



Análise de Dados Categóricos

(Relatório da Unidade Curricular no âmbito das Provas de Agregação)

Paulo de Jesus Infante dos Santos

Universidade de Évora
Escola de Ciências e Tecnologia
Departamento de Matemática

Outubro de 2025

Relatório da Unidade Curricular

Este relatório referente à unidade curricular (UC) *Análise de Dados Categóricos* é apresentado à Universidade de Évora no âmbito das Provas de Agregação em Matemática ao abrigo da alínea b) do artigo 5.º do Decreto-Lei n.º 239/2007, de 19 de junho, com atualização do Decreto-Lei n.º 64/2023, de 31 de julho.

Évora, 21 de Outubro de 2025

Paulo de Jesus Infante dos Santos

Resumo

A unidade curricular Análise de Dados Categóricos (MAT13608M) é da responsabilidade do Departamento de Matemática da Escola de Ciências e Tecnologia da Universidade de Évora. É obrigatória no Mestrado em Modelação Estatística e Análise de Dados, no Mestrado em Inteligência Artificial e Ciência de Dados e na Pós-Graduação em Epidemiologia, sendo também oferecida no Mestrado em Matemática. Tem 6 ECTS (156 horas: 45h contacto presencial + 4h tutoriais + 107h estudo autónomo), leciona-se no 2.º semestre em regime presencial e em Português (com apoio em Inglês quando necessário). Recomenda-se formação prévia em Probabilidades e Estatística e noções introdutórias de R.

Estrutura-se em três blocos: núcleo essencial, complementar aplicado e literacia alargada. O fio condutor são os Modelos Lineares Generalizados (GLM) para respostas categóricas (binárias, ordinais, multinomiais) e contagens, com extensões para zeros excessivos (modelos inflacionados de zeros, *hurdle*) e para dados correlacionados (Modelos de Equação de Estimação Generalizadas – GEE; e Modelos Mistos Lineares Generalizados – GLMM). O ensino conjuga aulas teórico-práticas com forte uso de R, projetos reprodutíveis com dados reais, *quizzes* individuais e apresentações críticas. A avaliação privilegia autenticidade e reprodutibilidade, centrando-se em trabalhos de grupo, incluindo também uma pequena componente individual.

A evidência empírica dos últimos três anos mostra taxas de aprovação elevadas e avaliações estudantis consistentes. Os planos de melhoria centram-se em reforçar a clareza na exposição dos conteúdos e na comunicação dos critérios avaliativos, aprofundar o apoio ao nivelamento inicial no *software* R e inferência estatística e diversificar os exemplos interdisciplinares usados nas aulas, consolidando a ligação entre teoria e aplicação. Em termos internacionais, a UC encontra-se alinhada com ofertas de universidades de referência, distinguindo-se pela reprodutibilidade, integração da ética e inteligência artificial e capacidade de adaptação a turmas heterogéneas.

Nota preliminar: motivação e trajeto científico do responsável

A docência e coordenação da unidade curricular Análise de Dados Categóricos assenta num percurso académico e científico em que a modelação estatística aplicada constituiu, desde o início, o eixo estruturante. A formação de base, sólida do ponto de vista teórico e fortemente orientada para aplicações, marcou de forma decisiva a forma de encarar a inferência em contextos reais.

Três princípios estatísticos atravessam toda a unidade curricular: o papel do desenho amostral, a adequabilidade do modelo aos dados e a avaliação do risco associado às decisões. A estes acrescem princípios pedagógicos igualmente centrais: a reprodutibilidade (relatórios transparentes com *scripts* de R reprodutíveis), a autenticidade (uso de dados reais e relevantes) e a responsabilização individual (*quizzes* e apresentações).

Do ponto de vista docente, a experiência acumula-se ao longo de mais de três décadas, desde monitor em 1992 até Professor Associado a partir de 2020, sempre em áreas de Probabilidade, Inferência e Modelação Estatística.

A motivação científica decorre em grande medida de projetos de investigação em que dados binários, ordinais, multinomiais e de contagem são a matéria-prima, aplicados em áreas como a saúde, ciências sociais, ecologia, zootecnia e segurança rodoviária. A forma como estes projetos são integrados no ensino, conferindo autenticidade e relevância social à UC, será também descrita neste relatório.

Em síntese, a Análise de Dados Categóricos é a unidade curricular onde confluem formação, investigação e docência, sendo por isso a escolha natural para estas Provas de Agregação.

Índice

1. Introdução, enquadramento e perfil de estudantes	7
1.1. Integração com investigação e sociedade	8
1.2. Pré-requisitos e nivelamento.....	10
2. Objetivos de aprendizagem e competências a adquirir	10
2.1. Objetivo Principal	10
2.2. Objetivos específicos.....	11
2.3. Competências desenvolvidas	13
2.4. Articulação com o plano de estudos	14
2.5. Ligação aos cursos e aos ODS	16
3. Conteúdos programáticos	17
3.1. Núcleo essencial	17
3.2. Complementar aplicado	19
3.3. Literacia alargada	20
3.4. Racional e coerência curricular	22
3.5. A pertinência de modelos de contagens em Análise de Dados Categóricos	22
3.6. Sumário dos Tópicos Programáticos abordados	23
4. Metodologia de ensino e diferenciação pedagógica	24
4.1. Estratégia pedagógica e alinhamento com os objetivos de aprendizagem.....	26
4.2. Acessibilidade e inclusão	27
5. Avaliação	28
5.1. Princípios orientadores e componentes	28
5.2. Avaliação individual e autorregulação	30
5.3. Síntese objetivos-avaliação	31
5.4. Feedback, calendário e integridade académica	33
5.5. Uso responsável de Inteligência Artificial Generativa	33
6. Planificação	34
7. Qualidade, evidência e melhoria contínua	37
8. Recursos, reprodutibilidade e inclusão.....	39
8.1. Recursos.....	39
8.2. Reprodutibilidade	40

8.3. Inclusão.....	40
9. Benchmarking e posicionamento.....	41
9.1. Panorama nacional (Portugal).....	41
9.2. Panorama internacional (Europa e Estados Unidos)	42
9.3. Posicionamento da UC: semelhanças e diferenças.....	43
9.4. Decisões curriculares	44
10. Considerações finais e plano de desenvolvimento futuro	45
11. Bibliografia.....	47
11.1. Bibliografia nuclear	48
11.2. Bibliografia complementar	48
11.3. Artigos aplicados	49
11.3a. Projetos de investigação e algumas publicações associadas.....	50
11.4. Fontes institucionais e pedagógicas	51
Anexo A – Catálogo de Bases de Dados (núcleo).....	54
Anexo B – Modelo para Declaração de Uso de Inteligência Artificial	56

1. Introdução, enquadramento e perfil de estudantes

A unidade curricular (UC) Análise de Dados Categóricos (ADC), anteriormente designada Análise Categórica de Dados, integra os planos de estudo de alguns ciclos de formação avançada da Universidade de Évora e desempenha, desde há mais de quinze anos, um papel estruturante na consolidação de competências em modelação estatística. Lecionei esta UC pela primeira vez em 2007/08, e fui o responsável também em 2008/09, 2009/10, 2010/11, 2012/13, 2014/15 e 2015/16, sob a designação inicial. Desde 2022/23, a UC passou a ter a designação atual, mantendo a essência do programa, mas com atualização de alguns tópicos programáticos. Em todas estas edições, tive a oportunidade de observar a evolução do perfil dos estudantes, dos contextos de aplicação e das exigências do mercado profissional, procurando a melhoria contínua da qualidade do ensino e das competências conferidas por esta UC.

O enquadramento nacional e internacional justifica a relevância desta UC. A análise de dados categóricos constitui, juntamente com a análise de regressão para variáveis contínuas e a análise de sobrevivência, um dos pilares do ensino avançado de estatística aplicada. Nas mais diversas áreas como a epidemiologia, a economia, a gestão, a sociologia, a psicologia, a biologia, as ciências da saúde e do desporto ou a inteligência artificial, muitos dos fenómenos observados são frequentemente descritos através de respostas binárias (sim/não, presença/ausência), ordinais (níveis de satisfação, estágios de doença), multinomiais (opções políticas, categorias de diagnóstico) ou de contagem (número de ocorrências de um evento). A literacia estatística neste domínio é hoje não apenas desejável, mas fundamental para qualquer profissional que opere em ambientes de dados reais e complexos.

No contexto da Universidade de Évora, a UC é lecionada a 4 cursos distintos: obrigatória no Mestrado em Modelação Estatística e Análise de Dados (MMEAD), no Mestrado em Inteligência Artificial e Ciência de Dados (MIACD), e na Pós-Graduação em Epidemiologia (PGE), sendo ainda oferecida como unidade curricular optativa no Mestrado em Matemática (MMAT). Tal traduz uma clara vocação interdisciplinar, refletindo tanto a natureza transversal dos métodos ensinados como a aposta institucional na formação estatística de excelência. Simultaneamente, obriga a um plano curricular que seja flexível, diferenciado e inclusivo, capaz de responder à heterogeneidade de perfis académicos, como acontece fundamentalmente no MMEAD. Com efeito, os estudantes que têm frequentado esta UC apresentam formações de base diversas: Desporto, Sociologia/Demografia, Economia e Gestão, Biologia/Ecologia, Ciências da Saúde/Epidemiologia, Engenharia Informática, Matemática e Matemática Aplicada à Economia e à Gestão. Esta heterogeneidade é, em simultâneo, um desafio pedagógico e uma oportunidade formativa. É um desafio, porque implica lidar com estudantes que chegam com níveis de literacia estatística muito distintos: alguns já dominam probabilidade e inferência, enquanto outros trazem apenas uns conceitos base, e muitas vezes superficiais, de estatística descritiva. É também uma oportunidade, porque permite criar um ambiente de aprendizagem interdisciplinar em que os exemplos discutidos ganham riqueza e relevância prática, pois os próprios estudantes

trazem problemas, contextos e linguagens de diferentes áreas científicas. Esta interação, devidamente orientada, potencia não só a compreensão dos modelos estatísticos, mas também a capacidade de comunicar resultados de forma clara a públicos diversificados, o que é uma competência indispensável.

A evolução recente da UC confirma a sua importância. Desde 2022/23, a unidade tem registado uma participação crescente: de turmas de 9 e 10 estudantes avaliados nos primeiros dois anos, passou a acolher 24 estudantes avaliados em 2024/25. A taxa de aprovação manteve-se elevada (100% nos dois primeiros anos e 88% no terceiro), evidenciando que é possível garantir sucesso académico mesmo perante turmas mais numerosas e diversificadas. Os resultados dos questionários pedagógicos aplicados aos estudantes (analisados mais adiante) corroboram esta visão, revelando apreciação elevada tanto da relevância da UC como da qualidade da docência.

Do ponto de vista estratégico, esta UC assume um papel distintivo no quadro dos Objetivos de Desenvolvimento Sustentável (ODS) a que a Universidade de Évora se encontra alinhada. Em particular, contribui para o ODS 4 – Educação de Qualidade, ao assegurar aprendizagem inclusiva, equitativa e relevante, com recurso a práticas pedagógicas inovadoras e a avaliação autêntica. Pela sua natureza aplicada, contribui também para o ODS 3 – Saúde e Bem-Estar, o ODS 5 – Igualdade de Género e o ODS 10 – Redução das Desigualdades, ao dotar futuros profissionais de ferramentas que permitem analisar e comunicar desigualdades e riscos com base em dados categóricos. Finalmente, articula-se com o ODS 9 – Indústria, Inovação e Infraestruturas, ao promover a literacia estatística e a competência analítica necessárias à inovação baseada em dados e à avaliação de políticas sustentáveis em contextos tecnológicos e sociais.

Consequentemente, a Análise de Dados Categóricos pode ser vista como uma UC já com alguma história, com relevância atual e fundamental no futuro. Pode dizer-se que é uma UC que combina tradição (lecionada desde 2007/08) com inovação: i) práticas pedagógicas inovadoras, como o recurso à revisão cruzada de código e trabalhos baseados em dados reais, que estimulam a participação ativa, a autonomia e a discussão e comunicação crítica; ii) integração da reprodutibilidade computacional, através da proposta de uso de relatórios anexando *scripts* de R, garantindo que cada análise pode ser replicada e auditada; iii) preocupação com equidade e calibração, pois para além da qualidade do ajustamento, valoriza-se a capacidade de identificar eventuais enviesamentos, garantir a inteligibilidade dos resultados para públicos não especialistas e assegurar que as probabilidades previstas estão devidamente calibradas.

A terminar, a UC justifica-se cientificamente por responder a necessidades essenciais de modelação estatística, e pedagogicamente por promover aprendizagens nucleares numa comunidade de estudantes heterogénea e interdisciplinar.

1.1. Integração com investigação e sociedade

A unidade curricular Análise de Dados Categóricos constitui o espaço natural onde confluem a formação académica, o percurso científico e a experiência docente do responsável. A motivação decorre de mais de três décadas de docência em Probabilidade,

Inferência e Modelação, articuladas com investigação aplicada. Esta experiência moldou a arquitetura da UC, garantindo que os tópicos lecionados refletem problemas reais e dados empíricos de elevada relevância social, reforçando a autenticidade e a motivação dos estudantes.

Nos últimos anos, foram coordenados ou integrados diversos projetos de impacto comunitário, que servem de suporte direto à UC. Entre eles destacam-se:

- i) os estudos sobre hábitos de atividade física e sono das crianças do 1.º ciclo de Évora, em parceria com a Câmara Municipal, ARS Alentejo e HESE, envolvendo modelação binária, ordinal e longitudinal;
- ii) o Diagnóstico Juvenil de Évora, em parceria com a Câmara Municipal de Évora, que originou relatórios e livros aplicando regressões logísticas e análise de perfis;
- iii) o projeto AdolesSer, em parceria com o Agrupamento de Centros de Saúde do Alentejo Central (ACES), centrado na educação para a saúde em saúde sexual e reprodutiva, com dados categóricos;
- iv) os estudos sobre envelhecimento ativo e qualidade de vida em seniores e adolescentes, em parceria com a Câmara Municipal de Évora, com dados ordinais e longitudinais;
- v) a caracterização da prática desportiva no concelho de Évora, em parceria com a Câmara Municipal de Évora, utilizando GLM, perfis, *machine learning* e amostragem complexa;
- vi) a epidemiologia clínica das feridas em contexto hospitalar e comunitário, em colaboração com a Sociedade Portuguesa de Feridas (ELCOS), recorrendo a modelos logísticos e de *machine learning*;
- vii) o projeto MOPREVIS – Modelação e Predição de Acidentes de Viação (FCT, 2019–2022), em parceria com a GNR, onde foram aplicados GLM, GEE, GLMM e integração com *machine learning* para previsão da gravidade dos acidentes e identificação de segmentos de maior risco.

Esta ligação entre investigação e ensino é reforçada por produção científica, que inclui livros, capítulos e artigos internacionais. Entre as obras de maior impacto estão o estudo sobre os determinantes da fecundidade em Portugal (FFMS, 2016), que, usando amostras complexas, exemplifica aplicações de regressões logísticas, ordinais, multinomiais e de processos de contagens em demografia; artigos em revistas internacionais de referência sobre segurança rodoviária, quedas em adultos, prevalência de feridas, modelação em ecologia marinha e saúde pública, e estudos recentes em áreas diversas como sustentabilidade e ciência alimentar, todos com aplicação de GLM, GLMM ou modelos de contagem. Esta produção não é listada como mérito curricular, mas para ser integrada na UC como material de apoio, estando disponível no Moodle para consulta e discussão pelos estudantes.

A incorporação destes projetos e publicações na UC tem três efeitos diretos:

- a)** Autenticidade pedagógica — os estudantes trabalham com dados reais, muitas vezes inéditos, aproximando a sala de aula dos contextos profissionais e de investigação.
- b)** Reforço metodológico — cada exemplo está alinhado com as famílias de modelos ensinadas (GLM, modelos de contagem, GEE, GLMM, ZIF), garantindo correspondência entre teoria e aplicação.

- c) Transferência de conhecimento — relatórios técnicos, publicações científicas e livros utilizados como material complementar demonstram como resultados estatísticos se traduzem em recomendações para políticas públicas, boas práticas profissionais e decisões informadas.

Em síntese, a Análise de Dados Categóricos constitui um espaço formativo que alia rigor científico, inovação pedagógica e aplicabilidade prática. A sua conceção assegura continuidade e consistência, independentemente do docente responsável, garantindo que os estudantes beneficiam sempre de uma aprendizagem exigente, atualizada e socialmente relevante.

1.2. Pré-requisitos e nivelamento

A diversidade de origens académicas dos estudantes torna inviável definir pré-requisitos em termos de unidades curriculares específicas. Em vez disso, estabelecem-se pré-requisitos em termos de aprendizagem prévia que permitem acompanhar a UC com sucesso:

- Domínio básico de probabilidades e inferência estatística (noções de variáveis aleatórias, estimadores, intervalos de confiança e testes de hipóteses);
- Experiência prévia, mesmo que introdutória, com regressão linear ou logística, incluindo interpretação de coeficientes e significância estatística;
- Competências elementares em R, incluindo a importação e organização de dados, utilização de funções básicas para estatística descritiva e regressão simples, e criação de representações gráficas.

Para estudantes que não cumpram integralmente estes requisitos, são disponibilizados materiais de nivelamento e sessões de apoio opcionais na Semana 0, com guias em R e exercícios de revisão. Em paralelo, a metodologia descrita na Secção 4 garante diferenciação pedagógica, permitindo ajustar o ritmo e o nível de apoio à heterogeneidade das turmas.

2. Objetivos de aprendizagem e competências a adquirir

A UC de Análise de Dados Categóricos foi concebida para garantir que os estudantes, independentemente da sua formação de base, adquiram um conjunto sólido de competências conceptuais, técnicas e transversais que os prepare para a prática profissional, para a investigação aplicada e para a comunicação de resultados a públicos diferenciados. Esta secção apresenta os objetivos da UC em duas dimensões complementares: (i) objetivos de aprendizagem específicos, que traduzem os conhecimentos e capacidades a adquirir, e (ii) competências desenvolvidas, que articulam saber e saber-fazer.

2.1. Objetivo Principal

O objetivo principal da ADC é capacitar os estudantes para especificar, ajustar, diagnosticar e interpretar modelos estatísticos adequados a dados categóricos e de contagem, compreendendo os pressupostos subjacentes, as limitações práticas e as implicações éticas

e sociais da sua utilização. Esta capacidade não se limita à dimensão técnica, incluindo também a comunicação crítica dos resultados, em linha com o recomendado internacionalmente em ciência de dados e com o compromisso institucional da Universidade de Évora com o ODS 4 – Educação de Qualidade.

2.2. Objetivos específicos

Os objetivos de aprendizagem da UC organizam-se em diferentes dimensões, que se complementam e estruturam o desenvolvimento de competências dos estudantes.

a) Compreender e aplicar medidas de associação, correlação e concordância entre variáveis categóricas

Os estudantes deverão dominar a construção e interpretação de tabelas de contingência, calculando e discutindo medidas como *razões de chances*, riscos relativos, ou V de Cramér, e compreender fenómenos de confundimento e paradoxos, como o de Simpson. Este objetivo assegura a ponte entre métodos descritivos e inferenciais e é especialmente relevante para estudantes provenientes das ciências sociais e da epidemiologia, onde as tabelas de associação constituem instrumentos de base. Além disso, os estudantes deverão conseguir calcular, avaliar a significância estatística e interpretar medidas de correlação quando em análise estão variáveis de natureza categórica ou ordinal, ou quando uma destas se relaciona com uma variável contínua. Neste âmbito, incluem-se coeficientes como o de Spearman, o gamma de Goodman-Kruskal, o bisserial, o bisserial por pontos e o tetracórico, cuja escolha depende da escala de medição e do problema em análise. Finalmente, deverão também compreender e aplicar medidas de concordância entre avaliadores ou instrumentos, como o coeficiente kappa de Cohen, o índice de concordância de Kendall e o kappa de Fleiss, bem como interpretar coeficientes de concordância intraclasse. Esta vertente é particularmente relevante em contextos de epidemiologia, ciências da saúde e psicométrica, onde a fiabilidade das classificações é condição essencial de validade científica.

b) Especificar, ajustar e interpretar modelos lineares generalizados (GLM)

O núcleo da UC reside no domínio dos GLM aplicados a respostas binárias, ordinais e multinomiais, constituindo a base metodológica para uma grande variedade de problemas práticos. Os estudantes deverão compreender como a escolha da função de ligação influencia a interpretação dos parâmetros e a adequação do modelo, explorando ligações logit, probit, *cloglog* e alternativas ordinais. Serão discutidos os princípios da estimação por máxima verosimilhança, bem como aspetos ligados à convergência, separação e identificação de modelos.

c) Modelar variáveis de contagem

A capacidade de ajustar modelos de regressão de Poisson e Binomial Negativa, de reconhecer a presença de sobredispersão e de incorporar *offsets* para modelar taxas constitui uma competência essencial no tratamento de dados de contagem. Estes modelos têm aplicações transversais em múltiplas áreas: em epidemiologia, para analisar o número de casos ou episódios clínicos; em ecologia, para modelar abundância de espécies; em desporto, para descrever eventos por unidade de tempo; em ciências

sociais, para avaliar frequência de comportamentos ou participação em atividades; em engenharias, para estudar falhas de equipamentos ou tráfego em redes; em economia e gestão, para quantificar transações ou pedidos de serviço; e em ciência de dados, como base para modelos probabilísticos de fenómenos discretos em larga escala. Este objetivo garante que os estudantes compreendem os limites do pressuposto de normalidade e dominam ferramentas adequadas para lidar com distribuições discretas caracterizadas por variância não constante, assegurando análises estatísticas robustas e interpretáveis em contextos reais.

d) Reconhecer e lidar com problemas de desbalanceamento e raridade de eventos

Em muitos contextos práticos, as respostas de interesse são raras ou desbalanceadas, como, por exemplo, acidentes rodoviários fatais, deteção de doenças pouco prevalentes, clientes que deixam de comprar ou de usar o serviço de uma empresa durante um certo período, fraude financeira, falhas em sistemas industriais ou eventos climáticos extremos. Os estudantes deverão compreender que a utilização acrítica de classificadores pode levar a modelos enviesados que favorecem a classe maioritária. Pretende-se que consigam diagnosticar este problema, discutir abordagens de correção e avaliar métricas adequadas, como curvas ROC e PR, área sob a curva ou *F1-score*. Este objetivo é particularmente relevante em saúde pública, ciências forenses, análise de risco e ciência de dados aplicada, onde o impacto de decisões erradas em eventos raros pode ter consequências significativas.

e) Integrar informação de amostras complexas

Grande parte da investigação empírica baseia-se em desenhos amostrais complexos, sejam eles estratificados, por conglomerados ou com ponderação, cuja estrutura deve ser reconhecida e incorporada na análise. Os estudantes devem compreender que ignorar o desenho amostral pode introduzir enviesamentos e comprometer a validade da inferência estatística. Mediante dados reais, é explorada a aplicação de métodos de estimação com pesos e replicação. Tal é fundamental em diferentes áreas: por exemplo, na epidemiologia e saúde pública, garante representatividade de populações; nas ciências sociais, assegura comparabilidade entre grupos; em estudos de mercado, permite extrapolar conclusões para populações-alvo. Tal permitirá aos estudantes ter conhecimentos exigidas em contextos de estatística oficial e estudos de grande escala.

f) Avaliar equidade e calibração dos modelos

Um modelo estatístico não deve ser avaliado somente pela sua precisão global. A capacidade de verificar se o desempenho é consistente entre diferentes subgrupos é essencial para assegurar equidade. Do mesmo modo, um modelo só é útil se as probabilidades preditas forem calibradas, ou seja, se refletirem de forma fiel a frequência observada dos eventos. A integração deste objetivo reforça também o contributo da UC para os ODS 5 e 10, ao dotar os futuros profissionais de ferramentas que permitem identificar e mitigar desigualdades sistemáticas em processos de decisão baseados em dados.

g) Produzir relatórios reprodutíveis e comunicar resultados de forma clara

A comunicação estatística exige não somente a apresentação de resultados, mas também a capacidade de documentar processos de forma transparente e reprodutível. Os estudantes deverão desenvolver competências na análise reprodutível de dados em R, com ênfase na documentação clara e na replicabilidade dos resultados. O uso de *Quarto* é apresentado de forma gradual e opcional, como extensão natural das boas práticas de integração entre texto, código e resultados. Para além da dimensão técnica, este objetivo inclui a comunicação eficaz: elaboração de relatórios acessíveis, produção de gráficos claros, apresentação oral estruturada e discussão crítica dos resultados. Esta competência é transversal a todas as áreas representadas no perfil de ingresso: desde relatórios técnicos em matemática aplicada e ciência de dados, passando por relatórios de investigação em saúde ou sociologia, até à produção de *dashboards* em gestão ou engenharia informática. Trata-se de um objetivo que prepara os estudantes para uma prática científica e profissional que valoriza transparência, rigor e credibilidade, pilares da ciência aberta e da tomada de decisão baseada em evidência científica.

Se os objetivos explicitam o que se espera que os estudantes sejam capazes de fazer ao longo e no final da unidade curricular, as competências traduzem esses objetivos em capacidades técnicas, científicas e transversais, articulando-os com a prática profissional e a formação interdisciplinar.

2.3. Competências desenvolvidas

Pretende-se que os estudantes adquiram competências em 3 domínios fundamentais: científico, técnico e transversal.

- i. Capacidade para, de forma crítica e autónoma, construir, ajustar e analisar diferentes modelos lineares generalizados (binário, ordinal, multinomial, Poisson, Binomial Negativa, modelos inflacionados de zeros), diagnosticando resíduos, sobredispersão e adequação da função de ligação.
- ii. Domínio dos princípios fundamentais da modelação estatística e das principais fases da modelação com GLM, incluindo especificação, estimação, validação, calibração e comunicação de resultados.
- iii. Capacidade de interpretar e analisar problemas envolvendo dados longitudinais e correlacionados, via extensões como GEE e GLMM, compreendendo os seus pressupostos e limitações.
- iv. Saber pesquisar e compreender literatura científica especializada, avaliando criticamente pressupostos e justificando a seleção de modelos adequados para dados categóricos.
- v. Capacidade de utilizar o *software* R para a análise de dados categóricos em contexto reprodutível, assegurando a organização e documentação do código, a gestão de dependências e, quando aplicável, a integração em relatórios dinâmicos em *Quarto*.

- vi. Desenvolvimento de pensamento crítico e autonomia na análise de dados, em paralelo com competências de comunicação escrita e oral para públicos técnicos e não técnicos, integrando preocupações éticas, de equidade e de transparência.
- vii. Capacidade de trabalhar colaborativamente em grupos heterogêneos, reconhecendo a interdisciplinaridade como oportunidade de aprendizagem, e para manter uma atitude de literacia científica e de aprendizagem ao longo da vida.

Estas competências não se esgotam no plano metodológico: têm impacto direto nos diferentes percursos formativos a que a UC serve, e encontram eco nos Objetivos de Desenvolvimento Sustentável (ODS), que enquadram o papel social da estatística e da ciência de dados. A UC contribui diretamente para os Objetivos de Desenvolvimento Sustentável, nomeadamente através da análise de dados categóricos para avaliar a equidade e a calibração de modelos em contextos de género e desigualdade.

Estas competências são avaliadas de forma coerente com os objetivos da UC, assegurando alinhamento construtivo: o que se ensina é o que se avalia (ver secção 5.4).

2.4. Articulação com o plano de estudos

A unidade curricular Análise de Dados Categóricos articula-se com diversas unidades curriculares dos planos de estudo dos Mestrados em Modelação Estatística, Inteligência Artificial e Ciência de Dados, Matemática e da Pós-Graduação em Epidemiologia, tanto na vertente metodológica como na aplicada. Esta articulação garante continuidade formativa, coerência pedagógica e aplicabilidade interdisciplinar.

a) Articulação vertical (fundamentos estatísticos e metodológicos):

- *Inferência Estatística* fornece os princípios de estimação e testes que sustentam os modelos logísticos, multinomiais e de contagens;
- *Análise de Dados com Software Estatístico* fornece práticas iniciais de análise de tabelas de contingência e regressões simples, que servem de base para ADC;
- *Amostragem e Sondagens* garante a correta utilização de planos amostrais e ponderações, essenciais em muitas aplicações reais da ADC;
- *Delineamento Experimental* introduz estruturas de dados e controlos de variabilidade que posteriormente são tratados com modelos categóricos.
- *Estatística Computacional* desenvolve algoritmos (MLE, EM, Reamostragem, MCMC) transversais ao ajustamento e diagnóstico dos modelos.

b) Articulação horizontal (mesma fase do plano de estudos):

- *Estatística de Dados Multivariados* complementa ADC ao lidar com dependências entre múltiplas variáveis e técnicas de redução de dimensão;
- *Séries Temporais* oferece extensões em contextos longitudinais, onde modelos categóricos com GEE/GLMM são aplicáveis;
- *Mineração de Dados* integra algoritmos de classificação e problemas de classes desbalanceadas, complementando a modelação estatística interpretável;
- *Econometria* reforça a ligação a modelos com variáveis dependentes discretas e de painel, próximos dos enquadrados em ADC;

- *Inteligência Artificial Aplicada e Tópicos de Inteligência Artificial e Ciência de Dados* permitem comparar ADC com métodos de aprendizagem supervisionada / não supervisionada.

c) Articulação aplicada (extensões naturais da ADC):

- *Análise de Sobrevida* generaliza a modelação de tempos até evento, próxima da lógica da regressão logística;
- *Modelação e Análise de Eventos Extremos* partilha problemas de fenómenos raros, análogos ao desafio de classes desbalanceadas;
- *Modelação Matemática em Biologia* aplica regressões de contagem e modelos populacionais, próximos das metodologias centrais de ADC;
- *Processos Estocásticos e Aplicações* fornece enquadramento teórico para modelos de Poisson.

d) Articulação no domínio da saúde e políticas públicas (Pós-Graduação em Epidemiologia – PGE):

- *Epidemiologia e Saúde Pública* aplica medidas de incidência, prevalência e fatores de risco, diretamente modeláveis com técnicas categóricas;
- *Epidemiologia Ambiental* utiliza regressões logísticas e multinomiais para avaliar determinantes ambientais;
- *Medições em Saúde: Questionários e Escalas* gera dados ordinais e nominais, objeto direto da ADC.

Para tornar essa articulação explícita, apresenta-se na Tabela 1 uma síntese, de modo sistemático, com as unidades curriculares relacionadas. Desta forma, evidencia-se que a ADC não é uma unidade curricular isolada, mas um nó estruturante na rede curricular, reforçando competências metodológicas transversais que suportam a análise empírica nos diferentes contextos.

Tabela 1 – Quadro-síntese da articulação da unidade curricular *Análise de Dados Categóricos* com outras unidades curriculares dos planos de estudos dos cursos a que é lecionada.

Categoria	Unidades Curriculares Relacionadas
Epidemiologia e Saúde	Epidemiologia e Saúde Pública; Epidemiologia Ambiental; Medições em Saúde: Questionários e Escalas
Estatística Aplicada	Amostragem e Sondagens; Inferência Estatística; Estatística Computacional; Estatística de Dados Multivariados
Modelação Avançada	Análise de Sobrevida; Modelação e Análise de Eventos Extremos; Modelação Matemática em Biologia
Inteligência Artificial e Ciência de Dados	Mineração de Dados; Inteligência Artificial Aplicada; Tópicos de Inteligência Artificial e Ciência de Dados

2.5. Ligação aos cursos e aos ODS

A importância dos conteúdos programáticos da ADC é transversal ao conjunto de estudantes que a frequentam, mas o seu impacto assume contornos distintos consoante o curso em que se insere.

No Mestrado em Modelação Estatística e Análise de Dados, a UC constitui a espinha dorsal da formação aplicada em modelação estatística. É com ela que os estudantes consolidam a capacidade de especificar, ajustar e validar modelos para variáveis categóricas, competências indispensáveis para a realização de dissertações de natureza aplicada e para o exercício profissional enquanto estatísticos nos mais diversos setores de atividade.

No Mestrado em Inteligência Artificial e Ciência de Dados, a ADC desempenha um papel de complemento crítico à aprendizagem automática. Ao oferecer uma perspetiva probabilística, interpretável e calibrada, a UC fornece as bases necessárias para explicar, validar e contextualizar modelos de *machine learning*. Assim, contribui para equilibrar a sofisticação algorítmica com a responsabilidade interpretativa, o que é particularmente valorizado em aplicações de IA em saúde, finanças ou políticas públicas.

No Mestrado em Matemática, reforça a formação teórica e inferencial. Ao introduzir rigorosamente os modelos lineares generalizados e as suas extensões, proporciona aos estudantes contacto com metodologias estatísticas clássicas que podem servir de ponto de partida para investigação fundamental em estatística, probabilidade e modelação matemática.

Na Pós-Graduação em Epidemiologia, a ADC disponibiliza instrumentos analíticos indispensáveis para a avaliação de riscos, incidências e desigualdades em saúde pública. Permite aos profissionais compreender e aplicar modelos adequados ao estudo de doenças, fatores de risco e contextos populacionais, fortalecendo a ligação entre estatística aplicada, epidemiologia e políticas de saúde.

No plano dos Objetivos de Desenvolvimento Sustentável (**ODS**), a UC contribui para vários domínios. Para além do já referido **ODS 4** – Educação de Qualidade, destaca-se o **ODS 3** – Saúde e Bem-Estar, pela formação de profissionais capazes de avaliar riscos e incidências em saúde; o **ODS 5** – Igualdade de Género e o **ODS 10** – Redução das Desigualdades, ao integrar práticas de avaliação da equidade dos modelos e de análise por subgrupos; e ainda o **ODS 9** – Indústria, Inovação e Infraestruturas, ao promover práticas modernas de ciência de dados (reprodutibilidade, transparência e inovação metodológica) que apoiam a construção de sistemas e serviços mais robustos e inclusivos.

Definidos os objetivos e as competências, importa agora responder a uma questão central: que conteúdos, em 45 horas, podem garantir que os estudantes atingem esses objetivos? A resposta traduz-se na organização do programa, que combina bases comuns, extensões aplicadas e literacia crítica, tal como se descreve no bloco seguinte.

3. Conteúdos programáticos

O programa de Análise de Dados Categóricos foi concebido para garantir uma progressão pedagógica que parte de conceitos descritivos fundamentais, evolui para modelos lineares generalizados (GLM) aplicados a diferentes tipos de respostas categóricas e culmina em extensões para situações de maior complexidade. Esta organização responde simultaneamente a duas exigências: (i) assegurar que todos os estudantes, independentemente do seu percurso académico prévio, adquirem as bases indispensáveis de modelação de dados categóricos, e (ii) expor os estudantes a temas recentes, ainda que numa lógica de literacia crítica, dada a limitação de 45 horas presenciais.

O programa está estruturado em três blocos complementares: núcleo essencial, complementar aplicado e literacia alargada. Cada bloco responde a diferentes níveis de profundidade e tem funções pedagógicas distintas.

3.1. Núcleo essencial

O núcleo essencial constitui a base comum a todos os estudantes e ocupa cerca de dez semanas (30 horas). Inicia-se com a construção e análise de tabelas de contingência, introduzindo medidas de associação, associação condicional, correlação e concordância entre variáveis categóricas, bem como a distinção entre associação estatística e causalidade. Nesta fase, os estudantes aprendem a ler e a comunicar razões de chance e riscos relativos e a utilizar testes formais criteriosamente: o qui-quadrado de Pearson e a razão de verosimilhanças (G^2) quando as condições assintóticas são cumpridas, o teste exato de Fisher em 2×2 (ou generalizações para dimensões superiores quando aplicável), e o teste de tendência de Cochran–Armitage quando há ordenação natural. É também aqui que se discutem os fenómenos de confundimento e o paradoxo de Simpson, justificando a estratificação e o uso do estimador e teste de Mantel–Haenszel (e a verificação de homogeneidade com Breslow–Day sempre que faça sentido), sem transformar estes tópicos num curso de inferência avançada: o objetivo é que o estudante reconheça quando estas ferramentas são necessárias e como as usar corretamente.

Para além da associação, trabalha-se a correlação adaptada à natureza das variáveis: Spearman aplicável a variáveis ordinais ou quantitativas ordenáveis, adequada para medir relações monotónicas (crescentes ou decrescentes) mesmo quando não lineares; gamma de Goodman-Kruskal para ordinais com leitura de pares concordantes/discordantes; phi/V de Cramér para nominais; e, em cenários mistos, bisserial/bisserial por pontos. Quando a estrutura latente o justifica, referem-se as versões tetracórica (e, por extensão, policórica para ordinais), sempre esclarecendo pressupostos e limitações. A concordância é abordada no essencial para problemas de fiabilidade: kappa de Cohen (nominal, dois avaliadores), Kendall (ordinais), kappa de Fleiss (múltiplos avaliadores) e coeficiente de correlação intraclass (ICC) para medidas repetidas, distinguindo consistência de acordo absoluto. Aqui o foco é operacional, selecionando o índice certo, estimando, apresentando o intervalo de confiança e interpretando no contexto do problema, sem derivar formalismos remetidos para leituras recomendadas.

Segue-se a introdução aos Modelos Lineares Generalizados (GLM), com um grande foco nas respostas binárias. Apresenta-se a teoria comum dos GLM: família exponencial, preditor linear, função de ligação, estimação por máxima verosimilhança e *deviance*. Pretende-se que os estudantes percebam que os modelos subsequentes partilham a mesma estrutura. Na regressão binária, comparam-se *logit*, *probit* e *cloglog*, esclarecendo as implicações interpretativas (razões de chance versus efeitos marginais e variação em probabilidade), e introduz-se o diagnóstico no nível adequado a esta fase: resíduos de deviance/Pearson, observações influentes e noções de separação/quase-separação (com solução prática referida, como penalização tipo Firth, a desenvolver mais tarde). Avança-se depois para modelos ordinais, com ênfase no modelo de *odds* proporcionais e em testes de proporcionalidade, explicitando alternativas quando o pressuposto falha; e, por fim, para modelos multinomiais com categoria de referência, enfatizando identificação e efeitos marginais na comunicação de resultados.

Ao longo desta parte, trabalha-se de forma sistemática a avaliação preditiva e a calibração, mas com bom senso na profundidade: os estudantes aprendem a distinguir entre discriminação — capacidade do modelo para separar corretamente as classes (curvas ROC e PR, AUC, sensibilidade, especificidade, precisão, F1, exatidão balanceada) — e calibração, que avalia a correspondência entre probabilidades previstas e observadas (Brier score, intercepto, declive). São ainda abordados testes de bondade de ajustamento, como Hosmer–Lemeshow e Cessie–van Houwelingen, discutidos com os devidos cuidados.

Em problemas com desbalanceamento da classe minoritária, discute-se o impacto que este fenómeno tem nas métricas globais, em particular, como uma elevada exatidão pode esconder o fraco desempenho na classe de interesse. Para mitigar este efeito no âmbito dos GLM, introduzem-se estratégias diretamente implementáveis, como o ajustamento do limiar de decisão e a utilização de pesos de classe (*case-weights*), que permitem corrigir a desproporção entre categorias na estimação. Estas soluções são trabalhadas em exemplos simples, de forma que os estudantes percebam a lógica e a sua aplicação prática em R. Metodologias de reestruturação da amostra (*resampling*), como SMOTE ou ROSE, são referidas somente de forma introdutória, ficando a exploração mais detalhada para a subsecção 3.3., dedicada à literacia alargada em métodos contemporâneos de classificação.

Introduz-se, ainda, uma nota de literacia sobre amostras complexas: quando os dados provêm de estratificação/conglomerados/ponderação, testes e erros-padrão usuais podem induzir decisões incorretas; por isso, mostra-se quando usar correções de desenho, como, por exemplo, Rao–Scott em tabelas, e quando recorrer a uma abordagem *design-based*, como, por exemplo, a função *svyglm* do pacote *survey* do R. O objetivo, nesta fase, é sensibilizar e preparar o terreno, não transformar a aula num curso de amostragem: a prática aprofundada surge nos trabalhos e em aulas dedicadas.

Os exemplos pedagógicos desta fase são realizados com bases de dados reais já consolidadas na UC, favorecendo a transferência e a motivação. A Base II – Atividade Física e Desportiva funciona como fio condutor: permite trabalhar medidas de associação e estratificação, aplicar regressão logística, ordinal e multinomial, e introduzir, com dados

reais, a questão dos pesos de calibração e da inferência em amostras complexas. Em saúde pública, a Base I – Feridas proporciona desfechos binários e ordinais, bem como contagens (número de feridas), possibilitando exercitar associações clínicas e demográficas, avaliar fiabilidade e discutir interpretação em contexto clínico. A Base III – Hábitos de Sono em Crianças reforça a aplicação de modelos binários, ordinais e multinomiais, o uso de efeitos marginais, e permite trabalhar com diagnóstico e avaliação de modelos (incluindo calibração e discriminação).

A procura um equilíbrio criterioso, assegurando domínio prático do essencial (tabelas, associação estratificada, GLM binário, ordinal e multinomial com diagnóstico e leitura de desempenho/calibração) serve de porta de entrada para tópicos que exigem maior maturidade (desbalanceamento e desenho complexo), sem dispersar tempo em formalismos excessivos. O objetivo é que, ao fim destas semanas, todos os estudantes escolham o modelo adequado, ajustem de acordo com os passos da modelação, validem pressupostos e interpretem efeitos.

3.2. Complementar aplicado

O bloco complementar aplicado ocupa cerca de cinco semanas (15 horas) do semestre e pretende ampliar o leque de ferramentas disponíveis, introduzindo modelos adequados a fenómenos de contagem e a situações em que os dados apresentam dependência temporal ou estrutural. Esta parte do programa assegura que os estudantes não ficam limitados a respostas binárias, ordinais ou multinomiais, mas adquirem competências para lidar com contextos onde a variável resposta é um número inteiro não negativo, frequentemente associado a taxas, bem como com dados onde as observações não são independentes.

Inicia-se com a regressão de Poisson, apresentada como a extensão natural dos modelos abordados anteriormente a dados de contagem. O modelo é explicado na sua lógica essencial: a ligação log garante que as taxas de ocorrência são sempre positivas, e a inclusão de um *offset* permite comparar grupos em termos de taxas ajustadas pela exposição. Os estudantes exploram exemplos concretos como o número de atropelamentos de vertebrados por troço do IP2 (Base de dados V).

Em seguida, aborda-se a limitação central da Poisson: o pressuposto de que a média é igual à variância. Na prática, muitos conjuntos de dados apresentam sobredispersão, quer devido a heterogeneidade não observada, quer devido a agregação espacial ou temporal. Para lidar com esta situação, apresentam-se diferentes estratégias. Uma primeira consiste em introduzir um parâmetro de dispersão na estimação dos erros-padrão, o que pode ser feito recorrendo ao modelo quasi-Poisson, preservando a estrutura da ligação *log* mas ajustando a variância observada. Esta abordagem é muitas vezes preferida na prática quando o objetivo principal é inferência e não modelação preditiva. Como alternativa, a Binomial Negativa é introduzida como solução natural, explicitando a sua flexibilidade em acomodar variâncias superiores à média por um parâmetro adicional. O foco não está na formulação matemática detalhada, mas sim em desenvolver uma literacia prática: saber diagnosticar sobredispersão, compreender quando uma correção simples é suficiente e

quando é preferível recorrer à Binomial Negativa, e interpretar corretamente as estimativas em cada caso.

Muitos fenómenos de contagem exibem um excesso de zeros relativamente ao esperado por Poisson ou Binomial Negativa. Exemplos típicos incluem o registo de atropelamentos no IP2 (vários troços sem ocorrências num dado período), a ocorrência de reinternamentos em saúde, o número de créditos solicitados por certos grupos populacionais ou o número de acidentes com vítimas mortais. Nesta fase, os estudantes são sensibilizados para a diferença entre zeros estruturais (verdadeiros “nunca ocorre”) e zeros ocasionais. Apresenta-se, assim, a lógica dos modelos inflacionados a zeros (ZIF) e dos modelos *hurdle*, que combinam um processo para zeros e outro para contagens positivas. Não se aprofunda a estimação, mas garante-se que os estudantes reconhecem estes cenários, compreendem as consequências de ignorar a estrutura de zeros e conseguem comunicar a racionalidade de utilizar tais modelos.

Vários são os contextos que envolvem observações não independentes: medidas repetidas no mesmo indivíduo, estudantes dentro de turmas, doentes dentro de hospitais. Para lidar com estas situações, introduzem-se duas abordagens complementares: as Equações de Estimação Generalizadas (GEE) e os Modelos Lineares Generalizados Mistos (GLMM).

As GEE são apresentadas como solução para estimar efeitos marginais robustos, permitindo obter inferências válidas mesmo em presença de correlação Intra grupo. Os GLMM, por sua vez, acrescentam efeitos aleatórios para capturar variação entre unidades, oferecendo uma leitura condicional e mais flexível. Os exemplos pedagógicos incluem: o estudo de prevalência de feridas (ULSAC/ELCOS), em que pacientes estão agrupados em unidades de saúde (Base de Dados 1).

O ensino destas metodologias não visa domínio técnico exaustivo, mas sim a compreensão dos seus fundamentos conceptuais e a capacidade crítica de escolher entre abordagens marginais (GEE) e condicionais (GLMM), sabendo identificar os contextos em que são mais adequados.

3.3. Literacia alargada

O terceiro bloco é operacionalizado através da Parte II do segundo trabalho prático, integrando-se diretamente no processo de avaliação sem consumir tempo adicional de aulas. O objetivo não é o domínio técnico aprofundado, mas sim a exposição crítica a tópicos avançados e emergentes, que marcam a investigação e a prática contemporâneas em modelação estatística e ciência de dados.

Nesta componente, os estudantes trabalham em grupos, exploram um tema específico, previamente atribuído pelo docente, produzindo uma apresentação curta, uma síntese dos principais conceitos e aplicações, e um exemplo em R (um script simples e útil, com anotações que o ajudam a entender). A atividade termina numa sessão de partilha em que os grupos apresentam uns aos outros os seus resultados, promovendo aprendizagem colaborativa, espírito crítico e literacia estatística em rede.

Os temas são escolhidos de forma dinâmica, variando de ano para ano, para se adaptarem ao perfil heterogéneo dos estudantes e ao contexto científico mais atual. No ano letivo 2024/25, a Parte II incluiu:

- Modelos de Regressão Logística Hierárquica (multinível)
- Modelos de Teoria da Resposta ao Item (TRI)
- Modelos Aditivos Generalizados (GAM) para respostas categóricas
- Modelos Lineares Mistos Generalizados (GLMM) para respostas categóricas
- Modelos de Zeros Inflacionados
- Random Forest
- XGBoost
- Support Vector Machines (SVM)
- Redes Neurais
- Modelos de Classes Latentes
- Modelos explicáveis e interpretáveis em classificação categórica (SHAP, LIME e *fairness*-equidade/justiça algorítmica).

Incluem-se ainda métodos de reamostragem (SMOTE, ROSE), a regressão logística de Firth para cenários com separação quase/completa ou amostras pequenas e modelos log-lineares, temática abordada em algumas unidades curriculares afins, como se verá na Secção 10.

Por um lado, procura-se garantir que os estudantes contactam com novas metodologias que são amplamente usadas em investigação e aplicações profissionais, mesmo que não dominem a sua implementação completa. Por outro lado, pretende-se ampliar horizontes, situando os modelos abordados, no núcleo e no complementar, num panorama mais vasto, mostrando que a estatística categórica se articula com métodos da aprendizagem automática, da psicometria, da modelação hierárquica e da explicabilidade.

Do ponto de vista pedagógico, esta componente tem várias mais-valias:

- i) permite consolidar a autonomia na procura, leitura e síntese de literatura científica;
- ii) desenvolve competências de comunicação científica, tanto escrita como oral, perante públicos técnicos e não técnicos;
- iii) estimula o pensamento crítico, ao obrigar os estudantes a identificar vantagens, limitações e pressupostos dos métodos estudados;
- iv) valoriza a aprendizagem colaborativa num ambiente interdisciplinar, onde cada grupo contribui para a formação coletiva.

Esta exposição crítica permite que os estudantes, em entrevistas de trabalho ou em contextos profissionais, reconheçam a terminologia e saibam situar as técnicas, mesmo que não dominem a sua implementação total. Paralelamente, o contexto de situações reais incentiva-os a tomar/simular decisões informadas, ponderando alternativas de modelação, implicações práticas e limitações dos resultados. Assim, a aprendizagem técnica transforma-se num exercício de raciocínio aplicado.

É importante sublinhar que esta abordagem não implica aumentar a carga horária nem dispersar o núcleo essencial da UC. Pelo contrário, constitui uma estratégia de avaliação inovadora, que transforma a avaliação em momento de aprendizagem, promovendo

literacia estatística avançada e preparando os estudantes para reconhecerem, no futuro, quando e como recorrer a metodologias mais sofisticadas.

3.4. Racional e coerência curricular

Os conteúdos programáticos foram organizados para corresponder diretamente aos objetivos da UC. A introdução com tabelas de contingência e medidas de associação/correlação/concordância garante literacia descritiva e inferencial de base. Segue-se a modelação por GLM binários, ordinais e multinomiais, que concretiza a competência central de análise de respostas categóricas. O bloco complementar expande a aplicação a dados de contagem, zeros em excesso e dados correlacionados, cobrindo cenários realistas de sobredispersão, taxas e hierarquias. Por fim, a literacia alargada introduz tópicos emergentes, promovendo autonomia crítica. Esta progressão evita redundâncias e garante que cada tópico serve um objetivo concreto: saber escolher, ajustar, diagnosticar, interpretar e comunicar modelos adequados a diferentes naturezas de resposta e desenhos de estudo.

Este percurso garante que todos os estudantes dominam um conjunto mínimo de ferramentas aplicáveis e que, ao mesmo tempo, ficam sensibilizados para os avanços metodológicos mais recentes. O programa é, por isso, simultaneamente realista (respeita as 45 horas de contacto) e inovador (expõe os estudantes a preocupações éticas, de reprodutibilidade e equidade que, do que conhecemos, ainda não são padrão no ensino nacional).

A UC contribui, ainda, para desenvolver uma perspetiva crítica sobre a equidade e a calibração de modelos. Por exemplo, discute-se como métricas de desempenho podem revelar enviesamentos e como a modelação estatística disciplinada pode ajudar a mitigá-los em contextos de saúde, género ou desigualdade social.

Entre todas as extensões consideradas no programa, a introdução dos modelos de contagem merece um destaque especial. Mais do que um complemento, constituem uma exigência científica e pedagógica, como se desenvolve de seguida.

3.5. A pertinência de modelos de contagens em Análise de Dados Categóricos

A inclusão de modelos de contagens nesta unidade curricular é uma exigência científica, pedagógica e ética. Uma parte substancial da evidência empírica com que trabalhamos em saúde pública, ecologia, segurança rodoviária, ciências sociais, economia, engenharia e desporto, entre outras áreas, assume a forma de números de ocorrências por unidade de tempo, espaço ou exposição: internamentos, episódios de doença, atropelamentos por troço e período, golos por jogo, falhas de equipamento por hora de operação, acidentes por semana numa via, entre muitos outros exemplos.

Reduzir estes fenómenos a desfechos binários (“houve / não houve”) implica perda de informação e, em muitos casos, decisões incorretas. A regressão de Poisson fornece o enquadramento probabilístico adequado, com interpretação direta em termos de razões de taxas/incidência (IRR), facilmente comunicáveis a decisores e profissionais. Na prática, o pressuposto de igual dispersão raramente se verifica. As ferramentas lecionadas permitem

reconhecer e corrigir sobredispersão, evitando erros-padrão subestimados e inferências ilusoriamente significativas. A possibilidade de incorporar *offsets* é outro ponto decisivo: comparar acidentes por milhão de veículo-quilômetro é diferente de comparar números brutos; o mesmo se aplica a hospitalizações por 1000 habitantes ou golos por minuto de jogo. Esta noção de taxa ajustada reforça a validade das análises e assegura comparações justas entre unidades, hospitais, regiões ou equipas.

Do ponto de vista pedagógico, os modelos de contagens prolongam naturalmente a gramática dos GLM introduzida no núcleo (mesma lógica de ligação *log*, *deviance* e estimação), enquanto consolidam competências transferíveis: diagnóstico de sobredispersão, escolha entre especificações alternativas, leitura de *deviance* e critérios de informação, e comunicação de efeitos em escala multiplicativa. Constituem também a ponte natural para extensões mais avançadas (modelos inflacionados a zeros, GEE/GLMM), mostrando aos estudantes que os problemas reais respeitam raramente a simplicidade dos manuais.

Existe, ainda, uma dimensão ética. Comparar números absolutos de eventos entre hospitais, escolas ou regiões, sem considerar exposição ou população em risco, pode perpetuar comparações injustas e enviesar políticas públicas. Ao ancorarem a análise no conceito de taxa, os modelos de contagem promovem equidade e transparência, alinhando-se diretamente com os ODS 3 (Saúde e Bem-Estar) e ODS 10 (Redução das Desigualdades).

Em síntese, a presença de modelos de contagem nesta UC garante aos estudantes uma competência de base, científica e profissionalmente indispensável, que aumenta a relevância da formação e os prepara para comunicar resultados com rigor, justiça e clareza em contextos aplicados.

Tendo delineado a progressão e justificado a presença de cada bloco, importa agora sintetizar os principais tópicos abordados, para ter um guia rápido, sem substituir a descrição detalhada já apresentada.

3.6. Sumário dos Tópicos Programáticos abordados

A título de síntese, os conteúdos da unidade curricular podem ser organizados nos seguintes tópicos principais, que servem de referência rápida e estruturada.

1. Associação e Associação condicional: tabelas de contingência, Razão de Chances (Odds Ratio - OR), Riscos Relativos (Risk Ratio - RR), V de Cramér; qui-quadrado, G^2 , Teste de Fisher, Teste de Cochran–Armitage; Teste de Mantel–Haenszel; Teste de Breslow–Day.
2. Correlação e concordância: Spearman; gamma de Goodman–Kruskal; phi/V de Cramér; bisserial/bisserial por pontos; tetracórica/policórica; kappa (Cohen, ponderado, Fleiss); Kendall W; ICC.
3. GLM para respostas categóricas: binária (logit, probit, cloglog), ordinal (*odds* proporcionais e alternativas), multinomial; diagnóstico (resíduos, influência); efeitos marginais; calibração (Brier, *calibration-in-the-large*, *slope*, gráficos; HL, Cessie–van Houwelingen); discriminação (ROC/PR, AUC, exatidão balanceada, precisão, F1).

4. Questões transversais: desbalanceamento (limiares e pesos de classe; *resampling* apenas referencial), amostras complexas (Rao–Scott; *svyglm*), validação (holdout, k-fold, *bootstrap*).
5. Contagens: Poisson, quasi-Poisson (parâmetro de dispersão) e Binomial Negativa; ZIF/hurdle (excesso de zeros); interpretação em taxas/IRR.
6. Dependência/longitudinal: GEE (efeitos marginais) e GLMM (efeitos aleatórios; leitura condicional).
7. Literacia alargada: Modelos Aditivos Generalizados (GAM), logística hierárquica, Teoria da Resposta ao Item (TRI), Modelos Lineares Mistos Generalizados (GLMM – respostas categóricas), Random Forest, XGBoost, SVM, Redes Neurais, Classes Latentes, SHAP/LIME/Fairness, SMOTE/ROSE, logística de Firth, Modelos Log-Lineares.

Os conteúdos programáticos organizam-se de forma progressiva, articulando fundamentos, modelos centrais, extensões aplicadas e literacia estatística. Esta lógica encontra-se resumida na Tabela 2, que sintetiza a articulação entre conteúdos e objetivos de aprendizagem.

Tabela 2. Estrutura dos conteúdos programáticos e respetivos objetivos de aprendizagem.

Bloco Temático	Objetivos de Aprendizagem Relacionados
Tabelas de contingência e medidas de associação.	Desenvolver literacia descritiva e inferencial de base.
GLM para respostas categóricas (binário, ordinal, multinomial).	Adquirir competências centrais de modelação estatística.
Modelos de contagem e dados correlacionados.	Alargar capacidade de análise a contextos realistas (taxas, sobredispersão, hierarquias).
Literacia alargada.	Desenvolver leitura crítica, comunicação científica e reflexão ética sobre o uso de modelos estatísticos e algoritmos.

O que ensinar está agora clarificado. Mas o sucesso da UC depende igualmente de como ensinar. O próximo bloco descreve a metodologia pedagógica adotada e as estratégias de gestão da heterogeneidade, sem as quais o programa não poderia ser concretizado de forma eficaz.

4. Metodologia de ensino e diferenciação pedagógica

Como referido anteriormente, a UC é frequentada por estudantes de formações muito diversas, das ciências exatas às sociais e da saúde. Esta heterogeneidade é, ao mesmo tempo, um desafio e uma oportunidade ao exigir um desenho pedagógico capaz de apoiar quem tem menor familiaridade estatística, sem desmotivar quem chega com uma preparação mais sólida. A estratégia de ensino combina rigor científico, flexibilidade didática e promoção da autonomia.

As aulas decorrem em regime teórico-prático, evitando a fragmentação entre exposição conceptual e aplicação computacional. Cada novo tópico é introduzido com uma explicação clara dos fundamentos estatísticos e, em seguida, explorado via exemplos concretos com bases de dados reais. Esta articulação garante que os estudantes percebem não somente a formulação dos modelos, mas também a sua utilidade prática e os limites da sua aplicação. O *software* R é a ferramenta central, utilizada em aula e nos trabalhos através de scripts reprodutíveis e documentados, promovendo boas práticas de ciência aberta.

A gestão da heterogeneidade de formação dos estudantes baseia-se em três pilares:

- a) Aprendizagem incremental e em espiral** - cada novo conceito retoma e aprofunda ideias já trabalhadas, evitando temas isolados e reforçando a compreensão. Por exemplo, ao transitar do *logit* binário para o ordinal, os estudantes reutilizam a lógica das razões de chance, mas analisam pressupostos adicionais como o das *odds* proporcionais e as suas alternativas. Com esta organização, os estudantes com menor experiência conseguem consolidar bases sólidas sem se sentirem perdidos, enquanto os mais avançados encontram novos níveis de profundidade e complexidade que os mantêm intelectualmente envolvidos. Na regressão de Poisson, retomam-se os princípios já aplicados na logística binária de forma que a estrutura conceptual aprendida num contexto binário seja reaproveitada para compreender fenómenos de contagem. De forma análoga, os diagnósticos de sobredispersão reaproveitam a intuição da leitura de resíduos e *deviance*, agora aplicada a novos cenários.
- b) Exemplos autênticos e diversificados** - as bases de dados escolhidas refletem a pluralidade dos cursos: saúde, fenómenos sociais, desporto e aplicações biológicas. Esta diversidade reforça a motivação e a transferência de conhecimento.
- c) Atividades práticas orientadas para a autonomia** - os estudantes são incentivados a trabalhar de forma independente, mas com apoio estruturado. Problemas práticos são resolvidos em aula, primeiro em conjunto e depois em pequenos grupos. Esta progressão fortalece a confiança e cria oportunidades para colaboração entre estudantes de diferentes origens, promovendo a literacia cruzada: quem domina a estatística ajuda quem tem mais experiência contextual, e vice-versa. A organização dos grupos de trabalho é intencionalmente heterogênea, reunindo estudantes de diferentes mestrados e formações de base, constituindo uma estratégia pedagógica inovadora para potenciar aprendizagem colaborativa e interdisciplinar.

A metodologia também integra a reprodutibilidade (relatórios que combinam texto, código e resultados) e discute a dimensão da equidade, sublinhando que escolhas estatísticas podem amplificar ou corrigir desigualdades.

Em termos de dinâmica de sala de aula, recorre-se frequentemente a estratégias de ensino ativo:

- i) pequenos *quizzes* em grupo, promovendo análise dos *outputs* (contribuindo para avaliação caso melhore a nota do estudante);

- ii) períodos de pensar sozinho, trocar ideias com um parceiro e depois discutir com a turma para incentivar a interação entre estudantes;
- iii) apresentações breves dos estudantes sobre temas avançados (no âmbito da literacia alargada, integradas na avaliação).

Estas práticas tornam as aulas mais participadas e reduzem a distância entre estudantes com diferentes perfis.

Finalmente, a gestão da heterogeneidade estende-se além da sala de aula, através de múltiplos canais de apoio. O docente mantém contacto próximo com os estudantes, com atendimento presencial e *online* em português ou inglês (quando necessário, para estudantes estrangeiros, repetindo até aulas). Para dúvidas rápidas e interação entre pares, existe um grupo *WhatsApp*, que funciona como fórum aberto e fomenta uma cultura colaborativa de entreajuda. Estão, ainda, disponíveis materiais de apoio diversificados, adaptados a diferentes perfis: *handouts* simplificados para consolidar conceitos-chave, *scripts* comentados passo a passo em R e referências bibliográficas que vão de introduções acessíveis a obras avançadas. Este ecossistema complementa as aulas e potencia a aprendizagem autónoma, permitindo que todos avancem ao seu ritmo.

Em síntese, a metodologia alia rigor conceptual, exemplos autênticos e aprendizagem ativa a uma gestão consciente da diversidade de perfis. Esta estratégia garante que a unidade curricular não se limita a transmitir técnicas estatísticas, mas constitui um verdadeiro espaço de formação sólida, equitativa e interdisciplinar, preparando os estudantes para enfrentar dados reais com espírito crítico, autonomia e sentido de responsabilidade.

4.1. Estratégia pedagógica e alinhamento com os objetivos de aprendizagem

A metodologia de ensino foi concebida para assegurar correspondência direta entre os objetivos de aprendizagem e as práticas pedagógicas. O ponto de partida são aulas teórico-práticas, em que a exposição dos fundamentos estatísticos é imediatamente acompanhada da aplicação em R com bases de dados reais de saúde, ciências sociais, ecologia ou desporto. Esta integração garante que os estudantes compreendem os modelos e pressupostos estatísticos, reconhecendo também à sua utilidade prática, limitações e implicações específicas.

O treino computacional, suportado por *scripts* reproduzíveis, promove autonomia e práticas de ciência aberta. Os projetos em grupo desenvolvem competências de análise, síntese e comunicação científica, exigindo clareza tanto na redação técnica como na interpretação para públicos não especialistas. Os *quizzes* individuais, realizados em aula com tempo limitado (15 a 20 minutos), verificam o entendimento pessoal e estimulam a autorregulação. As apresentações de literacia alargada cultivam pensamento crítico, síntese e comunicação oral. Em conjunto, estas metodologias operacionalizam os objetivos declarados: aplicar autonomamente modelos lineares generalizados a dados categóricos, interpretar resultados com rigor, diagnosticar pressupostos, comunicar conclusões de forma clara e agir de modo responsável e ético.

O alinhamento entre metodologias e objetivos é reforçado por três princípios:

- a) Aprendizagem ativa** – “aprender fazendo”, com tarefas autênticas baseadas em dados reais, onde o erro é parte do processo formativo;
- a) Avaliação formativa e *feedback* útil** – trabalhos acompanhados de rúbricas detalhadas e relatórios de *feedback*, identificando pontos fortes, lacunas e sugestões de melhoria;
- b) Atenção à diversidade** – práticas que garantem que todos os estudantes encontram oportunidades de progressão, sem reduzir a exigência académica.

O desenho curricular mostra-se robusto, exigente e responsivo à heterogeneidade das turmas, equilibrando rigor científico, aplicação prática e justiça pedagógica. As taxas de sucesso e os indicadores de comparabilidade dos últimos anos confirmam que os resultados de aprendizagem são atingidos e notificam planos de melhoria contínua.

Em síntese, a coerência entre metodologias e objetivos de aprendizagem é operacionalizada de forma clara e sistemática. A Tabela 3 resume esta correspondência, mostrando como cada prática pedagógica contribui diretamente para o desenvolvimento das competências nucleares da unidade curricular.

4.2. Acessibilidade e inclusão

A unidade curricular adota princípios de desenho universal da aprendizagem, assegurando condições equitativas para a participação e progressão de todos os estudantes. Entre as práticas incluem-se: disponibilização de *handouts* sintéticos, uso de linguagem clara nos enunciados, escolha de fontes legíveis e contraste adequado nas figuras. Sempre que necessário, são articulados ajustes razoáveis nos processos de avaliação em conjunto com os serviços competentes da Universidade de Évora. O objetivo é reduzir barreiras que possam limitar a aprendizagem, sem comprometer a exigência científica e académica.

Tabela 3. Metodologias de ensino e ligação direta aos objetivos de aprendizagem.

Metodologia / Atividade	Contributo para Objetivos de Aprendizagem
Aulas teórico-práticas com R e bases reais.	Compreender fundamentos estatísticos e aplicar em contextos reais.
Scripts reproduzíveis.	Promover autonomia e boas práticas de ciência aberta.
Projetos em grupo.	Desenvolver competências de análise, síntese e comunicação científica.
Quizzes individuais.	Assegurar compreensão pessoal e autorregulação.
Apresentações de literacia alargada.	Cultivar pensamento crítico, capacidade de síntese e comunicação oral.
Feedback estruturado em rúbricas.	Potenciar progressão contínua e melhoria do desempenho.

5. Avaliação

Se a metodologia de ensino define como os conteúdos são trabalhados e adaptados à diversidade dos estudantes, a avaliação é o momento em que esse processo se materializa. Na unidade curricular de Análise de Dados Categóricos, a avaliação não se limita a medir resultados: é um mecanismo pedagógico que consolida aprendizagens, estimula autonomia e promove literacia estatística.

Concebida como extensão natural da metodologia, a avaliação privilegia autenticidade, reprodutibilidade e comunicação, equilibrando trabalho colaborativo com momentos de responsabilização individual. Procura-se medir o que realmente importa que os estudantes saibam fazer com dados reais, garantindo simultaneamente justiça e transparência.

5.1. Princípios orientadores e componentes

A avaliação rege-se por princípios que asseguram consistência científica e pedagógica:

- a) **Alinhamento construtivo** – cada componente de avaliação mede diretamente os objetivos definidos na secção 2: modelar, diagnosticar, interpretar e comunicar resultados.
- b) **Autenticidade** – resolução de problemas com dados reais, relatórios reprodutíveis e critérios próximos da prática profissional em ciência de dados.
- c) **Feedback útil** – devolvido de forma rápida, acompanhado de rubricas claras que identificam pontos fortes e áreas de melhoria, promovendo aprendizagem contínua.
- d) **Equidade** – combinação de trabalhos em grupo com mecanismos de verificação individual, assegurando justiça e promovendo a colaboração.
- e) **Transparência** – regras, fórmulas de cálculo e condições de ponderação são comunicadas logo no início do semestre, reforçando confiança no processo avaliativo.

O regime de avaliação contínua assenta em dois trabalhos de grupo, que constituem a base da classificação, complementados por pequenos *quizzes* em Moodle.

O **Trabalho I**, com peso de 50% na nota final, centra-se na aplicação de modelos GLM a respostas categóricas. Os estudantes devem elaborar um relatório reprodutível baseado em dados reais, no qual se avaliam competências como a construção e interpretação de tabelas de contingência, a análise de associação condicional e a especificação de modelos GLM binários, ordinais e multinomiais. São ainda trabalhados diagnósticos de *outliers* e de observações influentes, métricas de discriminação e calibração, desenho de validação e tratamento do desbalanceamento. O corpo do relatório não deve ultrapassar 13 páginas, de modo a estimular a capacidade de síntese e a distinção entre o essencial e o acessório, conjugando rigor estatístico com conclusões inteligíveis para públicos não especializados. Em anexo devem constar as análises complementares que sustentam os resultados (por exemplo, verificação de pressupostos ou gráficos de resíduos), bem como o ficheiro de *script* completo, garantindo a replicabilidade do trabalho.

O **Trabalho II**, igualmente com peso de 50% na nota final, divide-se em duas componentes distintas. A **Parte I**, com peso de 25% da nota, é dedicada a modelos de contagens (Poisson, quasi-Poisson e Binomial Negativa), a modelos de zeros em excesso (quando abordados em

aula) e a modelos para dados correlacionados (GEE e/ou GLMM). O foco recai sobre a decisão de modelação, a interpretação dos efeitos e a clareza da comunicação. As exigências formais são idênticas às do primeiro trabalho. A **Parte II**, com peso de 25% na nota, destina-se à exploração de um tema avançado, atribuído pelo docente e preparado em grupo com base na bibliografia recomendada e em pesquisa autónoma. Os estudantes devem produzir uma apresentação de cerca de 20/25 minutos, num formato livre, que sintetize os conceitos fundamentais, destaque a aplicabilidade prática em contextos reais e inclua exemplos ilustrativos. Para além da apresentação oral, devem entregar o material produzido e um *script* em R comentado, que traduza na prática os principais aspetos discutidos. Esta componente culmina numa sessão de partilha em aula, onde os grupos apresentam e discutem os temas com os colegas, promovendo literacia alargada e aprendizagem colaborativa.

São ainda realizados quatro pequenos *quizzes* de interpretação de *outputs* em Moodle. Estes exercícios são curtos, focados e com tempo limitado (15 a 20 minutos no final das aulas), funcionando também como ferramentas formativas. São criados de modo que a sua correção possa valorizar não apenas a resposta final, mas também a capacidade de interpretar *outputs* e raciocinar com base em evidência. Constituem um incentivo à prática regular, recompensam o empenho dos estudantes e reforçam a vertente individual da avaliação. Têm a função de treinar leitura de *outputs*, diagnóstico e comunicação. Do nosso ponto de vista são sempre positivos (só ajudam).

- Período 1 (até ao Trabalho I): sendo $\bar{Q}_1$ a média dos *quizzes* deste período, a nota final do Trabalho I será dada por:

$$T_{1_final} = 0,8 \times T_1 + 0,2 \times \bar{Q}_1$$

- Período 2 (entre o Trabalho I e o final das aulas): sendo $\bar{Q}_2$ se a média dos *quizzes* deste período e T_{2P1} a nota da Parte I do Trabalho 2, a nota final do Trabalho II será dada por:

$$T_{2P1_final} = 0,8 \times T_{2P1} + 0,2 \times \bar{Q}_2,$$

A nota final do estudante em avaliação contínua será, pois, dada por

$$0,5 \times T_{1_final} + 0,25 \times T_{2P1_final} + 0,25 \times T_{2P2}$$

representando T_{2P2} a nota da Parte II do Trabalho 2.

O regime de avaliação final por exame prevê a realização de uma prova escrita, em época normal e em época de recurso, destinada aos estudantes que não realizam a avaliação contínua, que nela reprovam ou que pretendem melhorar/recuperar a classificação. Para assegurar a coerência entre os dois regimes, foi definido um mecanismo de valorização do trabalho desenvolvido ao longo do semestre: caso a média ponderada dos Trabalhos I e II seja superior à nota obtida no exame escrito, essa média contribui com 25% da classificação final, combinando-se com a nota do exame. Caso contrário, a classificação resulta exclusivamente da prova escrita.

De acordo com a ficha oficial da UC, a aprovação exige classificação final ≥ 10 valores, quer pelo regime contínuo (trabalhos, apresentações e *quizzes*), quer pelo regime de exame final.

5.2. Avaliação individual e autorregulação

A avaliação nesta unidade curricular assenta no trabalho de grupo, refletindo a realidade profissional na qual a modelação estatística raramente é uma atividade isolada. A regra geral é que a classificação dos trabalhos é comum a todos os elementos do grupo. Para garantir justiça e responsabilização, existem dois mecanismos de diferenciação individual, simples e transparentes.

Em primeiro lugar, os *quizzes* individuais realizados em Moodle incidem sobre a interpretação de *outputs* estatísticos. A classificação baseia-se na correção das respostas e na clareza da interpretação, funcionando como evidência objetiva de domínio individual. Estes momentos reforçam a responsabilidade de cada estudante na consolidação dos conhecimentos.

Em segundo lugar, na Parte II do Trabalho II, cada elemento do grupo deve participar ativamente na apresentação e discussão. Sempre que um estudante se destaque positivamente, ou revele falhas significativas, a classificação pode ser ajustada até ± 1 valor relativamente à nota do grupo. Este ajustamento é excecional e apenas aplicado quando existe evidência clara de desempenho distinto.

Complementarmente, integram-se exercícios de autoavaliação no final de cada aula, que não têm impacto direto na nota final, mas estimulam a autorregulação, permitindo aos estudantes acompanhar a evolução das suas competências, identificar fragilidades e planear melhorias.

Em síntese, a filosofia de avaliação da unidade curricular procura conciliar rigor científico, formação prática e equidade. A ênfase em dados reais, relatórios reproduzíveis e apresentações em grupo garante que os estudantes não se limitam a acumular classificações, mas desenvolvem competências transferíveis para contextos profissionais e académicos. Ao mesmo tempo, a integração de elementos individuais (*quizzes*, participação em apresentações e meta-reflexão) assegura que cada estudante é responsável pelo seu próprio percurso de aprendizagem.

A avaliação é, naturalmente, uma exigência do sistema, mas aqui vai muito além de um número: constitui o espaço em que se consolidam práticas, se corrigem fragilidades e se adquirem hábitos de trabalho que preparam os estudantes para enfrentar com solidez os desafios profissionais e de investigação associados à modelação de dados categóricos. Mais do que certificar, a avaliação nesta unidade curricular forma, orienta e reforça competências essenciais, enquanto habitua os estudantes a interagir criticamente com as ferramentas digitais que moldam o futuro na profissão.

5.3. Síntese objetivos-avaliação

A avaliação contínua baseia-se em critérios explícitos e previamente comunicados, assegurando correspondência direta entre objetivos da UC, competências trabalhadas e critérios de classificação:

- a) Trabalho I (50%)** – incide sobre GLM para respostas categóricas. Avalia enquadramento dos dados (10%); modelação e justificações (40%); diagnóstico e desempenho — resíduos, influência, calibração (incluindo *calibration-in-the-large* e slope), bondade de ajustamento e discriminação (25%); comunicação e reprodutibilidade em R (com possibilidade de integração em *Quarto*) (25%).
- b) Trabalho II – Parte I (25%)** – centrado em modelos de contagem e dados correlacionados. Avalia adequação do modelo (30%); interpretação (30%); diagnóstico e validação (20%); comunicação e reprodutibilidade em R (com possibilidade de integração em *Quarto*) (20%).
- c) Trabalho II – Parte II (25%)** – dedicado à literacia alargada e comunicação científica. Avalia rigor e profundidade conceptual (35%); demonstração prática em R (25%); comunicação oral e escrita (30%); e responsabilização individual em discussão (10%).
- d) Quizzes individuais (componente adicional)** – aferem interpretação de *outputs* estatísticos em tempo limitado, valorizando correção e clareza.

A Tabela 4 sintetiza a correspondência entre os objetivos operacionais de aprendizagem, os conteúdos/competências trabalhados e os critérios de avaliação definidos em 5.4, evidenciando que “o que se ensina é o que se avalia”.

Os trabalhos escritos seguem um formato profissional e conciso, limitado em extensão, centrado na interpretação dos resultados e na comunicação das conclusões principais. Todos os elementos que suportam a análise, tais como tabelas complementares, diagnósticos, scripts em R e outputs, são incluídos em anexo técnico, garantindo reprodutibilidade integral. Esta estratégia ensina os estudantes a distinguir entre o essencial e o acessório, desenvolvendo competências de síntese científica e de documentação de código. A utilização de *Quarto* é opcional e progressiva, sendo apresentada como boa prática para integração texto-código, mas sem constituir requisito obrigatório, atendendo à heterogeneidade de proficiências em R e à carga cognitiva associada. Assim, a UC assegura uma formação sólida em reprodutibilidade e comunicação científica, sem comprometer o equilíbrio pedagógico e o tempo disponível no semestre.

Tabela 4 — Alinhamento construtivo Objetivos–Conteúdos–Avaliação. Os critérios de avaliação listados correspondem exatamente aos definidos em 5.4 para o Trabalho I e II. Inclui-se ainda a contribuição opcional dos *quizzes* como elemento de reforço formativo.

Objetivo de aprendizagem (o estudante deve ser capaz de...)	Conteúdos/Competências trabalhados	Critério de avaliação	Instrumento
Interpretar associação em tabelas, incluindo OR e RR	Tabelas RxC; qui-quadrado, Fisher; Mantel–Haenszel; Breslow–Day; phi; V de Cramér	Trabalho I — Modelação e justificações	Relatório escrito
Avaliar correlação e concordância	Spearman; gamma de Goodman–Kruskal; bisserial; tetracórica/policórica; kappa; Kendall W; ICC	Trabalho I — Modelação e justificações	Relatório escrito
Ajustar e interpretar GLM binário, ordinal e multinomial	Regressão logística; odds proporcionais; modelos multinomiais	Trabalho I — Modelação e justificações	Relatório escrito
Realizar diagnóstico e avaliar desempenho de modelos	Resíduos; influência; calibração (<i>calibration-in-the-large</i> , slope, curvas de calibração); discriminação (ROC, PR, AUC)	Trabalho I — Diagnóstico e desempenho	Relatório escrito
Comunicar resultados de forma clara e reproduzível	Escrita científica; tabelas/gráficos; código R documentado	Trabalho I — Comunicação e reprodutibilidade	Relatório escrito + Código
Modelar contagens (Poisson, quasi-Poisson, NegBin, ZI, hurdle)	Modelos de contagem; sobredispersão; zeros excessivos	Trabalho II – Parte I — Adequação do modelo	Relatório escrito
Tratar dados correlacionados (dependência intra-unidade)	GEE, GLMM; efeitos aleatórios; correção marginal	Trabalho II – Parte I — Interpretação	Relatório escrito
Validar modelos com diagnóstico adequado	Resíduos; bondade de ajustamento; avaliação da cobertura dos intervalos de confiança/predição	Trabalho II – Parte I — Diagnóstico e validação	Relatório escrito
Comunicar ciência para públicos não especialistas	Escrita acessível; apresentação oral; literacia estatística	Trabalho II – Parte II — Comunicação oral e escrita	Apresentação oral
Demonstrar rigor conceptual e técnico	Conceitos avançados; fundamentação	Trabalho II – Parte II — Rigor e profundidade conceptual	Relatório/apresentação
Mostrar autonomia em R e responsabilização individual	Demonstração prática; discussão crítica	Trabalho II – Parte II — Demonstração prática em R + responsabilização individual	Relatório/apresentação
Consolidar competências básicas em tempo limitado	Questões de interpretação de <i>outputs</i>	<i>Quizzes</i>	Questionários online no Moodle

OR = razão de chances; RR = risco relativo; ROC/PR = curvas *receiver operating characteristic/ precision-recall*; ZI = *zero-inflated*; IC = intervalo de confiança; *svyglm* = regressão generalizada com desenho amostral.

5.4. Feedback, calendário e integridade académica

Após cada trabalho, os estudantes recebem nota e relatório detalhado, com pontos fortes, lacunas e sugestões de melhoria. Existe sempre a possibilidade de discussão presencial ou *online*.

O calendário é estruturado para equilibrar carga e progressão pedagógica: o Trabalho I é entregue 15 dias após a última aula sobre modelos multinomiais; o Trabalho II – Parte I três semanas após o final das aulas; a Parte II do Trabalho II é apresentada uma semana depois da entrega da Parte I. Os *quizzes* decorrem em dois momentos estratégicos: antes da entrega do Trabalho I e entre a sua entrega e o final das aulas.

A integridade académica é promovida desde o início: relatórios reprodutíveis, código comentado, citações completas e verificação de similaridade. Cada grupo deve declarar autoria e todos os *quizzes* são realizados em ambiente supervisionado. Sempre que necessário, ajustamentos avaliativos, como tempo adicional em *quizzes*, formatos alternativos de entrega ou adaptações no modo de apresentação, são naturalmente realizados. Estas adaptações visam remover barreiras específicas sem reduzir a exigência académica, garantindo condições equitativas para todos os estudantes.

5.5. Uso responsável de Inteligência Artificial Generativa

É esperado que os estudantes tenham acesso a ferramentas de IA generativa (ChatGPT, Gemini, Perplexity, entre outras). O seu uso é incentivado como apoio à aprendizagem, nunca como substituto do trabalho intelectual. Estas ferramentas podem ser úteis para clarificar dúvidas conceptuais, detetar erros de código ou ajudar na revisão de texto, mas não dispensam a compreensão crítica nem a autoria plena dos estudantes.

A estratégia assenta em três princípios: transparência (qualquer utilização deve ser declarada de forma explícita nos relatórios), responsabilização individual (cada estudante deve explicar e justificar as opções seguidas, mesmo quando apoiadas por IA) e reprodutibilidade (todos os resultados devem poder ser replicados a partir do código e dos dados entregues). *Quizzes* individuais, apresentações orais e pequenas variações do código submetido funcionam como mecanismos adicionais de verificação de autoria. O docente pode também solicitar micro-defesas orais (5–7 min) ou variações dirigidas do código para comprovar compreensão e capacidade de adaptação.

Reconhece-se, contudo, que estas ferramentas apresentam limitações relevantes: podem produzir respostas incorretas (*alucinações*), refletir enviesamentos dos dados de treino ou induzir dependência acrítica. Por isso, a sua utilização deve ser sempre mediada pelo conhecimento estatístico adquirido na UC, transformando potenciais riscos em oportunidades de aprendizagem crítica.

O presente relatório seguiu os mesmos princípios aqui defendidos. Foi utilizado apoio pontual de ferramentas de IA generativa exclusivamente para revisão linguística e edição textual, sendo todo o conteúdo científico, pedagógico e avaliativo da inteira responsabilidade do docente. Esta prática reforça a coerência entre as orientações transmitidas aos estudantes e a conduta seguida na própria elaboração deste documento.

6. Planificação

A planificação da unidade curricular foi desenhada de modo a respeitar a coerência pedagógica do programa e a assegurar que todos os blocos têm o espaço necessário para serem desenvolvidos com qualidade. O semestre estrutura-se em 15 semanas de contacto, cada uma com três horas de aulas presenciais, complementadas por trabalho autónomo dos estudantes. Assenta numa lógica de progressão pedagógica que conjuga rigor científico com realismo organizativo. Ainda que, na prática, em alguns anos as aulas tenham sido condicionadas por feriados ou outras interrupções, assume-se aqui o quadro de 15 semanas letivas, onde as últimas são dedicadas ao bloco complementar e à literacia alargada. Esta distribuição assegura que os conteúdos fundamentais não são sacrificados, e que a flexibilidade necessária é obtida deslocando tópicos menos nucleares para a parte final de literacia, quando necessário.

Na primeira semana, a aula inicia-se com a apresentação da UC, dos seus objetivos e do regime de avaliação. Este enquadramento ocupa somente parte da sessão, permitindo desde logo introduzir conteúdos do núcleo essencial. Inicia-se, assim, a construção e análise de tabelas de contingência, explorando medidas de associação e discussão crítica sobre a diferença entre associação estatística e causalidade. Esta abordagem inicial tem a dupla função de nivelar a turma, dado o seu carácter interdisciplinar, e de mergulhar os estudantes no primeiro contacto com dados reais.

Entre a segunda e a décima semana, desenvolve-se a totalidade do Núcleo Essencial. As três primeiras aulas deste ciclo são dedicadas a medidas de associação, correlação e concordância em variáveis categóricas. A partir da quarta semana, o foco desloca-se para os Modelos Lineares Generalizados (GLM) com respostas categóricas (binomial, ordinal e multinomial). O Trabalho I é lançado no decorrer da sexta semana e entregue na décima segunda semana, correspondendo ao momento em que os estudantes já consolidaram competências em tabelas de contingência e em GLM binários, ordinais e multinomiais. Os *quizzes* em Moodle, realizados em aula, funcionam em paralelo como treino individual de interpretação de *outputs*.

Entre a décima primeira e a décima quarta semanas decorre o Complementar Aplicado. Cerca de três semanas são dedicadas aos modelos de contagens e uma semana final aos modelos com dados correlacionados (GEE e GLMM). Também neste período funcionam *quizzes* no Moodle.

A décima quinta semana é reservada ao bloco de Literacia Alargada, já fora do período regular de aulas. Nesta sessão, os estudantes assumem protagonismo e apresentam os temas avançados que prepararam em grupo no âmbito da Parte II do Trabalho II.

Em cada uma das semanas do semestre, os estudantes são incentivados a complementar as 3 horas presenciais com cerca de 8 horas de estudo autónomo, distribuídas entre revisão do que foi lecionado, pequenos exercícios de compreensão e extensão, leitura de bibliografia, preparação de relatórios e realização de *quizzes*.

Além das 45 horas presenciais, a UC dispõe de 4 horas de tutorias formais, previstas no plano de estudos. Estas não substituem o regime de atendimento regular do docente

(metade das horas presenciais, conforme regulamento), mas funcionam como um recurso adicional e flexível. São tipicamente usadas em momentos críticos do semestre, por exemplo, antes da entrega de cada trabalho, para apoiar a interpretação de diagnósticos, a replicação em R ou a clarificação de decisões de modelação. Em alternativa, podem ser mobilizadas para pequenos grupos de nivelamento, quando a heterogeneidade da turma assim o exige.

A planificação é, assim, simultaneamente realista e flexível: realista porque respeita a experiência acumulada dos últimos anos, que mostra ser inviável lecionar com qualidade o núcleo essencial em menos de dez aulas; flexível porque prevê, desde o início, mecanismos de ajuste que permitem deslocar tópicos acessórios (ou que exijam maturidade adicional) para a literacia alargada quando imprevistos reduzem o número efetivo de sessões. Em qualquer cenário, o fio condutor mantém-se claro e estável: começar pelas fundações (tabelas, associação, correlação e concordância), avançar para os GLM (binário, ordinal, multinomial), estender a contagens (Poisson/quasi-Poisson/Binomial Negativa (BN), Zeros inflacionados (ZIF/hurdle) e dados correlacionados (GEE e GLMM) e concluir com uma abertura crítica a temas avançados.

Importa sublinhar que esta organização é adaptativa. O ritmo, a ênfase e a profundidade de cada tópico são ajustados ao grupo concreto de estudantes em cada ano, tendo em conta:

- i) a heterogeneidade da turma (proveniência disciplinar e objetivos profissionais);
- i) as dificuldades observadas na aprendizagem (por exemplo, maior tempo em associação estratificada e diagnóstico antes de entrar em modelos mais complexos);
- ii) os conhecimentos prévios em estatística e em R, com reforço inicial sempre que se detetam necessidades de nivelamento;
- iii) a evidência formativa recolhida nos *quizzes* em aula, que sinaliza lacunas a colmatar antes de avançar.

As bases de dados (algumas no Anexo A) podem ser substituídas por outras equivalentes se forem mais próximas do perfil da turma, sem comprometer os objetivos de aprendizagem. Quando o calendário real sofre estrangimentos (feriados, eventos institucionais), prioriza-se o núcleo essencial e deslocam-se tópicos acessórios para a literacia alargada, mantendo a coerência do fio GLM. Em turmas de maior dimensão, ajustam-se os tamanhos dos grupos e reforçam-se momentos assíncronos guiados (pontos formativos leves) para evitar sobrecarga presencial, mantendo o padrão de *feedback* e reprodutibilidade.

Esta flexibilidade não altera a carga global de trabalho nem os resultados de aprendizagem: as 45 horas presenciais, as 4 horas de tutorias e as cerca de 107 horas de estudo autónomo perfazem as 156 horas previstas para a UC, assegurando a correspondência com os 6 ECTS. O que muda é somente a distribuição interna de tempo e exemplos, sempre guiada pela evidência de aprendizagem e pelas necessidades do grupo, para que todos cheguem ao final do semestre com as competências nucleares dominadas e com uma visão atualizada e crítica da análise de dados categóricos.

Para tornar explícita a coerência entre os conteúdos programáticos, a carga horária e os 6 ECTS da unidade curricular, apresenta-se nas Tabelas 5 e 6, a distribuição aproximada de horas presenciais e de estudo autónomo por tópicos e pelos 3 grandes blocos de conteúdos.

Tabela 5 - Distribuição da carga de trabalho por blocos

Bloco	Conteúdos principais	Horas presenciais	Estudo autónomo	Total aproximado
Núcleo essencial (Semanas 1–10)	Tabelas de contingência, associação, correlação, concordância, GLM binário, ordinais e multinomiais	30h	69h	99h
Complementar aplicado (Semanas 11–14)	Modelos de contagens e de dados correlacionados	12h	28h	40h
Literacia alargada (Semana 15)	Apresentações e discussão de temas avançados (Parte II do Trabalho II)	3h	10h	13h
Tutorias	Apoio em momentos críticos (antes da entrega de trabalhos; dúvidas de diagnóstico/código)	4h	—	4h
Total	—	49h	107h	156h (6 ECTS)

Tabela 6 - Distribuição da carga de trabalho por conteúdos principais

Semana	Conteúdos principais	Horas presenciais	Estudo autónomo previsto
1	Apresentação da UC e regime de avaliação; introdução a tabelas de contingência; associação vs. causalidade	3h	5h
2	Medidas de associação em tabelas; qui-quadrado, OR, RR, V de Cramér	3h	5h
3	Correlação para variáveis ordinais/nominais; medidas de concordância (kappa, ICC)	3h	5h
4–8	GLM binário (logit, probit, cloglog); estimação, resíduos, separação, calibração, validação	15h (5x3h)	54h
9	GLM ordinais (odds proporcionais e alternativas)	3h	8h
10	GLM multinomiais (interpretação de coeficientes e probabilidades marginais)	3h	8h
11–12	Modelos de contagens: Poisson, quasi-Poisson, Binomial Negativa; diagnóstico de sobredispersão; offsets	6h (2x3h)	12h
13	Modelos de zeros em excesso (ZIF, hurdle)	3h	6h
14	Dados correlacionados: GEE (efeitos marginais) e GLMM (efeitos aleatórios)	3h	6h
15	Literacia Alargada: apresentações e discussão de temas avançados (Parte II do Trabalho II)	3h	12h

Naturalmente, os valores apresentados na Tabela 6 são apenas indicativos, em função do *background* estatístico, da familiaridade com R e da autonomia inicial de cada estudante. Servem, portanto, como referência média para a distribuição das horas presenciais e carga

de trabalho, e não devem ser vistos como estimativas rígidas. As horas de estudo autónomo resultam de estimativas conservadoras baseadas no histórico de turmas e na complexidade das tarefas (preparação dos trabalhos, leitura e exercícios). Os valores foram arredondados para múltiplos de hora, preservando o total de 156h/6 ECTS.

Esta estrutura reflete a importância relativa de cada componente, garantindo que os estudantes dispõem do tempo necessário para consolidar aprendizagens fundamentais, aplicar modelos em contextos diversos e desenvolver competências de síntese crítica.

Fica assim confirmada a consistência entre a carga de trabalho e os resultados de aprendizagem pretendidos: cerca de dois terços do esforço concentra-se no Núcleo Essencial, que constitui a base da UC; o Complementar Aplicado introduz competências transferíveis para contextos de contagem e dependência de dados; e a Literacia Alargada expõe os estudantes a temas emergentes, exigindo uma preparação autónoma relevante. As tutorias acrescentam flexibilidade e personalização, assegurando equidade no apoio a uma turma heterogénea. As taxas de sucesso registadas nos últimos anos e o *feedback* recolhido junto dos estudantes confirmam que esta calendarização, estruturada em blocos nucleares, complementares e de literacia alargada, é adequada para garantir rigor científico sem perder a dimensão formativa.

A planificação, complementada por esta distribuição da carga de trabalho, procura garantir o equilíbrio entre rigor científico, realismo pedagógico e sustentabilidade em termos de esforço do estudante. É nesta base que se articula o próximo bloco, dedicado à qualidade e melhoria contínua.

7. Qualidade, evidência e melhoria contínua

A caracterização dos métodos de avaliação da unidade curricular, descrita na secção anterior, ficaria incompleta sem a apresentação da evidência empírica resultante da sua aplicação nos últimos anos letivos. A reflexão pedagógica deve articular-se não apenas com o desenho metodológico, mas também com os resultados efetivos alcançados, tanto em termos de sucesso académico como de perceção dos estudantes relativamente à UC e ao desempenho docente.

Nos três anos mais recentes, a UC registou taxas de aprovação elevadas. Em 2022/23 e 2023/24, com turmas reduzidas (9 e 10 estudantes avaliados), a taxa de sucesso foi de 100%. Em 2024/25, perante um crescimento substancial da turma (24 estudantes avaliados), a taxa de aprovação situou-se em 88%. Este ligeiro decréscimo é considerado natural no contexto de maior heterogeneidade e de exigência mantida, revelando que a UC conseguiu escalar sem perda relevante de eficácia pedagógica.

Relativamente à avaliação da UC e do docente pelos estudantes, os resultados são sintetizados através do Índice de Comparabilidade (IC), métrica institucional usada pelo Gabinete de Planeamento e Garantia da Qualidade da Universidade de Évora. Este índice resulta da transformação das respostas dos estudantes aos questionários numa escala padronizada de quatro níveis: elevado (IC = 4), suficiente (IC = 2), insuficiente (IC = -2) e reduzido (IC = -4). Assim, valores positivos indicam avaliação favorável, sendo a proximidade a 4 sinal de desempenho muito positivo, enquanto valores negativos apontam

para necessidade de melhoria. Trata-se, pois, de uma métrica padronizada que permite comparar as respostas entre diferentes cursos e anos letivos.

Entre os anos letivos de 2022/23 e 2024/25 responderam aos questionários de avaliação 8, 13 e 9 estudantes, respetivamente. Apensar das reservas que os valores devem merecer (por exemplo, em 2023/24 responderam mais estudantes do que os que foram avaliados e em 2024/25 responderam menos de metade dos estudantes do que os que foram avaliados) considera-se esta ferramenta. Os questionários da UC evidenciam apreciações consistentes, com destaque para a importância atribuída à aquisição de competências relevantes para o futuro profissional (IC=3,14) e para a correspondência entre conteúdos lecionados e avaliação (IC=3,04). Também a adequação dos métodos de avaliação obteve valores positivos (IC=2,76), ainda que com margem para clarificação adicional de critérios. Já o nível de aquisição de conhecimentos (IC=2,48) e a acessibilidade da matéria face aos conhecimentos prévios (IC=2,41) constituem dimensões onde se identifica espaço para reforço, nomeadamente por meio de estratégias de nivelamento e de suporte pedagógico mais evidente.

Relativamente ao docente, os indicadores são mais fortes no domínio da matéria (IC=3,60), assiduidade e disponibilidade na plataforma (IC=3,59) e prontidão para esclarecer dúvidas (IC=3,45). Também o respeito demonstrado na relação pedagógica surge bem valorizado (IC=3,31). Como pontos a merecer atenção, surgem a clareza na apresentação (IC=2,69) e a perceção de equidade nas classificações (IC=2,76). O índice global da prestação docente, situado em IC=3,20, está acima da média institucional, evidenciando um reconhecimento sólido por parte dos estudantes.

Estes resultados alimentam o ciclo de melhoria contínua adotado pela Universidade de Évora: planear, executar, verificar (através dos questionários e taxas de aprovação) e atuar (ajustando metodologias e critérios).

No eixo da clareza expositiva, prevê-se o uso de mapas-conceito por tópico, folhas orientadoras para o estudo que liguem cada modelo ao anterior, e *outputs* de R com anotação mais completa, acompanhados de momentos de regulação do raciocínio no fecho de cada aula.

Quanto à equidade, a prioridade passa por reforçar a transparência na avaliação através da divulgação antecipada de um mapa das rubricas de avaliação, onde se explicitam os itens avaliados e a ponderação de cada critério. Esta prática, complementada com exemplos comentados de relatórios e o reforço dos *quizzes* individuais em aula como verificação contínua de aprendizagem, contribui para que a transparência já assegurada seja plenamente reconhecida e percecionada pelos estudantes.

Para potenciar a acessibilidade e o nível de aquisição, prevê-se a disponibilização de uma “semana zero” opcional de nivelamento em R e inferência estatística, além de materiais de apoio visuais e exercícios-ponte.

A monitorização faz-se por conversas abertas com os estudantes, análise semestral dos IC e das taxas de aprovação e registo de ajustes no Sistema de Informação Integrada da Universidade de Évora (SIIUE).

No plano institucional, a garantia da qualidade da UC é assegurada em dois níveis. Internamente, pela articulação entre metodologias, conteúdos e evidência pedagógica. Externamente, pelo enquadramento do Regulamento Académico da Universidade de Évora e pela supervisão do Conselho Pedagógico da Escola de Ciências e Tecnologia (CP-ECT), que aprova e homologa metodologias de avaliação e calendários. O júri da UC garante isenção e transparência nos processos de avaliação. O Sistema Interno de Garantia da Qualidade (SIGQ), acreditado pela A3ES, fornece instrumentos de monitorização sistemática do sucesso académico. Sempre que se identificam UCs com problemas de desempenho, são elaborados planos de melhoria acompanhados pelo CP-ECT e monitorizado pelo Conselho de Avaliação, que pode emitir recomendações às Comissões de Curso. Este enquadramento assegura que a UC não funciona isoladamente, mas integrada num sistema institucional de qualidade que reforça a credibilidade das práticas pedagógicas e garante convergência com padrões nacionais e internacionais.

Define-se como metas:

- i) elevar os IC de clareza expositiva e de perceção da equidade das classificações para $IC \geq 3,5$;
- ii) manter no domínio de matéria $IC \geq 3,5$;
- iii) estabilizar a taxa de aprovação em pelo menos 90% com turmas heterogéneas.

O progresso será monitorizado por análise semestral do IC por item, pelas taxas de aprovação e por conversa com os estudantes após a avaliação dos trabalhos.

A consolidação de uma cultura de qualidade depende não apenas de métricas e evidências institucionais, mas também dos recursos pedagógicos e tecnológicos que sustentam a prática diária. É neste enquadramento que se apresenta, de seguida, a forma como a UC garante reprodutibilidade, acessibilidade e inclusão.

8. Recursos, reprodutibilidade e inclusão

A planificação e a distribuição da carga de trabalho só cumprem plenamente a sua função quando apoiadas por um ecossistema de recursos adequados, por práticas que assegurem a reprodutibilidade das análises e por uma atenção constante à inclusão de todos os estudantes. Esta secção organiza-se em torno destes três pilares, que sustentam o ensino da unidade curricular e garantem que a experiência formativa vai além da mera transmissão de conteúdos.

8.1. Recursos

A UC assenta no *software* R, em regime *open-source*, com pacotes especializados que cobrem desde a análise de tabelas de contingência até modelos estatísticos avançados. Esta opção tecnológica não é apenas pragmática, pois garante acessibilidade sem custos adicionais, acompanha a evolução da comunidade científica internacional e assegura ligação direta ao mundo profissional e académico, onde é amplamente utilizado em estatística aplicada, ciência de dados e investigação interdisciplinar.

Para consolidar esta escolha, a UC disponibiliza um conjunto diversificado de recursos pedagógicos que respondem a diferentes perfis de estudantes. As bases de dados reais

(algumas descritas no Anexo A) permitem a aplicação prática em contextos substantivos variados, reforçando a transferibilidade das competências adquiridas. Em complemento, são fornecidos *handouts* simplificados para revisão rápida de conceitos, *scripts* comentados passo a passo que guiam da leitura dos dados à interpretação dos resultados, e uma bibliografia diferenciada, que inclui desde textos introdutórios até referências de nível avançado. Esta arquitetura de recursos facilita a aprendizagem autónoma e promove uma literacia estatística sustentável.

8.2. Reprodutibilidade

A reprodutibilidade constitui um princípio estruturante da UC e um critério de avaliação explícito. Todos os trabalhos dos estudantes devem integrar código, dados e instruções que permitam reconstruir a análise em ambiente independente. Este requisito converte o relatório de resultados num registo verificável e transparente, permitindo que cada etapa do processo seja revista, adaptada e ampliada.

Esta prática promove boas normas de ciência aberta e aproxima os estudantes dos padrões atuais de investigação e de mercado, onde a rastreabilidade das decisões analíticas é cada vez mais exigida. Na avaliação, a qualidade não é aferida apenas pelo produto final, mas também pela capacidade de replicação e clareza do processo, incentivando rigor técnico e responsabilidade científica. Para os estudantes, esta abordagem constitui simultaneamente um treino metodológico e uma garantia de que a aprendizagem não se limita ao “resultado certo”, mas se estende ao método transparente e verificável que o suporta.

8.3. Inclusão

O terceiro pilar é a inclusão, que assume particular relevância numa turma heterogénea como a desta UC. A diversidade de percursos disciplinares, níveis de literacia estatística e até de fluência em R exige uma gestão pedagógica intencional, que ofereça materiais e oportunidades diferenciadas, assegurando que todos podem aprender e participar com sucesso. A UC integra mecanismos de inclusão pedagógica explícitos:

- a) Materiais em diferentes níveis de complexidade, permitindo que estudantes com formação menos sólida em estatística acompanhem o essencial, sem descurar a profundidade exigida aos mais avançados.
- b) Apoio presencial (metade das horas letivas), que funciona de forma continuada ao longo do semestre, permitindo a clarificação de dúvidas e o acompanhamento próximo da evolução dos grupos e dos estudantes.
- c) A UC integra também 4 horas de apoio tutorial formal, distribuídas estrategicamente em momentos críticos do semestre, nomeadamente antes da entrega dos trabalhos, visando apoiar dificuldades individuais ou de pequenos grupos, oferecendo espaço adicional para discussão metodológica e revisão de código ou relatórios.
- d) Disponibilização bilingue de apoio (português e inglês), assegurando que estudantes internacionais não ficam em desvantagem.

- e) Ambiente colaborativo (grupo de *WhatsApp*, fóruns, Moodle, *quizzes* partilhados) que promove a entreajuda e transforma a diversidade da turma num recurso de aprendizagem coletiva.

A robustez dos recursos utilizados e o compromisso com práticas inclusivas conferem consistência interna à UC. Falta, contudo, situar esta experiência no panorama externo. O capítulo seguinte procede a essa comparação, posicionando a UC face a ofertas congéneres em contexto nacional e internacional.

9. Benchmarking e posicionamento

Esta secção compara, de forma crítica, unidades curriculares de mestrado em Portugal, Europa e Estados Unidos da América, centradas em Análise de Dados Categóricos e/ou no quadro dos Modelos Lineares Generalizados (GLM). Incluímos apenas unidades curriculares que cobrem pelo menos dois dos tópicos: GLM, regressão logística, modelos de contagem, GEE, GLMM ou ZIF. Quando disponível, registamos conteúdos nucleares, avaliação e software.

9.1. Panorama nacional (Portugal)

Para garantir comparabilidade e leitura rápida, a Tabela 7 sintetiza as unidades curriculares de mestrado nacionais mais próximas da Análise de Dados Categóricos, assinalando apenas os pontos em comum do núcleo metodológico (GLM com regressão logística e Poisson, sobredispersão) e, quando aplicável, a extensão a dados correlacionados (GEE/GLMM). Indicamos, ainda, de forma sucinta, o formato de avaliação e o *software* predominante. Como se pode observar, todas as ofertas listadas partilham o núcleo GLM (logística e Poisson, com atenção à sobredispersão) e utilizam R como software principal. Há variação no formato avaliativo: algumas mantêm exame como eixo (FCUL e FCUP), enquanto outras combinam avaliação contínua (mini-testes/trabalhos) com alternativa de exame (UMinho e NOVA-FCT).

Tabela 7 — Unidades curriculares nacionais próximas da Análise de dados Categóricos

Instituição (Curso)	UC (ECTS)	Núcleo GLM*	GEE/GLMM	Avaliação	Software
Universidade de Lisboa – FCUL (Bioestatística)	Modelos Lineares Generalizados (6)	✓	✓	Exame	R
Universidade de Lisboa – FCUL (Estatística e Investigação Operacional)	Modelos Estatísticos (6)	✓	✓ (GLMM)	Exame	R
Universidade do Minho (Estatística para Ciência de Dados)	Modelos Lineares Generalizados e Aplicações (7,5)	✓	—	Testes e Trabalhos	R
Universidade Nova - FCT (Estatística para a Saúde)	Modelos Lineares Generalizados (6)	✓	—	Mini-testes e um Trabalho individual	R
Universidade do Porto – FCUP (Estatística Computacional e Análise de Dados)	Modelos Estatísticos Avançados (6)	✓	✓ (GLMM)	Exame	R
Universidade de Coimbra (Biologia)	Tópicos Avançados em Bioestatística (6)	✓	—	Prova aplicada (100%)	R

* Núcleo GLM - inclui explicitamente GLM com regressão logística (binária/ordinal/multinomial, conforme o curso), modelos de contagem (Poisson/BN ou quasi-Poisson) e sobredispersão.

✓ = tópico explicitamente coberto; — = não explicitado.

9.2. Panorama internacional (Europa e Estados Unidos)

Para contextualizar Análise de Dados Categóricos no plano internacional, a Tabela 8 sintetiza unidades curriculares de mestrado (Europa e Estados Unidos) que cobrem explicitamente os Modelos Lineares Generalizados (GLM), regressão logística/Poisson e, em muitos casos, extensões para dados correlacionados (GEE/GLMM). A tabela destaca apenas os aspetos comuns relevantes quando disponíveis: conteúdos nucleares, créditos, avaliação e software.

O núcleo internacional coincide com o padrão nacional: logística (binária/ordinal), Poisson/log-linear e sobredispersão, com passagem frequente a dados correlacionados (GEE/GLMM). O R é dominante (especialmente na Europa), surgindo SAS/R em alguns contextos norte-americanos (p. ex., UF, Rutgers). As diferenças concentram-se no formato avaliativo (maior peso de exames finais em Glasgow, York, LSE, Oxford, face a projetos/trabalho aplicado na Escandinávia e em várias escolas norte-americanas), na ênfase em dados longitudinais (Uppsala, Estocolmo, NTNU) e na integração com Sobrevivência (York, LSE, Bergen).

Tabela 8 — Panorama internacional de unidades curriculares próximas da Análise de Dados Categóricos

Universidade (País)	UC (ECTS)	Núcleo GLM*	GEE/GLMM	Avaliação (essencial)	Software
Bergen (Noruega)	Generalised Linear Models (10)	✓	—	Oral exam	—
Uppsala (Suécia)	Generalised Linear Models (7,5)	✓	dados dependentes	Exame e Trabalhos	—
Estocolmo (Suécia)	Generalised Linear Models: Theory and Application (7,5)	✓	Longitudinais	Exame e Trabalho	R
Glasgow (Reino Unido)	Generalised Linear Models (10)	✓	—	Exame	—
Southampton (Reino Unido)	Generalised Linear Models (7,5)	✓	—	Exame e Trabalho	R
KU Leuven (Bélgica)	Generalised Linear Models (6)	✓	—	Exame (oral) e Projeto	R
NTNU (Noruega)	Generalised Linear Models (7,5)	✓	GLMM	Exame e Projeto	R
LSE (Reino Unido)	Generalised Linear Modelling and Survival Analysis (—)	✓	—	Exame	R
Oxford (Reino Unido)	Applied Statistics (—)	✓	—	Exame	R
York (Reino Unido)	Survival Analysis and Generalised Linear Models (—)	✓	—	Exame	R
Rutgers (EUA)	Generalised Linear Models (—)	✓	—	Exame, Trabalhos e Projeto	R/SAS
Leiden (Países Baixos)	Linear and Generalised Linear Models (6)	✓	—	Exame	R
Florida (EUA)	Categorical Data Analysis (—)	✓	GEE	Trabalhos de casa e Exames	R/SAS

* Núcleo GLM = regressão logística (binária/ordinal), modelos de contagem (Poisson/BN) e sobredispersão.
✓ = tópico explicitamente coberto; — = não explicitado.

9.3. Posicionamento da UC: semelhanças e diferenças

A Análise de Dados Categóricos cobre integralmente o núcleo internacionalmente consensual: GLM para respostas binárias, ordinais e multinomiais, modelos de contagem (Poisson, Binomial Negativa, quasi-Poisson) e tratamento da sobredispersão. Em linha com várias ofertas de referência, introduz ainda dados correlacionados (GEE/GLMM) em modo aplicado, assegurando contacto com cenários realistas como medições repetidas e estruturas em clusters.

No contexto nacional, a UC da Universidade de Évora partilha a mesma base de conteúdos que Lisboa, Porto, Minho e NOVA: núcleo GLM (logística e Poisson), atenção à sobredispersão, uso de R como *software* de trabalho e preservação da via de exame final. A inclusão de contagens e dados correlacionados numa UC semestral é sustentada por múltiplos exemplos (p. ex., Leiden, Uppsala, York, NTNU, UF/UNC). O desenho em blocos (núcleo essencial → complementar aplicado → literacia alargada) responde, com realismo, à heterogeneidade das turmas, preservando rigor e aplicabilidade. De tal abordagem, resulta uma unidade curricular alinhada com o núcleo nacional e internacional e adaptada ao ecossistema da Universidade de Évora, interdisciplinar, orientada a problemas reais e comprometida com práticas recentes de ciência de dados.

Distingue-se, contudo, por um conjunto de características que, embora pontualmente visíveis em algumas instituições, são raras quando combinadas de forma explícita e integrada:

- i) Reprodutibilidade e ciência aberta como princípio transversal (relatórios em R, código comentado, instruções de execução);
- ii) Avaliação com dados reais, organizada em projetos reprodutíveis, mas equilibrada com *quizzes* individuais curtos para garantir responsabilização e verificação contínua;
- iii) Literacia alargada, que vai além da técnica tradicional, articulando a ponte conceptual entre modelos log-lineares e GLM, a análise e discussão de estudos científicos recentes, a comunicação de resultados a públicos não especializados, e a extensão a tópicos contemporâneos como calibração e equidade preditiva, integração de métodos de machine learning interpretável; esta abordagem é ainda enquadrada pelos Objetivos de Desenvolvimento Sustentável, reforçando a ligação entre competência técnica, responsabilidade científica e relevância social;
- iv) Política explícita de uso responsável de IA, centrada em transparência, autoria demonstrável e auditabilidade do processo, ainda rara no panorama português e internacional.

Em síntese, a UC posiciona-se como plenamente alinhada com o núcleo nacional e internacional de ADC/GLM, mas projeta uma assinatura pedagógica própria: prática reprodutível, avaliação autêntica com responsabilização individual e integração explícita de ética e IA. Trata-se de um desenho realista para 45h presenciais, rigoroso nos fundamentos e relevante para a interdisciplinaridade dos mestrados que serve.

9.4. Decisões curriculares

Da leitura comparada (nacional e internacional) resultam aprendizagens claras. No plano nacional, confirmou-se a solidez e consistência do núcleo de ensino em GLM (logística e Poisson, com atenção à sobredispersão), bem como a adoção generalizada de R como *software* principal. Também se observou diversidade de práticas avaliativas: algumas instituições mantêm o exame como eixo central (FCUL, FCUP), enquanto outras combinam trabalhos e mini-testes (UMinho, NOVA-FCT). Estas constatações reforçaram a confiança no desenho da Análise de Dados Categóricos, que já se encontrava alinhada com o padrão nacional.

Da análise nacional e internacional resultam alguns ajustamentos incrementais que reforçam a coerência da UC, sem alterar a sua matriz:

- i) Ponte log-linear / GLM – sem transformar a UC em log-linear avançado (como em Uppsala), optou-se por manter esta ligação como literacia conceptual: explicitar a equivalência entre modelos log-lineares e GLM para contagens, sublinhando que os log-lineares podem ser vistos como um caso particular dos GLM. Esta abordagem surge no bloco de literacia alargada (Parte II do Trabalho II), em linha com a prática de Leiden e Uppsala.
- ii) Avaliação equilibrada – ao contrário de várias instituições (p. ex., York, Glasgow, LSE, Oxford) que privilegiam o exame escrito como eixo avaliativo, a Análise de Dados Categóricos valoriza a avaliação contínua com projetos reprodutíveis e dados reais. Para garantir equidade e responsabilização individual, incluem-se *quizzes* curtos em aula e, quando necessário, momentos individuais na literacia alargada. Mantém-se a possibilidade de exame final, em conformidade com os regulamentos.
- iii) Literacia científica crítica – expande-se a vertente de literacia alargada da Parte II do Trabalho II, dedicada à apresentação e discussão de temas avançados, para incluir, sempre que adequado, a análise crítica de artigos científicos recentes; cada grupo pode, para além da demonstração prática em R, comentar a aplicação de um modelo categórico num estudo publicado, identificando virtudes, limitações e alternativas metodológicas; esta prática promove literacia metodológica, consolida o raciocínio estatístico e reforça a capacidade de dialogar com a literatura de forma informada e responsável.

Nota sobre citações

As informações específicas de conteúdos, métodos, software e avaliação foram extraídas das páginas oficiais das unidades curriculares e guias de estudo das instituições referidas (links nos identificadores de citação). No final do relatório, estas fontes serão listadas na bibliografia em formato APA, com indicação da data de consulta.

10. Considerações finais e plano de desenvolvimento futuro

A Análise de Dados Categóricos consolidou-se, ao longo dos anos, como uma unidade curricular com identidade própria: rigorosa no enquadramento estatístico, profundamente aplicada, reprodutível do ponto de vista técnico e atenta às dimensões éticas e de equidade que hoje atravessam a ciência de dados. A articulação entre fundamentos (tabelas, associação, correlação e concordância), modelação no quadro dos GLM (binário, ordinal e multinomial) e extensões indispensáveis (contagens com Poisson/quasi-Poisson/Binomial Negativa, zeros em excesso, dados correlacionados com GEE/GLMM) mostrou ser pedagógica e cientificamente eficaz para um público heterogéneo. O facto de integrar de forma sistemática com dados reais, em R e com relatórios reprodutíveis, favorece a transferência para contextos profissionais e de investigação e cria hábitos saudáveis de trabalho: documentar processos, fundamentar escolhas, diagnosticar pressupostos e comunicar resultados com responsabilidade. A planificação realista e flexível permite

assegurar o domínio do núcleo essencial, sem abdicar de um espaço de literacia alargada onde os estudantes contactam, criticamente, com temas emergentes (interpretabilidade, *fairness*, TRI, GAM para respostas categóricas, entre outros).

Os resultados empíricos sustentam estas opções. Por um lado, registam-se taxas de aprovação consistentemente elevadas (com uma ligeira descida num ano de maior escala e heterogeneidade, compatível com um nível de exigência saudável). Por outro lado, os índices de comparabilidade situam-se acima da média institucional, com destaque para a perceção de relevância da UC e para o reconhecimento do domínio científico-pedagógico do docente e da sua disponibilidade para apoiar os estudantes.

As áreas de melhoria identificadas, sobretudo clareza expositiva e perceção de equidade avaliativa, já se encontram ancoradas num plano de ação explícito (diagramas conceptuais e roteiros temáticos; rúbricas e exemplos comentados), sob a lógica do plano de melhoria contínua.

Em termos de *benchmarking*, a comparação nacional e internacional confirmou o alinhamento com os conteúdos fundamentais (GLM, logística, contagens, sobredispersão) e mostrou que a unidade curricular de Análise de Dados Categóricos se posiciona de forma distintiva tanto no panorama português como internacional. Entre os diferenciais pedagógicos que a destacam contam-se: i) a ênfase estruturada na reprodutibilidade; ii) a avaliação com dados reais, equilibrada com momentos individuais de responsabilização; iii) a integração de literacia alargada; e iv) uma política explícita de uso responsável de IA, que privilegia a transparência e a autoria verificável: o foco não está em detetar o uso de IA, mas em garantir que cada estudante compreende, explica e reproduz os resultados que apresenta, pelo que desta forma substitui-se a vigilância pela responsabilização..

O plano de desenvolvimento futuro assenta em evolução incremental, ousada e realista, sem dilatar as 45 horas presenciais nem descaracterizar a UC. Por um lado, reforçar-se-á a ponte conceptual entre modelos log-lineares e GLM para contagens. Não se pretende abrir um novo bloco denso, mas dotar os estudantes de um quadro mental que clarifique a razão de determinados problemas multivariados de contagens poderem ser lidos tanto com parametrizações log-lineares como no léxico GLM. Este reforço será feito no âmbito da Literacia Alargada (Parte II do Trabalho II), com um guião curto, um exemplo mínimo em R e referências para estudo autónomo. Por outro lado, expandir-se-á a análise crítica de artigos científicos no contexto da Literacia Alargada. Esta prática promoverá literacia metodológica e reforçará a capacidade de dialogar com a literatura de forma informada.

No plano pedagógico, as prioridades são claras. A clareza expositiva continuará a ser trabalhada com “roteiros de aprendizagem” que fazem a ponte entre aulas (por exemplo, do *logit* binário para o ordinal e o multinomial; da Poisson para a quasi-Poisson e a Binomial Negativa; dos GLM “clássicos” para GEE/GLMM), e com *outputs* anotados que explicitam a leitura de resíduos, a identificação de observações influentes, a interpretação de OR/IRR e a avaliação de calibração. A perceção de equidade será reforçada pela transparência das rúbricas (com exemplos de boas respostas e erros comuns), pela declaração de contributos nos trabalhos e pela manutenção dos *quizzes* individuais em aula como mecanismo de responsabilização e autorregulação. Para mitigar desigualdades de preparação, surge uma

“semana zero” de nivelamento (R e inferência) e *handouts* visuais de referência rápida (ligações, métricas de desempenho e calibração, desenho de validação), garantindo acessibilidade sem abdicar de rigor.

A integração responsável da IA continuará a ser um eixo transversal da unidade curricular: incentiva-se o seu uso para apoio à aprendizagem, na clarificação de dúvidas, estruturação de ideias ou exploração de alternativas, sempre com transparência de utilização e domínio efetivo das decisões tomadas. Esta prática prepara os estudantes para um contexto profissional em que a literacia estatística e o pensamento crítico face a ferramentas automatizadas serão competências decisivas, sem comprometer o núcleo formativo da UC: aprender a escolher, ajustar, diagnosticar, interpretar e comunicar modelos de forma consciente e ética.

No plano da internacionalização e da diversidade de aplicações, promover-se-á, quando fizer sentido, problemas de outras áreas (marketing, ciência política, justiça, ambiente) e potenciar-se-á colaborações pontuais (partilha de bases de dados ou sessões abertas com convidados), preservando o tempo nuclear para o que é indispensável. Em paralelo, explorar-se-á pontos formativos leves (submissões intermédias muito curtas, focadas em uma decisão de modelação e respetiva justificação) para apoiar os grupos no percurso entre o lançamento e a entrega dos trabalhos, sem transformar a avaliação contínua num mosaico sobrecarregado.

Este plano mantém a UC fiel ao seu propósito e sustentável na carga de trabalho: o essencial continua a acontecer em sala e nos projetos com dados reais; o aprofundamento e a atualização metodológica têm espaço na literacia; e a melhoria contínua é dirigida por evidência (taxas de sucesso, IC por item, *feedback*) e responsabilização pública (rúbricas, exemplos comentados, FAQ de avaliação). Em síntese, a UC não se limita a certificar conhecimentos: forma práticas, consolida competências e cultiva hábitos de pensamento estatístico que perduram para lá do semestre. Ao terminar, os estudantes devem sair preparados para enfrentar dados categóricos do mundo real com rigor, com sentido crítico e com a capacidade de comunicar resultados de forma clara e útil para a decisão, e a unidade curricular permanece viva, evolutiva e alinhada com o melhor que se faz, nacional e internacionalmente, no ensino avançado de Estatística e Ciência de Dados.

11. Bibliografia

A bibliografia desta unidade curricular foi organizada para assegurar profundidade científica, atualidade metodológica e coerência pedagógica. A bibliografia foi selecionada para permitir um estudo em espiral: partir de um manual central (Agresti, Hosmer, McCullagh & Nelder), aprofundar temas com monografias direcionadas (Hilbe, Hardin & Hilbe, Wood, Zuur), consolidar com exemplos aplicados (artigos), e situar a UC no panorama comparado (fontes institucionais), assegurando a coerência com as orientações pedagógicas e regulatórias vigentes. Esta arquitetura bibliográfica suporta a identidade da UC: rigor conceptual, prática reproduzível, ética e relevância interdisciplinar. A bibliografia integra referências de base, obras complementares de aprofundamento e artigos aplicados,

assegurando a coerência entre fundamentos teóricos, prática metodológica e contacto com a investigação atual.

Em seguida apresenta-se:

- a) A **bibliografia nuclear** reúne os textos de referência indispensáveis para o estudo sistemático dos dados categóricos e dos modelos lineares generalizados.
- b) A **bibliografia complementar** aprofunda tópicos específicos (GEE, GLMM, ZIF, GAM, sobredispersão, contagens), oferecendo pontes para aplicações e para a prática em R.
- c) As **fontes institucionais** documentam programas de unidades curriculares homólogas em universidades portuguesas e internacionais de referência, suportando o benchmarking comparativo.
- d) As **referências pedagógicas e regulatórias** enquadram o alinhamento construtivo, a avaliação e a melhoria contínua.
- e) Incluem-se, ainda, **leituras aplicadas** (trabalhos do docente e coautores) que ilustram, com dados reais, vários métodos lecionados (logística, multinomial, Poisson/BN, GEE/GLMM), reforçando a transferência para contextos profissionais e de investigação.

11.1. Bibliografia nuclear

Agresti, A. (2018). *An introduction to categorical data analysis* (3rd ed.). Wiley.

Bilder, C. R., & Loughin, T. M. (2014). *Analysis of categorical data with R*. CRC Press.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman & Hall.

Turkman, M. A. A., & Silva, G. L. (2000). *Modelos lineares generalizados*. Edições SPE.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Springer.

11.2. Bibliografia complementar

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Wiley.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (2nd ed.). Cambridge University Press.

Dobson, A. J., & Barnett, A. (2018). *An introduction to generalized linear models* (4th ed.). CRC Press.

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: Models, methods and applications*. Springer.

Hardin, J. W., & Hilbe, J. M. (2018). *Generalized linear models and extensions* (4th ed.). Stata Press.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2017). *Applied survey data analysis* (2nd ed.). Chapman & Hall/CRC.

Hilbe, J. M. (2011). *Negative binomial regression* (2nd ed.). Cambridge University Press.

Lumley, T. (2010). *Complex surveys: A guide to analysis using R*. Wiley.

Sheskin, D. J. (2000). *Handbook of parametric and nonparametric statistical procedures* (2nd ed.). CRC Press.

- Steyerberg, E. W. (2019). *Clinical prediction models* (2nd ed.). Springer.
- Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, 17(1), 230.
- Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). CRC Press.
- Zuur, A. F., & Ieno, E. N. (2016). *Beginner's guide to zero-inflated models with R*. Highland Statistics.
- Zuur, A. F., & Ieno, E. N. (2019). *Beginner's guide to GLM and GLMM with R*. Highland Statistics.

11.3. Artigos aplicados

- Coelho, R., Infante, P., & Santos, M. N. (2013). Application of generalized linear models and generalized estimating equations to model at-haulback mortality of blue sharks captured in a pelagic longline fishery in the Atlantic Ocean. *Fisheries Research*, 145, 66–75. <https://doi.org/10.1016/j.fishres.2013.02.008>
- Coelho, R., Infante, P., & Santos, M. N. (2020). Comparing GLM, GLMM, and GEE modeling approaches for catch rates of bycatch species: A case study of blue shark fisheries in the South Atlantic. *Fisheries Oceanography*. Advance online publication. <https://doi.org/10.1111/fog.12462>
- Furtado, K. A. X., Infante, P., Sobral, A., Gaspar, P., Eliseu, G., & Lopes, M. (2020). Prevalence of acute and chronic wounds—with emphasis on pressure ulcers—in integrated continuing care units in Alentejo, Portugal. *International Wound Journal*, 17, 1002–1010. <https://doi.org/10.1111/iwj.13364>
- Infante, P., Jacinto, G., Afonso, A., Rego, L., Nogueira, P., Silva, M., Nogueira, V., Saias, J., Quaresma, P., Santos, D., Góis, P., & Manuel, P. R. (2023). Factors that influence the type of road traffic accidents: A case study in a district of Portugal. *Sustainability*, 15, 2352. <https://doi.org/10.3390/su15032352>
- Infante, P., Jacinto, G., Afonso, A., Rego, L., Nogueira, V., Quaresma, P., Saias, J., Santos, D., Nogueira, P., Silva, M., Pisco Costa, R., Góis, P., & Manuel, P. R. (2022). Comparison of statistical and machine-learning models on road traffic accident severity classification. *Computers*, 11, 80. <https://doi.org/10.3390/computers11050080>
- Pereira CL, Baptista F, Infante P. (2013). Men older than 50 yrs are more likely to fall than women under similar conditions of health, body composition, and balance. *American Journal of Physical Medicine & Rehabilitation*, 92 (12), p.1095-103. <https://doi.org/10.1097/PHM.0b013e31829b49eb>
- Pereira, C., Batista, F., & Infante, P. (2014). Role of Physical Activity in the Occurrence of Falls and Fall-related Injuries in Community-dwelling Adults over 50 Years Old. *Disability and Rehabilitation*, 36(2), 117–124. <https://doi.org/10.3109/09638288.2013.782355>
- Salgado, A., Moreira-Leite, B., Afonso, A., Infante, P., & Mata, P. (2023). Chocolates enriched with seaweed: Sensory profiling and consumer segmentation. *International Journal of Gastronomy and Food Science*, 32, 100689. <https://doi.org/10.1016/j.ijgfs.2023.100747>

Silva, F. G., Lamy, E., Infante, P., Conceição, C., Cerqueira, J. L., Ramalho, J. M., González-Cabrera, M., Caetano, P., Martins, L., Silva, S. R., Pereira, A., & Hernández-Castellano, L. E. (2025). Saliva Immunoglobulin Concentrations Are Associated with Colostrum Intake and with Serum Concentrations in Newborn Calves. *Animals*, 15(15), 2224. <https://doi.org/10.3390/ani15152224>

Kamińska, J. A., Dionísio, A., Infante, P., & Carrilho, R. (2025). More Is Still Not Enough—What Is Necessary and Sufficient for Happiness? *Sustainability*, 17(13), 6121. <https://doi.org/10.3390/su17136121>

Carreira, E., Serrano, J., Shahidian, S., Infante, P., Paniagua, L. L., Moral, F., Paixão, L., Gomes, C. P., de Castro, J. L., de Carvalho, M., & Pereira, A. F. (2025). Sustainable Intensification of the Montado Ecosystem: Evaluation of Sheep Stocking Methods and Dolomitic Limestone Application. *Sustainability*, 17(1), 363. <https://doi.org/10.3390/su17010363>

Nota: Estes artigos são disponibilizados no Moodle da UC como leituras aplicadas, ilustrando com dados reais vários métodos do programa e servem de base a discussões e a exercícios de literacia metodológica nas aulas e na Parte II do Trabalho II.

11.3a. Projetos de investigação e algumas publicações associadas

Os seguintes projetos de investigação, coordenados ou com participação ativa do corpo docente da UC, forneceram bases de dados e resultados aplicados que são integrados no ensino da Análise de Dados Categóricos. Estes projetos asseguram a ligação entre investigação, ensino e sociedade, reforçando a autenticidade e relevância prática da unidade curricular:

AdolesSer. Parceria entre a Universidade de Évora e a Associação de Centros de Saúde do Alentejo Central. Objetivos: promoção de competências socioemocionais em adolescentes, com foco em relações interpessoais, sexualidade e comportamentos de risco.

Caraterização da prática e da atividade física e desportiva no Concelho de Évora – Parceria entre a Universidade de Évora e a Câmara Municipal de Évora. Objetivos: Caraterizar a prática de atividade física e desportiva no concelho, atualizando os dados existentes (projeto de 2012), analisar fatores que influenciam padrões de prática em termos de intensidade, frequência e tipo de atividade, identificar determinantes da prática de atividade física e desportiva, e prever os níveis de prática de atividade física e desportiva no concelho.

Diagnóstico Juvenil do Concelho de Évora (PMJ). Parceria entre a Universidade de Évora e a Câmara Municipal de Évora. Objetivo: caracterizar jovens dos 15–29 anos para apoiar políticas municipais de juventude.

Hábitos de Atividade Física e do Sono das Crianças do 1.º Ciclo do Concelho de Évora. Parceria entre a Universidade de Évora com CME, ARS Alentejo, HESE e agrupamentos de escolas. Objetivos: caracterizar padrões de sono e atividade física, e monitorizar evolução após intervenções.

MOPREVIS – Modelação e Predição de Acidentes de Viação no Distrito de Setúbal (FCT, 2019–2022). Projeto financiado (DSAIPA/DS/0090/2018), em parceria com a GNR. Objetivos: identificar determinantes de ocorrência e gravidade de acidentes, construir modelos preditivos e desenvolver ferramentas digitais de apoio à decisão em tempo real.

Nosologia das Feridas. Projeto da Sociedade Portuguesa de Feridas (ELCOS) com o envolvimento das unidades de Saúde do Alentejo (UCC, UCSP, UCCI, USF e Hospitais). Objetivo: estudo epidemiológico de prevalência de feridas.

Qualidade de Vida em Idosos (WHOQOL) e em Adolescentes (KIDSCREEN-27). Parceria entre a Universidade de Évora e a Câmara Municipal de Évora. Objetivos: estudar as perceções de bem-estar e determinantes sociodemográficos da qualidade de vida.

Seniores Ativos. Parceria entre a Universidade de Évora, a Câmara Municipal de Évora e a Associação de Centros de Saúde do Alentejo Central. Objetivos: avaliar o impacto do programa municipal de atividade física em pessoas com 55 ou mais anos.

Estes projetos originaram livros que são usados como recursos pedagógicos na UC, incluindo:

- Infante, P., Pisco Costa, R., Afonso, A., Jacinto, G., Conde, J., & Policarpo, L. (2019). *Diagnóstico Juvenil: Os jovens em Évora dos 15 aos 29 anos*. CME/UÉ. <https://www.cm-evora.pt/wp-content/uploads/2020/07/PMJ-OsJovensEvora15-29anos.pdf>
- Infante, P., Pisco Costa, R., Afonso, A., Jacinto, G., Conde, J., & Policarpo, L. (2018). *Diagnóstico Juvenil: Jovens Estudantes na Universidade de Évora, Trabalhadores e Desempregados*. CME/UÉ. <https://www.cm-evora.pt/wp-content/uploads/2020/07/PMJ-UniversitariosTrabalhadoresDesempregados-online.pdf>
- Infante, P., Pisco Costa, R., Afonso, A., Jacinto, G., Conde, J., & Policarpo, L. (2018). *Diagnóstico Juvenil: Alunos do Ensino Secundário*. CME/UÉ. <https://core.ac.uk/download/pdf/186456853.pdf>
- Infante, P., Afonso, A., Carvalho, L., Engana, T. (2012). *A Atividade Física no Concelho de Évora*. CME/UÉ. https://www.cm-evora.pt/wp-content/uploads/2020/06/livro_atividade_fisica_concelho_evora.pdf
- Mendes, M. F., Infante, P., Afonso, A., Maciel, A., Ribeiro, F., Tomé, L. P., & Freitas, R. B. (2016). *Determinantes da Fecundidade em Portugal*. FFMS. https://ffms.pt/sites/default/files/2023-06/DetermFecunPort_Completo_Final_0.pdf

11.4. Fontes institucionais e pedagógicas

- Biggs, J., & Tang, C. (2011). *Teaching for quality learning at university* (4th ed.). Open University Press. [https://cetl.ppu.edu/sites/default/files/publications/-John Biggs and Catherine Tang- Teaching for Quali-BookFiorg.pdf](https://cetl.ppu.edu/sites/default/files/publications/-John%20Biggs%20and%20Catherine%20Tang-Teaching%20for%20Quality%20Learning-BookFiorg.pdf)
- Faculdade de Ciências da Universidade de Lisboa. *Modelos Lineares Generalizados*. Consultado em setembro de 2025, em <https://fenix.ciencias.ulisboa.pt/degrees/bioestatistica-564500436615179/disciplina-curricular/1127630778662943>

Faculdade de Ciências da Universidade de Lisboa. *Modelos Estatísticos*. Consultado em setembro de 2025, em <https://fenix.ciencias.ulisboa.pt/courses/mesta-2254879305243046/programa>

KU Leuven. *Generalised Linear Models*. Consultado em setembro de 2025, em <https://onderwijsaanbod.kuleuven.be/syllabi/e/G0A18BE?activetab=doelstellingen>

Leiden University. *Linear and Generalised Linear Models*. Consultado em setembro de 2025, em <https://studiegids.universiteitleidennl/en/courses/123873/linear-and-generalized-linear-models>

London School of Economics. *Generalised Linear Modelling and Survival Analysis*. Consultado em setembro de 2025, em https://www.lse.ac.uk/resources/calendar2023-2024/courseGuides/ST/2023_ST411.htm

Norwegian University of Science and Technology (NTNU). *Generalised Linear Models*. Consultado em setembro de 2025, em <https://www.ntnu.edu/studies/courses/TMA4315#tab=omEmnet>

Race, P. (2019). *The Lecturer's Toolkit: A Practical Guide to Assessment, Learning and Teaching* (5th ed.). Routledge. <https://doi.org/10.4324/9780429060205>

Rutgers, The State University of New Jersey. *Generalized Linear Models*. Consultado em setembro de 2025, em <https://sph.rutgers.edu/sites/default/files/2023-04/BIST0725Spring2022.pdf>

Stockholm University. *Generalised Linear Models: Theory and Applications*. Consultado em setembro de 2025, em https://utbildning.su.se/download/18.61cb70d01961e09c14212b3/1744447171592/st5601_Course%2520description%2520vt-25.pdf

Uppsala University. *Generalised Linear Models*. Consultado em setembro de 2025, em <https://www.uu.se/en/study/course?query=2ST075>

University of Bergen. *Generalised Linear Models*. Consultado em setembro de 2025, em <https://www4.uib.no/en/studies/courses/stat201>

Universidade de Coimbra. *Tópicos Avançados em Bioestatística*. Consultado em setembro de 2025, em https://apps.uc.pt/courses/PT/unit/91025/23461/2024-2025?common_core=true&type=ram&id=354

Universidade de Évora. *Regulamento Académico da Universidade de Évora*. Consultado em setembro de 2025, em <https://gesdoc.uevora.pt/api/pages/file/&id=400756>

Universidade de Évora. *Sistema Interno de Garantia da Qualidade (SIGQ)*. Consultado em setembro de 2025, em <https://www.uevora.pt/universidade/qualidade/Sistema-Interno-de-Garantia-da-Qualidade-da-Universidade-de-Evora>

University of Florida. *Categorical Data Analysis*. Consultado em setembro de 2025, em https://stat.ufl.edu/wp-content/uploads/sites/120/Johnson_STA-4504-Syllabus-Spring-2023.pdf

University of Glasgow. *Generalised Linear Models*. Consultado em setembro de 2025, em <https://www.gla.ac.uk/coursecatalogue/course/?code=STATS4043>

Universidade do Minho. *Modelos Lineares Generalizados e Aplicações*. Consultado em setembro de 2025, em

<https://www.ecum.uminho.pt/pt/Ensino/Paginas/mestrados/Mestrado-em-Estatistica-para-Ciencia-de-Dados.aspx>

Universidade Nova de Lisboa – FCT. *Modelos Lineares Generalizados*. Consultado em setembro de 2025, em

<https://guia.unl.pt/pt/2024/fct/program/1188/course/13288#subject>

University of Oxford. *Applied Statistics*. Consultado em setembro de 2025, em

<https://www.stats.ox.ac.uk/sites/default/files/2022-08/MSc%20and%20PGDip%20Handbook%202022-2023.pdf>

Universidade do Porto – FCUP. *Modelos Estatísticos Avançados*. Consultado em setembro de 2025, em

https://sigarra.up.pt/fcup/pt/ucurr_geral.ficha_uc_view?pv_ocorrencia_id=545951

University of Southampton. *Generalised Linear Models*. Consultado em setembro de 2025, em <https://www.southampton.ac.uk/courses/2026-27/modules/stat6135#overview>

University of York. *Survival Analysis and Generalised Linear Models*. Consultado em setembro de 2025, em

<https://www.york.ac.uk/students/studying/manage/programmes/module-catalogue/module/MAT00112M/latest>

Anexo A – Catálogo de Bases de Dados (núcleo)

Este catálogo ilustra o núcleo de bases de dados habitualmente trabalhadas na UC. Sempre que surjam novas bases de dados poderão ser integradas em novas edições da UC, mantendo-se, contudo, a diversidade temática e metodológica.

Base de Dados I – Saúde Pública (Feridas)

- **Fonte:** Inquéritos em Unidades de Saúde do Alentejo.
- **Tipo de dados:** Prevalência de feridas agudas/crónicas, variáveis clínicas e demográficas.
- **Finalidade didática:** Associações e correlações entre variáveis, regressão logística binária e ordinal, modelos de contagens.
- **Exemplos de variáveis:** Sexo, idade, patologias, circunstâncias adversas, presença de feridas, número de feridas, tipologia das feridas, localização das feridas, origem das feridas, duração das feridas.

Base de Dados II – Atividade Física e Desportiva

- **Fonte:** Inquérito à população residente no concelho de Évora (amostra probabilística 13–17 anos; amostra calibrada adultos 18 ou mais anos).
- **Tipo de dados:** Prática de atividade física, características sociodemográficas, estilos de vida, motivações e barreiras.
- **Finalidade didática:** Análises de associação, regressão logística/ordinal/multinomial, modelos de contagem, aplicação de métodos para amostras complexas.
- **Exemplos de variáveis:** Sexo, idade, escolaridade, freguesia de residência, prática de AF (sim/não), frequência semanal, modalidades, local de prática, motivação/barreiras, perceção de saúde, gasto mensal em AF, peso de calibração.

Base de Dados III – Hábitos de Sono em Crianças

- **Fonte:** Estudo longitudinal em escolas públicas do concelho de Évora (4 anos letivos consecutivos).
- **Tipo de dados:** Questionários parentais sobre hábitos e dificuldades de sono, variáveis sociodemográficas e familiares.
- **Finalidade didática:** Análise de associações, regressão logística binária, ordinal e multinomial, modelos de contagem, efeitos marginais, diagnóstico de modelos, calibração e discriminação.
- **Exemplos de variáveis:** Sexo, idade, escolaridade dos encarregados de educação, situação laboral, partilha de quarto, perceção parental de problemas de sono, índice clínico de perturbação do sono, subescalas do CSHQ, número de despertares, duração do sono.

Base de Dados IV – Pastagens e Maneio Pecuário

- **Fonte:** Ensaio experimental (2020–2021) na Herdade da Mitra, Universidade de Évora.
- **Tipo de dados:** Medições longitudinais e hierárquicas sobre altura e valor nutritivo das pastagens, comportamento de ovinos e compactação do solo, sob diferentes regimes de pastoreio (contínuo vs diferido) e aplicação de calcário dolomítico.
- **Finalidade didática:** Aplicação de modelos de efeitos mistos (GLMM) e GEE para dependência longitudinal; comparação de tratamentos em delineamentos fatoriais.
- **Exemplos de variáveis:** Altura da pastagem, proteína bruta (CP), fibra em detergente neutro (NDF), precipitação acumulada, temperatura média, animais por dia de pastoreio, índice de cone (compactação), localização de animais nos pontos de observação, tipo de manejo, aplicação de calcário.

Base de Dados V – Atropelamentos de Vertebrados no IP2

- **Fonte:** Estudo observacional (1995–1997) no troço Portalegre–Monforte (27 km) do IP2, com inspeções quinzenais a anfíbios atropelados.
- **Tipo de dados:** Contagens de atropelamentos georreferenciados, agregados por segmentos de 500 metros, com identificação da espécie e caracterização do meio envolvente através de cartografia detalhada e SIG.
- **Finalidade didática:** Introdução a modelos de contagem (Poisson, binomial negativa, modelos inflacionados de zeros e hurdle), discussão da escolha da unidade de exposição e ilustração de como covariáveis espaciais e ambientais explicam a variabilidade nos atropelamentos.
- **Exemplos de variáveis:** número de atropelamentos por segmento, distância ao Parque Natural, existência de montado com arbustos, distância a reservatórios de água, extensão dos cursos de água, área de terras abertas, existência de múltiplas culturas.

Nota de uso e licitude

Todas as bases de dados acima são utilizadas com autorização ou em regime de dados públicos/abertos quando aplicável. As versões distribuídas na UC são anonimizadas e preparadas para fins didáticos, assegurando reprodutibilidade e conformidade ética.

Anexo B – Modelo para Declaração de Uso de Inteligência Artificial

Com vista a assegurar transparência, responsabilização e reprodutibilidade, todos os trabalhos entregues devem incluir uma declaração explícita sobre o eventual uso de ferramentas de Inteligência Artificial (IA) generativa. O objetivo não é penalizar, mas valorizar a clareza e a reflexão crítica sobre o papel destas ferramentas no processo de aprendizagem.

Nota: O modelo apresentado foi igualmente aplicado na elaboração deste relatório, no qual a IA foi utilizada para revisão linguística e apoio à edição textual, sem qualquer intervenção no conteúdo científico ou pedagógico.

Declaração de Uso de Inteligência Artificial

1. Ferramenta(s) utilizada(s):
(ex.: ChatGPT, Gemini, Perplexity, Copilot, etc.)
2. Finalidade de utilização:
(ex.: revisão de linguagem, explicação de erro em R, sugestão de estrutura do relatório, alternativa de código, etc.)
3. Excerto(s) ou outputs relevantes gerados pela IA:
(copiar/colar ou descrever sucintamente)
4. Avaliação crítica do grupo/estudante:
 - O que foi aceite sem alterações?
 - O que foi modificado? Porquê?
 - O que foi rejeitado?
5. Conformidade com as regras da UC:
[] Confirmamos que a utilização de IA respeitou as normas da unidade curricular, servindo apenas como apoio e nunca como substituto do trabalho intelectual próprio.

Esta declaração contribui para reforçar a ética académica e garante que a avaliação reflete o domínio efetivo das competências pelos estudantes.

Assinatura(s): _____