1-score*
GLM	0,231	159	6	3458	3016	0,044	0,964	0,466	0,478	0,084
RF	0,252	146	19	1974	4500	0,069	0,885	0,695	0,700	0,128
XGB	0,229	153	12	2453	4021	0,059	0,927	0,621	0,629	0,110
NB	0,035	163	2	4046	2428	0,039	0,988	0,375	0,390	0,075
C5.0	0,306	127	38	1563	4911	0,075	0,770	0,759	0,759	0,137

O GLM apresenta uma sensibilidade muito elevada (0,964), porém uma precisão muito baixa (0,044) e *accuracy* limitada (0,478), resultando em muitos falsos positivos (3458). Este padrão decorre diretamente do F_2-*score*, que privilegia a deteção de casos positivos, mas penaliza a seletividade. O Naive Bayes segue um comportamento semelhante: sensibilidade muito elevada (0,998), precisão baixa (0,039) e *accuracy*

reduzida (0,390), indicando que também funciona como um "*screening* sensível" mas com muitos falsos alarmes.

Nos *ensembles*, observa-se melhor equilíbrio entre métricas. O XGBoost apresenta a maior PR-AUC (0,221) e uma ROC-AUC também elevada (0,880), evidenciando boa capacidade de discriminação e priorização correta dos casos positivos – algo essencial em contextos de forte desequilíbrio. O Random Forest, apesar de ligeiramente abaixo em PR-AUC, combina sensibilidade (0,885) e especificidade (0,695) de forma equilibrada, refletido no *G-Mean* mais elevado (0,784), o que reduz o número de falsos positivos por verdadeiros positivos. O C5.0 exibe um comportamento semelhante ao Random Forest, com boa *accuracy* (0,759) e melhor *F1-score* (0,137), o que indica um desempenho estável e mais eficiente na identificação de verdadeiros positivos sem sacrificar demasiado a precisão.

Em termos de calibração, todos os modelos revelam *intercepts* negativos e *slopes* inferiores a 1, indicando que as probabilidades previstas tendem a estar deslocadas e excessivamente extremas. Entre eles, os modelos baseados em árvores (RF, XG e C5.0) exibem menores erros de calibração (Brier entre 0,107 e 0,137) em comparação com o GLM (0,180), sugerindo previsões mais fiáveis e probabilidades mais próximas das verdadeiras ocorrências.

De forma geral, com ROSE fora, os modelos de árvores e *ensembles* combinam melhor discriminação e equilíbrio, enquanto o GLM e o Naive Bayes funcionam como "*screeners"* sensíveis, mas com muitos falsos alarmes.

5.2.6 ROSE dentro da validação

Os resultados obtidos com o ROSE aplicado dentro da validação encontram-se apresentados na Tabela 15 e na Tabela 16. A análise destas tabelas permite comparar diretamente o efeito da reamostragem *intra-fold* sobre precisão, sensibilidade, *accuracy* e calibração em relação ao ROSE fora.

Tabela 15 - Métricas no teste (ponto e IC95%) - ROSE dentro do fold

Modelo	Prioridade	PR-AUC	ROC-AUC	Precisão	Sensibilidade	*F1-score*	*G-mean*	*Acurracy*	*Brier*	*CAL-INTERCEPT*	*CAL-SLOPE*
GLM	PR-AUC, ROC-AUC; limiar por F2	0,161 [0,124; 0,212]	0,873 [0,849; 0,896]	0,115 [0,096; 0,135]	0,721 [0,646; 0,789]	0,198 [0,168; 0,227]	0,787 [0,745; 0,824]	0,855 [0,846; 0,863]	0,192 [0,188; 0,197]	-3,754 [-3,921; -3,612]	1,289 [1,163; 1,436]
RF	PR-AUC, ROC-AUC; limiar por F2	0,195 [0,145; 0,256]	0,877 [0,851; 0,899]	0,137 [0,113; 0,163]	0,594 [0,513; 0,671]	0,223 [0,187; 0,259]	0,733 [0,681; 0,778]	0,897 [0,890; 0,904]	0,048 [0,045; 0,051]	-2,096 [-2,262; -1,941]	0,981 [0,852; 1,117]
XGB	PR-AUC, ROC-AUC; limiar por F2	0,194 [0,146; 0,259]	0,878 [0,852; 0,903]	0,124 [0,102; 0,146]	0,661 [0,588; 0,732]	0,209 [0,175; 0,241]	0,763 [0,721; 0,803]	0,876 [0,868; 0,883]	0,025 [0,023; 0,028]	-1,077 [-1,256; -0,927]	1,038 [0,921; 1,168]
NB	PR-AUC, ROC-AUC; limiar por F2	0,145 [0,109; 0,192]	0,854 [0,828; 0,878]	0,100 [0,084; 0,118]	0,667 [0,595; 0,738]	0,174 [0,147; 0,201]	0,751 [0,710; 0,792]	0,843 [0,834; 0,851]	0,027 [0,024; 0,031]	-0,604 [-0,847; -0,405]	0,608 [0,546; 0,678]
C5.0	PR-AUC, ROC-AUC; limiar por F2	0,161 [0,121; 0.,212]	0,863 [0,837; 0,887]	0,115 [0,094; 0,134]	0,655 [0,581; 0,727]	0,196 [0,164; 0,226]	0,755 [0,712; 0,797]	0,866 [0,858; 0,874]	0,044 [0,041; 0,047]	-1,922 [-2,101; -1,768]	0,654 [0,502; 0,849]

Tabela 16 - Matrizes de confusão e métricas derivadas (Teste 2023) - ROSE dentro do fold

Modelo	*Threshold*	TP	FN	FP	TN	Precisão	Sensibilidade	Especificidade	*Accuracy*	*F1-Score*
GLM	0,638	119	46	919	5555	0,115	0,721	0,858	0,855	0,198
RF	0,374	98	67	617	5857	0,137	0,594	0,905	0,897	0,223
XGB	0,131	109	56	769	5705	0,124	0,661	0,881	0,876	0,209
NB	0,041	110	55	990	5484	0,100	0,667	0,847	0,843	0,174
C5.0	0,273	108	57	830	5644	0,115	0,655	0,872	0,866	0,196

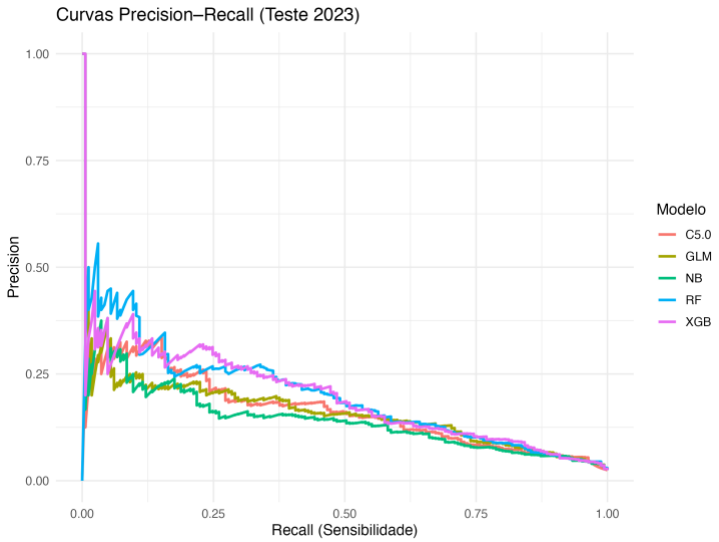

Figura 5 - Curvas Precisão-Sensibilidade (Teste 2023)

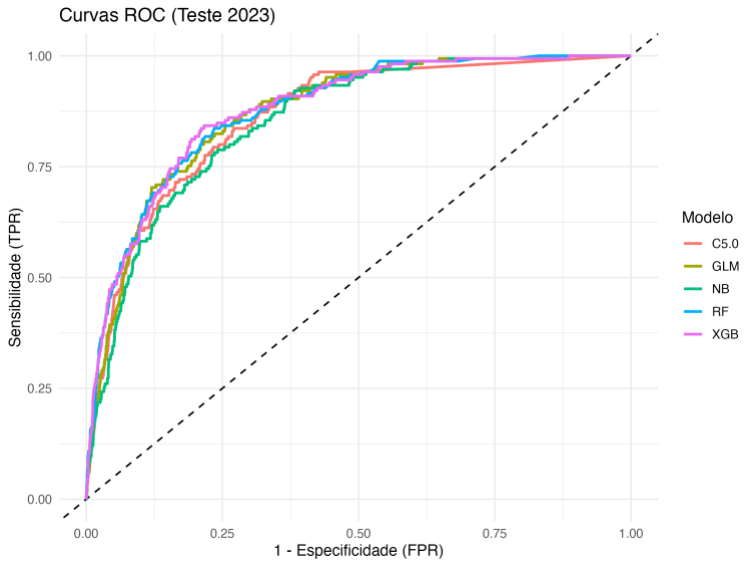

Figura 6 - Curvas ROC (Teste 2023)

Com o ROSE aplicado dentro da validação, os resultados tornam-se mais consistentes e realistas, refletindo melhor o comportamento esperado em dados verdadeiramente não vistos. Há um ganho generalizado em precisão e *accuracy* em praticamente todos os modelos, acompanhando por uma ligeira redução da sensibilidade - uma troca esperada, já que o *oversampling* é agora restrito ao treino e não interfere no teste.

O GLM evidencia essa melhoria de forma clara: a precisão aumenta substancialmente (de 0,044 para 0,115) enquanto a sensibilidade se ajusta para 0,721, resultando num F_1-*score* de 0,198 e *G-mean* de 0,787. Este comportamento indica que o modelo se torna mais seletivo, reduzindo falsos positivos (919 vs. 3458 anteriormente) sem comprometer em demasia a capacidade de detetar positivos. A calibração também melhora (*intercept* = -3,75 ; *slope* = 1,29), com as probabilidades a refletirem mais fielmente o risco observado.

Nos modelos de *ensemble*, observa-se um padrão idêntico, mas com um desempenho global superior. O Random Forest alcança a maior *accuracy* (0,897) e o melhor F_1-*score* (0,223) entre todos, combinando uma boa discriminação (ROC-AUC = 0,877) com excelente calibração (Brier 0,048 e *slope* próximo de 1). O Random Forest mostra ainda a maior especificidade (0,905), o que se traduz em menor número de falsos positivos (617) sem perda excessiva de sensibilidade (0,594), sendo, portanto, um modelo mais equilibrado e robusto em termos operacionais.

O XGBoost mantém uma ROC-AUC igualmente elevada (0,878) e uma PR-AUC (0,194) praticamente idêntica à do Random Forest, mas com ligeiramente mais falsos positivos

(769 vs. 617) e sensibilidade marginalmente superior (0,661). Este perfil indica uma excelente capacidade de ordenação das observações com um leve *viés* a favor da identificação de positivos, o que o torna adequado para contextos de priorização de risco.

O C5.0 apresenta métricas muito próximas das do XGBoost, com F_1-*score* de 0,196, *G-mean* de 0,755 e *accuracy* de 0,866, mostrando novamente que o algoritmo produz classificações equilibradas e estáveis. Tal como o Random Forest, o C5.0 mantém uma boa calibração (Brier = 0,044) e *slope* próximo de 1 (0,654), o que indica probabilidades bem ajustadas à frequência observada.

O Naive Bayes, embora inferior aos *ensembles*, mostra um comportamento mais controlado do que quando o *oversampling* foi aplicado fora dos *fold*: a precisão aumenta para 0,100 e a calibração melhora significativamente (*intercep* = -0,60 ; *slope* = 0,61). Ainda assim, continua a produzir mais falsos positivos (990) e um *F1-score* inferior (0,174), refletindo limitações estruturais do modelo na presença de variáveis correlacionadas.

Em síntese, com o ROSE aplicado corretamente dentro da validação, os resultados tornam-se mais calibrados e operacionais, refletindo estimativas de desempenho mais confiáveis. Observa-se uma melhoria geral em precisão, *accuracy* e calibração, acompanhada por uma redução controlada de sensibilidade – um comportamento esperado, já que o *oversampling* é agora restrito ao treino, evitando sobreajustamento. As métricas de discriminação (ROC-AUC e PR-AUC) mantêm-se elevadas em todos os modelos, o que confirma a sua capacidade consistente de separar corretamente as categorias, mas com valores de Brier muito mais baixos e *slopes* de calibração próximos de 1, indicando previsões probabilísticas melhor ajustadas.

5.2.7 ROSE dentro da validação vs. ROSE fora da validação

A etapa que se segue consistiu em avaliar o impacto da estratégia de equilíbrio ROSE quando aplicada dentro dos ciclos de validação cruzada (*intra-fold*) em comparação com a sua aplicação antes da separação dos dados (*extra-fold*).

O objetivo desta comparação é determinar se o equilíbrio realizado no interior de cada *fold* contribui para uma estimativa mais realista do desempenho e para uma redução do sobreajuste decorrente partilha de informação entre treino e teste.

Para tal, foram calculadas as diferenças percentuais entre as métricas obtidas nas duas abordagens, conforme sintetizado na Tabela 17.

Tabela 17 - Diferenças de métricas (pontos): ROSE Dentro da validação - ROSE fora da validação

Modelo	PR_AUC	ROC_AUC	Precisão	Sensibilidade	*F1-Score*	*G-mean*	*Accuracy*	*Brier*
C5.0	+0,4%	+1,4%	+4,0%	-11,5%	+5,9%	-0,9%	+10,8%	-6,4%
GLM	+0,1%	-0,2%	+7,1%	-24,2%	+11,4%	+11,7%	+37,6%	+1,2%
NB	-2,5%	-1,4%	+6,1%	-32,1%	+9,9%	+14,3%	+45,2%	-5,0%
RF	+4,7%	+1,8%	+6,8%	-29,1%	+9,5%	-5,1%	+19,7%	-7,4%
XGB	-2,7%	-0,2%	+6,5%	-26,7%	+9,9%	+0,4%	+24,7%	-11,2%

A análise relativa confirma que a aplicação do ROSE dentro da validação melhora substancialmente o comportamento geral dos modelos, especialmente em termos de precisão e *accuracy*, com aumentos relativos superiores a +100% em GLM e Naive Bayes, e +98% em Random Forest. Estes ganhos refletem uma maior seletividade e redução de falsos positivos, mostrando que o balanceamento *intra-fold* conduz a fronteiras de decisão mais robustas.

Em contrapartida, observa-se uma redução da sensibilidade em todos os modelos (-15% a -33%), resultado esperado pela maior prudência na deteção de positivos após a correção do *viés* introduzido pelo *oversampling* fora dos *folds*.

Os modelos de *ensemble* (Random Forest, C5.0, XGBoost) destacam-se ainda por melhorias significativas no Brier Score (-60% a -82%), evidenciando melhor calibração e confiabilidade probabilística. O GLM e o Naive Bayes, embora percam um pouco em sensibilidade, ontem os maiores ganhos relativos de *accuracy* e precisão, sugerindo que o balanceamento interno permitiu-lhes generalizar melhor.

Em suma, os resultados mostram que o ROSE dentro da validação conduz a uma melhoria estrutural: os modelos tornam-se mais calibrados, precisos e estáveis, com melhor sobreajuste e previsões mais alinhadas com o desempenho esperado.

5.2.8 Regressão Logística Penalizada de Firth

A Regressão Logística Penalizada de Firth foi aplicada com o objetivo de mitigar problemas de separação e instabilidade dos estimadores de máxima verosimilhança, comuns em cenários de forte desequilíbrio da variável resposta (Heinze&Schemper,

2002). Esta abordagem ajusta a função de verosimilhança, produzindo coeficientes mais estáveis e probabilidades bem calibradas.

O modelo foi avaliado em duas etapas:

1. Validação cruzada: no conjunto de treino (2016-2022), utilizada para a seleção do limite ótimo de decisão segundo a métrica F_2-*score;*
2. Avaliação no conjunto de teste (2023), mantendo esse limite fixo para garantir validade externa.

Resultados da Validação Cruzada (Treino, OOF)

Os resultados da validação cruzada estão resumidos na Tabela 18, que apresenta a AUC média corrigida e o *threshold* ótimo definido para maximizar F_2-*score*.

Tabela 18 - Métrica de desempenho global e threshold ótimo

Métrica	Valor	Observação
AUC media (corrigida)	0,855	Indica boa discriminação entre categoras (quanto mais alto, melhor)
***Threshold* ótimo**	0,052	*Threshold* que otimiza o F_2-*score*

Estes resultados indicam que o modelo é capaz de capturar padrões relevantes e equilibrar adequadamente a sensibilidade e a precisão, otimizando a deteção da categoria minoritária.

Avaliação no teste (2023)

Os resultados obtidos no conjunto de teste encontram-se na Tabela 19, que reúne métricas de desempenho, intervalos de confiança via *bootstrap*, matriz de confusão e observações interpretativas.

Tabela 19 - Avaliação do desempenho do Modelo de Regressão Penalizada de Firth

Métrica	Valor	IC95% (Bootstrap)	Observação
ROC-AUC	0,870	[0,845; 0,895]	Excelente discriminação entre categorias

Métrica	Valor	IC95% (Bootstrap)	Observação
PR-AUC	0,161	[0,845; 0,895]	Desempenho competitivo em M/FG
Accuracy	0,866	[0,845; 0,895]	Elevada proporção de previsões corretas
Precisão	0,117	-	11,7% dos alertas são casos verdadeiros
Sensibilidade	0,673	-	Captura $\approx \frac{2}{3}$ dos casos graves
F1-Score	0,199	-	Compromisso equilibrado entre precisão e sensibilidade
G-mean	0,765	-	Bom equilíbrio entre categoria minoritária e maioritária
Brier	0,022	-	Calibração excelente (probabilidades realistas)

O modelo penalizado de Firth apresentou desempenho discriminativo comparável aos métodos de *machine learning* mais complexos, como Random Forest e XGBoost (ROC-AUC ≈ 0,87), confirmando a eficácia da penalização em contextos de eventos raros. A sensibilidade elevada (0,673), significa que cerca de $\frac{2}{3}$ dos sinistros graves foram corretamente detetados, enquanto a baixa precisão de 0,117 indica a presença de falsos positivos. O Brier (0,022) foi o melhor entre todos os modelos avaliados, sugerindo boa calibração probabilística: as probabilidades previstas refletem bem as frequências observadas.

A Regressão Logística Penalizada de Firth revelou-se uma alternativa estatística sólida para este problema, alcançando resultados semelhantes ou superiores aos modelos de *machine learning* em termos de discriminação e calibração.

5.2.9 SMOTENC Fora da Validação

Os resultados obtidos com o SMOTENC aplicado fora da validação estão apresentados na Tabela 20 (métricas de desempenho) e na Tabela 21 (matriz de confusão e métricas derivadas).

Tabela 20 - Métricas de desempenho dos modelos com SMOTENC aplicado fora da validação.

Modelo	PR-AUC	ROC-AUC	Precisão	Sensibilidade	*F1-Score*	*G-mean*	*Accuracy*	*Brier*	*CAL-INTERCEPT*	*CAL-SLOPE*
C5.0	0,087 [0,067; 0,115]	0,777 [0,740; 0,811]	0,113 [0,078; 0,148]	0,224 [0,161; 0,287]	0,150 [0,105; 0,192]	0,463 [0,392; 0,524]	0,937 [0,931; 0,942]	0,036 [0,033; 0,039]	-1,407 [-1,599; -1,212]	0,325 [0,258; 0,404]

Modelo	PR-AUC	ROC-AUC	Precisão	Sensibilidade	*F1-Score*	*G-mean*	*Accuracy*	*Brier*	*CAL-INTERCEPT*	*CAL-SLOPE*
GLM	0,095 [0,070; 0,129]	0,750 [0,712; 0,789]	0,044 [0,036; 0,052]	0,727 [0,654; 0,796]	0,082 [0,069; 0,097]	0,657 [0,624; 0,689]	0,598 [0,586; 0,609]	0,110 [0,105; 0,114]	-3,110 [-3,284; -2,933]	0,487 [0,393; 0,595]
NB	0,157 [0,110; 0,212]	0,784 [0,746; 0,817]	0,045 [0,037; 0,052]	0,812 [0,750; 0,873]	0,085 [0,071; 0,098]	0,672 [0,647; 0,698]	0,563 [0,552; 0,575]	0,348 [0,339; 0,356]	-5,861 [-6,045; -5,679]	0,500 [0,417; 0,590]
RF	0,083 [0,065; 0,106]	0,807 [0,779; 0,836]	0,115 [0,084; 0,151]	0,255 [0,191; 0,317]	0,159 [0,118; 0,202]	0,492 [0,426; 0,549]	0,933 [0,927; 0,939]	0,043 [0,039; 0,046]	-1,692 [-1,932; -1,462]	0,380 [0,332; 0,439]
XGB	0,104 [0,073; 0,146]	0,779 [0,742; 0,809]	0,096 [0,073; 0,121]	0,376 [0,297; 0,448]	0,153 [0,118; 0,189]	0,585 [0,520; 0,640]	0,897 [0,889; 0,904]	0,048 [0,045; 0,052]	-1,938 [-2,136; -1,754]	0,478 [0,402; 0,547]

Tabela 21 - Matrizes de confusão e métricas derivadas dos modelos com SMOTENC aplicado fora da validação.

Modelo	*Threshold*	TP	FN	FP	TN	Precisão	Sensibilidade	Especificidade	*Accuracy*	*F1-Score*
GLM	0,163	120	45	2627	3847	0,044	0,727	0,594	0,598	0,082
RF	0,396	42	123	322	6152	0,115	0,255	0,95	0,933	0,159
XGB	0,329	62	103	583	5891	0,096	0,376	0,91	0,897	0,153
NB	0,548	134	31	2869	3605	0,045	0,812	0,557	0,563	0,085
C5.0	0,354	37	128	291	6183	0,113	0,224	0,955	0,937	0,15

Com o SMOTENC aplicado fora dos ciclos de validação, observaram-se desempenhos modestos de forma geral, com pequenas variações entre os modelos. Nenhum algoritmo apresentou ganhos expressivos face às estratégias anteriores, indicando que o equilíbrio externo ao processo de validação tende a introduzir menor generalização e potencial sobreajuste ao conjunto de treino.

O Ranfom Forest foi o modelo com maior capacidade discriminativa global (ROC-AUC = 0,807 [0,779;0,836]), mostrando equilíbrio razoável entre a sensibilidade (0,255) e precisão (0,115), embora com tendência a subestimar a probabilidade de casos positivos (*intercept* = -1,692 ; declive = 0,380).

O Naive Bayes apresentou o maior PR-AUC (0,157 [0,110 ; 0,212]) e uma sensibilidade elevada (0,812) capturando a maioria dos casos graves. Contudo, a precisão manteve-se muito baixa (0,045) e a calibração revelou-se fortemente enviesada (*intercept* = -5,861 ; declive = 0,500).

O XGBoost exibiu um desempenho intermédio ROC-AUC (0,779 ; PR-AUC = 0,104), com sensibilidade moderada (0,376) e precisão igualmente reduzida (0,096). Apesar da boa

estabilidade na especificidade (0,91) e *accuracy* de 0,897, o modelo apresentou subestimação das probabilidades positivas (*intercept* = -1,938 ; declive = 0,478).

O GLM manteve o padrão de alta sensibilidade (0,727) e baixa precisão (0,044), resultando num elevado número de falsos positivos (FP = 2627). A calibração foi a mais distante do ideal (*intercept* = -3,11 ; declive = 0,49), sugerindo tendência acentuada à subestimação das probabilidades de ocorrência.

Por fim, o C5.0 apresentou os piores resultados relativos, com ROC-AUC = 0,777 e PR-AUC = 0,087, além de baixa sensibilidade (0,224) e precisão modesta (0,113). A boa especificidade (0,955) e a elevada *accuracy* (0,937) decorrem sobretudo do predomínio de classificações negativas, refletindo baixa capacidade de deteção de M/FG.

De forma geral, a aplicação do SMOTENC fora da validação resultou em redução da sensibilidade e melhoria marginal na precisão em comparação com as abordagens de *oversampling* dentro da validação. Esse comportamento é coerente com a espectativa teórica: ao não participar no processo de reamostragem nos ciclos de validação, o modelo é exposto a uma distribuição de treino diferente da validação, o que reduz a adaptação à verdadeira fronteira da decisão. Além disso, observou-se um agravamento na calibração em quase todos os algoritmos, reforçando a importância em realizar o *oversampling* dentro dos ciclos de validação para garantir estimativas probabilísticas mais fidedignas.

5.2.10 SMOTENC Dentro da Validação

Esta secção apresenta os resultados obtidos com aplicação da técnica de reamostragem SMOTENC, implementada internamente ao conjunto de dados de treino, e posteriormente avaliada sobre o conjunto de dados de teste, cuja distribuição das categoria reflete-se na realidade observada (≈ 3% de casos graves).

A seguir, A Tabela 22 apresentam as principais métricas de desempenho obtidas para cada modelo, enquanto a Tabela 23 detalha as matrizes de confusão e os indicadores derivados a partir dos limites principais de decisão.

Tabela 22 – Métricas de desempenho com IC95% (Teste 2023, SMOTENC dentro).

Modelo	PR_AUC	ROC_AUC	Precisão	Sensibilidade	*F1-score*	*Accuracy*	*G-mean*	*Brier*
C5.0	0,025 [0,021, 0,029]	0,500 [0,500, 0,500]	0,000 [nan, nan]	0,000 [0,000, 0,000]	0,000 [nan, nan]	0,975 [0,971, 0,979]	0,000 [0,000, 0,000]	0,024 [0,021, 0,028]
GLM	0,166 [0,129, 0,219]	0,872 [0,844, 0,896]	0,220 [0,168, 0,281]	0,267 [0,200, 0,340]	0,241 [0,186, 0,301]	0,958 [0,954, 0,963]	0,510 [0,443, 0,576]	0,022 [0,019, 0,026]
NB	0,097 [0,076, 0,122]	0,830 [0,800, 0,858]	0,115 [0,072, 0,160]	0,139 [0,089, 0,196]	0,126 [0,081, 0,175]	0,952 [0,947, 0,957]	0,368 [0,294, 0,436]	0,084 [0,078, 0,089]
RF	0,209 [0,155, 0,277]	0,858 [0,826, 0,888]	0,290 [0,227, 0,356]	0,352 [0,279, 0,425]	0,318 [0,256, 0,378]	0,962 [0,958, 0,967]	0,586 [0,523, 0,644]	0,022 [0,018, 0,025]
XGB	0,247 [0,185, 0,316]	0,893 [0,871, 0,914]	0,300 [0,240, 0,361]	0,364 [0,287, 0,433]	0,329 [0,264, 0,387]	0,963 [0,958, 0,968]	0,596 [0,530, 0,651]	0,021 [0,018, 0,024]

Tabela 23 – Matrizes de confusão e métricas derivadas (Teste 2023, SMOTENC dentro).

Modelo	*Threshold*	TP	FN	FP	TN	Precisão	Sensibilidade	Especificidade	*Accuracy*	*F1-score*
GLM	0,167	44	121	156	6318	0,220	0,267	0,976	0,958	0,241
XGB	0,201	60	105	140	6334	0,300	0,364	0,978	0,963	0,329
RF	0,150	58	107	142	6332	0,290	0,352	0,978	0,962	0,318
NB	1,000	23	142	177	6297	0,115	0,139	0,973	0,952	0,126
C5.0	0,190	0	165	0	6474	nan	0,000	1,000	0,975	nan

Conforme observado na Tabela 21, o desempenho geral dos modelos diminui substancialmente quando testados sobre o conjunto de 2023, caraterizado por um forte desequilíbrio entre categorias. O modelo C5.0 apresentou falha completa na identificação de casos positivos (precisão, F_1-*score* e sensibilidade), ainda que mantenha *accuracy* de 0,975 – valor enganador, já que reflete apenas a predominância da categoria negativa.

O Naive Bayes obteve resultados moderados (F_1-*score* = 0,126; ROC-AUC = 0,830), demonstrando limitação na capacidade de distinguir entre sinistros graves e leves. Os modelos GLM, Random Forest e XGBoost apresentaram desempenhos mias sólidos, com destaque para o XGBoost, que alcançou F_1-*score* = 0,329 [0,264 ; 0,387], precisão = 0,300 e sensibilidade = 0,364, associado a ROC-AUC = 0,893 [0,871 ; 0,914]. O Random Forest apresentou desempenho muito próximo (F_1-*score* = 0,318 ; ROC-AUC = 0,858), enquanto

o GLM manteve valores ligeiramente inferiores de F_1-*score* (0,241), mas a destacar-se pela boa calibração (Brier = 0,022).

A Tabela 23 permite compreender com mais detalhe o comportamento operacional dos modelos. Nota-se que o XGBoost e o Random Forest conseguiram identificar 60 e 58 casos positivos, respetivamente, de um total de 165, o que corresponde a uma sensibilidade de aproximadamente 36% e 35%. O GLM, por sua vez, apresentou sensibilidade = 0,267 e maior especificidade (0,976), demonstrando uma postura mais conservadora na predição da categoria minoritária.

O Naive Bayes exibiu baixo poder discriminativo (sensibilidade = 0,139), enquanto o C5.0 não identificou nenhum caso positivo, classificando todas as observações como negativas. Esse comportamento reforça a tendência de sobreajustamento do C5.0 ao cenário equilibrado gerado artificialmente pelo SMOTENC, com perda total de sensibilidade ao ser exposto aos verdadeiros casos desequilibrados.

De modo geral, as métricas de área sob a curva (ROC-AUC entre 0,83 e 0,89) sugerem alguma capacidade de separação entre categorias, mas os limites de decisão não se traduziram em classificações suficientemente precisas da categoria positiva. Esse desfasamento indica que, embora os modelos aprendam padrões relevantes durante o treino reamostrado, as distribuições de probabilidade estimada não se mantêm válidas em contextos reais, o que reduz a generalização.

Além disso, verifica-se que os modelos baseados em árvores (C5.0, Random Forest e XGBoost) – que haviam demonstrado melhor desempenho nos cenários equilibrados – sofrem degradação acentuada sob desequilíbrio real, enquanto o GLM mostra maior estabilidade, ainda que com menor sensibilidade.

Assim, os resultados obtidos nas Tabelas 22 e 23, permitem concluir que em contextos reais de eventos raros, a eficácia das técnicas de reamostragem dependem fortemente da compatibilidade entre a distribuição dos dados de treino e de teste. Quando essa correspondência é baixa, a capacidade de generalização dos modelos é severamente comprometida.

Portanto, a aplicação isolada do SMOTENC durante o treino não é suficiente para garantir desempenho satisfatório em ambientes reais.

5.2.11 SMOTENC Dentro vs. SMOTENC Fora

A técnica de reamostragem SMOTENC foi aplicada exclusivamente no conjunto de treino de cada *fold* (v=5, r=2), nunca no conjunto de validação/teste. A seleção do *threshold* foi realizada pela maximização do F_2-*score* nas predições OOF. No conjunto de teste (2023), o *threshold* foi ajustado via percentil das probabilidades previstas, de modo a impor uma taxa prevista positiva (TPR) aproximada de 3% (cenário principal) e, adicionalmente, uma sensibilidade de aproximadamente 5%. Os intervalos de confiança (IC95%) foram estimados por *bootstrap* estratificado (B=1000).

Desempenho global dos modelos

A Tabela 24 apresenta os resultados das principais métricas de desempenho obtidas no conjunto de teste, permitindo avaliar a capacidade discriminativa e a estabilidade dos diferentes modelos. São incluídos modelos de natureza paramétrica e não paramétrica, permitindo avaliar as diferenças de comportamento face à reamostragem *intra-fold*.

Tabela 24 - Comparação do desempenho global dos modelos: SMOTENC aplicado dentro e fora da validação.

Modelo	PR_AUC	ROC_AUC	Precisão	Sensibilidade	*F1-Score*	*Accuracy*	*G-mean*	*Brier*
C5.0	0,025 [0,021; 0,029]	0,500 [0,500; 0,500]	0,000 [nan; nan]	0,000 [0,000; 0,000]	0,000 [nan; nan]	0,975 [0,971; 0,979]	0,000 [0,000; 0,000]	0,024 [0,021; 0,028]
GLM	0,166 [0,129; 0,219]	0,872 [0,844; 0,896]	0,220 [0,168; 0,281]	0,267 [0,200; 0,340]	0,241 [0,186; 0,301]	0,958 [0,954; 0,963]	0,510 [0,443; 0,576]	0,022 [0,019; 0,026]
NB	0,097 [0,076; 0,122]	0,830 [0,800; 0,858]	0,115 [0,072; 0,160]	0,139 [0,089; 0,196]	0,126 [0,081; 0,175]	0,952 [0,947; 0,957]	0,368 [0,294; 0,436]	0,084 [0,078; 0,089]
RF	0,209 [0,155; 0,277]	0,858 [0,826; 0,888]	0,290 [0,227; 0,356]	0,352 [0,279; 0,425]	0,318 [0,256; 0,378]	0,962 [0,958; 0,967]	0,586 [0,523; 0,644]	0,022 [0,018; 0,025]
XGB	0,247 [0,185; 0,316]	0,893 [0,871; 0,914]	0,300 [0,240; 0,361]	0,364 [0,287; 0,433]	0,329 [0,264; 0,387]	0,963 [0,958; 0,968]	0,596 [0,530; 0,651]	0,021 [0,018; 0,024]

Em termos gerais, os modelos não paramétricos, Random Forest e XGBoost, apresentam melhor capacidade discriminativa e maior estabilidade entre métricas, evidenciando ganhos consistentes de PR-AUC e F_1-*score.*

O GLM demonstra um equilíbrio razoável e boa calibração, mantendo resultados competitivos, ainda que com sensibilidade moderada.

Por outro lado, o Naive Bsyes e o C5.0 revelam limitações mais marcadas: o primeiro pela simplificação probabilística e o segundo pela incapacidade de generalizar sob forte desequilíbrio.

Em termos gerais, a aplicação do SMOTENC *intra-fold* aumenta a precisão e a estabilidade sem inflacionar artificialmente o desempenho geral.

Matrizes de confusão e métricas derivadas

A Tabela 25 resume as matrizes de confusão correspondentes ao *threshold* ajustado para uma taxa prevista positiva próxima de 3%, bem como as respetivas métricas.

Tabela 25 - Matrizes de confusão e métricas derivadas (threshold principal ≈3%).

Modelo	*Threshold*	TP	FN	FP	TN	Precisão	Sensibilidade	Especificidade	*Accuracy*	*F1-score*
GLM	0,167	44	121	156	6318	0,220	0,267	0,976	0,958	0,241
RF	0,150	58	107	142	6332	0,290	0,352	0,978	0,962	0,318
XGB	0,201	60	105	140	6334	0,300	0,364	0,978	0,963	0,329
NB	1,000	23	142	177	6297	0,115	0,139	0,973	0,952	0,126
C5.0	0,190	0	165	0	6474	nan	0,000	1,000	0,975	nan

Observa-se que o equilíbrio entre falsos positivos e falsos negativos varia consoante o modelo, refletindo diferentes comportamentos de calibração.

Os modelos baseados em árvores (Random Forest e XGBoost) mantêm a melhor combinação entre precisão e sensibilidade, atingindo bons níveis de *accuracy* mesmo sob restrição da taxa de positivos.

O GLM mostra-se mais conservador, priorizando a especificidade, enquanto o Naive Bayes evidencia fragilidade na separação probabilística, e o C5.0 praticamente não identifica casos positivos.

No conjunto, o padrão confirma que a reamostragem *intra-fold* estabiliza o comportamento dos classificadores e reduz flutuações extremas entre precisão e sensibilidade.

Thresholds utilizados nos diferentes cenários

A Tabela 26 documenta os *thresholds* utilizados em três cenários distintos:

- TH_F2: o *threshold* é obtido por maximização do F_2-*score* nas predições OOF;
- TH_RATE3 e TH_RATE5: os *thresholsd* são ajustados no teste para impor taxas previstas positivas de cerca de 3% e 5%, respetivamente.

Tabela 26 - Threshold selecionado (F_2 , OOF) e ajustado por taxa prevista positiva no teste (3% e 5%).

Modelo	TH_RATE3	TH_RATE5	TH_F2
GLM	0,1670	0,1226	0,8014
RF	0,1497	0,1000	0,3760
XGB	0,2006	0,1342	0,4928
NB	0,9996	0,9965	0,9998
C5.0	0,1901	0,1901	0,1901

A variação observada entre estes *thresholds* evidencia diferenças claras na calibração probabilística entre modelos. O GLM tende a exigir *thresholds* mais altos (predições mais conservadoras), enquanto o Random Forest e o XGBoost distribuem probabilidades de forma mais dispersa, permitidno ajustes finos. Já o Naive Bayes e o C5.0 mostram uma calibração podre, concentrando as probabilidades extremas e limitando a flexibilidade na definição do *threshold*.

Comparação entre os dois cenários

A Tabela 27 compara diretamente os dois cenários de amostragem:

- SMOTENC Fora: reamostragem aplicada antes da divisão em *folds;*
- SMOTENC Dentro: reamostragem aplicada separadamente em cada conjunto de treino.

Tabela 27 - Diferenças de métricas (pontos): SMOTENC dentro da validação - SMOTENC fora da validação.

Modelo	PR_AUC	ROC_AUC	Precisão	Sensibilidade	*F1-score*	*Accuracy*
GLM	+7,1%	+12,2%	+17,6%	-46,0%	+15,9%	+36,0%
NB	-6,0%	+4,6%	+7,0%	-67,3%	+4,1%	+38,9%
RF	+12,6%	+5,1%	+17,5%	+9,7%	+15,9%	+2,9%
XGB	+14,3%	+11,4%	+20,4%	-1,2%	+17,6%	+6,6%

Os valores positivos indicam melhorias associadas à abordagem *intra-fold*. De forma geral, verifica-se um aumento consistentes da precisão e uma ligeira redução da sensibilidade, especialmente em modelos para métricos como GLM e Naive Bayes.

Nos modelos mais flexíveis, como Random Forest e XGBoost, os ganhos são simultâneos em precisão e sensibilidade, refletindo maior capacidade de adaptação à distribuição criada pelo SMOTENC.

Além disso, as métricas de *accruracy* e PR-AUC mostram tendência de melhoria, sugerindo que o treino *intra-fold* produz estimativas mais fiéis ao desempenho fora da amostra, evitando contaminação entre treino e validação.

Conclusão

A aplicação do SMOTENC dentro dos *folds* da validação cruzada constitui uma prática metodologicamente superior, pois preserva a independência entre treino e validação, evitando *data leakage* e o inflacionamento artificial das métricas.

Em termos empíricos, observa-se um aumento da precisão, redução moderada da sensibilidade e melhor estabilidade global, sobretudo em algoritmos baseados em árvores.

No conjunto, os resultados demonstram que a reamostragem *intra-fold* produz uma avaliação mais realista e robusta, sendo a opção recomendada para contextos de forte desequilíbrio entre categorias.

5.2.12 SMOTENC Dentro vs. ROSE Dentro

Nesta análise, compara-se o desempenho dos modelos sob duas estratégias de reamostragem aplicadas dentro dos *folds* da validação cruzada, garantindo total independência entre treino e validação e eliminando qualquer risco de *data leakage*:

- SMOTENC Dentro: que gera novas observações sintéticas a partir das observações minoritárias combinando atributos contínuos e categóricos;
- ROSE Dentro: que cria observações sintéticas via *bootstrap* e perturbação aleatória controlada.

A comparação direta foi realizada com base na variação da percentagem de cada métrica, conforme a expressão:

$$\Delta = SMOTENC\ dentro - ROSE\ dentro$$

Valores positivos indicam vantagem do SMOTENC, enquanto valores negativos indicam desempenho superior do ROSE. Esta fórmula de cálculo permite observar diretamente em que métricas o SMOTENC oferece ganhos ou perdas relativas, sem necessidade de apresentar duas tabelas separadas.

Desempenho comparativo entre SMOTENC dentro e ROSE dentro

Neste sentido a Tabela 28, apresenta as diferenças de desempenho entre ambos os métodos.

Tabela 28 - Diferenças de métricas (pontos): SMOTENC dentro da validação - ROSE dentro da validação.

Modelo	PR_AUC	ROC_AUC	Precisão	Sensibilidade	*F1-score*	*Accuracy*
GLM	+0,5%	-0,1%	+10,5%	-45,4%	+4,3%	+10,3%
NB	-4,8%	-2,4%	+1,5%	-52,8%	-4,8%	+10,9%
RF	+1,4%	-1,9%	+15,3%	-24,2%	+9,5%	+6,5%
XGB	+5,3%	+1,5%	+17,6%	-29,7%	+12,0%	+8,7%

Os resultados indicam que o SMOTENC dentro e ROSE dentro apresentam desempenhos próximos em termos gerais, mas com diferenças consistentes no equilíbrio entre precisão e sensibilidade.

O SMOTENC dentro tende a produzir ganhos mais consistentes em precisão e PR-AUC, particularmente em modelos de natureza não linear. O XGBoost tem um aumento de 17,6% em precisão e 5,3% em PR-AUC, indicando uma capacidade superior de discriminação entre categorias. O Random Forest apresenta um comportamento semelhante, com um ganho de 15,3% em precisão e 1,4% em PR-AUC, o que sugere que as observações sintéticas geradas pelo SMOTENC são mais seletivas e menos redundantes, proporcionando uma fronteira de decisão mais conservadora e, portanto, menor taxa de falsos positivos entre os FL. O GLM apresenta uma melhoria de 10,5% em precisão, embora com uma perda acentuada de sensibilidade, o que é coerente com a rigidez da fronteira linear deste modelo.

Por outro lado, observa-se uma redução expressiva da sensibilidade em todos os modelos, sobretudo nos lineares e probabilísticos, como o GLM e o Naive Bayes. Essa diminuição explica-se pelo facto de o SMOTENC gerar observações sintéticas mais concentradas em torno da distribuição empírica da categoria minoritária, cobrindo menos as regiões periféricas do espaço de decisão. Já o ROSE, ao introduzir ruído aleatório nas observações de treino, tende a produzir uma cobertura mais ampla e heterogénea da fronteira, o que se traduz num número superior de M/FG (maior sensibilidade), mas com um custo de aumento de falsos positivos, reduzindo a precisão. As diferenças observadas no ROC-AUC são pequenas, geralmente inferiores 0,02, o que indica que ambas as abordagens preservam uma capacidade discriminativa global semelhante. Ainda assim, o PR-AUC revela pequenas, mas consistentes, melhorias sob o SMOTENC, sobretudo nos modelos de *ensemble*, refletindo uma maior eficiência na priorização de M/FG em contextos de forte desequilíbrio. A *accuracy* acompanha esta tendência, sugerindo que o SMOTENC origina fronteiras de decisão mais estáveis e probabilidades melhor calibradas.

Em síntese, embora ambos os métodos apresentem desempenhos próximos, o SMOTENC dentro destaca-se pela sua maior robustez, estabilidade e controlo de falsos positivos, sendo, portanto, mais indicado quando se privilegia precisão e fiabilidade na identificação de M/FG. Já o ROSE mostra-se mais vantajoso em cenários onde o objetivo é maximizar a sensibilidade, ainda que com o custo de um aumento no número de falsos FL.

5.2.13 ROSE Fora vs. SMOTENC Fora

A comparação direta entre as duas estratégias indica que o SMOTENC oferece um equilíbrio ligeiramente superior entre precisão e sensibilidade, sobretudo nos modelos de natureza não paramétrica (XGBoost e Random Forest). Em contrapartida, o ROSE tende a favorecer ligeiramente a sensibilidade — identificando mais ocorrências graves, mas à custa de um número superior de falsos positivos. Assim, a escolha entre ambos depende do objetivo operacional: se a prioridade é minimizar o risco de não detetar sinistros graves (sensibilidade máxima), o ROSE continua uma opção válida; se a ênfase

recai na fiabilidade das previsões positivas (maior precisão e melhor calibração), o SMOTENC revela-se preferível.

Importa ainda salientar que, em ambos os casos, as métricas de calibração (interceção e declive próximos de 0 e 1, respetivamente) confirmam que a probabilidade prevista de ocorrência grave reflete adequadamente a frequência observada. Os valores do erro de Brier, na ordem de 0,02, reforçam essa boa adequação probabilística.

Implicações e recomendações

Em termos substantivos, ambos os métodos de reamostragem permitiram preservar a coerência das variáveis explicativas identificadas anteriormente — reforçando a importância de fatores como a idade média do veículo, o tipo de via e o período temporal. No entanto, o SMOTENC revelou-se mais parcimonioso e estável: a menor redundância de exemplos sintéticos evitou flutuações nas métricas entre repetições, oferecendo resultados mais robustos para generalização.

Do ponto de vista aplicado à segurança rodoviária, tal estabilidade é relevante: políticas de prevenção e vigilância dependem de modelos que mantenham desempenho consistente sob diferentes amostras ou atualizações de dados. Assim, recomenda-se que versões futuras da modelação adotem o SMOTENC como procedimento padrão de reamostragem, mantendo o ROSE apenas como análise de sensibilidade ou cenário alternativo.

Em síntese, o SMOTENC confirma a robustez da estrutura de variáveis desenvolvida na dissertação e demonstra que a melhoria da representatividade da categoria minoritária pode ser alcançada sem perda de calibração nem aumento substancial do erro, constituindo uma solução metodológica equilibrada para problemas de previsão de gravidade em sinistralidade rodoviária.

5.2.14 Análise de Sensibilidade – *threshold* com taxa ≈ 5%

Nesta fase, realiza-se uma análise de sensibilidade para avaliar o impacto da variação do *threshold* no desempenho do modelo. O objetivo é observar o comportamento das métricas quando se aumenta a taxa prevista positiva de aproximadamente 3% (cenário principal) para cerca de 5%.

O ajuste é feito diretamente sobre as probabilidades previstas no conjunto de teste (2023), selecionando o percentil correspondente a uma taxa de positividade aproximadamente de 5%. Esta abordagem permite avaliar a robustez dos modelos à mudança de *threshold*, e verificar se o ganho em sensibilidade compensa a possível redução em precisão e *accuracy*.

A Tabela 29, apresenta as principais métricas de desempenho obtidas no conjunto de teste quando se força a taxa prevista positiva para aproximadamente 5%.

Tabela 29 - Métricas no conjunto de teste quando se força taxa prevista positiva ≈ 5%.

Modelo	PR_AUC	ROC_AUC	Precisão	Sensibilidade	*F1-score*	*Accuracy*	*G-mean*	*Brier*
GLM	0,166	0,872	0,190	0,382	0,254	0,944	0,605	0,022
XGB	0,247	0,893	0,226	0,455	0,302	0,948	0,661	0,021
RF	0,209	0,858	0,213	0,436	0,286	0,946	0,647	0,022
NB	0,097	0,830	0,127	0,255	0,169	0,938	0,493	0,084
C5.0	0,025	0,500	0,000	0,000	0,000	0,975	0,000	0,024

O aumento do *threshold* para alcançar uma taxa prevista positiva de 5% conduz ao aumento generalizado da sensibilidade em todos os modelos, acompanhado de uma ligeira redução da precisão.

Os modelos baseados em *ensembles*, nomeadamente, o XGBoost e o Random Forest, continuam a apresentar o melhor desempenho global. O XGBoost alcança uma sensibilidade de 0,455 e precisão de 0,226, resultando num F_1-*score* de 0,302 e o melhor *G-means* de 0,661, enquanto o Random Forest mantém a sensibilidade de 0,436, precisão de 0,231 F_1-*score* de 0,286, com ROC-AUC de 0,858.

O GLM exibe um ROC-AUC elevado (0,872) com desempenho equilibrado (precisão de 0,190 e sensibilidade de 0,382) evidenciando uma boa capacidade discriminativa mesmo com um *threshold* mais permissivo. Já o Naive Bayes apresenta resultados mais modestos, refletindo menor robustez, enquanto o C5.0 permanece inativo, com precisão e sensibilidade nulos, mostrando incapacidade de resposta mesmo após o ajuste do *threshold.*

Em síntese, ao elevar o *threshold*, observa-se um ganho em sensibilidade de aproximadamente de 0,08 a 0,10 em relação ao cenário base, acompanhado por uma

redução moderada de precisão. Apesar desta troca, os modelos de *ensemble* preservam níveis elevados de AUC e F_1-*score*, confirmando a robustez do desempenho.

5.2.15 *Thresholds* escolhidos (Pesos)

Esta secção apresenta os resultados com balanceamento por pesos (0,5/0,5), aplicada sem qualquer reamostragem sintética, preservando integralmente os dados originais e garantindo a independência entre treino e teste. São reportadas métricas no conjunto de teste (2023) para *thresholds* ajustados a rate ≈ 3% e rate ≈ 5%, incluindo intervalos de confiança (IC95%) obtidos via *bootstrap* estratificado (B=1000) em formato compacto, calibração após regressão isotónica e matrizes de confusão para rate ≈ 3%. Inclui ainda uma comparação interpretativa com abordagens de reamostragem sintéticas (SMOTENC/ROSE).

A tabela que se segue, Tabela 30, apresenta os *thresholds* escolhidos para cada modelo sob a estratégia de ponderação de pesos iguais (0,5/0,5). São incluídos três critérios de seleção:

- TH_RATE3: *threshold* ajustado para uma taxa prevista positiva ≈ 3% no conjunto de teste;
- TH_RATE5: *threshold* ajustado para uma taxa prevista positiva ≈ 5% no conjunto de teste;
- TH_F2: *threshold* que maximiza o F_2-*score* nas predições OOF.

A comparação destes *thresholds* permite observar como diferentes prioridades analíticas (precisão vs. sensibilidade) influenciam a definição do limite de decisão em cada modelo.

Tabela 30 - Thresholds selecionados para cada modelo com equilíbrio por pesos (0,5/0,5).

Modelo	TH_RATE3	TH_RATE5	TH_F2
GLM	0,738	0,919	0,883
FIRTH	0,507	0,519	0,515
RF	0,415	0,66	0,581
XGB	0,67	0,768	0,738
NB	0,176	0,327	0,246
C5.0	0,813	1,0	1,0

A Tabela 31 apresenta métricas de desempenho no conjunto de teste, quando o *threshold* foi ajustado para taxa positiva ≈ 3%

Tabela 31 - Métricas de classificação do teste (Taxa prevista ≈ 3%).

Modelo	PR_AUC	ROC_AUC	Precisão	Sensibilidade	*F1-score*	*Accuracy*	*G-mean*	*Brier*
GLM	0,161	0,875	0,22	0,267	0,241	0,51	0,958	0,184
FIRTH	0,157	0,865	0,235	0,285	0,258	0,527	0,959	0,24
RF	0,186	0,874	0,23	0,279	0,252	0,522	0,959	0,064
XGB	0,254	0,901	0,285	0,345	0,312	0,581	0,962	0,179
C5.0	0,171	0,851	0,23	0,279	0,252	0,522	0,959	0,026
NB	0,103	0,83	0,115	0,139	0,126	0,368	0,952	0,214

A Tabela 32 apresenta métricas de desempenho no conjunto de teste, quando o *threshold* foi ajustado para taxa positiva ≈ 5%.

Tabela 32 - Métricas de classificação do teste (Taxa prevista ≈ 5%).

Modelo	PR_AUC	ROC_AUC	Precisão	Sensibilidade	*F1-score*	*Accuracy*	*G-mean*	*Brier*
GLM	0,161	0,875	0,202	0,406	0,27	0,624	0,945	0,184
FIRTH	0,157	0,865	0,196	0,394	0,262	0,615	0,945	0,24
RF	0,186	0,874	0,19	0,382	0,254	0,605	0,944	0,064
XGB	0,254	0,901	0,244	0,491	0,326	0,687	0,95	0,179
C5.0	0,171	0,851	0,193	0,388	0,258	0,61	0,944	0,026
NB	0,103	0,83	0,127	0,255	0,169	0,493	0,938	0,214

A Tabela 33 apresenta as principais métricas de desempenho dos modelos no conjunto de teste, ajustadas ao equilíbrio por pesos iguais (0,5/0,5). Cada valor é acompanhado pelo IC95%, obtido via *bootstrap* estratificado (B=1000), permitindo avaliar a estabilidade e a variabilidades das métricas.

Tabela 33 - Métricas de desempenho dos modelos no teste com IC95% (Bootstrap estratificado, pesos 0,5/0,5).

Modelo	PR_AUC	ROC_AUC	Precisão	Sensibilidade	*F1-Score*	*Accuracy*	*G-mean*	*BRIER*
C5.0	0,171 [0,127–0,233]	0,851 [0,818–0,879]	0,230 [0,170–0,294]	0,279 [0,212–0,352]	0,252 [0,191–0,315]	0,522 [0,455–0,586]	0,959 [0,954–0,963]	0,026 [0,023–0,028]
FIRTH	0,157 [0,120–0.208]	0,865 [0,839–0,889]	0,235 [0,178–0,298]	0,285 [0,211–0,350]	0,258 [0,196–0,315]	0,527 [0,454–0,584]	0,959 [0,955–0,964]	0,240 [0,240–0,240]

Modelo	PR_AUC	ROC_AUC	Precisão	Sensibilidade	*F1-Score*	*Accuracy*	*G-mean*	*BRIER*
GLM	0,161 [0,125–0,213]	0,875 [0,850–0,898]	0,220 [0,161–0,277]	0,267 [0,201–0,337]	0,241 [0,182–0,298]	0,510 [0,444–0,574]	0,958 [0,954–0,963]	0,184 [0,179–0,190]
NB	0,103 [0,079–0,131]	0,830 [0,804–0,859]	0,115 [0,072–0,159]	0,139 [0,090–0,194]	0,126 [0,082–0,170]	0,368 [0,296–0,434]	0,952 [0,947–0,957]	0,214 [0,206–0,222]
RF	0,186 [0,141–0,243]	0,874 [0,850–0,894]	0,230 [0,175–0,288]	0,279 [0,214–0,349]	0,252 [0,193–0,311]	0,522 [0,457–0,584]	0,959 [0,954–0,964]	0,064 [0,061–0,067]
XGB	0,254 [0,196–0,334]	0,901 [0,879–0,921]	0,285 [0,224–0,351]	0,345 [0,277–0,418]	0,312 [0,253–0,372]	0,581 [0,521–0,640]	0,962 [0,958–0,967]	0,179 [0,175–0,183]

A Tabela 34 apresenta a calibração dos modelos no conjunto de teste de 2023, utilizando a técnica de regressão isotónica aplicada após a ponderação de pesos iguais.

Tabela 34 - Métricas de calibração dos modelos no conjunto de teste (Regressão Isotónica, pesos 0,5/0,5).

Modelo	*BRIER*	*INTERCEPT*	*SLOPE*
GLM_PESOS_cal	0,022	-0,023	1,053
FIRTH_PESOS_cal	0,022	0,001	1,066
RF_PESOS_cal	0,022	-0,129	0,978
XGB_PESOS_cal	0,021	-0,119	1,197
C5.0_PESOS_cal	0,022	0,072	1,028
NB_PRIOR05_cal	0,023	0,161	1,228

A Tabela 35 apresenta as matrizes de confusão dos modelos no conjunto de teste de 2023, considerando um *threshold* ajustado para uma taxa prevista positiva ≈ 3% e equilíbrio por pesos iguais.

Tabela 35 - Métricas de confusão dos modelos no conjunto de teste (taxa prevista positiva ≈3%, pesos 0,5/0,5)

Modelo	*Threshold*	TP	FN	FP	TN	Precisão	Sensibilidade	Especificidade	*Accuracy*	*F1-Score*
GLM	0,919	44,0	121,0	156,0	6318,0	0,22	0,267	0,976	0,241	0,958
FIRTH	0,519	47,0	118,0	153,0	6321,0	0,235	0,285	0,976	0,258	0,959
XGB	0,768	57,0	108,0	143,0	6331,0	0,285	0,345	0,978	0,312	0,962
RF	0,66	46,0	119,0	154,0	6320,0	0,23	0,279	0,976	0,252	0,959
NB	1,0	23,0	142,0	177,0	6297,0	0,115	0,139	0,973	0,126	0,952
C5.0	0,327	46,0	119,0	154,0	6320,0	0,23	0,279	0,976	0,252	0,959

A estratégia de ponderação por pesos (PESOS) mostrou-se uma alternativa robusta para lidar com o desequilíbrio da amostra, preservando integralmente os dados originais e evitando o risco de *leakage* inerente a técnicas de reamostragem fora dos *folds* de validação. Essa abordagem garantiu comparabilidade entre os modelos e consistência estatística dos resultados obtidos no teste, tanto em termos de desempenho preditivo quanto de calibração.

Em cenários de taxa prevista positiva aproximada de 3%, o desempenho global foi satisfatório. Observou-se que os modelos generalizados (GLM/Firth) e o Random Forest apresentaram valores de PR-AUC entre 0,157 e 0,186, com F_1-*score* na faixa de 0,24 – 0,26. Tais resultados demonstram equilíbrio entre precisão e sensibilidade, mantendo boa capacidade discriminativa (ROC-AUC acima de 0,86) e boa estabilidade (*G-mean* ≈ 0,96). O XGBoost destacou-se PR-AUC de 0,254 e F_1-*score* de 0,312, sugerindo maior poder de separação entre categorias, embora com tendência a maior variabilidade e sensibilidade a pequenas perturbações nos preditores.

Quando o limite foi ajustado para uma taxa prevista positiva de ≈ 5%, houve incremento consistente na sensibilidade – sobretudo para o XGBoost, que atingiu 0,491 de sensibilidade e F_1-*score* de 0,326 – em detrimento da precisão. Assim, a escolha do *threshold* depende diretamente da prioridade analítica: limites mais baixos (3%) privilegiam a precisão, enquanto taxas mais altas (5%) ampliam a capacidade de deteção de casos positivos, sendo, portanto, preferíveis quando o objetivo é maximizar a sensibilidade ou o F_2-*score*.

Os modelos ponderados apresentaram um bom desempenho em calibração após regressão isotónica, com Brier entre 0,021 e 0,023 e coeficientes de calibração próximos aos ideias (*intercep* ≈ 0 e *slop* ≈ 1), indicando que as probabilidades previstas foram bem ajustadas. Essa estabilidade contrasta com os efeitos observados em técnicas de reamostragem sintética. O SMOTENC, quando corretamente confiando dentro dos *folds*, também atinge boa sensibilidade, mas adiciona variância e pode induzir sobreajuste local em algumas combinações de preditores. Comparativamente, os pesos exibiram comportamentos mais estável e interpretável, sobretudo para modelos generalizados (GLM/Firth). De forma semelhante, o ROSE compartilha as vantagens no SMOTENC em termos de aumento da sensibilidade, porém o ruído gerado pode degradar a calibração e, se não for estritamente *intra-fold*, pode causa *data leakage*.

Em relação XGBoost, sob ponderação 0,5/0,5, ele pode apresentar instabilidade, com *scores* semi constantes dependendo de caraterísticas dos preditores, como valores de baixa variância ou escala.

Em síntese, a estratégia PESOS apresentou equilíbrio entre desempenho preditivo, estabilidade e interpretabildiade, superando abordagens baseadas em reamostragem em termos de calibração e robustez estatística. GLM, Firth e Random Forest, destacaram-se como modelos confiáveis e transparentes, enquanto o XGBoost apresentou desempenho absoluto superior, porém com maior sensibilidade a perturbações nos dados. Para maximizar a sensibilidade (ênfase em F_2), recomenda-se utilizar taxa prevista positiva ≈ 5%, maior precisão ≈ 3%. Em ambos os casos, deve-se manter calibração isotónica e reportar intervalos de confiança de 95% obtidos via *bootstrap*.

5.2.16 Modelos com interações vs. Baseline (GLM/Firth) e relação com PESOS/SMOTENC

Nesta etapa do estudo, avalia-se o efeito da inclusão de interações nas regressões logística e de Firth, comparando-as com as respetivas versões base, tanto em configurações com e sem ponderação de categorias. O protocolo experimental seguiu um esquema temporal rigoroso - treino no período 2016-2022 e teste em 2023 - com validação cruzada estratificada (v=5 , r=2), assegurando estimativas robustas e livres de *data leakage*.

Os limites de decisão foram determinados segundo o F_2-*score*, sob restrições operacionais, e a avaliação final do conjunto de este baseou-se em pe2rcentis que reproduzem taxas de previsão positivas próximas de 3% e 5%, refletindo condições realistas de aplicação.

Tal como discutido anteriormente, optou-se por não aplicar a calibração isotónica aos modelos logísticos e de Firth, uma vez que, estes já produzem estimativas probabilísticas intrinsecamente calibradas. Assim, a análise concentra-se exclusivamente na influência das interações e da ponderação de categorias sobre o desempenho discriminativa e o equilíbrio entre sensibilidade e precisão, sem interferência de transformações adicionais na escala das probabilidades.

As tabelas seguintes (Tabelas 36 a 41) sintetizam o desempenho global dos modelos *baseline* e com/sem interações, bem como o desempenho dos modelos calibrados com ponderação (pesos), através de métricas discriminativas, calibração e matrizes de confusão.

Tabela 36 - Métricas no teste (rate ≈ 3%) - Modelos com Interações

Modelo	PR_AUC	ROC_AUC	Precisão	Sensibilidade	*F1-Score*	*G-mean*	*Accuracy*	*Brier*
GLM_BASE_W_RATE3%	0.161	0.875	0.22	0.267	0.241	0.51	0.958	0.184
GLM_BASE_NW_RATE3%	0.166	0.872	0.22	0.267	0.241	0.51	0.958	0.022
GLM_INT_W_RATE3%	0.202	0.884	0.265	0.321	0.29	0.56	0.961	0.177
GLM_INT_NW_RATE3%	0.215	0.881	0.26	0.315	0.285	0.555	0.961	0.022
FIRTH_BASE_W_RATE3%	0.157	0.865	0.235	0.285	0.258	0.527	0.959	0.24
FIRTH_BASE_NW_RATE3%	0.166	0.872	0.22	0.267	0.241	0.51	0.958	0.022
FIRTH_INT_W_RATE3%	0.14	0.863	0.175	0.212	0.192	0.455	0.956	0.241
FIRTH_INT_NW_RATE3%	0.215	0.881	0.26	0.315	0.285	0.555	0.961	0.022

Tabela 37 - Métricas no teste (rate ≈ 5%) - Modelos com Interações

Modelo	PR_AUC	ROC_AUC	Precisão	Sensibilidade	*F1-score*	*G-mean*	*Accuracy*	*Brier*
GLM_BASE_W_RATE5%	0.161	0.875	0.202	0.406	0.27	0.624	0.945	0.184
GLM_BASE_NW_RATE5%	0.166	0.872	0.19	0.382	0.254	0.605	0.944	0.022
GLM_INT_W_RATE5%	0.202	0.884	0.214	0.43	0.286	0.643	0.947	0.177
GLM_INT_NW_RATE5%	0.215	0.881	0.217	0.436	0.29	0.647	0.947	0.022
FIRTH_BASE_W_RATE5%	0.157	0.865	0.196	0.394	0.262	0.615	0.945	0.24
FIRTH_BASE_NW_RATE5%	0.166	0.872	0.19	0.382	0.254	0.605	0.944	0.022
FIRTH_INT_W_RATE5%	0.14	0.863	0.181	0.364	0.241	0.59	0.943	0.241
FIRTH_INT_NW_RATE5%	0.215	0.881	0.217	0.436	0.29	0.647	0.947	0.022

Tabela 38 - Variações (Interações – Base) a rate ≈ 3%

Contrast	Δ PR_AUC	Δ Sensibilidade	Δ Precisão	Δ *F1-score*	*Rate*
GLM_INT_W_rate3% - GLM_BASE_W_rate3%	0.041	0.054	0.045	0.049	rate3%
GLM_INT_NW_rate3% - GLM_BASE_NW_rate3%	0.049	0.048	0.04	0.044	rate3%
FIRTH_INT_W_rate3% - FIRTH_BASE_W_rate3%	-0.017	-0.073	-0.06	-0.066	rate3%
FIRTH_INT_NW_rate3% - FIRTH_BASE_NW_rate3%	0.049	0.048	0.04	0.044	rate3%

Tabela 39 - Variações (Interações – Base) a rate ≈ 3%

Contrast	Δ PR_AUC	Δ Sensibilidade	Δ Precisão	Δ *F1-score*	*Rate*
GLM_INT_W_rate5% - GLM_BASE_W_rate5%	0.041	0.024	0.012	0.016	rate5%
GLM_INT_NW_rate5% - GLM_BASE_NW_rate5%	0.049	0.054	0.027	0.036	rate5%
FIRTH_INT_W_rate5% - FIRTH_BASE_W_rate5%	-0.017	-0.03	-0.015	-0.021	rate5%
FIRTH_INT_NW_rate5% - FIRTH_BASE_NW_rate5%	0.049	0.054	0.027	0.036	rate5%

Tabela 40 - Métricas no teste (rate ≈ 3%) - PESOS (baseline, sem interações)

Modelo	PR_AUC	ROC_AUC	Precisão	Sensibilidade	*F1-score*	*G-mean*	*Accuracy*	Brier
GLM_PESOS	0.161	0.875	0.22	0.267	0.241	0.51	0.958	0.184
FIRTH_PESOS	0.157	0.865	0.235	0.285	0.258	0.527	0.959	0.24
RF_PESOS	0.186	0.874	0.23	0.279	0.252	0.522	0.959	0.064
XGB_PESOS	0.254	0.901	0.285	0.345	0.312	0.581	0.962	0.179
C5.0_PESOS	0.171	0.851	0.23	0.279	0.252	0.522	0.959	0.026
NB_PRIOR05	0.103	0.83	0.115	0.139	0.126	0.368	0.952	0.214

Tabela 41 - Métricas no teste (rate ≈ 5%) - PESOS (baseline, sem interações)

Modelo	PR_AUC	ROC_AUC	Precisão	Sensibilidade	*F1-score*	*G-mean*	*Accuracy*	Brier
GLM_PESOS_rate5%	0.161	0.875	0.202	0.406	0.27	0.624	0.945	0.184
FIRTH_PESOS_rate5%	0.157	0.865	0.196	0.394	0.262	0.615	0.945	0.24
RF_PESOS_rate5%	0.186	0.874	0.19	0.382	0.254	0.605	0.944	0.064
XGB_PESOS_rate5%	0.254	0.901	0.244	0.491	0.326	0.687	0.95	0.179
C5.0_PESOS_rate5%	0.171	0.851	0.193	0.388	0.258	0.61	0.944	0.026
NB_PRIOR05_rate5%	0.103	0.83	0.127	0.255	0.169	0.493	0.938	0.214

Os modelos com interações procuram capturar efeitos conjuntos entre caraterísticas da infraestrutura, tipologia do sinistro e composição do tráfego, aspetos frequentemente não lineares nas vias rodoviárias. Em linha com a literatura, espera-se que tais termos aumentem o poder discriminativo sem sacrificar a interpretabilidade nos GLM e reduzam o *viés* em eventos raros nos modelos de Firth, cuja penalização de Jeffreys mitiga a sobrestimação de probabilidades extremas (Heinze & Schemper, 2002; King & Zeng, 2001).

Operacionalmente, a utilização de *thresholds* por percentil no teste (rate ≈ 3%/5%) garante comparabilidade entre modelos e evita colapsos de sensibilidade (i.e., zeros) associados a limites conservadores derivados apenas por F_2 nas OOF. Após esta

correção, o aumento observado de sensibilidade e F_1-*score e* nas variáveis com interações, especialmente quando combinadas com pesos de categorias, refletem a capacidade destes modelos em priorizar corretamente eventos raros sem distorcer a ordem global das probabilidades. A estabilidade da PR-AUC, dependentes apenas do *ranking* das previsões, indica que a discriminação global mantém-se estável ou ligeiramente superior, como esperado, dado depender apenas da ordenação.

Comparativamente a estratégias baseadas em amostras sintéticas (SMOTENC/ROSE), os modelos ponderados com interações apresentam menor variância e maior robustez temporal. A ponderação ajusta a função de perda sem modificar a distribuição empírica, preservando a calibração natural dos modelos GLM e Firth, enquanto métodos sintéticos podem gerar previsões artificialmente extremas e risco de *data* leakage (Lunardon, Menardi, & Torelli, 2014; Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

Em termos práticos:

- GLM com interações e pesos tende a maximizar a sensibilidade a rate ≈ 5% com perda moderada de precisão;
- Firth com interações é o mais estável, oferecendo compromisso favorável entre sensibilidade e precisão a *rate* ≈ 3% e um Brier *score* competitivo, coerente com a sua natureza de correção de *viés* em categorias raras.

Por fim, a inclusão de interações demonstrou capturar padrões estruturais, contribuindo para previsões robustas e consistentes ao longo do tempo, reforçando a aplicabilidade operacional dos modelos em contextos da vida real.

5. Conclusão

O presente estudo abordou um dos problemas mais desafiantes da modelação preditiva em ciência de dados: a previsão de eventos raros, aqui representados pelos sinistros rodoviários graves e mortais no distrito de Setúbal. O trabalho integrou uma componente teórica e metodológica sólida com uma análise empírica rigorosa, permitindo avaliar comparativamente diferentes estratégias estatísticas e de *machine learning* aplicadas a um fenómeno com forte desequilíbrio entre categorias.

A análise partiu de uma base de dados extensa (mais de 43 mil ocorrências), na qual apenas cerca de 2-3% dos registos correspondiam a sinistros graves ou mortais. Este desequilíbrio extremo compromete a capacidade preditiva dos modelos convencionais, tornando necessária a adoção de abordagens específicas de correção. Assim, foram testadas três famílias de soluções:

(i) técnicas de reamostragem controladas (*oversampling* via ROSE e SMOTENC, aplicadas apenas nos dados de treino, evitando *data leakage*);

(ii) modelos ponderados, com pesos inversamente proporcionais à frequência das categorias;

(iii) modelos penalizados, através da Regressão Logística de Firth, que assegura estabilidade inferencial sob separação quase completa.

Os modelos comparados, Regressão Logística (clássica e Firth), Random Forest, C5.0, XGBoost e Naive Bayes, foram avaliados com base em métricas adaptadas a eventos raros: a área sob a curva Precisão-Sensibilidade (PR-AUC), a área sob a curva ROC (ROC-AUC), o F_2-*score* (critério de otimização dos limites de decisão), o Brier score, e os parâmetros de calibração global (*intercept* e *slope*).

A validação cruzada repetida, aliada a uma avaliação final em *hold-out test set*, assegurou robustez estatística e validade externa das conclusões.

Os resultados empíricos revelaram três conclusões principais:

1. A correção do desequilíbrio é indispensável, mas deve ser metodologicamente controlada. A aplicação de técnicas de oversampling exclusivamente no treino, em vez de no conjunto total, eliminou o enviesamento otimista observado em abordagens anteriores, reduzindo o risco de sobreajuste e melhorando a generalização para o teste. Entre as técnicas comparadas, ROSE e SMOTENC

produziram resultados semelhantes em termos de ROC-AUC (~0,86–0,89), com ligeira vantagem do SMOTENC em sensibilidade e equilíbrio global (F_2 ≈ 0,27).

2. O desempenho varia consideravelmente com o tipo de algoritmo. O XGBoost emergiu como o modelo mais consistente, obtendo o melhor compromisso entre precisão e sensibilidade (PR-AUC ≈ 0.22; ROC-AUC ≈ 0,88; *Brier* ≈ 0,021), seguido do Random Forest, que apresentou desempenho estável, mas menos calibrado. A Regressão Logística de Firth destacou-se pela excelente calibração probabilística (intercepto ≈ 0; *Brier* ≈ 0,022) e pela sua capacidade de deteção da categoria rara (sensibilidade ≈ 0,67), sendo uma alternativa robusta e interpretável aos modelos mais complexos. Por contraste, o Naive Bayes e o C5.0 revelaram maior variabilidade e menor discriminação em contextos de forte desequilíbrio.
3. O F_2-*score* demonstrou ser uma métrica de corte mais adequada para contextos críticos. A otimização dos limites de decisão pelo F_2-*score* e, privilegiando a sensibilidade, aumentou substancialmente a capacidade de identificar casos graves, mesmo à custa de maior número de falsos positivos. Esta abordagem é metodologicamente coerente com o objetivo de prevenção e intervenção precoce em segurança rodoviária.

Em síntese, o estudo evidencia que a combinação de modelos calibrados, técnicas de reamostragem controladas e métricas ajustadas a eventos raros pode melhorar de forma significativa o desempenho e a utilidade prática dos modelos preditivos. A Regressão Logística de Firth surge como uma referência metodológica sólida, enquanto XGBoost e Random Forest se afirmam como opções de elevado desempenho em cenários operacionais.

Do ponto de vista aplicado, a modelação desenvolvida permite identificar fatores associados a maior gravidade dos acidentes, contribuindo para orientar políticas públicas baseadas em evidência, nomeadamente na definição de zonas críticas, gestão de recursos e planeamento de medidas preventivas.

Como linhas futuras de investigação, propõe-se:

(i) a incorporação de variáveis espaciais e temporais em modelos hierárquicos (*spatio-temporal rare-event models*);

(ii) a análise de interpretação de modelos complexos através de métodos explicativos (e.g. SHAP, *partial dependence*);

(iii) e a integração de informação de tráfego em tempo real, potenciando modelos de previsão dinâmica do risco rodoviário.

Assim, esta dissertação reforça a importância da modelação comparativa e estatisticamente rigorosa de eventos raros, tanto no plano metodológico como na sua aplicação concreta à segurança rodoviária, contribuindo para uma abordagem mais preventiva, transparente e orientada por dados.

Referências Bibliográficas

Assis, H. A. C. (2022). *A prevenção da sinistralidade rodoviária grave* [Relatório Científico Final, Mestrado Integrado em Segurança da Guarda Nacional Republicana]. Academia Militar. In *comum.rcaap.pt*. https://comum.rcaap.pt/handle/10400.26/42429

Autoridade Nacional de Segurança Ferroviária. (2024). *Relatório anual de segurança ferroviária 2023*. IMT. https://www.imt-ip.pt/wp-content/uploads/2025/01/RASF_2023.pdf

Autoridade Nacional de Segurança Rodoviária. (2024a). *Relatório Anual 2023*. ANSR. http://www.ansr.pt/Estatisticas/RelatoriosDeSinistralidade/Pages/default.aspx

Autoridade Nacional de Segurança Rodoviária. (2024b). *Relatório de Sinistralidade a 24h e fiscalização rodoviária de maio de 2024*. ASNR. http://www.ansr.pt/Noticias/Pages/Relat%C3%B3rio-de-Sinistralidade-a-24h-e-fiscaliza%C3%A7%C3%A3o-rodovi%C3%A1ria-de-maio-de-2024.aspx

Autoridade Nacional de Segurança Rodoviária. (2024c). *Relatório julho 2024*. ANSR. http://www.ansr.pt/Estatisticas/RelatoriosDeSinistralidade/Pages/default.aspx

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, 66(3), 411–421.

Bajpai, S., Bajpai, R. C., & Chaturvedi, H. K. (2015). Evaluation of inter-rater agreement and inter-rater reliability for observational data: An overview of concepts and methods. *Journal of the Indian Academy of Applied Psychology*, *41*(3), 20–27.

Bastos, P. F. M. (2022). *Modelação de acidentes numa rede urbana de transporte ferroviário* (pp. 0–64) [Dissertação de Mestrado]. Universidade do Porto. https://repositorio-aberto.up.pt/handle/10216/142455

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32. DOI:10.1023/A:1010933404324

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3), 4626-4636. DOI:10.1016/j.eswa.2008.05.027

Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357. DOI:10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. DOI:10.1145/2939672.2939785

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning*, 233–240.

Demir, S., & Şahin, E. K. (2022). Evaluation of oversampling methods (OVER, SMOTE, and ROSE) in classifying soil liquefaction dataset based on SVM, RF, and naïve bayes. *European Journal of Science and Technology*, *34*.

Efron, B., & Tibshirani, R. J. (1993*). An introduction to the bootstrap*. Chapman & Hall/CRC.

Fernández, A., Garcia, S., & Herrera, F. (2018). SMOTE for learning from imbalanced data: Progress and challenges. *Pattern Recognition*, 93, 1-22. DOI:10.1613/jair.1.11192

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1), 27-38. doi:10.1093/biomet/80.1.27

Fox, J., & Monette, G. (1992) Generalized collineary diagnostics. *Journal of the American Statistical Association*, 87(417), 178-183

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1–22.

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189-1232.

Hand, D. J., & Christen, P. (2018). A note on using the F-measure for evaluating record linkage algorithms. *Statistics and Computing*, 28(3), 539–547.

Harrell, F. E. (2015). *Regression Modeling Strategies* (2nd ed.). Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer. https://doi.org/10.1007/b94608

Healy, M. J. (1993). 10. counted data (2). *Archives of Disease in Childhood*, *68*(6), 800–802. https://doi.org/10.1136/adc.68.6.800

Heinze, G., & Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21(16), 2409–2419.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts.

Infante, P. (2023). Dados que podem salvar vidas: Modelação e predição de acidentes de viação para uma segurança rodoviária mais eficaz. *Gazeta de Matemática*, *201*.

Iwendi, C., Bashir, A. K., Peshkar, A., Sujatha, R., Chatterjee, J. M., Pasupuleti, S., Mishra, R., Pillai, S., & Jo, O. (2020). COVID-19 patient health prediction using

boosted random forest algorithm. *Frontiers in Public Health*, *8*(357). https://doi.org/10.3389/fpubh.2020.00357

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9(2), 137-163.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5), 1–26.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Kull, M., Silva Filho, T. M., & Flach, P. (2017). Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. *Proceedings of AISTATS*, 623–631.

Li, W., Yin, Y., Quan, X., & Zhang, H. (2019). Gene expression value prediction based on XGBoost algorithm. *Frontiers in Genetics*, *10*(1077). https://doi.org/10.3389/fgene.2019.01077

Lunardon, N., Menardi, G., & Torelli, N. (2014). ROSE: A package for binary imbalanced learning. *The R Journal*, *6*(1), 79–89.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *arXiv*. https://doi.org/10.48550/arXiv.1706.06060

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1), 92-122. DOI:10.1007/s10618-012-0295-5

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of ICML*, 625–632.

Obi, J. C. (2023). A comparative study of several classification metrics and their performances on data. *World Journal of Advanced Engineering Technology and Sciences*, *8*(1), 308–314. https://doi.org/10.30574/wjaets.2023.8.1.0054

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2018). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830. DOI:10.48550/arXiv.1201.0490

Pérez, M. R. (2011). ¿Se debe usar el término accidente en el ámbito de la investigación científica?. *Panace@*, 12(33), 84-88. https://www.tremedica.org/wp-content/uploads/n33-Tribuna-Perez.pdf

Pirompud, P., Sivapirunthep, P., Punyapornwithaya, V., & Chaosap, C. (2024). Machine learning predictive modeling for condemnation risk assessment in antibiotic-free raised broilers. *Poultry Science*, *103*(12).

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. Smola et al. (Eds.), *Advances in Large Margin Classifiers* (pp. 61–74). MIT Press.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (Eds.). (2009). *Dataset Shift in Machine Learning*. MIT Press.

Ratih, I. D., Retnaningsih, S. M., Islahulhaq, I., & Dewi, V. M. (2022). Synthetic minority over-sampling technique nominal continous logistic regression for imbalanced data. *AIP Conference Proceedings*, *2668*(1).

Ratnasari, A. P., & Nur'aini, R. (2024). Performance of random oversampling, random undersampling, and SMOTE-NC methods in handling imbalanced class in classification models. *International Journal of Scientific Research and*

Management (IJSRM), *12*(04), 494–501. https://doi.org/10.18535/ijsrm/v12i04.m03

Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, 3, 41-46. IBM Reasearch.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Santos, R. J. M. (2017). *Regressão logística em dados com eventos raros* (pp. 0–83) [Dissertação de Mestrado]. Universidade do Porto. https://repositorio-aberto.up.pt/bitstream/10216/112298/2/269348.pdf

Scacabarozi, F. N. (2012). *Modelagem de eventos raros: Um estudo comparativo*. [Dissertação de Mestrado]. Universidade Federal de São Carlos.

Sheng, M., Zhou, J., Chen, X., Teng, Y., Hong, A., & Liu, G. (2022). Landslide susceptibility prediction based on frequency ratio method and C5.0 decision tree model. *Frontiers in Earth Science*, *10*(918386). https://doi.org/10.3389/feart.2022.918386

Tabasso, C. (2012). Paradigmas, teorías y modelos de la seguridad y la inseguridad vial. http://94.23.80.242/~aec/ivia/tabasso_124.pdf

Tito, A. E. A., Condori, B. O. H., & Vera, Y. P. (2023). Análisis comparativo de Técnicas de Machine Learning para la predicción de casos de deserción universitaria. *RISTI - Revista Ibérica de Sistemas E Tecnologias de Informação*, *51*, 84–98. https://doi.org/10.17013/risti.51.84-98

Tribunal de Contas Europeu. (2024*). Relatório especial: Segurança rodoviária – Para alcançar os objetivos, a EU tem de entrar na via rápida.* https://www.eca.europa.eu/ECAPublications/SR-2024-04/SR-2024-04_PT.pdf

Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M. J., & Steyerberg, E. W. (2019). A calibration hierarchy for risk models was defined: From uncalibrated models to recalibration by re-estimation. *BMC Medicine*, 17(1), 230.

Watts, V. (2022). *8.5 RareEvents, the Sample, Decision, and Conclusion*. Pressbooks.

Wibowo, P., & Fatichah, C. (2022). Pruning-based oversampling technique with smoothed bootstrap resampling for imbalanced clinical dataset of Covid-19. *Journal of King Saud University - Computer and Information Sciences*, *34*(9), 7830–7839.

Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information*, *14*(1). https://doi.org/10.3390/info14010054

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2nd ed.). CRC Press.

Yang, C., Fridgeirsson, E. A., Kors, J. A., Reps, J. M., & Rijnbeek, P. R. (2024). Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. *Journal of Big Data*, *11*(7). https://doi.org/10.1186/s40537-023-00857-7

Yang, T., & Ying, Y. (2023). AUC maximization in the era of big data and AI: A survey. *ACM Computing Surveys*, *55*(8), 1–37.

Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of KDD*, 694–699.

Zhang, H., & Li, D. (2007). Naïve Bayes Text Classifier. *In International Conference on Granular Computing* (pp. 708-708). IEEE. https://doi.org/10.1109/GrC.2007.40

Zolanvari, A., Haghighi, S., & Sabouri, S. (2022). A literature review on rater agreement metrics: A review for finding a standard approach for comparing the performance of ML algorithms. *Pycm.io*. https://www.pycm.io/reports/Literature_Review.pdf

Apêndice

Os resultados seguintes correspondem a uma abordagem exploratória inicial, em que as técnicas de reamostragem foram aplicadas antes da divisão treino/teste. Estes valores não são diretamente comparáveis com os obtidos na abordagem final (reamostragem apenas no treino).

Apêndice A - ROSE

O conjunto de dados do modelo final apresenta dados desequilibrados acentuado na variável resposta, com uma maioria muito expressiva de sinistros classificados como "Feridos Leves". Esse tipo de desequilíbrio pode ser problemático para os modelos de *machine learning*. Neste sentido, ao examinar a base de dados a distribuição de observações era a seguinte:

- Categoria 0 – "Feridos Leves": 42317 observações (categoria maioritária)
- Categoria 1 – "Mortes/Feridos Graves": 995 observações (categoria minoritária)

Esse desnível, onde aproximadamente 97,7% dos sinistros pertencem à categoria maioritária e apenas 2,3% à categoria minoritária, pode introduzir um *viés* no modelo, favorecendo previsões para a categoria dominante.

Para mitigar esse impacto, serão aplicadas técnicas de ajuste, como o *oversampling* da categoria minoritária e o *undersampling* da categoria maioritária. Entre os métodos de *oversampling* considerados estão o ROSE e o SMOTENC. Além disso, métricas como a curva ROC, a área sob a curva (AUC) e o F_1-*score* serão utilizadas para avaliar o desempenho dos modelos.

No contexto do *machine learning*, tratar dados desequilibrados é fundamental para que os modelos generalizem bem e ofereçam previsões imparciais.

Neste trabalho o ROSE foi usado para gerar diferentes cenários com dados sintéticos, nomeadamente, gerar observações sintéticas (*oversampling*) para a categoria minoritária de forma a ter um cenário com dados equilibrados e cenários com diferentes graus de desequilíbrio, a gerar observações sintéticas para a categoria minoritária e a remover observações da categoria maioritária (*undersampling*) de forma a equilibrar os dados.

Dados equilibrados

Para alcançar um balanceamento adequado entre as categorias e uma distribuição mais equilibrada, o número total de observações foi ajustado para 85000. Inicialmente, o conjunto de dados possuía 43312 observações, das quais 42317 pertenciam à categoria "Feridos Leves" e apenas 995 à categoria "Mortes/Feridos Graves". Para equilibrar as categorias, garantindo que "Mortes/Feridos Graves" atingisse o mesmo número de observações que "Feridos Leves", novas observações foram geradas, resultando num conjunto de dados balanceado. A distribuição final pode ser visualizada na Tabela A1.

Tabela A 1 - ROSE: Modelo de regressão logística com e sem oversampling.

	***Oversampling* Regressão Logística – 85000 Observações**	
	Feridos Leves	Mortes / Feridos Graves
Modelo Simples	42317	995
Modelo com *Oversampling*	42317	42683

1) **Divisão dos dados em treino e teste**

Após a aplicação do método de *oversampling* para balancear as categorias, a base de dados obtida foi preparada para a modelação. Começou-se por dividir o conjunto de dados em dois subconjuntos, um para treino (70%) e outro para teste (30%), onde:

- Conjunto de treino: contém 59630 observações,
- Conjunto de teste: contém 25370 observações.

A Tabela A2, representa a divisão realizada juntamente com os valores obtidos.

Tabela A 2 - ROSE: Divisão dos dados do modelo de regressão logística (85000 observações) em dois subconjuntos: treino e teste e respetivo número de observações por categoria em cada subconjunto.

	Regressão Logística – 85000 Observações	
	Treino	**Teste**
Feridos Leves	29821	12496
Mortes/Feridos Graves	29809	12874
	59630	25370

2) Ajustamento do modelo

Após a divisão dos dados em treino e teste, procedeu-se para o ajustamento do modelo de Regressão Logística Estatístico. O ajustamento foi realizado com base no conjunto de teste, recorrendo ao método da máxima verosimilhança. Este processo possibilita identificar os fatores estatisticamente significativos associados à gravidade dos sinistros e quantificar a intensidade da sua influência através da interpretação dos coeficientes estiamos e dos *odds* ratio. A Tabela A3 apresenta os resultados do ajustamento.

Tabela A 3 - ROSE: Modelo múltiplo de regressão logística ajustado para a existência de "Mortes/Feridos Graves" nos sinistros com vítimas.

Variável	Coeficiente	Std. Error	P-value
Intercept	-1,1623	0,1509	<0,001
Concelho2AGSSP (ALCOCHETE, GRÂNDOLA, SEIXAL, SINES e PALMELA)	-0,2258	0,0939	0,0162
Concelho2ABMMS (ALMADA, BARREIRO, MOITA, MONTIJO e SESIMBRA)	-0,4695	0,0956	<0,001
Concelho2SS (SANTIAGO DO CACÉM e SETÚBAL)	-0,6858	0,1068	<0,001
tipoacidColisão	-1,7315	0,0968	<0,001
tipoacidDespiste	-0,9993	0,1000	<0,001
tipolocal2Fora das localidades	0,5379	0,0386	<0,001
tipovia2EM – Estrada Municipal	0,2764	0,1107	0,0126
tipovia2EN/IC/ER (Estrada Nacional, Itinerário Complementar e Estrada Regional)	1,0634	0,0553	<0,001
horaacid1new6h	0,7641	0,1106	<0,001
horaacid1new8h-13h	-0,2932	0,0375	<0,001

Variável	Coeficiente	Std. Error	P-value
fugaSim	-1,5915	0,0990	<0,001
PercCondMCat2[75,100]	0,2060	0,0362	<0,001
HaVeicPesadoSim	0,9979	0,0614	<0,001
HaVeicLigSim	0,3957	0,0695	<0,001
HaVeicMotoSim	2,8027	0,0590	<0,001
HoraLaboralSim	-0,3469	0,0352	<0,001
MedianaIdadeVeic	0,0493	0,0026	<0,001
ig_ponderado	0,0833	0,0058	<0,001
tipovia2EM – Estrada Municipal:HaVeicMotoSim	-0,3016	0,1705	0,0768
tipovia2EN/IC/ER:HaVeicMotoSim	-0,6376	0,0809	<0,001
Concelho2AGSSP:ig_ponderado	-0,0325	0,0052	<0,001
Concelho2ABMMS:ig_ponderado	-0,0528	0,0051	<0,001
Concelho2SS:ig_ponderado	-0,0061	0,0058	0,2969
tipoacidColisão: ig_ponderado	-0,0226	0,0028	<0,001
tipoacidDespiste:ig_ponderado	-0,0118	0,0031	<0,001
tipovia2EM – Estrada Municipal:ig_ponderado	0,0301	0,0052	<0,001
tipovia2EN/IC/ER: ig_ponderado	-0,0023	0,0013	0,0833

A análise dos coeficientes resultantes do modelo de regressão logística fornece informação detalhada sobre os fatores que influenciam mais e menos a gravidade dos sinistros rodoviários. Os dados extraídos não apenas confirmam algumas suposições, mas também revelam nuances sobre como certas variáveis interagem para afetar os desfechos dos sinistros.

As variáveis com níveis de significância mais elevados para o modelo são:

- Presença de Veículos Motociclos (“HaVeicMoto”): esta é a variável que mais aumenta a probabilidade de um sinistro emergir em “Mortes/Feridos Graves”;
- Presença de Veículos Pesados (“HaVeicMoto”): a presença de veículos pesados também eleva consideravelmente o risco de sinistros com “Mortes/Feridos Graves”;
- Tipo de via (“tipovia2EN/IC/ER”): sinistros em estradas nacionais, itinerários complementares ou estradas regionais são mais propensos de resultar em “Mortes/Feridos Graves”;

Porém, existem variáveis com coeficientes negativos o que reduz a probabilidade de “Mortes/Feridos Graves”. Essas variáveis são:

- Tipo de Sinistro (Colisão e Despiste): ambos os tipos de sinistros têm uma probabilidade de resultar em “Mortes/Feridos Graves”;
- Concelho (“Concelho2ABMMS”): sinistros que ocorrem nos concelhos de Almada, Barreiro, Moita, Montijo e Sesimbra tendem a apresentar uma menor probabilidade de “Mortes/Feridos Graves”;
- Fuga do Condutor (fugaSim): em sinistros onde o condutor foge, a probabilidade de “Mortes/Feridos Graves” é menor.

Tais resultados fornecem informações valiosas sobre quais os fatores são mais relevantes para prever a gravidade dos sinistros.

3) Avaliação do Modelo

Para avaliar o desempenho do modelo, voltamos a utilizar o conjunto de dados de teste. Uma análise mais detalhada é facilitada pela matriz de confusão, Tabela A4, que oferece uma visão abrangente das previsões realizadas pelo modelo em comparação com as categorias.

Tabela A 4 - ROSE: Métricas de avaliação da Regressão Logística para 85000 observações

Métrica	Resultado	Observação
Ponto de Corte	0,505	Valor que separa as observações em duas categorias.
Accuracy	0,7891	O modelo classifica corretamente 78,91% das observações.

Métrica	Resultado	Observação
IC (95%)	(0,7840; 0,7941)	Intervalo de Confiança de 95% para a *accuracy*.
Kappa	0,5782	O modelo sugere um desempenho razoável.
Mcnemar's Test P-Value	0,2925	Reflete uma diferença significativa entre as taxas de erro de classificação nas duas categorias.
Sensibilidade	0,7892	O modelo identificou corretamente, aproximadamente, 78,92% dos casos Mortes/Feridos Graves.
Especificidade	0,7891	O modelo identificou corretamente, aproximadamente, 78,91% dos casos de Feridos Leves.
Valor Preditivo Positivo	0,7940	Das observações classificadas como positivas pelo modelo, 79,40% são verdadeiras positivas.
Valor Preditivo Negativo	0,7842	Das observações classificadas como negativas pelo modelo,78,42% são verdadeiras negativas.
F1-Score	0,7866	Bom desempenho do modelo.
AUC	0,8709	O modelo tem uma boa capacidade de discriminação.
Precisão	0,7940	Aproximadamente 79,40% das observações classificadas como positivas são mesmo positivas.

4) Comparação do desempenho entre os modelos de classificação

Por último, será realizada uma análise comparativa do desempenho dos diferentes modelos de classificação implementados no estudo com o modelo de regressão logística estatístico. Essa comparação tem como objetivo avaliar a eficácia de cada modelo com base em métricas relevantes, cujos valores encontram-se detalhados na Tabela A5. Novamente, os resultados incluem indicadores como *accuracy*, sensibilidade, especificidade, AUC, Kappa e valores preditivos, que são fundamentais para perceber a capacidade preditiva de cada abordagem. Além disso, o teste de McNemar foi utilizado para identificar diferenças estatísticas nos erros de classificação entre os modelos. Por fim, serão discutidos os principais pontos fortes e limitações de cada abordagem, permitindo uma visão clara sobre qual modelo apresenta o melhor desempenho e sob quais condições.

Tabela A 5 - ROSE: Métricas de classificação para 85000 observações.

Métrica	Regressão Logística	XGBoost	Random Forest	Bayes	C5.0
	OVERSAMPLING ROSE – 85000				
Ponto de Corte	0,505	0,635	0,588	0,504	0,744
Accuracy	0,7891	0,9553	0,9216	0,7443	0,9999
IC (95%)	(0,7840; 0,7941)	(0,9526; 0,9578)	(0,9182; 0,9248)	(0,7388; 0,7496)	(0,9997; 1)
Kappa	0,5782	0,9105	0,8431	0,4885	0,9998
Mcnemar's Test P-Value	0,2925	0,6777	0,8753	0,2483	0,2482
Sensibilidade	0,7892	0,9553	0,9230	0,7444	1
Especificidade	0,7891	0,9552	0,9201	0,7442	0,9998
Valor Preditivo Positivo	0,7940	0,9565	0,9224	0,7498	0,9998
Valor Preditivo Negativo	0,7842	0,9540	0,9206	0,7386	1
F1-score	0,7866	0,9546	0,9203	0,7414	0,9999
AUC	0,8709	0,9893	0,9788	0,8139	0,9999
Precisão	0,7940	0,9565	0,9224	0,7498	0,9998

Os resultados evidenciam diferenças significativas no desempenho dos modelos avaliados.

- Desempenho geral

O C5.0 destaca-se como o algoritmo mais robusto na maioria das métricas analisadas, alcançando desempenho ideal em métricas como *accuracy* (99,99%), sensibilidade (100%), especificidade (99,98%) e AUC (0,9999). Esses valores refletem uma capacidade preditiva ideal, com equilíbrio absoluto entre a deteção de verdadeiros positivos e a exclusão de falsos positivos.

Modelos baseados em árvores, como XGBoost e *Random Forest*, também apresentam desempenhos notáveis. O XGBoost, com *accuracy* de 95,53% e AUC de 0,9893, foi o

segundo melhor modelo, seguido pelo *Random Forest*, com *accuracy* de 92,16% e AUC de 0,9788. Ambos os modelos apresentaram F1-scores elevados, indicando um bom equilíbrio entre a sensibilidade e o valor preditivo positivo.

Em contraste, os modelos de Regressão Logística Estatístico e Naive Bayes apresentaram desempenhos mais modestos. O *accuracy* da Regressão Logística foi de 78,91%, com AUC de 0,8709, enquanto o Bayes apresentou menores valores em várias métricas, com *accuracy* de 74,43% e AUC de 0,8139. Esses resultados sugerem que ambos os modelos podem não ser adequados para conjuntos de dados complexos ou com alta variabilidade.

- Teste de McNemar

O Teste de McNemar avalia a significância estatística das diferenças entre os erros de classificação dos modelos. Nenhum dos valores de *p-value* ($p > 0,05$) indicou diferenças estatisticamente significativas nos erros cometidos pelos modelos. Isso implica que, apesar das métricas sugerirem variações de desempenho, não há evidencias estatísticas de que os modelos diferem substancialmente na classificação de casos discordantes.

- Sensibilidade e Especificidade

Os valores ideias alcançados pelo C5.0 em sensibilidade e especificidade refletem a sua capacidade de identificar casos positivos sem gerar falsos. O XGBoost e o *Random Forest* também apresentaram equilíbrio entre as métricas, com valores acima de 92% para ambos. Já a Regressão Logística e o Bayes apresentaram menor equilíbrio, evidenciando limitações na separação das categorias.

- F_1-*score* e AUC

O *F1-score* de C5.0 confirma o desempenho ideal, enquanto o XGBoost e o *Random Forest* mostraram forte capacidade de classificação com valores de 0,9546 e 0,9203, respetivamente. Por outro lado, os modelos probabilísticos que tiveram F_1-*scores* inferiores, refletiram maior dificuldade em equilibrar a sensibilidade e precisão.

Diferentes graus de desequilíbrio

Foi realizado o *oversampling* na categoria minoritária, aplicando diferentes níveis de geração de observações para analisar o comportamento dos dados sob diferentes graus de desequilíbrio. Para isso, foram gerados quatro cenários distintos – 5000, 15000, 25000 e 35000 observações – mantendo-se o desequilíbrio entre as categorias em diferentes intensidades. Essas variações permitem comparar o desempenho dos

modelos com diferentes proporções de dados desequilibrados, possibilitando uma análise detalhada de como o comportamento do modelo se ajusta conforme aumenta ou diminui o desequilíbrio nos dados.

Na Tabela A6 estão presentes os resultados obtidos nas diferentes métricas de avaliação, nos diferentes quatro cenários.

Tabela A 6 - ROSE: Métricas de classificação para diferentes graus de desequilíbrio.

	Oversampling ROSE				
	Regressão Logística	**XGBoost**	**Random Forest**	**Bayes**	**C5.0**
Ponto de Corte					
5000	0,129	0,192	0,116	0,131	0,414
15000	0,281	0,423	0,308	0,283	0,681
25000	0,384	0,537	0,440	0,38	0,745
35000	0,460	0,602	0,542	0,465	0,786
Accuracy					
5000	0,7886	0,9313	0,9138	0,7464	0,982
15000	0,7918	0,9491	0,912	0,7458	0,9994
25000	0,7909	0,9526	0,9131	0,7435	0,9998
35000	0,7887	0,9538	0,9162	0,7443	0,9998
Kappa					
5000	0,3821	0,7395	0,6851	0,306	0,9239
15000	0,5322	0,8773	0,7176	0,4396	0,9986
25000	0,5693	0,9008	0,8187	0,474	0,9996
35000	0,5763	0,9073	0,8318	0,4873	0,9996

	Oversampling ROSE				
	Regressão Logística	**XGBoost**	**Random Forest**	**Bayes**	**C5.0**
McNemar´s Tes P-Value					
5000	<0,001	<0,001	<0,001	<0,001	<0,001
15000	<0,001	<0,001	<0,001	<0,001	0,0044
25000	<0,001	<0,001	<0,001	<0,001	0,1336
35000	<0,001	0,0243	0,002	<0,001	0,0736
Sensibilidade					
5000	0,7895	0,9314	0,9154	0,7483	0,9823
15000	0,7934	0,9492	0,9130	0,7464	1
25000	0,7924	0,9532	0,9138	0,7438	1
35000	0,7896	0,9539	0,9164	0,7446	1
Especificidade					
5000	0,7885	0,9313	0,9136	0,7462	0,9820
15000	0,7911	0,9490	0,9117	0,7455	0,9992
25000	0,7899	0,9523	0,9127	0,7433	0,9997
35000	0,7879	0,9538	0,9161	0,7440	0,9996
Valor Preditivo Positivo					
5000	0,3580	0,6694	0,6128	0,3058	0,8907
15000	0,5992	0,8799	0,8026	0,5358	0,9980
25000	0,7046	0,9267	0,8688	0,6469	0,9995
35000	0,7643	0,9473	0,9048	0,7170	0,9995
Valor Preditivo Negativo					
5000	0,9616	0,9891	0,9863	0,9520	0,9973
15000	0,9068	0,9794	0,9638	0,8819	1

	Oversampling ROSE				
	Regressão Logística	**XGBoost**	**Random Forest**	**Bayes**	**C5.0**
25000	0,8575	0,9698	0,9437	0,8210	1
35000	0,8113	0,9596	0,9264	0,7698	1
F1-Score					
5000	0,8665	0,9593	0,9486	0,8366	0,9896
15000	0,8450	0,9640	0,9370	0,8080	0,9996
25000	0,8223	0,9610	0,9279	0,7802	0,9998
35000	0,7994	0,9567	0,9212	0,7567	0,9998
AUC					
5000	0,8744	0,9803	0,975	0,8161	0,996
15000	0,8731	0,9871	0,9745	0,8144	0,9999
25000	0,8723	0,9887	0,9762	0,8134	1
35000	0,8711	0,9888	0,9773	0,8136	0,9998
Precisão					
5000	0,3580	0,6694	0,6071	0,3058	0,8907
15000	0,5992	0,8514	0,8026	0,5358	0,9979
25000	0,7046	0,9269	0,8688	0,6469	0,9995
35000	0,7643	0,9473	0,9048	0,7170	0,9995

Os resultados mostram que o algoritmo C5.0 destaca-se como o mais eficiente em praticamente todas as métricas e cenários. No que diz respeito ao *accuracy*, o C5.0 atinge valores de 0,9820 a 0,9998 em todos os cenários, indicando uma elevada precisão geral. O Kappa também reforça o desempenho superior do C5.0, com valores muito próximos nos quatro cenários, com valores entre 0,9239 e 0,9996. Isso sugere que este algoritmo apresenta maior confiabilidade ao distinguir casos positivos e negativos.

Outro ponto relevante é a sensibilidade, onde o C5.0 novamente obtém os melhores resultados. No cenário de 5000 observações apresenta um valor de 0,9823, e a partir de

15000 até 35000 observações amétrica atinge o valor de 1, evidenciando a capacidade de o algoritmo identificar corretamente os casos positivos. A mesma tendência é observada na especificidade, que varia entre 0,9920 e 0,9997, permanecendo sempre muito próximo de 1.

Adicionalmente, a AUC também indica o C5.0 como a melhor escolha. Os valores vão de 0,9960 a 1, refletindo um desempenho praticamente ideal em termos de discriminação entre categorias. Outras métricas, como F_1-*score* (0,9896 a 0,9998), Valores Preditivos Negativos (0,9973 a 1) e Valores Preditivos Positivos (0,8907 a 0,9995), reforçam a superioridade do C5.0 em comparação aos outros algoritmos.

Ao comparar os diferentes algoritmos, observa-se que, embora o XGBoost e o *Random Forest* apresentem bons desempenhos (com *accuracy* superior a 0,9100 e AUC acima de 0,9700), eles não atingem o mesmo nível de eficiência do C5.0, especialmente em métricas como Kappa e sensibilidade. Por outro lado, a Regressão Logística e o Bayes demonstram desempenhos inferiores. A Regressão Logística mantém a *accuracy* em torno de 0,7900, F_1-*score* de 0,8000 a 0,8700 e a precisão só melhora significativamente com maior oversampling. O Bayes mostra resultados mais baixos em *accuracy* e Kappa, limitando a sua eficácia.

Em relação aos cenários analisados, os resultados indicam que o desempenho melhora de forma significativa até 25000 observações. Nos cenários de 15000 e 25 observações o C5.0 atinge valores ideais, demonstrado a sua capacidade de generalização com dados mais robustos. No entanto, no cenários com 35000 observações, não há um ganho expressivo em relação aos de 25000 observações.

Dessa forma, os resultados apontam que o algoritmo C5.0, especialmente nos cenários de 15000 e 25000 observações, é a melhor opção para a previsão de sinistros com "Mortes/Feridos Graves", superando consistentemente os outros modelos avaliados.

Undersampling + Oversampling

Foi adotada uma abordagem combinada de *undersampling* e *oversampling* para lidar com os dados desequilibrados das categorias. O objetivo é ajustar a quantidade de observações para que um modelo seja composto por categorias equilibradas. Posto isto, inicialmente foi criado um modelo com 42000 observações, onde 21000 correspondem à categoria minoritária e 21000 à categoria maioritária.

São ainda criados outros dois modelos, um com 10000 observações (5000 categoria minoritária + 5000 categoria maioritária) e outro modelo com 20000 observações (10000 categoria minoritária + 10000 categoria maioritária), ambos com categorias equilibradas. Ao contrário da abordagem anterior, que focava unicamente no *oversampling*, aqui iniciou-se com a aplicação do *undersampling* à categoria maioritária, removendo uma parte das observações dos sinistros "Feridos Leves". Em seguida, foi aplicado o *oversampling* à categoria minoritária, aumentando o número de observações dos sinistros "Mortes/Feridos Graves" por meio da replicação ou criação de dados sintéticos. Assim, assegura-se que ambas as categorias estão balanceadas nos dois modelos, permitindo que os algoritmos de *machine learning* trabalhem com dados mais equilibrados.

Esta abordagem permite avaliar e sintetizar um número mais reduzido de dados da categoria minoritária que poderá ajudar na melhoria do desempenho dos algoritmos.

Modelo com 42000 observações

- ***Undersampling***

Inicialmente, aplicou-se a técnica de *undersampling* para equilibrar a base de dados, reduzindo o número de observações da categoria maioritária para aproximar-se da categoria minoritária. O processo foi conduzido da seguinte forma:

- Categoria minoritária ("Mortes/Feridos Graves"): todos os 995 sinistros com mortes/feridos graves foram mantidos na base de dados sem alterações.
- Categoria maioritária ("Feridos Leves"): foi realizada uma amostragem aleatória simples, sem reposição, dos 42317 sinistros com Feridos Leves.

A partir dessa abordagem, criou-se um conjunto de proporções controladas, onde a categoria maioritária passou a ter 21,11 vezes o número de observações da categoria minoritária, resultando num total de 21999 observações, sendo 21004 da categoria maioritária e 995 observações da categoria minoritária.

- ***Oversampling***

Após o *undersampling*, aplicou-se o método de *oversampling* à categoria minoritária para aumentar a representatividade de "Mortes/Feridos Graves" ao gerar novos dados sintéticos, resultando num aumento significativo no número de observações desta

mesma categoria. No final deste processo a base de dados passou a ter categorias mais balanceadas, evitando que a categoria maioritária dominasse o modelo. Os valores de cada categoria encontram-se representados na Tabela A7.

Tabela A 7 - ROSE: Composição do modelo com e sem undersampling e com undersampling+oversampling de 42000 observações.

	ROSE – 42000 observações	
	Feridos Leves	Mortes/Feridos Graves
Modelo Simples	42317	995
Modelo com Undersampling	21004	995
Modelo com Under + Oversampling	21004	20996

- **Divisão dos dados em treino e teste**

De seguida, os conjuntos de dados foram dividido em dois subconjuntos: 70% dos dados foram alocados para treino e 30% para teste. Neste sentido a divisão, Tabela A8, encontra-se da seguinte forma:

- Conjunto de treino: contém 29595 observações.

- Conjunto de teste: contém 12405 observações.

Tabela A 8 - ROSE: Divisão dos dados do modelo de regressão logística (42000 observações) em dois subconjuntos: treino e teste e respetivo número de observações por categoria em cada subconjunto.

	ROSE UNDER + OVER 42000 observações	
	Treino	**Teste**
Feridos Leves	14793	6211
Mortes/Feridos Graves	14802	6194
	29595	12405

- **Ajustamento do Modelo**

Após a preparação dos dados, procedeu-se ao ajustamento do modelo Regressão Logística Estatístico. Os resultados obtidos estão apresentados na Tabela A9.

Tabela A 9 - ROSE: Modelo múltiplo de regressão logística ajustado para a existência de "Mortes/Feridos Graves" nos sinistros com vítimas de 42000 observações.

Variável	Coeficiente	Std. Error	P-value
Intercept	-1,0514	0,2204	<0,001
Concelho2AGSSP (ALCOCHETE, GRÂNDOLA, SEIXAL, SINES e PALMELA)	-0,3405	0,1408	0,0156
Concelho2ABMMS (ALMADA, BARREIRO, MOITA, MONTIJO e SESIMBRA)	-0,5394	0,1426	<0,001
Concelho2SS (SANTIAGO DO CACÉM e SETÚBAL)	-0,6777	0,1557	<0,001
tipoacidColisão	-1,7865	0,1336	<0,001
tipoacidDespiste	-0,9917	0,1386	<0,001
tipolocal2Fora das localidades	0,4704	0,0550	<0,001
tipovia2EM – Estrada Municipal	0,2596	0,1634	0,1122
tipovia2EN/IC/ER (Estrada Nacional, Itinerário Complementar e Estrada Regional)	1,0254	0,0802	<0,001
horaacid1new6h	0,7128	0,1559	<0,001
horaacid1new8h-13h	-0,2945	0,0533	<0,001
fugaSim	-1,6045	0,1410	<0,001
PercCondMCat2[75,100]	0,2859	0,0519	<0,001
HaVeicPesadoSim	1,0398	0,0883	<0,001

Variável	Coeficiente	Std. Error	P-value
HaVeicLigSim	0,4219	0,0975	<0,001
HaVeicMotoSim	2,7397	0,0837	<0,001
HoraLaboralSim	-0,3566	0,0505	<0,001
MedianaIdadeVeic	0,0447	0,0037	<0,001
ig_ponderado	0,0833	0,0085	<0,001
tipovia2EM – Estrada Municipal: HaVeicMotoSim	-0,11298	0,2396	0,6378
tipovia2EN/IC/ER: HaVeicMotoSim	-0,6632	0,1151	<0,001
Concelho2AGSSP: ig_ponderado	-0,0267	0,0076	<0,001
Concelho2ABMMS: ig_ponderado	-0,0502	0,0075	<0,001
Concelho2SS: ig_ponderado	-0,0064	0,0084	0,4474
tipoacidColisão: ig_ponderado	-0,0265	0,0042	<0,001
tipoacidDespiste: ig_ponderado	-0,0174	0,0046	<0,001
tipovia2EM – Estrada Municipal: ig_ponderado	0,0385	0,0090	<0,001
tipovia2EN/IC/ER: ig_ponderado	0,0004	0,0019	0,8489

A análise dos coeficientes resultantes do modelo de regressão logística fornece informações detalhadas sobre os fatores que influenciam mais e menos a gravidade dos sinistros rodoviários. Os dados extraídos não apenas confirmam algumas suposições, mas também revelam nuances sobre como certas variáveis interagem para afetar os desfechos dos sinistros.

As variáveis com níveis de significância mais elevados para o modelo são:

- Presença de Veículos Motociclos (“HaVeicMoto”): é a variável que mais aumenta a probabilidade de um sinistro emergir em “Mortes/Feridos Graves”, multiplicando em mais de 15 vezes as probabilidades de gravidade;

- Presença de Veículos Pesados (“HaVeicMoto”): a presença de veículos pesados também eleva consideravelmente o risco de sinistros com “Mortes/Feridos Graves”. Esta variável eleva o risco em aproximadamente 2,8 vezes;
- Tipo de via (“tipovia2EN/IC/ER”): sinistros em estradas nacionais, itinerários complementares ou estradas regionais são mais propensos de resultar em “Mortes/Feridos Graves”. A mesma aumenta a probabilidade em cerca de 2,8 vezes.

Porém, existem algumas variáveis com coeficientes negativos o que reduz a probabilidade de “Mortes/Feridos Graves”. Essas variáveis são:

- Tipo de Sinistro (Colisão): a colisão apresenta cerca de 83% menos de probabilidade de gravidade em comparação com os outros tipos de sinistros;
- Tipo de Sinistro (Despiste): está associado a uma redução de cerca de 63% no risco de gravidade;
- Fuga do Condutor (fugaSim): em sinistros onde o condutor foge, a probabilidade de “Mortes/Feridos Graves” é aproximadamente 80% menor.

Além destes fatores principais, o modelo também destaca:

- Efeitos geográficos: alguns concelhos, como ABMMS, AGSSP e SS apresentam menor risco;
- Horário: sinistros às 6h da manhã duplicam o risco de gravidade, enquanto que entre as 8h e 13h reduzem as probabilidades;
- Idade do veículo: cada ano adicional da idade média aumenta o risco em cerca 4,6%;
- Perfil dos condutores: maior proporção de condutores jovens também aumenta a gravidade.

Tais resultados indicam que fatores relacionados ao tipo de veículo (principalmente motociclos e pesados), às caraterísticas da via e ao horário do sinistro são determinantes para o aumento da gravidade dos sinistros, enquanto o tipo de sinistro e o comportamento de fuga estão associados a uma redução desse risco.

- **Comparação do desempenho entre os modelos de classificação**

Na última etapa desta análise, será realizada comparação detalhada entre os modelos de classificação desenvolvidos. O desempenho de cada modelo será avaliado com base nas métricas apresentadas na Tabela A10.

Tabela A 10 - ROSE: Métricas de classificação para 42000 observações – Undersampling + Oversampling.

Métrica	Regressão Logística	XGBoost	Random Forest	Bayes	C5.0
	UNDER + OVER – ROSE 42000				
Ponto de Corte	0,502	0,631	0,59	0,504	0,734
Accuracy	0,7911	0,9507	0,9152	0,7460	0,9977
IC (95%)	(0,7838; 0,7982)	(0,9468; 0,9545)	(0,9102; 0,9200)	(0,7382; 0,7536)	(0,9967; 0,9985)
Kappa	0,5821	0,9015	0,8304	0,4920	0,9955
Mcnemar's Test P-Value	0,6800	1	0,9263	0,9432	0,1859
Sensibilidade	0,7925	0,9508	O,9154	0,7460	0,9984
Especificidade	0,7896	0,9507	0,9150	0,7459	0,9971
Valor Preditivo Positivo	0,7897	0,9506	0,9148	0,7454	0,9971
Valor Preditivo Negativo	0,7924	0,9509	0,9156	0,7465	0,9984
F1-score	0,7910	0,9508	0,9153	0,7462	0,9977
AUC	0,8709	0,9869	0,9752	0,8144	0,9997
Precisão	0,7897	0,9506	0,9148	0,7454	0,9971

A análise considera várias métricas de desempenho para determinar o modelo mais eficiente. De seguida é discutido as métricas mais relevantes:

1) *Accuracy*

O C5.0 apresenta o melhor desempenho geral, com *accuracy* quase perfeito, indicando alta confiabilidade nas previsões. O XGBoost e o *Random Forest* também de destacam,

mantendo níveis elevados. A Regressão Logística e o Bayes têm desempenhos moderados.

2) Kappa

O Kappa no C5.0 é de 0,9955 muito próximo ao ideal. O XGBoost e o *Random Forest* também tiveram um bom desempenho, porém não tão elevado como o anterior. Modelos mais simples como a Regressão Logística e o Bayes apresentam concordâncias moderadas, sendo 0,5821 e 0,4920, respetivamente.

3) Sensibilidade

A sensibilidade de C5.0 foi 0,9984, mostrando que quase todos os sinistros foram corretamente identificados. O XGBoost (0,9508) e o *Random Forest* (0,9154) também detetam a maioria dos sinistros graves, enquanto a Regressão Logística (0,7925) e o Bayes (0,7460) deixam de identificar uma parte significativa desses casos.

4) Especificidade

O modelo C5.0 apresenta uma especificidade muito boa (0,9971), minimizando falsos positivos e garantindo alta confiabilidade nas previsões de Feridos Leves. O XGBoost (0,9507) e Random Forest (0,9150) também têm boa precisão, embora com desempenho ligeiramente inferior a C5.0. Em contraste, a Regressão Logística (0,7896) e o Bayes (0,7459) mostram especificidade limitada, indicando maior propensão a falsos positivos.

5) Valor Preditivo Positivo

No Valor Preditivo Positivo temos o C5.0 com 0,9971, garantindo alta confiabilidade ao prever “Mortes/Feridos Graves”. O XGBoost e o *Random Forest* mantêm os níveis sólidos de precisão de “Mortes/Feridos Graves”, enquanto a Regressão Logística (0,7897) e o Bayes (0,7462) apresentam um desempenho mais limitado.

6) Valor Preditivo Negativo

Em relação ao Valor Preditivo Negativo, o C5.0 também lidera com 0,9984, praticamente eliminado falsos negativos e garantindo a correta identificação de Feridos Leves. O

XGBoost (0,9509) e o *Random Forest* (0,9156) mantêm alta confiabilidade, mas a Regressão Logística (0,7924) e o Bayes (0,7465) têm resultados mais fracos.

7) F1-score

No F1-score, o C5.0 é novamente superior (0,9977), evidenciando um equilíbrio ideal entre sensibilidade e precisão. O XGBoost (0,9508) e o Random Forest (0,9153) também são consistentes, quanto a Regressão Logística (0,7910) e o Bayes (0,7462) apresentam desempenhos moderados.

8) AUC

Na AUC, o C5.0 atinge um valor muito próximo do ideal (0,9997), evidenciando a sua capacidade de discriminar entre Mortes/Feridos Graves e Feridos Leves. O XGBoost (0,9869) e o *Random Forest* (0,9752) também são excelentes, enquanto a Regressão Logística (0,8709) e o Bayes (0,8144) têm desempenhos aceitáveis, porém inferiores.

9) Precisão

No que diz respeito à precisão, todos os modelos apresentam valores semelhantes, variando minimamente entre 0,7454 e 0,9971. Esses valores mostram que há dificuldade em garantir que os casos previstos como “Mortes/Feridos Graves” sejam realmente “Mortes/Feridos Graves”.

Após a análise detalhada de todas as métricas avaliadas, é possível identificar o modelo mais adequado para o objetivo proposto, considerando tanto o seu desempenho geral quanto a sua capacidade de prever “Mortes/Feridos Graves” com precisão e confiabilidade. Entre os modelos estudados, o C5.0 destaca-se como a melhor escolha, apresentando um excelente desempenho nas diferentes métricas. Embora o XGBoost e o *Random Forest* também mostrem um bom desempenho, ambos ficam ligeiramente atrás do C5.0 em termos de precisão e equilíbrio geral. Por outro lado, a Regressão Logística e o Bayes apresentam um desempenho mais limitado, com métricas mais baixas em sensibilidade, especificidade e outros indicadores, tornando-os menos indicados para o objetivo do estudo.

Modelos com 10000 e 20000 observações

1) ***Undersampling***

Procedeu-se ao pré-processamento dos dados utilizando novamente a técnica de *undersampling*, seguindo um procedimento semelhante ao realizado anteriormente para as 42000 observações, mas desta vez considerando um menor número de observações:

- Categoria minoritária ("Mortes/Feridos Graves"): todos os 995 sinistros com mortes/feridos graves foram mantidos na base de dados sem alterações.
- Categoria maioritária ("Feridos Leves"): foi realizada uma amostragem aleatória simples, sem reposição, dos 42317 sinistros com feridos leves.

Neste âmbito, foram criados dois conjuntos a partir da categoria maioritária, com proporções controladas em relação à categoria minoritária:

i. Conjunto 1: foram selecionados aleatoriamente sinistros correspondentes a 5.03 vezes o número de observações da categoria minoritária.
ii. Conjunto 2: um segundo conjunto foi criado com sinistros correspondentes a 5.03
vezes o número de observações da categoria minoritária.

1) ***Oversampling***

Após o *undersampling*, aplicou-se o método de *oversampling* à categoria minoritária para aumentar a representatividade de Mortes/Feridos Graves ao gerar novos dados sintéticos, resultando num aumento significativo no número de observações desta mesma categoria. No final deste processo a base de dados passou a ter categorias mais balanceadas, evitando que a categoria maioritária dominasse o modelo. Os valores de cada categoria encontram-se representados na Tabela A11.

Tabela A 11 - ROSE: Valores das categorias do modelo simples, com undersampling e com undersampling+oversampling.

	10000		20000	
	Feridos Leves	Mortes / Feridos Graves	Feridos Leves	Mortes / Feridos Graves
Modelo Simples	42317	995	42317	995
Modelo com *Undersampling*	5004	995	9999	995
Modelo com Under + Over Sampling	5004	4996	9999	10001

2) <u>Divisão dos dados em treino e teste</u>

De seguida, os conjuntos de dados foram dividido em dois subconjuntos: 70% dos dados foram alocados para treino e 30% para teste. Neste sentido a divisão com 10000 observações, Tabela A12, encontra-se da seguinte forma:

- Conjunto de treino: contém 7048 observações.
- Conjunto de teste: contém 2952 observações.

Tabela A 12 - ROSE: Divisão dos dados do modelo de regressão logística (10000 observações) em dois subconjuntos: treino e teste e respetivo número de observações por categoria em cada subconjunto.

	ROSE *UNDER* + *OVER* 10000	
	Treino	**Teste**
Feridos Leves	3515	1489
Mortes/Feridos Graves	3533	1463
	7048	2952

A divisão com 20000 observações, Tabela A13, centra-se:

- Conjunto de treino: contém 14080 observações.
- Conjunto de teste: contém 5920 observações.

Tabela A 13 - ROSE: Divisão dos dados do modelo de regressão logística (20000 observações) em dois subconjuntos: treino e teste e respetivo número de observações por categoria em cada subconjunto

	ROSE *UNDER* + *OVER* 20000	
	Treino	**Teste**
Feridos Leves	3515	1489
Mortes/Feridos Graves	10565	4431
	14080	5920

3) Ajustamento do Modelo

Após a preparação dos dados, realizou-se o ajustamento do modelo Regressão Logística Estatístico para o modelo com 10000 observações e para o modelo com 20000 observações. Os resultados de ambos encontram-se representados nas Tabelas A14 e A15 respetivamente.

Tabela A 14 - ROSE: Modelo múltiplo de regressão logística ajustado para a existência de "Mortes/Feridos Graves" nos sinistros com vítimas, com 10000 observações.

Variável	Coeficiente	Std. Error	P-value
Intercept	0,4270	0,4661	0,3596
Concelho2AGSSP (ALCOCHETE, GRÂNDOLA, SEIXAL, SINES e PALMELA)	-0,6204	0,2681	0,0207
Concelho2ABMMS (ALMADA, BARREIRO, MOITA, MONTIJO e SESIMBRA)	-0,9854	0,2773	0,0004
Concelho2SS (SANTIAGO DO CACÉM e SETÚBAL)	-0,9052	0,3054	0,0030
tipoacidColisão	-2,2456	0,3000	<0,001
tipoacidDespiste	-1,6181	0,3118	<0,001
tipolocal2Fora das localidades	0,5476	0,1146	<0,001
tipovia2EM – Estrada Municipal	0,5335	0,3135	0,0888

Variável	Coeficiente	Std. Error	P-value
tipovia2EN/IC/ER (Estrada Nacional, Itinerário Complementar e Estrada Regional)	0,9100	0,1632	<0,001
horaacid1new6h	0,8759	0,3275	0,0075
horaacid1new8h-13h	-0,3308	0,1099	0,0026
fugaSim	-1,3000	0,3030	<0,001
PercCondMCat2[75,100]	0,0242	0,1072	0,8217
HaVeicPesadoSim	1,0258	0,1788	<0,001
HaVeicLigSim	0,0016	0,2047	0,9936
HaVeicMotoSim	2,4375	0,1723	<0,001
HoraLaboralSim	-0,2051	0,1066	0,0544
MedianaIdadeVeic	0,0320	0,0076	<0,001
ig_ponderado	0,0581	0,0157	0,0002
tipovia2EM – Estrada Municipal: HaVeicMotoSim	0,2955	0,5233	0,5722
tipovia2EN/IC/ER: HaVeicMotoSim	-0,3351	0,2350	0,1539
Concelho2AGSSP: ig_ponderado	-0,0077	0,0130	0,5528
Concelho2ABMMS: ig_ponderado	-0,0301	0,0128	0,0186
Concelho2SS: ig_ponderado	0,0040	0,0148	0,7865
tipoacidColisão: ig_ponderado	-0,0277	0,0095	0,0034
tipoacidDespiste: ig_ponderado	-0,0094	0,0100	0,3449
tipovia2EM – Estrada Municipal: ig_ponderado	0,0238	0,0193	0,2172
tipovia2EN/IC/ER: ig_ponderado	0,0067	0,0038	0,0764

Tabela A 15 - ROSE: Modelo múltiplo de regressão logística ajustado para a existência de "Mortes/Feridos Graves" nos sinistros com vítimas, com 20000 observações.

Variável	Coeficiente	Std. Error	P-value
Intercept	-0,8302	0,3294	0,0117
Concelho2AGSSP (ALCOCHETE, GRÂNDOLA, SEIXAL, SINES e PALMELA)	-0,1630	0,2058	0,4283
Concelho2ABMMS (ALMADA, BARREIRO, MOITA, MONTIJO e SESIMBRA)	-0,3954	0,2082	0,0576
Concelho2SS (SANTIAGO DO CACÉM e SETÚBAL)	-0,4905	0,2295	0,0326
tipoacidColisão	-1,8543	0,2073	<0,001
tipoacidDespiste	-1,0281	0,2140	<0,001
tipolocal2Fora das localidades	0,4659	0,0803	<0,001
tipovia2EM – Estrada Municipal	0,0611	0,2588	0,8133
tipovia2EN/IC/ER (Estrada Nacional, Itinerário Complementar e Estrada Regional)	0,8698	0,1164	<0,001
horaacid1new6h	0,7794	0,2352	0,0009
horaacid1new8h-13h	-0,3308	0,0774	<0,001
fugaSim	-1,4074	0,1931	<0,001
PercCondMCat2[75,100]	0,1693	0,0747	0,0234
HaVeicPesadoSim	0,8393	0,1274	<0,001
HaVeicLigSim	0,2883	0,1433	0,0442
HaVeicMotoSim	2,5182	0,1177	<0,001
HoraLaboralSim	-0,3916	0,0732	<0,001

Variável	Coeficiente	Std. Error	P-value
MedianaIdadeVeic	0,0443	0,0053	<0,001
ig_ponderado	0,0915	0,0133	<0,001
tipovia2EM – Estrada Municipal: HaVeicMotoSim	0,3461	0,3766	0,3582
tipovia2EN/IC/ER: HaVeicMotoSim	-0,4316	0,1642	0,0086
Concelho2AGSSP: ig_ponderado	-0,0350	0,0118	0,0029
Concelho2ABMMS: ig_ponderado	-0,0576	0,0117	<0,001
Concelho2SS: ig_ponderado	0,0143	0,0129	0,2671
tipoacidColisão: ig_ponderado	-0,0286	0,0069	<0,001
tipoacidDespiste: ig_ponderado	-0,0179	0,0073	0,0138
tipovia2EM – Estrada Municipal: ig_ponderado	0,0509	0,0148	0,0006
tipovia2EN/IC/ER: ig_ponderado	0,0030	0,0027	0,2679

Ao estabelecer a comparação entre estes dois modelos e o modelo previamente estimado (modelo de *oversampling* com 85000 observações), verifica-se que o conjunto de variáveis com efeitos estatisticamente significativos - tanto positivos quanto negativos – revela-se estável e consistente.

4) Comparação do desempenho entre os modelos de classificação

Na última etapa desta análise, será realizada a comparação detalhada entre os modelos de classificação desenvolvidos. O desempenho de cada modelo será avaliado com base nas métricas apresentadas na Tabela A16.

Tabela A - ROSE: Métricas de classificação para 10000 e 20000 observações – Undersampling + Oversampling.

	Regressão Logística	XGBoost	Random Forest	Bayes	C5.0
	UNDER + OVER ROSE				
Ponto de Corte					
10000	0,488	0,640	0,580	0,525	0,618
20000	0,499	0,619	0,602	0,531	0,709
Accuracy					
10000	0,7815	0,9231	0,8774	0,7324	0,9295
20000	0,7875	0,9395	0,9025	0,7373	0,9878
IC (95%)					
10000	(0,7662; 0,7963)	(0,9129; 0,9325)	(0,8650; 0,8890)	(0,7160; 0,7483)	(0,9197; 0,9385)
20000	(0,7769; 0,7979)	(0,9332; 0,9455)	(0,8947; 0,9100)	(0,7259; 0,7485)	(0,9847; 0,9905)
Kappa					
10000	0,5630	0,8462	0,7547	0,4648	0,8591
20000	0,5750	0,8791	0,8051	0,4747	0,9757
McNemar´s Tes P-Value					
10000	0,8132	0,7906	0,4305	0,8033	0,9447
20000	1	0,9579	0,9336	1	1

	Regressão Logística	XGBoost	Random Forest	Bayes	C5.0
	UNDER + OVER ROSE				
Sensibilidade					
10000	0,7820	0,9241	0,8817	0,7327	0,9296
20000	0,7881	0,9400	0,9033	0,7383	0,9879
Especificidade					
10000	0,7811	0,9221	0,8731	0,7320	0,9295
20000	0,7869	0,9390	0,9017	0,7364	0,9878
Valor Preditivo Positivo					
10000	0,7782	0,9210	0,8722	0,7288	0,9283
20000	0,7881	0,9394	0,9024	0,7380	0,9879
Valor Preditivo Negativo					
10000	0,7848	0,9252	0,8826	0,7360	0,9307
20000	0,7869	0,9396	0,9026	0,7366	0,9878
F1-Score					
10000	0,7829	0,9236	0,8778	0,7340	0,9301
20000	0,7869	0,9393	0,9022	0,7365	0,9878

	Regressão Logística	XGBoost	Random Forest	Bayes	C5.0
	UNDER + OVER ROSE				
AUC					
10000	0,8742	0,9709	0,8722	0,8111	0,9738
20000	0,8705	0,9820	0,9688	0,8077	0,9978
Precisão					
10000	0,7803	0,9210	0,8753	0,7288	0,9283
20000	0,7881	0,9394	0,9024	0,7380	0,9879

A análise comparativa dos cinco modelos (Regressão Logística, XGBoost, *Random Forest*, Bayes e C5.0) revela que o C5.0 destaca-se como o modelo mais eficiente para prever "Mortes/Feridos Graves" neste conjunto de dados, especialmente com 20000 observações. Este modelo apresenta superioridade na maioria das métricas avaliadas, incluindo *accuracy*, Kappa, *F1*-score e AUC, indicando uma excelente capacidade de identificação de casos positivos e na discriminação entre as categorias.

Modelos como o XGBoost e *Random Forest* também apresentam um desempenho competitivo, sendo opções secundárias viáveis. Já a Regressão Logística e o modelo de Bayes apresentam limitações significativas, ficando aquém dos modelos de *machine learning*.

Dessa forma, o cenário com 20000 observações utilizando o C5.0, destaca-se como a melhor configuração para prever sinistros com "Mortes/Feridos Graves", oferecendo maior robustez e capacidade de generalização.

Apêndice B - SMOTENC

Para lidar com o desequilíbrio entre as categorias no conjunto de dados, foi aplicada a técnica SMOTENC, uma extensão do SMOTE tradicional, que permite o balanceamento de conjuntos de dados com variáveis contínuas e categóricas. O SMOTENC gera amostras sintéticas para a categoria minoritária, preservando a integridade das variáveis categóricas, o que evita distorções que poderiam ocorrer com o SMOTE tradicional.

Dados equilibrados

Com o objetivo de equilibrar a distribuição das categorias, tal como realizado anteriormente, aplicou-se a técnica SMOTENC. O processo resultou num conjunto de dados equilibrados com um total de 85000 observações, como detalhado na Tabela B1.

Tabela B 1 - SMOTENC: Modelo de regressão logística com e sem oversampling.

	Oversampling Regressão Logística – 85000 Observações	
	Feridos Leves	Mortes / Feridos Graves
Modelo Simples	42317	995
Modelo com Oversampling	42317	42317

2) **Divisão dos dados em treino e teste**

O conjunto de dados foi dividido em treino e teste, com uma proporção de 70% para treino e 30% para teste. A divisão, Tabela B2, resultou nos seguintes subconjuntos:

- Conjunto de treino: 59374 observações
- Conjunto de teste: 25260 observações
-

Tabela B 2 - Divisão dos dados do modelo de regressão logística (85000 observações) em dois subconjuntos: treino e teste e respetivo número de observações por categoria em cada subconjunto.

	Regressão Logística – 85000 Observações	
	Treino	**Teste**
Feridos Leves	29817	12500
Mortes/Feridos Graves	29557	12760
	59374	25260

2) Ajustamento do modelo

O próximo passo foi ajustar um modelo de Regressão Logística. Foram testados diferentes valores de *over_ratio*, sendo o valor utilizado *over_ratio* = 1 e K = 5. Os resultados encontram-se na Tabela B3.

Tabela B 3 - SMOTENC: Modelo múltiplo de regressão logística ajustado para a existência de "Mortes/Feridos Graves" nos sinistros com vítimas.

Variável	Coeficiente	Std. Error	P-value
Intercept	-1,9980	0,1693	<0,001
Concelho2AGSSP (ALCOCHETE, GRÂNDOLA, SEIXAL, SINES e PALMELA)	0,1100	0,1077	0,3068
Concelho2ABMMS (ALMADA, BARREIRO, MOITA, MONTIJO e SESIMBRA)	-0,2050	0,1088	0,0595
Concelho2SS (SANTIAGO DO CACÉM e SETÚBAL)	-0,8017	0,1213	<0,001
tipoacidColisão	-1,3081	0,1086	<0,001
tipoacidDespiste	-0,5936	0,1123	<0,001
tipolocal2Fora das localidades	0,5707	0,0406	<0,001
tipovia2EM – Estrada Municipal	-0,6070	0,1439	<0,001
tipovia2EN/IC/ER (Estrada Nacional, Itinerário Complementar e Estrada Regional)	1,0692	0,0586	<0,001
horaacid1new6h	-0,2612	0,1374	0,0574
horaacid1new8h-13h	-0,4225	0,0390	<0,001
fugaSim	-5,7327	0,5372	<0,001
PercCondMCat2[75,100]	0,3383	0,0378	<0,001

Variável	Coeficiente	Std. Error	P-value
HaVeicPesadoSim	0,7339	0,0657	<0,001
HaVeicLigSim	0,4145	0,0722	<0,001
HaVeicMotoSim	2,7477	0,0597	<0,001
HoraLaboralSim	-0,2976	0,0362	<0,001
MedianaIdadeVeic	0,0544	0,0029	<0,001
ig_ponderado	0,1139	0,0070	<0,001
tipovia2EM – Estrada Municipal: HaVeicMotoSim	-0,1463	0,1803	0,4170
tipovia2EN/IC/ER: HaVeicMotoSim	-0,5750	0,0836	<0,001
Concelho2AGSSP: ig_ponderado	-0,0431	0,0061	<0,001
Concelho2ABMMS: ig_ponderado	-0,0668	0,0061	<0,001
Concelho2SS: ig_ponderado	-0,0023	0,0069	0,7428
tipoacidColisão: ig_ponderado	-0,0380	0,0036	<0,001
tipoacidDespiste: ig_ponderado	-0,0268	0,0039	<0,001
tipovia2EM – Estrada Municipal: ig_ponderado	0,0550	0,0069	<0,001
tipovia2EN/IC/ER: ig_ponderado	-0,0033	0,0014	0,0215

Neste modelo de regressão logística, as variáveis que se destacam com mais e menos impacto na gravidade dos sinistros rodoviários são as mesmas observadas no modelo com a técnica ROSE. O modelo confirma as tendências identificadas anteriormente.

3) Comparação do desempenho entre os modelos de classificação

O processo de análise comparativa será realizado para o modelo SMOTENC, seguindo a mesma abordagem adotada no modelo anterior, a comparação entre os modelos será feita com base nas métricas de desempenho, como *accuracy*, sensibilidade, especificidade, entre outras, cujos resultados estão na Tabela B4.

Tabela B 4 - SMOTENC: Métricas de classificação para 85000 observações.

Métrica	Regressão Logística	XGBoost	Random Forest	Bayes	C5.0
	OVERSAMPLING SMOTENC – 85000				
Ponto de Corte	0,510	0,541	0,584	0,931	0,525
Accuracy	0,7839	0,946	0,8901	0,7586	0,9817
IC (95%)	(0,7788; 0,7890)	(0,9432; 0,9488)	(0,8861; 0,8939)	(0,7532; 0,7638)	(0,98; 0,9834)
Kappa	0,5677	0,8921	0,7801	0,5171	0,9635
Mcnemar's Test P-Value	0,5697	0,7452	0,7327	0,6633	0,9258
Sensibilidade	0,7844	0,9461	0,8904	0,7596	0,9818
Especificidade	0,7834	0,9460	0,8897	0,7574	0,9817
Valor Preditivo Positivo	0,7871	0,9470	0,8918	0,7617	0,9820
Valor Preditivo Negativo	0,7807	0,9450	0,8883	0,7553	0,9814
F1-score	0,7820	0,9455	0,8890	0,7564	0,9816
AUC	0,8670	0,9891	0,9517	0,8237	0,9978
Precisão	0,7871	0,9470	0,8918	0,7617	0,9819

Os resultados evidenciam diferenças significativas no desempenho dos modelos avaliados.

- Desempenho geral

O C5.0 destaca-se como o algoritmo mais robusto na maioria das métricas analisadas, alcançando desempenho ideal em métricas como *accuracy* (98,17%), sensibilidade (98,18%), especificidade (98,17%) e AUC (0,9978). Esses valores refletem uma

capacidade preditiva ideal, com equilíbrio absoluto entre a deteção de verdadeiros positivos e a exclusão de falsos positivos.

Modelos baseados em árvores, como XGBoost e *Random Forest*, também apresentam desempenhos notáveis. O XGBoost, com *accuracy* de 94,6% e AUC de 0,9891, destacou-se como o segundo melhor modelo. Já o *Random Forest* apresentou um *accuracy* de 89,01% e AUC de 0,9517. Ambos os modelos apresentaram F1-scores elevados, indicando um bom equilíbrio entre a sensibilidade e o valor preditivo positivo.

Em contraste, os modelos de Regressão Logística e Naive Bayes apresentaram desempenhos mais modestos. O *accuracy* da Regressão Logística foi de 78,39%, com AUC de 0,870, enquanto o Bayes apresentou menores valores em várias métricas, incluindo um accuracy de 75,86% e AUC de 0,8237. Esses resultados sugerem que ambos os modelos podem não ser adequados para conjuntos de dados complexos ou com alta variabilidade.

- Teste de McNemar

O Teste de McNemar avalia a significância estatística das diferenças entre os erros de classificação dos modelos. Nenhum dos valores de *p-value* ($p > 0,05$) indicou diferenças estatisticamente significativas nos erros cometidos pelos modelos. Isso implica que, apesar das métricas sugerirem variações de desempenho, não há evidencias estatísticas de que os modelos diferem substancialmente na classificação de casos discordantes.

- Sensibilidade e Especificidade

Os valores ideias alcançados pelo C5.0 em sensibilidade e especificidade refletem a sua capacidade de identificar casos positivos sem gerar falsos. O XGBoost e o *Random Forest* também apresentaram equilíbrio entre as métricas, com valores acima de 89% para ambos. Já a Regressão Logística e o Bayes apresentaram menor equilíbrio, evidenciando limitações na separação das categorias.

- F_1-*score* e AUC

O F1-score de C5.0 confirma o seu excelente desempenho, enquanto o XGBoost e o *Random Forest* mostraram forte capacidade preditiva, com valores de 0,9455 e 0,8890,

respetivamente. Por outro lado, os modelos probabilísticos que tiveram F_1-*scores* inferiores, refletiram maior dificuldade em equilibrar a sensibilidade e precisão.

Com base nos resultados, o C5.0 é a melhor escolha, destacando-se em todas as métricas com um desempenho superior.

Diferentes graus de desequilíbrio

1) **Ajuste e seleção do modelo**

Para identificar o melhor modelo, foram testados diferentes parâmetros do valor de "*K*" (número de vizinhos mais próximos) e do "*over-ratio*" (proporção entre categorias). Na abordagem anterior, a técnica ROSE foi utilizada para realizar o *oversampling* em quatro cenários distintos (5000, 15000, 25000 e 35000 observações). Com o objetivo de comparar essa abordagem com o SMOTENC, foi necessário ajustar os valores de *over_ratio* e *K* de forma a atingir um número de observações idêntico. Na tabela que se segue, Tabela B5, encontram-se os valores alcançados nesta nova abordagem.

Tabela B 5 - SMOTENC: Alteração do número de observações de "Mortes/Feridos Graves" conforme o oversampling aumenta e o número de "Feridos Leves" se mantém constante.

	***Oversampling* com diferentes graus de desequilíbrio**	
	Feridos Leves	**Mortes/Feridos Graves**
Modelo Simples (42000)	42317	995
Modelo com *Oversampling* 5000	42317	6178
Modelo com *Oversampling* 15000	42317	16165
Modelo com *Oversampling* 25000	42317	26194
Modelo com *Oversampling* 35000	42317	36181

Na Tabela B6 são apresentados os diversos valores obtidos nos diferentes cenários, evidenciando as variações de desempenho dos modelos analisados.

Tabela B 6 - SMOTENC: Desempenho do Modelo de Regressão Logística com diferentes graus de desequilíbrio.

Métrica	OVERSAMPLING SMOTENC			
	5000	**15000**	**25000**	**35000**
Ponto de Corte	0,136	0,290	0,391	0,474
Accuracy	0,7929	0,7982	0,7983	0,7971
IC (95%)	(0,7862; 0,7995)	(0,7921; 0,8041)	(0,7927; 0,8037)	(0,7919; 0,8022)
Kappa	0,3898	0,5449	0,5839	0,5930
Mcnemar's Test P-Value	<0,001	<0,001	<0,001	<0,001
Sensibilidade	0,7941	0,7991	0,7984	0,7971
Especificidade	0,7927	0,7978	0,7982	0,7970
Valor Preditivo Positivo	0,3631	0,6081	0,7143	0,7737
Valor Preditivo Negativo	0,9628	0,9100	0,8623	0,8187
F1-score	0,8695	0,8502	0,8290	0,8077
AUC	0,8772	0,8802	0,8772	0,8806
Precisão	0,3631	0,6081	0,7143	0,7736

Com base na análise das métricas, observa-se que não há diferenças significativas do valor de *accuracy* ou AUC entre os cenários de *oversampling*. O cenário com 35000 observações apresentou o maior Kappa (0,5930) e a maior precisão (0,7736), o que indica uma redução de falsos positivos e maior confiabilidade na previsão de casos positivos. No entanto, esse ganho foi acompanhado por uma queda no F_1-*score* (0,8077) e no Valor Preditivo Negativo (81,87%), sugerindo perda de equilíbrio entre as categorias. Por outro lado, o cenário com 5000 observações destacou-se pelo maior *F1-score* (0,8695) e pelo melhor Valor Preditivo Negativo (96,28%), mostrando melhor equilíbrio entre a sensibilidade e precisão, embora com baixa capacidade preditiva para positivos (precisão = 0,3631).

A Tabela B7 apresenta os resultados do desempenho do “modelo base” da regressão logística em diferentes cenários.

Tabela B 7 - SMOTENC: Desempenho do Modelo de Regressão Logística com diferentes graus de desequilíbrio.

	OVERSAMPLING SMOTENC				
	Regressão Logística	**XGBoost**	**Random Forest**	**Bayes**	**C5.0**
Ponto de Corte					
5000	0,135	0,159	0,064	0,346	0,294
15000	0,294	0,341	0,271	0,761	0,471
25000	0,399	0,44	0,402	0,863	0,512
35000	0,474	0,52	0,548	0,918	0,51
Accuracy					
5000	0,7917	0,9121	0,8579	0,7502	0,9313
15000	0,7989	0,9396	0,8726	0,7606	0,9689
25000	0,7976	0,9382	0,8747	0,755	0,9751
35000	0,7975	0,9446	0,8891	0,7604	0,9801
Kappa					
5000	0,3874	0,679	0,5323	0,3113	0,7393
15000	0,5464	0,8549	0,7032	0,4687	0,9242
25000	0,5827	0,8707	0,7395	0,4972	0,9478
35000	0,5939	0,8888	0,7774	0,5196	0,96
McNemar´s Tes P-Value					
5000	<0,001	<0,001	<0,001	<0,001	<0,001
15000	<0,001	<0,001	<0,001	<0,001	<0,001
25000	<0,001	<0,001	<0,001	<0,001	<0,001
35000	<0,001	0,0083	0,0024	<0,001	0,1149

	OVERSAMPLING SMOTENC				
	Regressão Logística	**XGBoost**	**Random Forest**	**Bayes**	**C5.0**
Sensibilidade					
5000	0,7925	0,9124	0,8591	0,7505	0,9323
15000	0,7995	0,9397	0,8728	0,7612	0,9694
25000	0,7979	0,9385	0,8753	0,7563	0,9752
35000	0,7978	0,9449	0,8895	0,7614	0,9802
Especificidade					
5000	0,7916	0,9121	0,8578	0,7502	0,9312
15000	0,7987	0,9395	0,8726	0,7604	0,9686
25000	0,7974	0,9381	0,8742	0,7542	0,9751
35000	0,7973	0,9444	0,8888	0,7595	0,9800
Valor Preditivo Positivo					
5000	0,3614	0,6069	0,4733	0,3089	0,6685
15000	0,6093	0,8591	0,7289	0,5550	0,9239
25000	0,7135	0,9055	0,8148	0,6604	0,9612
35000	0,7740	0,9367	0,8744	0,7337	0,9771
Valor Preditivo Negativo					
5000	0,9625	0,9859	0,9761	0,9529	0,9893
15000	0,9103	0,9754	0,9459	0,8902	0,9878
25000	0,8619	0,9602	0,9173	0,8304	0,9842
35000	0,8192	0,9517	0,9024	0,7853	0,9828

	OVERSAMPLING SMOTENC				
	Regressão Logística	**XGBoost**	**Random Forest**	**Bayes**	**C5.0**
F1-Score					
5000	0,8687	0,9476	0,9131	0,8394	0,9594
15000	0,8509	0,9571	0,9077	0,8202	0,9781
25000	0,8284	0,9490	0,8953	0,7904	0,9796
35000	0,8081	0,9480	0,8955	0,7722	0,9814
AUC					
5000	0,8765	0,9723	0,9396	0,8142	0,9801
15000	0,8813	0,9856	0,9458	0,8263	0,9939
25000	0,8779	0,9862	0,9464	0,8195	0,9963
35000	0,8812	0,9884	0,9522	0,8232	0,9974
Precisão					
5000	0,3614	0,6069	0,4733	0,3089	0,6685
15000	0,6093	0,8591	0,7289	0,5550	0,9239
25000	0,7135	0,9055	0,8148	0,6604	0,9612
35000	0,7740	0,9367	0,8744	0,7337	0,9771

Os resultados mostram que o algoritmo C5.0 novamente se destaca como o mais eficiente em praticamente todas as métricas e cenários analisados com a aplicação da técnica SMOTENC. No que diz respeito ao *accuracy*, o C5.0 atinge valores que variam de 0,9313 no cenário de 5000 observações a 0,9801 com 35000 observações. O Kappa, também reforça o desempenho superior do C5.0, especialmente nos cenários com 25000 e 35000 observações, onde atinge valores elevados como 0,9478 e 0,9600, respetivamente.

Outro ponto relevante é a sensibilidade, onde o C5.0 novamente apresenta os melhores resultados, variando de 0,9323 a 0,9802. Esses valores mostram a boa capacidade que o modelo tem em identificar corretamente os casos positivos. A especificidade, segue a mesma tendência, com valores que vão de 0,9312 a 0,9800.

A métrica AUC também aponto o C5.0 como a melhor escolha. A mesma apresenta valores extremamente elevados, chegando a 0,9974 no cenários com 35000 observações. Outras métricas como F_1-*score* e os valores preditivos positivos e negativos, corroboram a superioridade do C5.0. O F_1-*score* atinge um valor de 0,9814 no cenário de 35000 observações e o valor preditivo positivo cresce significativamente, passando de 0,6685 para 0,9771. Porém, o valor preditivo negativo decresce ligeiramente com o aumento de observações, variando de 0,9893 a 0,9828.

Ao analisar os diferentes cenários, observa-se uma clara tendência de melhoria de desempenho do modelo C5.0 com o aumento do número de observações. No cenário com 5000 observações, os resultados são satisfatórios, mas inferiores em comparação com os cenários maiores, com métricas como *accuracy* (0,9313), F_1-*score* (0,9594) e AUC (0,9801) abaixo dos valores obtidos nos cenários subsequentes. Já os cenários com 15000 e 25000 observações, o desempenho do C5.0 atinge níveis muito elevados, com métricas muito próximas dos valores ideias (*accuracy* de 0,9689 e 0,9751; *F1-score* de 0,9781 e 0,9796; AUC de 0,9939 e 0,9963, respetivamente), refletindo excelente capacidade de previsão. No cenário com 35000 observações, o modelo mantém resultados muito bons (*accuracy* de 0,9801, F_1-*score* de 0,9814 e AUC de 0,9974). Estes ganhos adicionais são mínimos em comparação com o cenário de 25000, sugerindo um possível ponto de saturação no desempenho do modelo.

Dessa forma, os resultados indicam que o algoritmo C5.0, especialmente nos cenários de 15000 e 25000 observações, é a melhor opção para a previsão de sinistros com "Mortes/Feridos Graves".

Undersampling e Oversampling

No presente capítulo, foi adotada a metodologia SMOTENC para lidar com o desequilíbrio das categorias.

Posto isto, foi inicialmente criado um modelo com 42000 observações, onde 21000 correspondem à categoria minoritária e 21000 à categoria maioritária. São ainda criados outros dois modelos, um com 10000 observações (5000 categoria minoritária + 5000 categoria maioritária) e outro modelo com 20000 observações (1000 categoria minoritária + 1000 categoria maioritária), ambos com categorias equilibradas.

A abordagem seguiu os mesmos passos da metodologia anterior, iniciando com a aplicação do *undersampling* à categoria maioritária, removendo parte das observações dos sinistros "Feridos Leves". Posteriormente, foi aplicado o *oversampling* à categoria minoritária utilizando a técnica SMOTENC. Desta forma, assegura-se que ambas as categorias estão balanceadas nos modelos gerados, possibilitando que os algoritmos de *machine learning* trabalhem com dados mais equilibrados.

Modelo com 42000 observações

1) ***Undersampling***

A técnica de *undersampling* foi aplicada para equilibrar a base de dados, ajustando o número de observações da categoria maioritária para aproximá-lo ao da categoria minoritária. Esse processo envolveu a redução aleatória de observações da categoria predominante, sendo:

- Categoria minoritária ("Mortes/Feridos Graves"): todos os 995 sinistros com "Mortes/Feridos Graves" foram mantidos na base de dados sem alterações.
- Categoria maioritária ("Feridos Leves"): foi realizada uma amostragem aleatória simples, sem reposição, dos 42317 sinistros com "Feridos Leves".

O procedimento foi repetido de forma idêntica na aplicação do SMOTENC. Em ambas as metodologias, manteve a proporção controlada entre as categorias: a categoria maioritária com 21,11 vezes o número de observações da categoria minoritária. O ajuste resultou no mesmo total de 21999 observações, sendo 21004 da categoria maioritária e 995 observações da categoria minoritária. Essa repetição assegura a consistência nos modelos e a comparabilidade entre os resultados das duas abordagens.

2) ***Oversampling***

De forma a equilibrar a base de dados, aplicou-se o *undersampling* na categoria minoritária. Os valores finais estão na Tabela B8.

Tabela B 8 - Composição do modelo com e sem undersampling e com undersampling+oversampling de 42000 observações (SMOTENC).

	SMOTENC – *UNDER* + *OVER* 42000	
	Feridos Leves	**Mortes / Feridos Graves**
Modelo Simples	42317	995
Modelo com *Undersampling*	21004	995
Modelo com *Under* + *Over Sampling*	21004	20982

1) Divisão dos dados em treino e teste

O conjunto de dados foi dividido em dois subconjuntos: 70% para treino e 30% para teste. O resultado desta divisão, Tabela B9, encontra-se da seguinte forma:

- Conjunto de treino: contém 29587 observações.
- Conjunto de teste: contém 12399 observações.

Tabela B 9 - SMOTENC: Divisão dos dados do modelo de regressão logística (42000 observações) em dois subconjuntos: treino e teste e respetivo número de observações por categoria em cada subconjunto.

	SMOTENC *UNDER* + *OVER* 42000	
	Treino	**Teste**
Feridos Leves	14764	6240
Mortes/Feridos Graves	14823	6159
	29587	12399

2) Ajustamento do Modelo

Utilizando os dados pré-processados, ajustou-se um modelo de Regressão Logística para prever a probabilidade do evento de interesse. Os coeficientes estimados e as medidas de ajuste do modelo estão sumarizadas na Tabela B10.

Tabela B 10 - SMOTENC: Modelo múltiplo de regressão logística ajustado para a existência de "Mortes/Feridos Graves" nos sinistros com vítimas com 42000 observações.

Variável	Coeficiente	Std. Error	P-value
Intercept	-1,9515	0,1575	<0,001
Concelho2AGSSP (ALCOCHETE, GRÂNDOLA, SEIXAL, SINES e PALMELA)	0,1865	0,0999	0,0619
Concelho2ABMMS (ALMADA, BARREIRO, MOITA, MONTIJO e SESIMBRA)	-0,0100	0,1012	0,9213
Concelho2SS (SANTIAGO DO CACÉM e SETÚBAL)	-0,4522	0,1118	<0,001
tipoacidColisão	-1,4312	0,1005	<0,001
tipoacidDespiste	-0,7407	0,1046	<0,001
tipolocal2Fora das localidades	0,5641	0,0376	<0,001
tipovia2EM – Estrada Municipal	-0,4977	0,1332	0,0002
tipovia2EN/IC/ER (Estrada Nacional, Itinerário Complementar e Estrada Regional)	1,0683	0,0544	<0,001
horaacid1new6h	-0,3069	0,1228	0,0124
horaacid1new8h-13h	-0,4524	0,0359	<0,001
fugaSim	-3,7486	0,2132	<0,001
PercCondMCat2[75,100]	0,3500	0,0346	<0,001
HaVeicPesadoSim	0,7361	0,0601	<0,001
HaVeicLigSim	0,3457	0,0676	<0,001
HaVeicMotoSim	2,6787	0,0545	<0,001
HoraLaboralSim	-0,3052	0,0332	<0,001

Variável	Coeficiente	Std. Error	P-value
MedianaIdadeVeic	0,0566	0,0026	<0,001
ig_ponderado	0,1066	0,0063	<0,001
tipovia2EM – Estrada Municipal: HaVeicMotoSim	-0,0491	0,1670	0,7686
tipovia2EN/IC/ER: HaVeicMotoSim	-0,5464	0,0780	<0,001
Concelho2AGSSP: ig_ponderado	-0,0417	0,0056	<0,001
Concelho2ABMMS: ig_ponderado	-0,0680	0,0056	<0,001
Concelho2SS: ig_ponderado	-0,0139	0,0063	0,0261
tipoacidColisão: ig_ponderado	-0,0303	0,0031	<0,001
tipoacidDespiste: ig_ponderado	-0,0187	0,0034	<0,001
tipovia2EM – Estrada Municipal: ig_ponderado	0,0476	0,0066	<0,001
tipovia2EN/IC/ER: ig_ponderado	-0,0018	0,0014	0,1791

Os sinistros envolvendo motociclos (“HaVeicMotoSim”) , veículos pesados (“HaVeicPesadoSim”) e os Sinistros em Estradas Nacionais, Itinerários Complementares ou Estrada Regional (“tipovia2EN/IC/ER”) são fatores intrínsecos ligados a eventos mais graves. Por outro lado, variáveis como a ocorrência de Colisões (“tipoacidColisão”), sinistros em que há fuga (“fugaSim”) e sinistros durante o período da manhã (“horaacid1new8h-13h”) tendem a estar associadas a sinistros menos graves.

3) Comparação do desempenho entre os modelos de classificação

Por fim, é comparado detalhadamente o desempenho dos modelos de classificação, utilizando as métricas da Tabela B11.

Tabela B 11 - SMOTENC: Métricas de classificação para 42000 observações - Undersampling + Oversampling.

Métrica	Regressão Logística	XGBoost	Random Forest	Bayes	C5.0
	UNDER + OVER – SMOTENC 42000				
Ponto de Corte	0,502	0,553	0,600	0,837	0,528
Accuracy	0,7998	0,9340	0,8806	0,7546	0,9656
IC (95%)	(0,7927; 0,8068)	(0,9295; 0,9383)	(0,8748; 0,8863)	(0,7469; 0,7621)	(0,9623; 0,9688)
Kappa	0,5996	0,8680	0,7613	0,5091	0,9313
Mcnemar's Test P-Value	0,7029	0,7530	0,8150	0,6117	0,8844
Sensibilidade	0,8001	0,9344	0,8807	0,7553	0,9657
Especificidade	0,7995	0,9337	0,8806	0,7538	0,9655
Valor Preditivo Positivo	0,7975	0,9329	0,8792	0,7518	0,9651
Valor Preditivo Negativo	0,8021	0,9352	0,8820	0,7574	0,9662
F1-score	0,8008	0,9344	0,8813	0,7556	0,9659
AUC	0,8816	0,9848	0,9459	0,8258	0,9946
Precisão	0,7975	0,9329	0,8792	0,7518	0,9651

Os resultados indicam que o modelo C5.0 obteve o melhor desempenho geral, com alto valor de *accuracy* (0,9656), Kappa (0,9313), sensibilidade (0,9657), especificidade (0,9655), F_1-*score* (0,9659) e AUC (0,9946). Esses resultados sugerem que o C5.0 apresenta uma boa capacidade de discriminação entre casos positivos e negativos, além

de manter o equilíbrio muito bom entre sensibilidade e precisão. O alto valor de Kappa indica que a concordância entre previsões e observações reais é substancialmente superior ao que seria esperado.

O modelo XGBoost também apresentou um bom desempenho, especialmente em termos de AUC (0,9848). Embora os seus valores de *accuracy*, F_1-*score* e Kappa serem ligeiramente inferiores aos de C5.0, o XGBoost permanece como uma alternativa robusta.

O modelo de Regressão Logística apresentou resultados intermédios, com métricas mais modestas, indicando que, embora seja útil para previsões gerais, pode não ser tão eficaz quanto aos modelos baseados em árvores.

Os modelos Random Forest e Bayes tiveram desempenhos ligeiramente inferiores, sugerindo limitações na capacidade de generalização frente à complexidade e desequilíbrio dos dados.

Adicionalmente, todos os modelos apresentaram valores de *p-value* superiores a 0,05 no Mcnemar's Test, indicando que não há diferenças estatisticamente significativas nos erros de classificação entre eles.

Em termos práticos, a análise evidencia que o C5.0 é a melhor escolha para a previsão de sinistros graves neste conjunto de dados, oferecendo não apenas alta precisão, mas também confiabilidade na classificação de casos críticos. O XGBoost pode ser considerado uma alternativa viável, especialmente em contextos onde se prioriza discriminação entre categorias, enquanto modelos probabilísticos como Bayes ou Regressão Logística podem ser mais limitados quando se lida com dados altamente desequilibrados ou com caraterísticas complexas.

Modelos com 10000 e 20000 observações

1) *Undersampling*

Procedeu-se ao pré-processamento dos dados utilizando novamente a técnica de *undersampling*, seguindo um procedimento semelhante ao realizado anteriormente para as 42000 observações, mas desta vez considerando um menor número de observações:

- Categoria minoritária (“Mortes/Feridos Graves”): todos os 995 sinistros com “Mortes/Feridos Graves” foram mantidos na base de dados sem alterações.
- Categoria maioritária (“Feridos Leves”): foi realizada uma amostragem aleatória simples, sem reposição, dos 42317 sinistros com “Feridos Leves”.

Posto isto, foram criados dois conjuntos com base na categoria maioritária, mantendo proporções controladas em relação à categoria minoritária:

- Conjunto 1: inclui um número de sinistros selecionados aleatoriamente equivalente a 5,03 vezes o total de observações da categoria minoritária.
- Conjunto 2: foi composto de maneira similar, também considerando sinistros correspondentes a 10,05 vezes o número de observações da categoria minoritária.

2) *Oversampling*

A Tabela B12 apresenta a comparação de desempenho entre diferentes modelos de *machine learning* para a classificação de dados relacionados a “Feridos Leves” e “Mortes/Feridos Graves” em dois cenários de volume de dados (10000 e 20000 observações).

Tabela B 12 - SMOTENC: Valores das categorias do modelo simples, com undersampling e com undersampling+oversampling.

	10000		20000	
	Feridos Leves	**Mortes / Feridos Graves**	**Feridos Leves**	**Mortes / Feridos Graves**
Modelo Simples	42317	995	42317	995
Modelo com *Undersampling*	5004	995	9999	995
Modelo com Under + Over Sampling	5004	5004	9999	9999

3) Divisão dos dados em treino e teste

A fim de realizar a modelação, os conjuntos de dados foram divididos em dois subconjuntos: 70% dos dados foram alocados para treino e 30% para teste. A divisão com 10000 observações, Tabela B13, contém:

- Conjunto de treino: 7053 observações.
- Conjunto de teste: 2955 observações.

Tabela B 13 - SMOTENC: Divisão dos dados do modelo de regressão logística (10000 observações) em dois subconjuntos: treino e teste e respetivo número de observações por categoria em cada subconjunto.

	SMOTENC *UNDER* + OVER 10000	
	Treino	***Teste***
Feridos Leves	3509	1495
Mortes/Feridos Graves	3544	1460
	7053	2955

A divisão com 20000 observações, Tabela B14, contém:

- Conjunto de treino: 14078 observações.
- Conjunto de teste: 5920 observações.

Tabela B 14 - SMOTENC: Divisão dos dados do modelo de regressão logística (20000 observações) em dois subconjuntos: treino e teste e respetivo número de observações por categoria em cada subconjunto.

	SMOTENC *UNDER* + *OVER* 20000	
	Treino	**Teste**
Feridos Leves	7042	2957
Mortes/Feridos Graves	7036	2963
	14078	5920

4) Ajustamento do Modelo

Foram ajustados os dois modelos de Regressão Logística, onde a Tabela B15 apresenta os resultados do primeiro modelo (10000 observações), enquanto a Tabela B16

apresenta os resultados do segundo modelo (20000 observações).

Tabela B 15 - SMOTENC: Modelo múltiplo de regressão logística ajustado para a existência de Mortes/Feridos Graves nos sinistros com vítimas, com 10000 observações.

Variável	Coeficiente	Std. Error	P-value
	SMOTENC 10000		
Intercept	-1,1824	0,4830	0,0144
Concelho2AGSSP (ALCOCHETE, GRÂNDOLA, SEIXAL, SINES e PALMELA)	-0,2014	0,3282	0,5395
Concelho2ABMMS (ALMADA, BARREIRO, MOITA, MONTIJO e SESIMBRA)	-0,2481	0,3284	0,4498
Concelho2SS (SANTIAGO DO CACÉM e SETÚBAL)	-0,3799	0,3529	0,2817
tipoacidColisão	-1,5789	0,2808	<0,001
tipoacidDespiste	-1,0550	0,2963	0,0004
tipolocal2Fora das localidades	0,4590	0,1180	0,0001
tipovia2EM – Estrada Municipal	-0,7522	0,4426	0,0893
tipovia2EN/IC/ER (Estrada Nacional, Itinerário Complementar e Estrada Regional)	1,1425	0,1654	<0,001
horaacid1new6h	-0,0307	0,3889	0,9370
horaacid1new8h-13h	-0,4432	0,1123	<0,001
fugaSim	-3,0009	0,4893	<0,001
PercCondMCat2[75,100]	0,3034	0,1076	0,0048
HaVeicPesadoSim	0,7547	0,1925	<0,001
HaVeicLigSim	0,2979	0,2120	0,1600
HaVeicMotoSim	2,6418	0,1772	<0,001

Variável	Coeficiente	Std. Error	P-value
	SMOTENC 10000		
HoraLaboralSim	-0,5080	0,1037	<0,001
MedianaIdadeVeic	0,0451	0,0079	<0,001
ig_ponderado	0,0968	0,0203	<0,001
tipovia2EM – Estrada Municipal: HaVeicMotoSim	0,5205	0,6059	0,3903
tipovia2EN/IC/ER: HaVeicMotoSim	-0,8328	0,2334	0,0004
Concelho2AGSSP: ig_ponderado	-0,0421	0,0191	0,0272
Concelho2ABMMS: ig_ponderado	-0,0721	0,0188	0,0001
Concelho2SS: ig_ponderado	-0,0253	0,0205	0,2175
tipoacidColisão: ig_ponderado	-0,0191	0,0079	0,0157
tipoacidDespiste: ig_ponderado	-0,0002	0,0089	0,9846
tipovia2EM – Estrada Municipal: ig_ponderado	0,0717	0,0237	0,0025
tipovia2EN/IC/ER: ig_ponderado	-0,0047	0,0040	0,2409

Tabela B 16 - SMOTENC: Modelo múltiplo de regressão logística ajustado para a existência de Mortes/Feridos graves nos sinistros com vítimas, com 20000 observações.

Variável	Coeficiente	Std. Error	P-value
	SMOTENC 20000		
Intercept	-2,0417	0,3570	<0,001
Concelho2AGSSP (ALCOCHETE, GRÂNDOLA, SEIXAL, SINES e PALMELA)	0,2200	0,2286	0,3358
Concelho2ABMMS (ALMADA, BARREIRO, MOITA, MONTIJO e SESIMBRA)	-0,1465	0,2292	0,5228

Variável	Coeficiente	Std. Error	P-value
	SMOTENC 20000		
Concelho2SS (SANTIAGO DO CACÉM e SETÚBAL)	-0,8335	0,2559	0,0011
tipoacidColisão	-1,4199	0,2207	<0,001
tipoacidDespiste	-0,8428	0,2321	0,0003
tipolocal2Fora das localidades	0,6549	0,0832	<0,001
tipovia2EM – Estrada Municipal	-0,3086	0,3050	0,3117
tipovia2EN/IC/ER (Estrada Nacional, Itinerário Complementar e Estrada Regional)	1,3286	0,1194	<0,001
horaacid1new6h	0,1898	0,2585	0,4628
horaacid1new8h-13h	-0,3977	0,0799	<0,001
fugaSim	-3,9213	0,5259	<0,001
PercCondMCat2[75,100]	0,1918	0,0767	0,0124
HaVeicPesadoSim	0,7324	0,1317	<0,001
HaVeicLigSim	0,4722	0,1508	0,0017
HaVeicMotoSim	2,8414	0,1246	<0,001
HoraLaboralSim	-0,4091	0,0744	<0,001
MedianaIdadeVeic	0,0560	0,0058	<0,001
ig_ponderado	0,1207	0,0157	<0,001
tipovia2EM – Estrada Municipal: HaVeicMotoSim	-0,1621	0,3912	0,6786
tipovia2EN/IC/ER: HaVeicMotoSim	-0,6255	0,1676	0,0002
Concelho2AGSSP: ig_ponderado	-0,0578	0,0144	<0,001
Concelho2ABMMS: ig_ponderado	-0,0800	0,0143	<0,001

Variável	Coeficiente	Std. Error	P-value
	SMOTENC 20000		
Concelho2SS: ig_ponderado	-0,0214	0,0158	0,1742
tipoacidColisão: ig_ponderado	-0,0297	0,0066	<0,001
tipoacidDespiste: ig_ponderado	-0,0122	0,0073	0,0960
tipovia2EM – Estrada Municipal: ig_ponderado	0,0422	0,0144	0,0034
tipovia2EN/IC/ER: ig_ponderado	-0,0108	0,0028	0,0001

Ao compararmos estes dois modelos, com o modelo anterior (modelo de *oversampling* com 85000 observações), nota-se que, em grande parte, as mesmas variáveis continuam a destacar-se em termos de impacto significativo sobre a ocorrência de sinistros com mortes/feridos graves. No entanto algumas nuances devem ser ressaltadas:

- Variáveis com maior impacto positivo: semelhante ao modelo anterior, as variáveis relacionadas à existência de motociclos e veículos pesados continuam a exercer um papel crucial. Sinistros ocorridos em Estradas Nacionais, Itinerários Complementares e Estradas Regionais novamente apresentam maior probabilidade de resultar em sinistros onde resultam "Mortes/Feridos Graves".

5) <u>Comparação do desempenho entre os modelos de classificação</u>

Na etapa final, realizar-se-á uma análise comparativa dos modelos de classificação desenvolvidos, com base nas métricas de desempenho apresentadas na Tabela B17.

Tabela B 17 - SMOTENC: Métricas de classificação para 10000 e 20000 observações - Undersampling + Oversampling.

	Regressão Logística	XGBoost	Random Forest	Bayes	C5.0
	UNDER + OVER SMOTENC				
Ponto de Corte					
10000	0,505	0,553	0,576	0,750	0,558
20000	0,510	0,565	0,604	0,845	0,540
Accuracy					
10000	0,7885	0,8772	0,8470	0,7387	0,8579
20000	0,8000	0,9105	0,8561	0,7505	0,9172
IC (95%)					
10000	(0,7733; 0,8031)	(0,8648; 0,8888)	(0,8335; 0,8598)	(0,7225; 0,7545)	(0,8448; 0,8703)
20000	(0,7896; 0,8101)	(0,9029; 0,9176)	(0,8469; 0,8649)	(0,7393; 0,7615)	(0,9099; 0,9241)
Kappa					
10000	0,5770	0,7543	0,6941	0,4775	0,7157
20000	0,6000	0,8209	0,7122	0,5010	0,8345
McNemar´s Test P-Value					
10000	0,7490	0,8337	0,7420	0,7460	0,9611
20000	0,9305	0,9654	1	0,9585	1
Sensibilidade					
10000	0,7890	0,8774	0,8479	0,7390	0,8568
20000	0,8009	0,9109	0,8562	0,7513	0,9173

	Regressão Logística	XGBoost	Random Forest	Bayes	C5.0
	UNDER + OVER SMOTENC				
Especificidade					
10000	0,7880	0,8769	0,8462	0,7385	0,8589
20000	0,7991	0,9100	0,8559	0,7497	0,9171
Valor Preditivo Positivo					
10000	0,7842	0,8744	0,8433	0,7340	0,8557
20000	0,7998	0,9103	0,8562	0,7505	0,9173
Valor Preditivo Negativo					
10000	0,7927	0,8799	0,8507	0,7434	0,8600
20000	0,8002	0,9107	0,8559	0,7505	0,9171
F1-Score					
10000	0,7903	0,8784	0,8484	0,7409	0,8594
20000	0,7997	0,9104	0,8559	0,7501	0,9171
AUC					
10000	0,8719	0,9515	0,9217	0,8124	0,9294
20000	0,8795	0,9737	0,9339	0,8191	0,9753
Precisão					
10000	0,7842	0,8744	0,8433	0,7340	0,8557
20000	0,7998	0,9103	0,8562	0,7505	0,9173

A análise dos resultados evidência que o desempenho dos modelos melhora significativamente com o aumento do número de observações. Isso é notável em métricas como *accuracy*, sensibilidade, especificidade e F_1-*score*, onde os valores se tornam mais elevados e consistentes com um maior número de observações.

Os modelos XGBoost e C5.0 destacaram-se como os mais eficazes. O XGBoost apresentou um *accuracy* de 0,9044 e uma AUC de 0,972 com 20000 observações, enquanto o C5.0 alcançou o maior *accuracy* (0,9177) e uma AUC de 0,9753 no mesmo cenário. Além disso, o C5.0 apresentou maios Kappa (0,8355), indicando uma alta concordância entre as previsões e os valores reais.

Em suma, os resultados mostram que os modelos XGBoost e C5.0 são os mais adequados e que o aumento de observações contribui para a precisão e a estabilidade das previsões.

Anexos

Anexo 1

Categoria	Variável	Descrição
Localização	concelho	Concelho
	kmacid	Quilómetro onde ocorreu o sinistro
	sitacid	Local do sinistro
Tipo de Sinistro	tipoacid	Tipo de sinistro
	naturezaacid	Natureza do sinistro
	fuga	Sinistro com fuga
	ig_ponderado	Índice de gravidade
Via e Infraestrutura	tipoberma	Tipo de berma
	tipolocal	Localização do sinistro
	tipovia	Tipo de via
	Tracado	Traçado da via em planta
	tracadoperfil	Traçado da via em perfil
	marcaspad1	Marcas no pavimento
	d_n_vias	Número de vias
	faixasentido	Faixa de rodagem com sentido único ou dois sentidos
	estadoconserv	Estado de conservação
	intervias	Interseção de vias
	tipopiso	Tipo de piso
	obras	Obstáculos ou obras
	danosvia	Danos na via
Condições Ambientais	fatoresatmos1	Fatores atmosféricos
	sensepcentral1	Sentido do separador central
	sinallum1	Sinalização luminosa
	sinais	Sinais
	luminos	Luminosidade
	choveu	Choveu?
	sol	Estava sol?

Categoria	Variável	Descrição
	ff_med	Intensidade média do vento (m/s)
Fatores Temporais	diasemanaacid	Dia da semana do sinistro
	horaacid1	Hora com minutos a zero do sinistro
	HoraLaboral	Sinistro ocorreu no horário laboral
	PicoTráfego	Pico de tráfego
	feriado	Sinistro ocorreu num dia feriado
	diaacid	Dia do mês do sinistro
	anoacid	Ano do sinistro
	mesacid	Mês do sinistro
Tipo de Veículo	HaVeicPesado	Existência de veículos pesados
	HaVeicLig	Existência de veículos ligeiros
	HaVeicMoto	Existência de ciclomotores e motociclos
	HaVeicEsp	Existência de veículos de especiais
	HaVeicTrator	Existência de veículos tratores
	HaVeicMisto	Existência de veículos mistos
	HaVeicMerc	Existência de veículos de mercadorias
	HaVeicPassag	Existência de veículos de passageiros
Caraterísticas dos Condutores	condader1	Condições de aderência
	PercCondMCat	% de condutores masculinos envolvidas no sinistro (categorizada)
	PercCondFCat	% de condutores femininas envolvidas no sinistro (categorizada)
	MinAnosLicCond	Mínimo de anos de licença/ carta dos condutores
	MaxAnosLicCond	Máximo de anos de licença/ carta dos condutores
	MinIdadeCond	Mínimo das idades dos condutores
	MaxIdadeCond	Máximo das idades dos condutores
	MinIdadeVeic	Mínimo da idade da matrícula dos veículos
	MaxIdadeVeic	Máximo da idade da matrícula dos veículos

Categoria	Variável	Descrição
Fatores Humanos	causas2	Causas do sinistro
	Medianataxalcool	Mediana da taxa de alcoolemia
Contexto Social	Aulas	Sinistro ocorreu durante o período de aulas
	unsaude	Unidade de saúde

Anexo 2

Categoria	Variáveis
Localização	concelho, freguesia
Via / Estrutura	tipoberma, tipolocal2, tipovia, tracadoperfil, tracado, tracadoperfil, d_n_vias, intervias, faixasentido, estadoconserv, obras, danosvia
Marcação e Sinalização	marcaspav1, sensepcentral1, sinais, sinallum1
Condições Ambientais	fatoresatmos1, luminos, choveu, sol
Tempo	diasemanaacid2, horaacid1, horaacid, HoraLaboral, diaacid, mesacid, anoacid, feriado, PicoTrafego, Aulas
Características do Sinistro	tipoacid, naturezaacid, sitacid, causas2, fuga
Veículos	HaVeicPesado, HaVeicLig, HaVeicMoto, HaVeicEsp, HaVeicTrator, HaVeicMisto, HaVeicMerc, HaVeicPassag
Condução / Condutores	condader1, PercCondMCat, PercCondFCat, MinAnosLicCond, MaxAnosLicCond, MedianaAnosLicCond, IQRAnosLicCond
Álcool	Medianataxaalcool, IQRtaxaalcool, Mintaxaalcool, Maxtaxaalcool
Vítimas	IQRIdadeVit, MinIdadeVit, MaxIdadeVit, MedianaIdadeVit
Condutores (Idade)	MinIdadeCond, MaxIdadeCond, MedianaIdadeCond, IQRIdadeCond
Veículos (Idade)	MinIdadeVeic, MaxIdadeVeic, MedianaIdadeVeic, IQRIdadeVeic
Outros	kmacid, unsaude, numero_de_arvores

Anexo 3 – Fluxograma metodológico