

INSTITUTO
SUPERIOR
DE CONTABILIDADE
E ADMINISTRAÇÃO
DO PORTO
POLITÉCNICO
DO PORTO



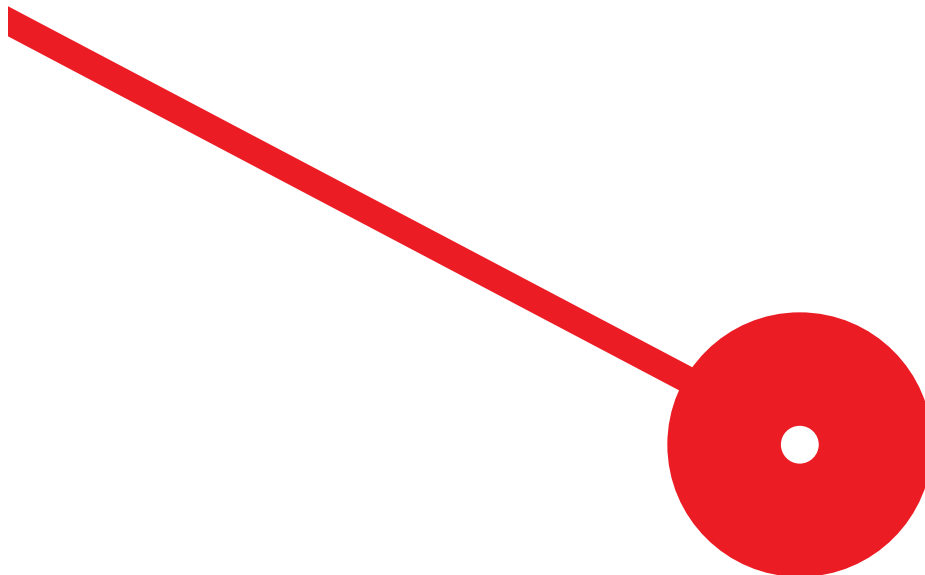
MESTRADO
BUSINESS INTELLIGENCE AND ANALYTICS

Integrating *Machine Learning* And *Business Intelligence* For Student Dropout Prediction And Decision Support in Vocational Education

Guilherme Machado dos Santos

06/2026

Guilherme Machado dos Santos. Integrating *Machine Learning* and *Business Intelligence* For
Student Dropout Prediction And Decision Support in Vocational Education
06/2026



INSTITUTO
SUPERIOR
DE CONTABILIDADE
E ADMINISTRAÇÃO
DO PORTO
POLITÉCNICO
DO PORTO

M

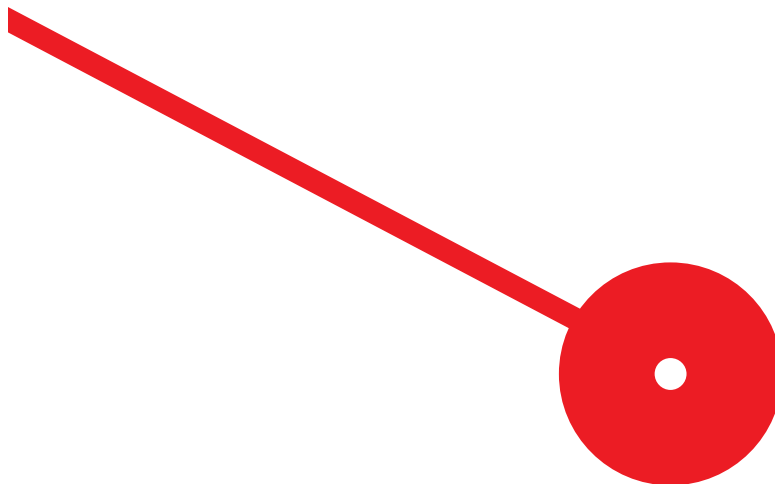
MESTRADO
BUSINESS INTELLIGENCE AND ANALYTICS

Integrating *Machine Learning* And *Business Intelligence* For Student Dropout Prediction And Decision Support in Vocational Education

Guilherme Machado dos Santos

Dissertação de Mestrado apresentada ao Instituto Superior de Contabilidade e Administração do Porto para a obtenção do grau de Mestre em *Business Intelligence* and Analytics sob orientação da Professora Doutora Maria José Angélico Gonçalves e da Professora Doutora Célia Talma Gonçalves.

Guilherme Machado dos Santos. *Integrating Machine Learning And Business Intelligence For Student Dropout Prediction And Decision Support in Vocational Education*
06/2026



Aos jovens empreendedores que
acreditam na educação como plataforma
de mudança para um mundo melhor.

Agradecimentos

Gostaria de começar agradecendo às minhas orientadoras, pelo acompanhamento, pela disponibilidade e pelas contribuições fundamentais para o desenvolvimento deste trabalho. Agradeço também à Profitecla pela colaboração e abertura que tornaram possível a realização desta investigação.

Mas estes agradecimentos não seriam completos se se limitassem apenas ao percurso académico.

Quando me mudei do Brasil para Portugal, em 2023, para realizar um intercâmbio durante o último semestre da minha graduação, não imaginava que aquela decisão daria início a um dos períodos mais transformadores da minha vida. O que começou como uma oportunidade académica acabou se tornando uma experiência muito maior, marcada por descobertas, desafios, crescimento profissional e, principalmente, crescimento pessoal.

Ao longo desses anos, tive o privilégio de trabalhar na Escola 42 (Porto e Lisboa). Foi nesse ambiente que pude desenvolver interesses que sempre me acompanharam, a educação, a tecnologia, a inovação e a construção de projetos de impacto. Mais do que uma experiência profissional, foi um espaço onde encontrei pessoas que acreditaram em mim, me acolheram e me mantiveram motivado ao longo dessa jornada paralela de estudos e trabalho.

Sou especialmente grato à Mariane Bittencourt Fagundes, Mariana Lemos Martins, à Rita Bertrand Franco e à Maria Sottomayor. Cada uma, à sua maneira, tornou esta jornada mais leve, mais divertida e mais significativa. Obrigado pela amizade, pela confiança, pelos incentivos, pelas conversas e por todos os momentos compartilhados ao longo desse caminho, vocês me ajudaram a construir um sentimento de pertencimento em um país que ainda aprendia a chamar de casa.

Também guardo com muito carinho a experiência vivida no Livensa Living e os amigos ali feitos. Conviver diariamente com pessoas de diferentes culturas, histórias e perspectivas ampliou minha visão de mundo e tornou essa experiência ainda mais rica. Muitas das memórias mais felizes desses anos nasceram ali.

À minha família, à minha tia Dominique e aos meus amigos no Brasil, agradeço por permanecerem presentes mesmo à distância. Obrigado por acreditarem em mim, por apoiarem minhas escolhas, pelas mensagens, pelas visitas e pelo carinho constante ao longo dessa jornada. Obrigado por me darem raízes fortes e, ao mesmo tempo, a liberdade

para sonhar. Em muitos momentos, vocês foram o lembrete de que a distância não diminui os laços que realmente importam.

Ao longo desse período, aprendi que crescer nem sempre acontece nos momentos mais visíveis. Muitas vezes, acontece no silêncio da adaptação, na saudade que acompanha a distância, na coragem de recomeçar, na construção de novos vínculos e na capacidade de encontrar pertencimento longe de casa.

Ao olhar para esta dissertação, vejo muito mais do que um trabalho concluído. Vejo os lugares que conheci, as pessoas que cruzaram o meu caminho, os desafios que me transformaram, as escolhas e abdições que tornaram esta jornada possível e a vida que construí durante este período.

Talvez por isso seja impossível separar esta dissertação da experiência que vivi em Portugal. Ambas cresceram juntas. E, embora este trabalho chegue agora ao seu fim, tudo aquilo que ele representa continuará comigo pelos caminhos que ainda estão por vir.

Resumo:

A previsão do abandono escolar tem registado avanços significativos no ensino, beneficiando da disponibilidade de dados comportamentais e da utilização de modelos preditivos de elevada precisão. Contudo, no ensino profissional, onde se concentram estudantes com maior vulnerabilidade socioeconómica e risco de insucesso, a investigação permanece escassa e as instituições dispõem de reduzidos mecanismos de apoio à identificação precoce de alunos em risco. Neste contexto, a presente dissertação propõe, desenvolve e valida um *framework* de *Educational Analytics* explicável, integrado numa arquitetura de *Business Intelligence*, destinado à previsão do risco de não conclusão e ao apoio à tomada de decisão em instituições de ensino profissional. A investigação estrutura-se em torno de três objetivos: caracterizar os perfis de sucesso e insucesso escolar, desenvolver modelos preditivos para diferentes momentos do percurso académico e disponibilizar informação analítica que suporte intervenções educativas. Para o efeito, foi adotada uma abordagem de *Design Science Research*, operacionalizada através da metodologia CRISP-DM, recorrendo a seis anos letivos de dados administrativos da Profitecla – Escola Profissional, correspondentes a 1 086 estudantes do polo do Porto. Os resultados permitiram identificar três perfis distintos de estudantes, com diferenças estatisticamente significativas nas taxas de conclusão e nos padrões de permanência ao longo do percurso escolar. Os modelos desenvolvidos evidenciaram elevado desempenho preditivo, tanto na análise retrospectiva como na previsão realizada no momento da matrícula, permitindo identificar precocemente estudantes com maior probabilidade de abandono. A utilização de técnicas de Inteligência Artificial Explicável possibilitou interpretar os principais fatores associados ao risco de não conclusão, enquanto a solução analítica implementada em Microsoft Power BI disponibilizou *dashboards* interativos para monitorização dos indicadores e apoio à decisão. A solução foi validada por responsáveis institucionais da escola profissional, que reconheceram a sua utilidade, facilidade de utilização e potencial para apoiar processos de intervenção precoce junto dos estudantes em risco, evidenciando a aplicabilidade prática do *framework* desenvolvido. Como principal contributo, a investigação demonstra que a integração de *Business Intelligence*, *Machine Learning*, Inteligência Artificial Explicável e *Learning Analytics* constitui uma abordagem eficaz para apoiar a gestão académica em contextos de ensino profissional, mesmo quando apenas estão disponíveis dados administrativos. O *framework* proposto fornece um sistema preditivo, interpretável e

validado em contexto real, demonstrando o potencial destas tecnologias para apoiar estratégias de prevenção do abandono escolar e promover uma tomada de decisão mais informada nas instituições de ensino.

Palavras chave: *Educational Data Mining*. Previsão de Abandono. *Business Intelligence*. *Machine Learning*. Inteligência Artificial Explicável. Ensino Profissional

Abstract:

Dropout prediction has achieved significant progress in education, benefiting from the availability of behavioural data and the use of high-precision predictive models. However, in vocational education, where students with greater socioeconomic vulnerability and higher risk of failure are concentrated, research remains scarce and institutions have limited mechanisms to support the early identification of at-risk students. In this context, this dissertation proposes, develops, and validates an explainable Educational Analytics framework, integrated into a Business Intelligence architecture, designed to predict non-completion risk and support decision-making in vocational education institutions. The research is structured around three objectives: characterising profiles of academic success and failure, developing predictive models for different moments of the academic pathway, and providing analytical information to support educational interventions. To this end, a Design Science Research approach was adopted, operationalised through the CRISP-DM methodology, drawing on six academic years of administrative data from Profitecla – Escola Profissional, corresponding to 1,086 students from the Porto campus. The results identified three distinct student profiles, with statistically significant differences in completion rates and in patterns of retention throughout the academic pathway. The developed models showed high predictive performance, both in retrospective analysis and in prediction at the time of enrolment, enabling the early identification of students with a higher probability of dropout. The use of Explainable Artificial Intelligence techniques made it possible to interpret the main factors associated with non-completion risk, while the analytical solution implemented in Microsoft Power BI provided interactive dashboards for monitoring indicators and supporting decision-making. The solution was validated by institutional stakeholders from the vocational school, who recognised its usefulness, ease of use, and potential to support early intervention processes for at-risk students, demonstrating the practical applicability of the developed framework. As a main contribution, the research demonstrates that the integration of Business Intelligence, Machine Learning, Explainable Artificial Intelligence, and Learning Analytics constitutes an effective approach to supporting academic management in vocational education contexts, even when only administrative data are available. The proposed framework provides a predictive, interpretable system validated in a real-world context, demonstrating the

potential of these technologies to support dropout-prevention strategies and promote more informed decision-making in educational institutions.

Keywords: Educational Data Mining. Dropout Prediction. Business Intelligence. Machine Learning. Explainable Artificial Intelligence. Vocational Education

Índice geral

Capítulo I – Introdução	1
1 Introdução	2
1.1 Problema de Investigação	4
1.2 Objetivos Gerais e Específicos	6
1.3 Contributos do Trabalho	7
1.3.1 Contributos académicos	8
1.3.2 Contributos para a Profitecla	8
1.4 Estrutura da Dissertação	9
Capítulo II – Enquadramento Teórico e Trabalho Relacionado	10
2 Enquadramento Teórico e Trabalho Relacionado	11
2.1 Enquadramento, Metodologia e Critérios da Revisão	11
2.2 O Ensino Profissional como Contexto de <i>Educational Analytics</i> ..	12
2.3 Disciplinas e Níveis de <i>Analytics</i> em Educação	13
2.4 <i>Machine Learning</i> para a Previsão de Abandono e Sucesso Escolar	14
2.5 Aprendizagem não supervisionada (<i>Clustering</i>) na Definição de	
Perfis de Estudantes	17
2.6 Inteligência Artificial Explicável (<i>eXplainable AI</i>)	19
2.7 De Preditivo a Prescritivo: Sistemas de Apoio à Decisão e BI	20
2.8 Considerações Éticas e de Equidade	22
2.9 Síntese, Lacunas e Posicionamento da Solução	22
Capítulo III – Metodologia	24
3 Metodologia	25
3.1 Enquadramento Metodológico	25
3.2 Metodologia de Investigação (<i>Design Science Research</i>)	25
3.2.1 Fases da DSR Aplicadas à Investigação	26
3.2.2 Relação entre DSR e CRISP-DM	28

3.3	CRISP-DM: Justificação e Fases Aplicadas ao Projeto	28
3.4	Considerações Éticas, Privacidade e Proteção de Dados	34
3.5	Ferramentas e Ambiente Tecnológico	35
Capítulo IV – Implementação.....		38
4	Implementação.....	39
4.1	Breve Descrição dos Dados	39
4.1.1	Problemas de Qualidade dos Dados	41
4.1.2	Estratégias de Correção e Validação Aplicadas	42
4.2	Preparação dos Dados.....	43
4.2.1	Construção da Variável-Alvo	43
4.2.2	Seleção e Tratamento de Variáveis.....	44
4.2.3	Criação de Novas <i>Features</i>	44
4.2.4	Tratamento de Valores Omissos, <i>Outliers</i> e Codificação.....	45
4.2.5	Divisão Treino/Teste e Validação Cruzada	46
4.3	<i>Clustering</i> de Perfis de Alunos	46
4.3.1	<i>Pipeline</i> de Pré-Processamento e Seleção do Algoritmo.....	47
4.3.2	Determinação do Número de <i>Clusters</i>	47
4.3.3	Perfis Identificados	48
4.3.4	Refinamento dos Perfis de Risco	49
4.4	Modelos de Previsão de Conclusão	50
4.4.1	Seleção dos Algoritmos e Configuração do <i>Pipeline</i>	50
4.4.2	M1a — Modelo Retrospectivo	52
4.4.3	M1b — Modelo Prospetivo	53
4.4.4	Análise Comparativa e <i>Deployment</i>	56
4.5	Interpretação dos Modelos com Inteligência Artificial Explicável	57
4.5.1	Interpretação do Modelo Retrospectivo (M1a)	58
4.5.2	Interpretação do Modelo Prospetivo (M1b).....	58

4.5.3	Comparação entre Modelos e Limitações da XAI.....	59
4.6	Análise de Sobrevida: Tempo até ao Abandono.....	60
4.6.1	Fundamentação	60
4.6.2	Curvas de Sobrevida e Validação Estatística	61
4.7	Sistema de Apoio à Decisão: Arquitetura e <i>Dashboards</i>	62
4.7.1	Arquitetura de Dados	62
4.7.2	Ecrã Inicial e <i>Dashboard</i> de Monitorização de Risco	63
4.7.3	<i>Dashboard</i> de Perfis de Alunos.....	65
4.7.4	<i>Dashboard</i> de Captação e Recrutamento.....	66
Capítulo V – Resultados e Discussão		69
5	Resultados e Discussão.....	70
5.1	Caracterização da Sessão de Avaliação	70
5.2	Resultados Quantitativos	71
5.3	Resultados Qualitativos	72
5.4	Triangulação e Validade da Avaliação	74
5.5	Discussão sobre a Transferibilidade e a Implementação.....	75
Capítulo VI – Conclusões.....		76
6	Conclusões.....	77
6.1	Síntese da Investigação.....	77
6.2	Resposta às Questões de Investigação	77
6.3	Contributos da Investigação	79
6.4	Limitações	80
6.5	Direções para Investigação Futura	81
6.6	Reflexão Final.....	81
Referências bibliográficas		83
Apêndices.....		93
Apêndice I – Instrumento de Avaliação e Estrutura da Sessão.....		94

Apêndice II – Publicações.....	100
--------------------------------	-----

Índice de Figuras

Figura 1 — Fases da Design Science Research aplicadas à investigação (adaptado de Peffers et al., 2007)	26
Figura 2 — Fases do CRISP-DM e carácter iterativo do processo (adaptado de Chapman et al., 2000)	29
Figura 3 — Elbow Method e Silhouette Score	48
Figura 4 — Síntese visual dos três perfis de alunos identificados por K-Means ($k = 3$)	48
Figura 5 — Evolução das médias de absentismo ao longo dos três anos curriculares, por cluster	49
Figura 6 — Comparação das principais métricas entre o Modelo Retrospectivo e o Modelo Prospetivo	52
Figura 7 — Matriz de confusão do Modelo Retrospectivo (M1a) no conjunto de teste	53
Figura 8 — Curvas de fiabilidade e Brier Score nas três configurações	55
Figura 9 — Evolução de F1, Recall e Precision da Classe 0 em função do threshold, e curva ROC do M1b	55
Figura 10 — Comparação entre o risco à matrícula (M1b) e o risco atual (M1a) por curso	57
Figura 11 — Direção e magnitude dos efeitos das features no M1a	58
Figura 12 — Importância global das features do M1b	59
Figura 13 — Importância média ($ SHAP $) das features comuns aos dois modelos	60
Figura 14 — Curvas de Kaplan-Meier por cluster, com intervalos de confiança de 95%	61
Figura 15 — Arquitetura de Dados do Sistema	63
Figura 16 — Ecrã inicial do sistema, com indicadores globais e acesso aos dashboards de monitorização, perfis e captação	64
Figura 17 — Dashboard operacional de monitorização de risco	65
Figura 18 — Dashboard de perfil de alunos	66
Figura 19 — Dashboard de captação e recrutamento	67

Índice de Tabelas

Tabela 1 — Síntese de trabalhos sobre previsão de abandono e sucesso académico	17
Tabela 2 — Síntese de trabalhos sobre clustering e perfis de estudantes.....	18
Tabela 3 — Síntese de trabalhos que aplicam XAI à previsão de sucesso e abandono.....	20
Tabela 4 — Ferramentas e Bibliotecas utilizadas.....	36
Tabela 5 — Dicionário de variáveis: Fich1 (Perfil do aluno)	40
Tabela 6 — Dicionário de variáveis: Fich2 (Desempenho académico anual)....	40
Tabela 7 — Comparação de algoritmos de clustering ($k = 3$).....	47
Tabela 8 — Caracterização dos clusters ($k = 3$)	48
Tabela 9 — Métricas de desempenho no conjunto de teste (M1a e M1b)	52
Tabela 10 — Comparação de estratégias de reamostragem (M1b, validação cruzada)	53
Tabela 11 — Sistema de semáforo de risco (M1b)	56
Tabela 12 — Semáforo de risco para os 396 alunos ativos (M1b vs M1a).....	56
Tabela 13 — Matriz de transição M1b $\rightarrow$ M1a.....	57
Tabela 14 — Taxa de conclusão por tipo de escola de origem (M1b)	59
Tabela 15 — Resultados do questionário de avaliação (N=4, escala 1–5).....	71

Lista de abreviaturas

AIRR — Antecipação, Inclusão, Responsividade e Reflexividade

AUC — *Area Under the Curve*

AUC-ROC — *Area Under the Receiver Operating Characteristic Curve*

BI — *Business Intelligence*

BIRCH — *Balanced Iterative Reducing and Clustering using Hierarchies*

CEF — Cursos de Educação e Formação

CRISP-DM — *Cross-Industry Standard Process for Data Mining*

DAX — *Data Analysis Expressions*

DBSCAN — *Density-Based Spatial Clustering of Applications with Noise*

DGEEC — *Direção-Geral de Estatísticas da Educação e Ciência*

DSR — *Design Science Research*

EDM — *Educational Data Mining*

EDUSTAT — Sistema de Estatísticas da Educação

EFP — Ensino e Formação Profissional

EQAVET — *European Quality Assurance in Vocational Education and Training*

ERP — *Enterprise Resource Planning*

ETL — *Extract, Transform, Load*

FEDS — *Framework for Evaluation in Design Science*

GMM — *Gaussian Mixture Models*

KDD — *Knowledge Discovery in Databases*

KNN — *K-Nearest Neighbors*

LA — *Learning Analytics*

LIME — *Local Interpretable Model-agnostic Explanations*

LMS — *Learning Management System*

LSTM — *Long Short-Term Memory*

M1a — Modelo Retrospectivo

M1b — Modelo Prospetivo

ML — *Machine Learning*

PCA — *Principal Component Analysis*

RGPD — Regulamento Geral sobre a Proteção de Dados

RNN — *Recurrent Neural Networks*

ROC — *Receiver Operating Characteristic*

SEMMA — *Sample, Explore, Modify, Model, Assess*

SHAP — *SHapley Additive exPlanations*

SMOTE — *Synthetic Minority Over-sampling Technique*

SMOTE-ENN — *Synthetic Minority Over-sampling Technique with Edited Nearest Neighbours*

SPPA — *Student Performance Prediction and Action*

SVM — *Support Vector Machine*

TAM — *Technology Acceptance Model*

TVET — *Technical and Vocational Education and Training*

XAI — *Explainable Artificial Intelligence*

XGBoost — *Extreme Gradient Boosting*

CAPÍTULO I – INTRODUÇÃO

1 Introdução

Nas últimas duas décadas, o ensino e formação profissional (EFP) tornaram-se um dos principais instrumentos de combate ao abandono escolar em Portugal. A generalização dos cursos profissionais às escolas públicas a partir de 2006, em conjunto com o alargamento da escolaridade obrigatória até aos 18 anos em 2009, alterou profundamente o sistema educativo em Portugal. Os dados confirmam esta mudança. A taxa de abandono escolar precoce, que mede a percentagem de jovens entre os 18 e os 24 anos que deixam o ensino sem concluir o secundário, desceu de cerca de 23% em 2011 para 8,1% em 2023, valor já abaixo da média da União Europeia (9,5%) e da meta europeia para 2030 (Eurostat, 2024; DGEEC, 2024). Atualmente, o ensino profissional representa cerca de 40% das matrículas no secundário, com a meta nacional a apontar para os 55% (Martins et al., 2025).

O estudo da Fundação Francisco Manuel dos Santos sobre a expansão dos cursos profissionais mostra que a taxa de retenção e abandono no sistema educativo caiu de 39% em 2000 para menos de 10% em 2023. Mostra também que 72% dos alunos que concluem o ensino profissional e não prosseguem estudos encontram emprego no prazo de um a dois anos, contra 56% dos que concluem a via científico-humanística (Martins et al., 2025). A via profissional cumpre, portanto, a sua promessa de qualificação, inclusão social e empregabilidade.

É neste ponto que se situa o problema desta dissertação. O contributo do ensino profissional para reduzir o abandono a nível nacional não significa que as instituições que o ministram estejam livres do desafio da retenção. O abandono que ocorre dentro do próprio curso, quando o aluno se matricula, mas não chega a concluir, continua a ser um problema crítico e pouco estudado, distinto do indicador nacional de abandono precoce. As estatísticas agregadas mostram o efeito sistémico da existência da via profissional, mas não tornam visível a atrição que cada instituição enfrenta ao longo dos três anos de duração dos seus cursos. É justamente este nível, o da instituição e o do aluno individual, que esta dissertação analisa.

A relevância do problema acentua-se pelo perfil dos alunos que o ensino profissional acolhe. À semelhança do resto da Europa, a via profissional recebe sobretudo jovens com maior vulnerabilidade socioeconómica e percursos escolares fragmentados, muitas vezes marcados por retenções ou pela transição da via científico-humanística. Em contextos urbanos, acolhe ainda uma parte significativa de alunos com trajetórias

migratórias. A teoria de retenção de Tinto (1975, 1993) e a sua extensão a alunos não-tradicionais por Bean e Metzner (1985) mostram que a decisão de abandonar resulta de um processo gradual de desintegração académica e social, mediado pelo sentido de pertença do aluno, pela qualidade da sua relação com docentes e colegas e pelo seu compromisso com a instituição. Embora formulados originalmente para o ensino superior, os mecanismos centrais desta teoria foram reconhecidos como pertinentes em contextos educativos distintos, incluindo populações não-tradicionais com características semelhantes às que o ensino profissional acolhe (Bean & Metzner, 1985). Identificar cedo, e com fundamento, os alunos em maior risco é por isso a condição que permite que o acompanhamento chegue a quem mais precisa, e a tempo de intervir.

A instituição estudada nesta dissertação é a Profitecla, a qual foi fundada em 1989 ao abrigo de contrato-programa com o Ministério da Educação, conta com mais de três décadas de atividade e milhares de alunos formados. Integra o Grupo Rumos, um dos principais operadores nacionais de formação profissional, e está organizada numa rede de polos em várias cidades do país, entre as quais o Porto, onde fica a sede, Lisboa, Braga, Guimarães, Barcelos, Coimbra e Viseu. A sua oferta abrange áreas como Tecnologias de Informação, Saúde, Hotelaria e Restauração, Comunicação, Gestão e Turismo. A estratégia do grupo assenta em adaptar a oferta às necessidades do tecido empresarial de cada região e em garantir qualidade certificada pelo quadro europeu EQAVET (*European Quality Assurance in Vocational Education and Training*). Este referencial compromete a instituição com ciclos regulares de planeamento, monitorização e melhoria da sua atividade, o que torna a Profitecla um contexto adequado à introdução de ferramentas analíticas de apoio à decisão, dado existir já uma cultura de monitorização de resultados. A investigação centra-se no polo do Porto, embora a solução desenvolvida tenha sido pensada para poder ser replicada nos restantes polos.

Do ponto de vista dos dados, a Profitecla opera, como a generalidade das instituições de ensino profissional, com um sistema de gestão administrativa (*Enterprise Resource Planning*, ERP escolar) e sem plataforma de gestão de aprendizagem (*Learning Management System*, LMS). Por esse motivo, não existem dados comportamentais detalhados, como registos de acesso ou de participação online, que têm sido o principal motor dos modelos preditivos de alto desempenho no ensino superior. A fonte de dados disponível resume-se aos registos administrativos e académicos, designadamente o perfil de entrada, a assiduidade, os módulos em atraso, a escola de origem e os resultados de conclusão e empregabilidade. Longe de ser uma limitação acidental do estudo, esta

restrição reflete a condição real em que a maioria das instituições de ensino profissional trabalha.

1.1 Problema de Investigação

A aplicação de métodos de *Machine Learning* à previsão do abandono escolar avançou muito na última década, com modelos capazes de identificar alunos em risco com elevada precisão quando alimentados por dados comportamentais ricos (Matz et al., 2023; Bañeres et al., 2023). No entanto, quase toda esta investigação foi conduzida no ensino superior, onde existem LMS, avaliação contínua e métricas de admissão padronizadas. As instituições de ensino profissional secundário, que operam apenas com sistemas administrativos, permanecem muito sub-representadas na literatura de *Learning Analytics* e *Educational Data Mining* (Fernández-García et al., 2021).

A esta lacuna metodológica junta-se outra, de natureza organizacional. Revisões recentes mostram que a análise preditiva domina a investigação em educação, com cerca de 60% dos estudos, enquanto a analítica prescritiva, que liga a previsão à recomendação de ação, representa menos de 15% (Asha & Joshith, 2025). O resultado é uma desconexão frequente entre o resultado técnico de um modelo e a decisão que ele deveria informar. Prever o risco de abandono tem pouco valor prático se esse resultado não puder ser traduzido posteriormente em decisões concretas, compreensíveis por gestores e diretores de curso sem formação em ciência de dados.

Por baixo de ambas as lacunas está um obstáculo mais amplo, que é a dificuldade das instituições de ensino em inovar e em adotar uma cultura de decisão orientada por dados (*data-driven decision making*). A literatura identifica de forma recorrente as barreiras que travam esta transição, entre elas a baixa literacia em dados, a resistência à mudança, a infraestrutura tecnológica limitada, a dispersão dos dados em silos e as preocupações com a privacidade (Hora, Bouwma-Gearhart & Park, 2017; Gaftandzhieva et al., 2023). Estas barreiras são ainda mais fortes no ensino profissional, onde os recursos técnicos são limitados e as equipas raramente incluem perfis especializados em análise de dados. Mostrar que é possível superar este obstáculo num caso concreto, construindo um sistema útil, interpretável e operável por não-especialistas a partir de dados administrativos já existentes, tem valor para além da instituição estudada. Cria um modelo que pode ser replicado por outras escolas e contribuir para qualificar a gestão do ensino profissional.

A literatura em *Learning Analytics* mostra que o impacto dos sistemas preditivos depende, antes de mais, da forma como a informação de risco é incorporada nas práticas institucionais e na decisão pedagógica (Gašević, Dawson & Siemens, 2015). Intervenções dirigidas e oportunas, baseadas em modelos de risco, podem aumentar as taxas de retenção quando corretamente integradas nos fluxos de trabalho dos docentes e das equipas de acompanhamento (Jayaprakash et al., 2014). O modelo mais eficaz é aquele em que a tecnologia apoia o processo de tomada de decisão do profissional, num esquema designado por *human-in-the-loop*, libertando diretores de curso, psicólogos e professores da tarefa difícil de adivinhar quem está em dificuldade para que possam concentrar o acompanhamento em quem dele mais necessita. A concretização plena deste potencial extravasa o âmbito técnico desta dissertação e depende de decisões organizacionais que a produção dos artefactos aqui apresentados visa informar, a capacitação das equipas em literacia de dados, a eventual afetação de recursos especializados ao acompanhamento individualizado e o desenho de protocolos de intervenção. Estas questões são identificadas ao longo do trabalho como linhas de investigação e desenvolvimento futuro.

A convergência destas dimensões, a saber, a lacuna do ensino profissional como contexto de dados restritos, o défice de tradução da previsão em ação e o desafio de adotar uma cultura orientada por dados, conduz ao problema central desta investigação:

Como integrar *Business Intelligence* e *Machine Learning* num sistema de apoio à decisão estratégica para uma instituição de ensino profissional que opera apenas com dados administrativos, de modo a apoiar tanto a retenção de alunos como a estratégia de captação de novos alunos?

Este problema concretiza-se em três questões de investigação, organizadas numa progressão que vai da caracterização ao prognóstico e deste à ação.

Q1 — Quais os fatores associados à conclusão do curso no ensino profissional? Esta questão pertence à análise descritiva e diagnóstica. Pretende-se identificar, com recurso a dados históricos, recorrendo também a *Machine Learning* (aprendizagem não supervisionada) e à Inteligência Artificial Explicável (*Explainable AI*), os perfis de alunos associados a diferentes resultados de conclusão, definidos pelas suas características de entrada, comportamento e trajetória escolar. Identificar perfis distintos é a condição para que qualquer intervenção possa ser direcionada e diferenciada.

Q2 — É possível prever, com um desempenho aceitável, o risco de não conclusão ao longo do percurso do aluno? A previsão de risco não se faz num único momento. No ato da matrícula só estão disponíveis as variáveis de entrada, o que limita

a precisão possível, mas à medida que o aluno avança acumulam-se dados de comportamento e desempenho que tornam a previsão mais fiável (Fernández-García et al., 2021). Por essa razão, a questão é respondida por dois modelos complementares. O modelo prospetivo estima a probabilidade de não conclusão apenas a partir das variáveis conhecidas no momento da matrícula, permitindo uma triagem precoce ainda antes do início das aulas. O modelo retrospectivo reavalia esse risco incorporando os indicadores de desempenho e assiduidade que se acumulam ao longo do ano, o que produz uma estimativa mais exata em fases mais avançadas do percurso. O objetivo não se esgota em saber se cada modelo é suficientemente bom em termos absolutos; procura sobretudo demonstrar como a combinação dos dois permite acompanhar o risco de forma contínua ao longo do curso.

Q3 — Como traduzir os resultados dos modelos em *dashboards* prescritivos úteis para a gestão? Esta questão trata da última etapa da análise, que consiste em converter resultados técnicos em instrumentos de apoio à decisão utilizáveis por gestores com diferentes níveis de literacia em dados. Um semáforo de risco, um mapa de escolas de origem por probabilidade de sucesso ou um painel de monitorização de alunos ativos são exemplos de artefactos que colocam desafios próprios de design da informação, e que esta dissertação procura resolver.

1.2 Objetivos Gerais e Específicos

No âmbito do paradigma de *Design Science Research* (DSR) (Hevner et al., 2004), que enquadra esta investigação como um processo de construção e avaliação de artefactos, os objetivos organizam-se em torno de dois eixos estratégicos complementares, cada um com implicações diretas para a decisão na Profitecla.

Eixo 1 — Perfis de Sucesso (conclusão, prosseguimento de estudos e empregabilidade). O objetivo geral deste eixo consiste em caracterizar empiricamente os perfis de sucesso e de insucesso dos alunos da Profitecla, com base em seis anos letivos de dados (2018/19 a 2023/24). Os objetivos específicos são:

(i) Construir uma tipologia de perfis de alunos recorrendo à aprendizagem não-supervisionada (*clustering*), comparando métodos alternativos e justificando a escolha final com métricas objetivas e com argumentos de adequação ao caso de uso da Profitecla;

(ii) Desenvolver dois modelos preditivos complementares, um prospetivo, com as variáveis disponíveis na matrícula, e um retrospectivo, com dados do percurso académico, que em conjunto permitem acompanhar o risco ao longo do curso;

(iii) Interpretar os modelos com SHAP (*SHapley Additive exPlanations*), medindo a contribuição de cada variável para o risco de não conclusão e documentando como a importância dos preditores se altera entre o momento da matrícula e as fases mais avançadas;

(iv) Analisar o momento do abandono com análise de sobrevivência (Kaplan-Meier), identificando o período de maior risco de desistência e as suas implicações para o calendário das intervenções;

(v) Aplicar o modelo prospetivo aos alunos ativos no ano letivo 2024/25, gerando uma estratificação de risco atualizada que serve de base ao sistema de apoio à decisão.

Eixo 2 — Estratégia de Captação (escolas de origem e perfis de entrada). O objetivo geral deste eixo consiste em produzir evidência quantitativa que apoie o recrutamento e a seleção de novos alunos, identificando que escolas de origem e que perfis de entrada se associam a maior probabilidade de sucesso. Os objetivos específicos são:

(i) Caracterizar as escolas de origem dos alunos da Profitecla pela taxa histórica de conclusão e pela probabilidade de sucesso estimada pelo modelo prospetivo, identificando padrões geográficos e de tipo de ensino;

(ii) Analisar a relação entre o tipo de escola de origem e o desempenho académico posterior, explorando a hipótese, sustentada pela análise SHAP, de que o tipo de escola de origem é um dos preditores mais relevantes no modelo prospetivo;

(iii) Traduzir estas análises em instrumentos de apoio à captação, concretamente num mapa estratégico de escolas e num perfil de candidato associado a maior probabilidade de sucesso, úteis nos processos de candidatura e divulgação.

A articulação entre os dois eixos materializa-se no sistema de *dashboards* desenvolvido em Power BI, o artefacto central desta dissertação, cuja arquitetura segue o modelo multidimensional de Kimball, com extensão do tipo Data Vault na camada analítica de *Machine Learning*.

1.3 Contributos do Trabalho

Os contributos desta dissertação situam-se em dois planos, o da produção académica orientada para o estado da arte em *Learning Analytics* e *Educational Data Mining*, e o da intervenção aplicada na Profitecla.

1.3.1 Contributos académicos

Do ponto de vista científico, a dissertação apresenta três contributos principais. O primeiro consiste em propor e validar um *framework* de análise educacional modular e explicável, que integra análise não supervisionada (*clustering*), análise preditiva em dois momentos distintos da trajetória do aluno, interpretabilidade por SHAP e análise de sobrevivência, desenhado para contextos de dados restritos, ou seja, instituições sem LMS, sem notas contínuas e sem métricas de admissão padronizadas. Demonstrar que é possível construir modelos válidos e úteis apenas com dados administrativos contribui para uma literatura que tem documentado a sub-representação do ensino profissional nos estudos de *Learning Analytics* (Fernández-García et al., 2021).

O segundo contributo prende-se com a interpretabilidade dos modelos. A dissertação documenta como a importância dos modelos preditivos se altera entre dois horizontes temporais, com as variáveis demográficas e socioeconómicas a dominarem na matrícula e as variáveis de comportamento e desempenho a ganharem peso à medida que se acumulam dados do percurso. Esta evidência sustenta a ideia de que as características de origem do aluno funcionam como indicadores temporários de falta de informação, e não como determinantes fixos do resultado, uma distinção com implicações éticas diretas para o design das intervenções.

O terceiro contributo ajuda a fechar a lacuna entre análise preditiva e prescritiva, ao mostrar num caso aplicado como os resultados de modelos de *Machine Learning* podem ser integrados numa arquitetura de *Business Intelligence* que torna a informação utilizável por não-especialistas. Esta integração responde diretamente ao desafio, muito reconhecido na literatura, da transição das instituições de ensino para uma cultura orientada por dados (Hora, Bouwma-Gearhart & Park, 2017; Gaftandzhieva et al., 2023).

1.3.2 Contributos para a Profitecla

Para a Profitecla, a dissertação entrega quatro instrumentos de apoio à decisão diretamente utilizáveis. O primeiro é um semáforo de risco de não conclusão para os alunos ativos, que os classifica em três categorias: verde (baixo risco), amarelo (risco intervencionável) e vermelho (risco elevado), com base nos dois modelos preditivos. O segundo é um painel de monitorização operacional, que permite acompanhar o risco por curso e aluno ao longo do percurso académico. O terceiro é um mapa estratégico de captação, que identifica as escolas de origem e cursos mais associados ao sucesso na Profitecla. O quarto é uma análise de perfis históricos, que orienta ações de intervenções

estratégicas, a divulgação e a seleção de candidatos com base em probabilidades de conclusão e outros atributos.

Para além do seu carácter técnico, estes artefactos representam uma passagem da gestão intuitiva para a gestão orientada por dados, numa instituição que até agora não dispunha de qualquer sistema formal de análise preditiva do percurso dos alunos. O valor destes instrumentos realiza-se quando os profissionais que acompanham os alunos os incorporam nas suas práticas de acompanhamento, desde a sinalização precoce até à conversa individualizada ou ao plano de recuperação.

1.4 Estrutura da Dissertação

A dissertação está organizada em seis capítulos. O Capítulo I (Introdução) apresenta o contexto, o problema, a lacuna e os objetivos. O Capítulo II (Enquadramento Teórico e Trabalho Relacionado) mapeia o estado da arte do *Business Intelligence* e do *Machine Learning* no Ensino Profissional, com a identificação da lacuna que o trabalho preenche. O Capítulo III (Metodologia) descreve detalhadamente o processo de *Design Science Research* e a implementação do CRISP-DM. O Capítulo IV (Implementação) apresenta os resultados da análise preditiva recorrendo a *Machine Learning*, bem como o *design* e a construção do Sistema de *Business Intelligence*. O Capítulo V (Resultados e Discussão) discute a avaliação do artefacto pela Profitecla e o seu impacto organizacional. Por fim, o Capítulo VI (Conclusões) resume as contribuições e propõe vias para investigações futuras.

CAPÍTULO II – ENQUADRAMENTO TEÓRICO E TRABALHO RELACIONADO

2 Enquadramento Teórico e Trabalho Relacionado

2.1 Enquadramento, Metodologia e Critérios da Revisão

A presente revisão analisa e sintetiza o estado da arte na aplicação de *analytics* em educação, abrangendo o *Business Intelligence* (BI), o *Educational Data Mining* (EDM) e o *Learning Analytics* (LA), com ênfase na transição de modelos preditivos para prescritivos e no apoio à decisão institucional. O objetivo é fundamentar a questão de investigação central da dissertação, identificando as lacunas metodológicas e organizacionais que justificam o desenvolvimento de um artefacto de *Educational Analytics* no contexto específico do ensino profissional.

A metodologia adotada corresponde a uma revisão narrativa crítica, informada por princípios de sistematicidade para assegurar transparência e reduzir o viés de seleção. As fontes primárias foram identificadas em motores de busca académica (Clarivate, Consensus e Google Scholar), privilegiando a indexação em bases de dados de referência como a Scopus e a Web of Science. A estratégia de pesquisa combinou termos como *Learning Analytics*, *Predictive Analytics*, *Student Dropout*, *Prescriptive Analytics*, *Vocational Education* e *Higher Education*, com restrição preferencial a publicações entre 2020 e 2025 para garantir atualidade tecnológica. Contudo, foram também incluídos estudos anteriores considerados relevantes para o enquadramento conceptual e teórico do tema.

Os critérios de inclusão privilegiaram artigos empíricos e revisões sistemáticas publicados em revistas com fator de impacto elevado, bem como estudos que abordam não só a componente técnica, mas também a implementação organizacional, as questões éticas e a intervenção pedagógica. Esta amplitude justifica-se pela natureza do problema, que não se esgota na construção de um modelo preditivo e exige a consideração do ciclo completo, desde a análise até à ação institucional.

Uma observação metodológica condiciona toda a leitura que se segue. O estudo empírico desta dissertação decorre num ambiente de dados restrito, característico do ensino profissional, ao passo que a literatura disponível foi produzida maioritariamente em instituições de ensino superior dotadas de infraestruturas digitais ricas. Por essa razão, a revisão lê a literatura predominante pela lente da sua transferibilidade para contextos de dados limitados, distinguindo de forma explícita o que é generalizável daquilo que depende de fontes de dados indisponíveis no setor profissional.

2.2 O Ensino Profissional como Contexto de *Educational Analytics*

O ensino e a formação profissional (em terminologia internacional, *Technical and Vocational Education and Training*, TVET) constituem uma via educativa orientada para a prática, caracterizada pela especialização técnica, pela organização modular dos planos de estudo, pela existência de formação em contexto de trabalho e por um alinhamento estrito com as necessidades do mercado de emprego. Esta configuração distingue o ensino profissional do ensino superior em dois aspetos relevantes para a análise de dados. O resultado terminal não se esgota na conclusão académica, estendendo-se à empregabilidade, e a progressão é avaliada à escala do módulo, o que torna o número de módulos em atraso um indicador de risco mais informativo do que uma média global de classificações.

A população servida por estas instituições apresenta, em Portugal e na generalidade do espaço europeu, uma vulnerabilidade socioeconómica elevada, trajetórias escolares prévias fragmentadas e, com frequência, uma transição direta do ensino obrigatório sem uma identidade vocacional consolidada (Schmid & Garrels, 2021). São condições sob as quais a identificação precoce de alunos em risco encerra o maior potencial de valor, precisamente porque a margem para intervenção atempada é limitada.

O ambiente de dados deste contexto difere de forma substantiva daquele que sustenta a maioria da literatura. As instituições de ensino profissional operam tipicamente sobre sistemas administrativos do tipo ERP, sem plataformas de gestão de aprendizagem (*Learning Management Systems*, LMS) que gerem registos contínuos de interação e sem uma métrica padronizada de desempenho cumulativo. O espaço de variáveis disponível restringe-se, assim, a dados administrativos e académicos. Esta restrição é simultaneamente uma limitação metodológica e um traço de representatividade: um modelo validado nestas condições estabelece um patamar de desempenho conservador e transferível para a maioria das instituições do setor, que partilham a mesma escassez de infraestrutura digital. A escassez de dados comportamentais, que noutros contextos emergem como preditores poderosos do sucesso (Matz et al., 2023), define, portanto, a fronteira do que é possível modelar no ensino profissional.

Apesar da amplitude da literatura de LA e EDM, a quase totalidade dos estudos empíricos foi conduzida em instituições de ensino superior com ambientes digitais ricos. As instituições de ensino profissional, e em particular as que funcionam sem integração

de LMS, permanecem sub-representadas, deixando em aberto a questão de saber se os enquadramentos desenvolvidos para ambientes de dados abundantes se mantêm aplicáveis sob condições de dados restritos. A consequência é que as instituições com maior necessidade de capacidade de alerta precoce são as menos servidas pela investigação existente (Gedrimiene et al., 2019). É esta a lacuna que a presente dissertação pretende colmatar.

2.3 Disciplinas e Níveis de *Analytics* em Educação

No campo da análise de dados educacionais, os termos *Educational Data Mining* e *Learning Analytics* são frequentemente usados de forma intercambiável, mas correspondem a campos com focos, objetivos e públicos distintos. O *Learning Analytics* centra-se na medição, recolha, análise e reporte de dados sobre os alunos e os seus contextos, com o objetivo de compreender e otimizar a aprendizagem, orientando-se para o *actionable insight*, isto é, uma conclusão que possa informar diretamente uma intervenção (Clow, 2013, citado em Klenke et al., 2022). O *Educational Data Mining* é um campo mais amplo, focado na aplicação de técnicas de *data mining* a dados educacionais, com objetivos que incluem a investigação dos processos de aprendizagem e a descoberta de padrões e estruturas no domínio do conhecimento (Romero & Ventura, 2020; Palacios et al., 2021; Hung et al., 2020; Yağcı, 2022). O *Teaching Analytics* reconceptualiza o LA com foco explícito no corpo docente, fornecendo ao professor análises sobre o seu próprio ensino e sobre o envolvimento dos alunos (Ndukwe & Daniel, 2020). A distinção fundamental reside no propósito e na audiência: o LA orienta-se para a intervenção em tempo real junto dos *stakeholders* diretos da aprendizagem, enquanto o EDM se foca na análise *post-hoc* e no desenvolvimento de modelos por parte de investigadores.

A operacionalização destas disciplinas em ambiente institucional apoia-se, de forma crescente, em arquiteturas de *Business Intelligence*. O BI replica no contexto educativo as infraestruturas tradicionalmente usadas no mundo empresarial para extrair, transformar, armazenar e analisar dados de múltiplos sistemas de informação, materializando-se em repositórios centralizados (*Data Warehouses*), processos de *Extract, Transform, Load* (ETL) e camadas de visualização através de *dashboards* e relatórios (Villegas-Ch et al., 2020). O BI constitui assim a base operacional que transporta o resultado analítico até ao decisor não técnico, função que será central no artefacto desenvolvido nesta dissertação.

A modelação da camada analítica pode seguir paradigmas distintos consoante a natureza dos dados a integrar. A modelação multidimensional de Kimball, em particular o esquema em estrela, otimiza a estrutura para a consulta e o reporte e constitui a referência para o núcleo estável do *Data Warehouse* (Kimball & Ross, 2013). Quando a esse núcleo é necessário acrescentar atributos derivados, voláteis ou de proveniência heterogénea, como os resultados de modelos preditivos sujeitos a recalibração periódica, a metodologia Data Vault oferece uma alternativa complementar (Linstedt & Olschimke, 2015). O seu conceito de satélite permite anexar a uma entidade central um conjunto de atributos descritivos com historização, entendida como o registo versionado das alterações ao longo do tempo, e com linhagem de dados própria, que assegura a rastreabilidade da respetiva proveniência, isolando a informação derivada do núcleo dimensional sem comprometer a sua integridade. Esta separação adequa-se a sistemas que combinam dados administrativos estáveis com modelos preditivos que evoluem ao longo do tempo, e fundamenta a opção, detalhada no Capítulo 3, de estender o esquema em estrela do artefacto com uma tabela satélite dedicada às saídas dos modelos preditivos de risco.

A aplicação de *analytics* em educação organiza-se numa taxonomia clássica de complexidade e valor crescentes. A análise descritiva responde à pergunta sobre o que aconteceu, sumariando e visualizando dados históricos, com aplicação típica em *dashboards* de progresso (Ndukwe & Daniel, 2020). A análise diagnóstica procura compreender por que motivo um resultado ocorreu, analisando, por exemplo, a relação entre a baixa frequência às aulas e o insucesso (Kassarnig et al., 2017; Newman-Ford et al., 2008; Ancheta et al., 2021; Etemadpour et al., 2020). A análise preditiva antecipa o que poderá acontecer, construindo modelos sobre dados históricos para prever resultados como o desempenho final ou o risco de abandono escolar, e constitui o nível mais investigado na literatura de LA e EDM (Alturki et al., 2022; Costa-Mendes et al., 2022; Matz et al., 2023). A análise prescritiva, finalmente, recomenda ações concretas para otimizar os resultados desejados, fechando o ciclo entre a previsão e a intervenção. A implementação do nível preditivo recorre predominantemente a técnicas de *Machine Learning*, exploradas na secção seguinte.

2.4 *Machine Learning* para a Previsão de Abandono e Sucesso Escolar

O *Machine Learning* (ML) posiciona-se como a principal ferramenta tecnológica para a realização da análise preditiva em educação. O seu valor reside na capacidade dos

algoritmos para identificar padrões complexos e não lineares em grandes volumes de dados, prevendo resultados relevantes como o sucesso acadêmico, o risco de abandono escolar e a retenção, e permitindo uma alocação mais eficiente dos recursos de apoio (Alturki et al., 2022).

A literatura explorou um amplo espectro de algoritmos, que podem ser organizados por complexidade e abordagem. Os classificadores tradicionais, como *Decision Tree*, *Support Vector Machine*, *K-Nearest Neighbors*, *Naïve Bayes* e *Logistic Regression*, são valorizados pela interpretabilidade e pela eficiência computacional, surgindo com frequência como referência em estudos comparativos (Aremu et al., 2022; Brahim et al., 2022; Palacios et al., 2021). Os métodos de *ensemble*, que combinam múltiplos modelos mais simples para reduzir a variância e melhorar a generalização, incluem o *Random Forest*, o *AdaBoost* e o *Extreme Gradient Boosting* (XGBoost), e demonstram em numerosos estudos um desempenho superior aos classificadores individuais na previsão do abandono (Chen & Guestrin, 2016; Costa-Mendes et al., 2022; Petrović et al., 2024). As redes neurais e as arquiteturas de *deep learning*, incluindo as *Recurrent Neural Networks* e as *Long Short-Term Memory*, são aplicadas para capturar padrões sequenciais no comportamento dos alunos, como as interações com o LMS ao longo do tempo (Waheed et al., 2020). A aprendizagem não supervisionada, por seu turno, é utilizada sobre dados não rotulados para descobrir estruturas latentes, sendo objeto da secção 2.5.

A eficácia de qualquer modelo depende da qualidade e da relevância das *features* utilizadas. A literatura converge num conjunto de categorias de variáveis: dados demográficos e socioeconómicos, como a idade na matrícula, o género e o estatuto socioeconómico; histórico académico prévio, incluindo as classificações do ensino anterior e a nota de admissão; desempenho durante o percurso, traduzido em classificações, frequência e módulos em atraso; e dados comportamentais, como o envolvimento com o LMS (Alturki et al., 2022, Costa-Mendes et al., 2022; Matz et al., 2023). Esta última categoria, embora poderosa em contexto de ensino superior, encontra-se pouco representada no ensino profissional, o que reforça a importância das restantes.

O momento em que a previsão é realizada condiciona de forma decisiva as *features* disponíveis e a precisão alcançável. Uma previsão efetuada no momento da matrícula dispõe apenas de variáveis de entrada, como o perfil socioeconómico e a escola de origem, ao passo que uma previsão efetuada ao longo do percurso incorpora a trajetória comportamental e académica acumulada (Fernández-García et al., 2021). Esta distinção

temporal é central para a presente dissertação, que articula um modelo retrospectivo, alimentado pela trajetória completa, com um modelo prospetivo, restrito às variáveis disponíveis na entrada. O modelo prospetivo sustenta, em particular, o segundo eixo do estudo, dedicado ao apoio à captação com base na escola de origem, dado que a literatura confirma o valor preditivo das características sociodemográficas disponíveis no momento da admissão (Aulck et al., 2016; Matz et al., 2023).

Dois desafios metodológicos recorrentes condicionam a fiabilidade dos modelos. O primeiro é a qualidade dos dados, dado que registos incompletos, inconsistentes ou ruidosos comprometem o desempenho das previsões (Staneviciene et al., 2024; Batool et al., 2022; Fischer et al., 2020). O segundo é o desbalanceamento de classes, comum nos problemas de abandono, em que o grupo dos alunos que saem é minoritário e tende a ser ignorado pelos modelos. Técnicas de reamostragem, como o *Synthetic Minority Over-sampling Technique* (SMOTE) e as suas variantes híbridas, são empregues para mitigar este enviesamento (Chawla et al., 2002, Batista et al., 2004; Mduma, 2023).

A avaliação do desempenho recorre a um conjunto de métricas estabelecidas. A taxa de exatidão mede a proporção de previsões corretas, mas não é fiável sob classes desbalanceadas. A matriz de confusão detalha verdadeiros e falsos positivos e negativos. A precisão e a revocação capturam, respetivamente, a fiabilidade das previsões positivas e a cobertura dos casos reais, existindo entre ambas um *trade-off*. O *F1-Score*, média harmónica das duas, equilibra-as numa métrica única, particularmente útil sob desbalanceamento. A *Area Under the Curve* (AUC) avalia a capacidade de discriminação entre classes e é insensível à sua distribuição, o que a torna robusta neste domínio (Fawcett, 2006; Matz et al., 2023). O Brier Score (Brier, 1950), finalmente, avalia a calibração das probabilidades previstas, isto é, o grau em que uma probabilidade prevista de 0,7 corresponde de facto a uma frequência observada próxima de 70%, complementando as métricas de discriminação com uma medida da fiabilidade das estimativas.

A Tabela 1 sintetiza um conjunto de trabalhos representativos da previsão de abandono e sucesso, evidenciando a predominância de contextos de ensino superior e de ambientes de dados ricos.

Tabela 1 — Síntese de trabalhos sobre previsão de abandono e sucesso acadêmico

Estudo	Contexto educativo	Técnica / Modelo	Tipo de dados	Principal contributo
Aulck et al. (2016)	Ensino superior	Regressão e <i>ensemble</i>	Admissão e sociodemográfico	Previsão de abandono a partir de dados de entrada
Costa-Mendes et al. (2022)	Ensino superior	<i>Ensemble</i>	Histórico e demográfico	Superioridade de <i>ensemble</i> ; alerta para <i>knowledge bias</i>
Matz et al. (2023)	Ensino superior	ML supervisionado	Sociodemográfico e <i>engagement</i> via <i>app</i>	Valor preditivo de dados comportamentais e de entrada
Fernández-García et al. (2021)	Ensino básico/secundário	ML por estádios	Administrativo e socioeconómico	Previsão faseada ao longo do percurso
Petrović et al. (2024)	Ensino superior	<i>Ensemble</i> com explicabilidade	Académico e comportamental	Integração de XAI na previsão de abandono
Mduma (2023)	Ensino secundário	ML com balanceamento	Administrativo	Comparação de técnicas de balanceamento de classes
Santos et al. (2026, este estudo)	Ensino profissional	XGBoost (dual-horizonte)	Administrativo, sem LMS	Previsão explicável sob dados restritos

2.5 Aprendizagem não supervisionada (*Clustering*) na Definição de Perfis de Estudantes

A aprendizagem não supervisionada permite descobrir estruturas latentes em dados não rotulados, segmentando os alunos em grupos com perfis de comportamento ou desempenho semelhantes sem recorrer a uma variável-alvo previamente definida. Esta abordagem responde diretamente ao primeiro objetivo desta dissertação, dedicado à identificação de perfis de alunos, e complementa a previsão supervisionada ao oferecer uma leitura estrutural da população estudantil (Nafuri et al., 2022; Choi, 2024).

O *k-means* é um algoritmo de clustering particional que atribui iterativamente cada observação ao grupo cujo centróide lhe é mais próximo, recalculando os centróides até à convergência (MacQueen, 1967). O algoritmo *k-means* é o método mais difundido neste domínio, pela sua eficiência e pela capacidade de atribuir novos alunos a um *cluster* existente sem necessidade de retreino, propriedade relevante para aplicações prospetivas e para a expansão a novas *cohorts*. A sua aplicação envolve a determinação prévia do número de grupos, frequentemente apoiada no método do cotovelo sobre a soma dos erros quadráticos, e a interpretação dos centróides resultantes para a construção de perfis nomeados (Omar et al., 2020; Umargono et al., 2020; Syakur et al., 2018; Pamungkas et

al., 2024; Durachman, 2025). Alternativas como o *clustering* hierárquico e os Gaussian Mixture Models, ambos comparados empiricamente nesta investigação (Secção 4.3.1), oferecem compromissos distintos entre rigidez das fronteiras, robustez ao ruído e custo computacional; o DBSCAN e o BIRCH, também discutidos em estudos comparativos da literatura, não foram aplicados neste trabalho (Pecuchova & Drlik, 2024).

A qualidade da segmentação é aferida por métricas de validação interna, das quais as mais utilizadas são o coeficiente de *Silhouette*, o índice de Davies-Bouldin e o índice de Calinski-Harabasz. Estas métricas avaliam a coesão intra-grupo e a separação inter-grupo de forma independente da existência de rótulos, e a sua convergência reforça a confiança na estrutura adotada (Pecuchova & Drlik, 2024). Para além da sua função descritiva, o *clustering* pode ainda alimentar a fase supervisionada; a literatura demonstra que a incorporação dos rótulos de grupo como informação adicional melhora o desempenho dos modelos preditivos a jusante, ligando as duas tarefas analíticas numa arquitetura integrada (Alvarez-Garcia et al., 2024; Pecuchova & Drlik, 2024).

Os perfis assim obtidos sustentam estratégias de intervenção diferenciada, ao distinguir grupos que requerem respostas institucionais distintas, desde o apoio académico ao acompanhamento socioeconómico. A coerência temporal desses perfis, ou seja, a verificação de que os grupos identificados se traduzem em trajetórias de permanência distintas ao longo do tempo, pode ainda ser validada através de análise de sobrevivência, tema retomado no Capítulo 3. A Tabela 2 sintetiza trabalhos representativos.

Tabela 2 — Síntese de trabalhos sobre clustering e perfis de estudantes

Estudo	Contexto educativo	Técnica de clustering	Validação / Abordagem	Principal contributo
Nafuri et al. (2022)	Ensino superior	<i>k-means</i>	Análise de perfis	Segmentação de alunos por padrões de desempenho
Choi (2024)	Ensino superior	Não supervisionado	Perfis comportamentais	Caracterização de grupos de risco
Syakur et al., 2018	Ensino superior	<i>k-means</i>	Método do cotovelo	Perfis para diagnóstico de insucesso
Alvarez-Garcia et al. (2024)	Ensino superior	<i>k-means</i> ponderado com PCA	Validação interna	Integração de <i>clustering</i> e classificação
Pecuchova & Drlik (2024)	Ensino superior	DBSCAN, BIRCH, GMM	<i>Silhouette</i> , Davies-Bouldin, Calinski-Harabasz	<i>Clustering</i> de rastros digitais melhora a previsão precoce

Santos et al. (2026, este estudo)	Ensino profissional	<i>k-means</i>	Três métricas internas e Kaplan-Meier	Perfis de risco interpretáveis sob dados administrativos
--------------------------------------	------------------------	----------------	--	--

2.6 Inteligência Artificial Explicável (*eXplainable AI*)

Os modelos de melhor desempenho na previsão de abandono escolar, sobretudo os métodos de *ensemble* e as redes neurais, são intrinsecamente opacos, produzindo previsões precisas sem revelar os fatores que as determinam. Esta opacidade constitui uma barreira prática à adoção em contexto educativo, onde diretores de curso, orientadores e equipas de apoio necessitam não apenas de uma classificação de risco, mas da compreensão dos fatores que a motivam, condição para desenhar respostas adequadas (Khosravi et al., 2022). A *Explainable Artificial Intelligence* (XAI) responde a esta necessidade, tornando os modelos transparentes através da atribuição de um valor de contribuição a cada *feature*.

Entre as técnicas disponíveis, o SHAP (*SHapley Additive exPlanations*) emergiu como a abordagem dominante, ao fornecer simultaneamente rankings globais de importância das variáveis e explicações ao nível da instância individual, que revelam como cada variável influencia a previsão para um aluno específico (Lundberg & Lee, 2017). O LIME e as medidas de importância de *features* baseadas no próprio modelo constituem alternativas complementares. A aplicação de XAI à previsão de desempenho e abandono tem-se consolidado como prática de referência, associada ao reforço da confiança dos *stakeholders* e à viabilização da adoção institucional (Ahmed et al., 2025; Petrović et al., 2024; Liu et al., 2025).

Os estudos que aplicam SHAP em previsão educativa assumem elevada relevância para este trabalho. As variáveis financeiras e socioeconómicas tendem a dominar a explicação no momento da matrícula, quando ainda não existem dados comportamentais, ao passo que os indicadores de desempenho académico e de envolvimento as substituem como preditores mais influentes à medida que a trajetória se acumula (Ahmed et al., 2025). Esta evolução temporal da relevância dos preditores sugere uma interpretação institucionalmente significativa: o contexto socioeconómico funciona como *proxy* da escassez de informação no início do percurso, e não como determinante fixo do resultado, distinção com implicações éticas e práticas diretas no desenho das intervenções. A explicabilidade adquire, por esta via, uma dupla função, simultaneamente operacional, ao orientar a ação, e ética, ao permitir a deteção de enviesamentos, conforme se desenvolve

na secção 2.7. A Tabela 3 sintetiza os trabalhos que aplicam XAI à previsão de sucesso e abandono escolar.

Tabela 3 — Síntese de trabalhos que aplicam XAI à previsão de sucesso e abandono

Estudo	Contexto educativo	Modelo preditivo	Técnica de explicabilidade	Principal contributo
Lundberg & Lee (2017)	Transversal (fundacional)	Agnóstico	SHAP	Formulação unificada da interpretação de modelos
Khosravi et al. (2022)	Educação	Diversos	Revisão de XAI em educação	Enquadramento da interpretabilidade para decisão escolar
Petrović et al. (2024)	Ensino superior	<i>Ensemble</i>	SHAP	Deteção de fatores de risco e mitigação de viés
Liu et al., 2025	Educação	<i>Ensemble</i>	SHAP	Explicabilidade para predição de resultados
Ahmed et al. (2025)	Educação	ML supervisionado	SHAP / <i>feature importance</i>	Evolução temporal da relevância dos preditores
Santos et al. (2026, este estudo)	Ensino profissional	XGBoost	SHAP (<i>TreeExplainer</i>)	Explicação dual-horizonte sob dados administrativos

2.7 De Preditivo a Prescritivo: Sistemas de Apoio à Decisão e BI

A investigação focada na análise preditiva atingiu maturidade técnica e demonstra hoje uma notável capacidade de identificar alunos em risco. Este sucesso expôs, contudo, uma limitação de fundo: a capacidade de prever um problema não garante a sua resolução. A abordagem puramente preditiva termina no diagnóstico, sem oferecer um caminho claro para a ação, e a passagem da previsão à intervenção informada define a fronteira atual do campo (Alalawi et al., 2025).

Três barreiras dificultam esta transição. A primeira é o *last-mile problem*, que descreve a dificuldade de traduzir as previsões em intervenções adotadas em larga escala e com melhorias mensuráveis, existindo frequentemente um fosso entre o *insight* analítico e a sua implementação prática (Alalawi et al., 2025). A segunda é a escassez de conhecimento causal, dado que a maioria dos estudos se apoia em modelos correlacionais, o que dificulta o desenho de intervenções com garantia de efeito (Klenke et al., 2022). A terceira é a adoção limitada das intervenções baseadas em *learning analytics*,

condicionada pela falta de infraestrutura, pela resistência à mudança e pela complexidade de integração nos processos existentes (Alalawi et al., 2025).

A resposta a estas barreiras passa pela análise prescritiva, nível mais avançado da taxonomia, que recomenda ações específicas para otimizar resultados. No ensino profissional, isto traduz-se em sistemas que, a partir das previsões, sugerem intervenções personalizadas, fornecem recomendações às equipas pedagógicas ou orientam a alocação de recursos. A operacionalização destas recomendações depende de artefactos de *software* bem desenhados, que funcionam como interface entre os modelos analíticos e os utilizadores finais. A literatura descreve várias categorias de tais artefactos. Os Sistemas de Alerta Precoce (*Early Warning Systems*) identificam alunos em risco numa fase inicial, permitindo intervenção atempada (Ali et al., 2023; Bañeres et al., 2023). O projeto *Course Signals* da Universidade de Purdue, exemplo pioneiro, utilizava um sistema de semáforos para classificar e comunicar o risco a alunos e professores (Arnold & Pistilli, 2012, citado em Alalawi et al., 2025). Os *dashboards* de gestão e os *Teaching Analytics Dashboards* permitem monitorizar indicadores-chave e refletir sobre a prática pedagógica (Ndukwe & Daniel, 2020). As plataformas integradas, como o *Student Performance Prediction and Action* (SPPA), reúnem previsão, gestão de intervenções e avaliação numa única ferramenta de autosserviço (Alalawi et al., 2025).

A eficácia destes artefactos não é uma questão puramente técnica. Os fatores críticos de sucesso incluem a usabilidade e a interpretabilidade das visualizações por utilizadores não técnicos, o alinhamento com os processos e a cultura institucional através do envolvimento dos *stakeholders* no desenho, e a existência de um ciclo de *feedback* que permita avaliar o impacto e refinar continuamente o sistema (Ali et al., 2023; Ndukwe & Daniel, 2020; Villegas-Ch et al., 2020). Do ponto de vista arquitetural, este ciclo apoia-se nas infraestruturas de BI descritas na secção 2.3, em que um *Data Warehouse* organizado segundo metodologias como a de Kimball, suporta a integração de dados de múltiplos sistemas e a sua disponibilização em camadas de visualização (Villegas-Ch et al., 2020). É também nesta camada de apoio à decisão que se concretiza o segundo objetivo desta dissertação: a análise agregada dos padrões associados à escola de origem e performance por curso e perfis históricos informa as estratégias institucionais de captação, transformando o resultado preditivo em orientação operacional para a admissão.

2.8 Considerações Éticas e de Equidade

A utilização de dados de alunos para fundamentar decisões que afetam o seu percurso acarreta responsabilidades éticas que condicionam a legitimidade de qualquer iniciativa de *learning analytics*. A recolha de dados levanta questões de privacidade e consentimento, agravadas em contextos regulados pelo *General Data Protection Regulation*, que exige transparência quanto à finalidade do tratamento (Karunaratne, 2021; Liu & Khalil, 2023; Sun, 2023; Jones et al., 2023). A propriedade dos dados permanece em debate, com implicações sobre os direitos de acesso e utilização, e a monitorização contínua dos alunos pode ser percebida como vigilância, erodindo a confiança institucional quando os processos são opacos (Danaher et al., 2023).

A preocupação mais premente é o viés algorítmico, que ocorre quando um modelo produz resultados que discriminam sistematicamente certos grupos com base em características como a etnia, o género ou o estatuto socioeconómico (Baker & Hawn, 2021). Um caso particular é o *knowledge bias*, segundo o qual o conhecimento prévio incorporado pelo criador do modelo, designadamente através da seleção de *features*, pode introduzir e perpetuar preconceitos existentes (Costa-Mendes et al., 2022). A explicabilidade descrita na secção 2.6 assume aqui um papel de mitigação, ao permitir que especialistas analisem os fatores que mais influenciam as previsões e detetem padrões discriminatórios (Petrović et al., 2024). A literatura propõe ainda enquadramentos processuais para guiar um desenvolvimento responsável, como o *framework* AIRR, assente nos princípios de antecipação, inclusão, responsividade e reflexividade, com aplicação direta ao desenho de sistemas de alerta precoce (Ali et al., 2023). As implicações éticas operacionais deste estudo, incluindo a pseudonimização, o enquadramento no artigo 22.º do RGPD e a manutenção da decisão humana no ciclo, são detalhadas no Capítulo 3.

2.9 Síntese, Lacunas e Posicionamento da Solução

A revisão mapeou a trajetória de maturação do *analytics* em educação, do ensino superior ao ensino profissional. O percurso parte da abundância de dados gerada pela digitalização, abundância que em larga medida não alcança a realidade das instituições de ensino profissional, atravessa o domínio técnico dos modelos de *Machine Learning* para fins preditivos e de *clustering* para a definição de perfis, e culmina na constatação de que a previsão, isoladamente, é insuficiente. A fronteira atual reside na transição para

uma abordagem prescritiva, suportada por artefactos de apoio à decisão e condicionada por fatores sociotécnicos e éticos.

Desta análise emergem cinco lacunas. A primeira é a escassez de conhecimento causal, decorrente do predomínio de modelos correlacionais, que limita o desenho de intervenções com efeito garantido (Klenke et al., 2022). A segunda é o desafio da generalização e validação, dado que a maioria dos modelos é desenvolvida e testada num único contexto institucional (Quille & Bergin, 2022). A terceira é o problema da adoção e do impacto real, em que o *last-mile problem* e a medição rigorosa do retorno das intervenções permanecem por resolver (Alalawi et al., 2025). A quarta é a integração insuficiente da equidade, com atenção limitada ao efeito diferencial das ferramentas sobre populações distintas (Baker & Hawn, 2021). A quinta é a escassez de investigação em contextos de ensino profissional, ainda incipiente face à dimensão do setor (Gedrimiene et al., 2019).

Para além destas lacunas, identifica-se uma lacuna de integração que as atravessa. A literatura tende a tratar a previsão (EDM), a explicabilidade (XAI) e o apoio à decisão (BI) como contribuições separadas, raramente articuladas num único sistema operacional. São poucos os trabalhos que integram a modelação preditiva, a interpretabilidade e a camada prescritiva de BI num artefacto coerente e implantável, e essa escassez é ainda mais acentuada no ensino profissional com dados administrativos restritos. É precisamente esta lacuna integrada que a presente dissertação procura preencher.

A solução proposta neste trabalho de dissertação, procura responder a esta lacuna ao desenvolver e validar um *framework* de *educational analytics* explicável, especificamente desenhado para o ensino profissional em ambiente de dados restrito. O artefacto integra, num sistema único, a definição de perfis de risco por *clustering*, a previsão de não conclusão em dois horizontes temporais, a explicabilidade por SHAP e uma camada de apoio à decisão em *Business Intelligence*. O primeiro horizonte é prospetivo e estima o risco de não conclusão logo à entrada, recorrendo apenas às variáveis conhecidas no momento da matrícula; o segundo é retrospectivo e revê essa estimativa ao longo do percurso, à medida que o desempenho académico do aluno passa a estar disponível. A arquitetura dual-horizonte serve simultaneamente os dois eixos do estudo, ao sustentar a intervenção sobre alunos em risco e a orientação das estratégias de captação a partir da escola de origem. A validação do artefacto junto dos responsáveis institucionais, documentada no Capítulo 5, dirige-se diretamente ao problema da adoção, complementando o rigor técnico com a aferição da utilidade percebida em contexto real.

CAPÍTULO III – METODOLOGIA

3 Metodologia

Este capítulo descreve as seleções metodológicas que orientaram a conceção, o desenvolvimento e a avaliação do artefacto produzido nesta investigação.

3.1 Enquadramento Metodológico

A presente investigação tem como objetivo principal a conceção, desenvolvimento e avaliação de um sistema de análise para apoiar a previsão da não conclusão e a tomada de decisão institucional no ensino profissional. O carácter construtivo da investigação, centrado na produção de um artefacto com utilidade operacional verificável, determina a escolha metodológica a dois níveis complementares.

Em termos metodológicos adotou-se o Design Science Research (DSR) (Hevner et al., 2004), que fornece o enquadramento epistemológico adequado para estudos que criam e avaliam artefactos em contexto real. Ao nível do processo de construção do artefacto, foi implementado o *Cross-Industry Standard Process for Data Mining* (CRISP-DM) (Chapman et al., 2000), que operacionaliza a fase de conceção e desenvolvimento através de um ciclo iterativo orientado ao problema institucional. A DSR e o CRISP-DM articulam-se de forma hierárquica: a primeira define o enquadramento de investigação, o porquê e o para quê da construção do artefacto, enquanto o segundo operacionaliza a execução técnica dentro da fase de Design e Desenvolvimento. O presente capítulo descreve ambos os níveis, fundamenta as escolhas efetuadas e documenta a sua articulação ao longo do projeto.

3.2 Metodologia de Investigação (*Design Science Research*)

A *Design Science Research* (DSR) constitui um paradigma de investigação em Sistemas de Informação cujo propósito central é a produção de conhecimento científico através da criação e avaliação de artefactos, entendidos como construtos, modelos, métodos ou instanciações que respondem a problemas concretos e verificáveis (Hevner et al., 2004). Ao contrário da investigação behaviorista, centrada na explicação e previsão de fenómenos existentes, a DSR orienta-se para a criação e avaliação de artefactos capazes de os transformar (Hevner et al., 2004). Esta orientação é particularmente adequada em contextos de investigação aplicada onde o produto da investigação é simultaneamente um contributo científico e um instrumento com utilidade institucional direta.

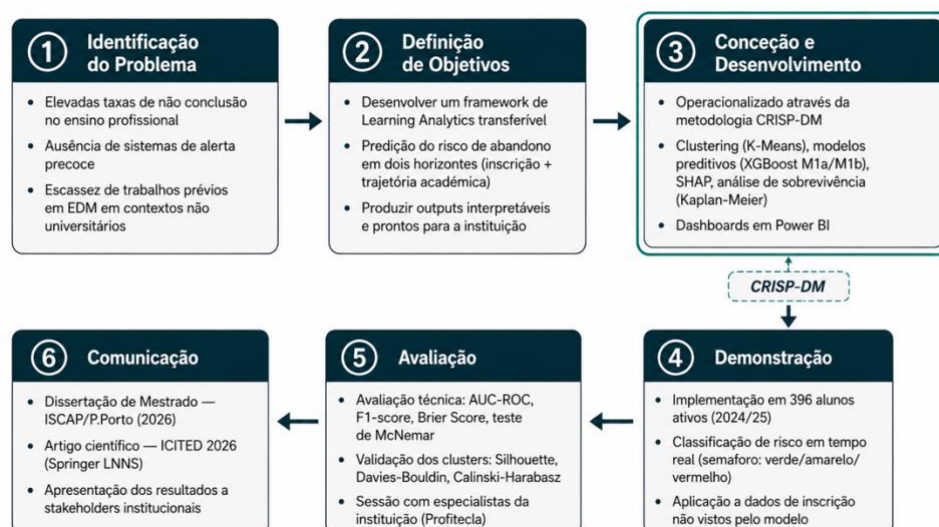
A aplicação da DSR ao domínio da educação tem vindo a consolidar-se como abordagem metodológica legítima em estudos que desenvolvem sistemas de apoio à decisão educacional. Fahd e Miah (2023) demonstram a adequação da DSR, combinada com metodologias de investigação baseada em design próprias do contexto educativo, para o desenvolvimento de sistemas analíticos de suporte à aprendizagem em larga escala, argumentando que o DSR complementa abordagens educativas ao fornecer perspectivas multiparadigmáticas para a construção de conhecimento através da pesquisa de pragmatismos sociais.

A adoção da DSR como metodologia de investigação neste estudo justifica-se por três razões. O resultado principal da investigação é um artefacto (um *framework* preditivo que integra modelos de *machine learning*, mecanismos de explicabilidade e *dashboards* de apoio à decisão), e não uma teoria ou descrição de fenómenos. O artefacto é desenvolvido e avaliado com dados reais de uma instituição portuguesa de ensino profissional, garantindo a relevância prática exigida pela DSR. A metodologia orienta ainda explicitamente a investigação para a transferibilidade do conhecimento produzido, alinhando-se com o objetivo de disponibilizar um *framework* replicável para outras instituições de ensino profissional com características semelhantes.

3.2.1 Fases da DSR Aplicadas à Investigação

A presente investigação segue o processo DSR formalizado por Peffers et al. (2007), estruturado em seis fases operacionais. O mapeamento das seis fases ao projeto é apresentado na Figura 1.

Figura 1— Fases da Design Science Research aplicadas à investigação (adaptado de Peffers et al., 2007)



Fase 1 — Identificação do Problema: O problema central reside na retenção de alunos no ensino profissional português, combinada com a ausência de mecanismos sistemáticos de intervenção precoce e a escassez de estudos de *Educational Data Mining* adaptados a contextos institucionais com dados administrativos limitados, sem registos de interação digital com plataformas de e-learning.

Fase 2 — Definição dos Objetivos: A solução deve produzir um *framework* de *Learning Analytics* capaz de: (i) classificar alunos por nível de risco em dois horizontes temporais distintos, no momento de matrícula e ao longo do percurso académico; (ii) fornecer explicações interpretáveis das previsões de risco de não conclusão, acessíveis a gestores sem formação técnica em *Machine Learning*; e (iii) apresentar os resultados num formato operacionalmente utilizável para apoio à decisão institucional. A transferibilidade do *framework* para outros polos da mesma instituição e para outras organizações de ensino profissional constitui um objetivo adicional.

Fase 3 — Conceção e Desenvolvimento: Esta fase é operacionalizada através da metodologia CRISP-DM, detalhada na secção 3.3. O artefacto desenvolvido integra técnicas de *clustering* (K-Means), análise preditiva supervisionada (XGBoost), explicabilidade algorítmica (SHAP), análise de sobrevivência (Kaplan-Meier) e um sistema de *Business Intelligence* integrador de todo o sistema para visualização e suporte à decisão (recorrendo à tecnologia Power BI).

Fase 4 — Demonstração: O artefacto é aplicado operacionalmente a 396 alunos ativos no ano letivo 2024/25, utilizando exclusivamente variáveis disponíveis no momento de matrícula, sem *outcome* observado. Esta aplicação constitui a demonstração da capacidade do *framework* para classificar risco de abandono em tempo real, em condições idênticas às de utilização institucional efetiva.

Fase 5 — Avaliação: O enquadramento FEDS (*Framework for Evaluation in Design Science*), proposto por Venable et al. (2016), orienta o planeamento da avaliação de artefactos em Design Science Research ao distinguir dois paradigmas, a avaliação artificial, realizada em condições controladas e centrada nas propriedades técnicas do artefacto, e a avaliação naturalista, conduzida com utilizadores reais no contexto de utilização efetiva. Seguindo este enquadramento, a avaliação do artefacto integra duas componentes complementares: (i) avaliação técnica, conduzida através de métricas de desempenho dos modelos preditivos (AUC-ROC, F1-score, Brier Score), validação estatística dos *clusters* (Silhouette, Davies-Bouldin, Calinski-Harabasz) e testes de

significância estatística (McNemar); e (ii) a avaliação de utilidade percebida foi conduzida numa sessão estruturada de demonstração e avaliação realizada a 9 de junho de 2026, com a participação de quatro responsáveis da Profitecla (Apêndice I). O instrumento, baseado nos construtos do Technology Acceptance Model (Davis, 1989) e nos critérios de avaliação de artefactos de Venable et al. (2016), incluiu onze itens de escala Likert (1–5) organizados em quatro dimensões, utilidade percebida, facilidade de utilização, relevância e transferibilidade, e avaliação global, complementados por três questões abertas. Os resultados são apresentados e discutidos no Capítulo 5.

Fase 6 — Comunicação: A disseminação dos resultados da investigação é conduzida em dois registos complementares. No plano científico, os resultados são comunicados através da presente dissertação de mestrado e de um artigo submetido ao ICITED 2026 (Springer LNNS). No plano institucional, os resultados são apresentados aos responsáveis da Profitecla no âmbito da sessão de avaliação (Apêndice I) referida na fase anterior, promovendo a adoção informada do *framework* e a identificação de desenvolvimentos futuros.

3.2.2 Relação entre DSR e CRISP-DM

A relação entre DSR e CRISP-DM é de complementaridade hierárquica, a DSR define o enquadramento de investigação, enquanto o CRISP-DM operacionaliza o processo de construção do artefacto dentro da fase 3. Esta articulação permite que os resultados sejam avaliados simultaneamente pelo seu rigor metodológico enquanto investigação em Sistemas de Informação (DSR) e pela sua reprodutibilidade técnica enquanto processo de *Data Mining* (CRISP-DM). A fase de *Deployment* do CRISP-DM corresponde às fases de Demonstração e Avaliação da DSR, fechando o ciclo de investigação com evidência operacional recolhida em contexto real.

3.3 CRISP-DM: Justificação e Fases Aplicadas ao Projeto

O Cross-Industry Standard Process for *Data Mining* (CRISP-DM) foi desenvolvido como um processo de referência independente de domínio e de ferramenta para projetos de *Data Mining*, estruturando o trabalho analítico em seis fases iterativas orientadas ao problema de negócio (Chapman et al., 2000). A sua adoção como *framework* dominante em projetos de análise de dados aplicada é amplamente documentada na literatura: Schröer et al. (2021), numa revisão sistemática abrangendo 14 anos de publicações científicas, confirmam o CRISP-DM como o processo mais utilizado

em estudos de *Data Mining* e *Machine Learning*, destacando a sua flexibilidade e adequação a contextos com requisitos institucionais explícitos.

A seleção do CRISP-DM em detrimento de abordagens alternativas, nomeadamente o processo SEMMA (*Sample, Explore, Modify, Model, Assess*) e o *Knowledge Discovery in Databases* (KDD), fundamenta-se em três características que o tornam particularmente adequado ao contexto desta investigação. O CRISP-DM inclui a compreensão do negócio como fase autónoma, forçando a articulação entre os objetivos analíticos e os requisitos operacionais da instituição antes de qualquer manipulação de dados. Num estudo cujos resultados se destinam a apoiar decisões de gestão por parte de utilizadores sem formação técnica, este alinhamento prévio é determinante. A natureza iterativa e não linear do processo, com possibilidade de visitar fases anteriores à medida que os resultados evoluem, reflete a realidade dos projetos de análise aplicada, onde descobertas na modelação frequentemente implicam revisões na preparação de dados. Por fim, a independência do CRISP-DM em relação a ferramentas e tecnologias específicas facilita a documentação reproduzível do processo, condição relevante para a transferibilidade do *framework* a outros contextos institucionais. A Figura 2 representa as seis fases do CRISP-DM e a sua natureza cíclica.

Figura 2 — Fases do CRISP-DM e carácter iterativo do processo (adaptado de Chapman et al., 2000)



Fase 1 — Compreensão do Negócio (*Business Understanding*): A fase de compreensão do negócio estabelece o enquadramento institucional da investigação e

traduz os objetivos organizacionais em questões analíticas operacionalizáveis. Esta fase foi conduzida através de reuniões e trocas de informação com a equipa responsável da Profitecla, nomeadamente com os interlocutores diretos da instituição parceira, que identificaram as principais necessidades e as limitações do sistema de monitorização existente. Desta interação resultou a caracterização do problema principal, a ausência de mecanismos sistemáticos de antecipação da não conclusão do curso e de fundamentação das decisões de captação em evidência histórica.

Com base nesse processo de levantamento de requisitos, foram definidos dois eixos analíticos complementares: a predição de risco de não conclusão do curso em dois horizontes temporais distintos, e a caracterização das escolas de origem associadas a diferentes perfis de sucesso académico. A definição operacional de não conclusão, enquanto variável-alvo da investigação, foi igualmente estabelecida nesta fase com base nos estados académicos registados no Sistema de Informação da instituição, excluindo os alunos em situação de transição ativa cujo outcome definitivo permanece indeterminado no momento da análise. Esta decisão operacional, tomada em articulação com os responsáveis institucionais, determinou a composição da amostra disponível para modelação e garantiu que a definição adotada correspondesse à noção de não conclusão do curso operacionalmente relevante para a Profitecla.

Fase 2 — Compreensão dos Dados (*Data Understanding*): A fase de compreensão dos dados envolveu a análise exploratória da fonte de dados disponibilizada pela Profitecla, constituída por registos administrativos do Polo do Porto abrangendo o período de 2018/19 a 2023/24. Os dados estão organizados ao nível do aluno-ano letivo, integrando três categorias de variáveis: características de perfil de entrada, registos de desempenho académico ao longo do percurso, e outcomes de empregabilidade pós-conclusão.

A exploração inicial dos dados permitiu identificar um conjunto de problemas de qualidade, incluindo inconsistências em campos de datas, duplicações de registos e variáveis com elevada percentagem de valores em falta, cujo tratamento detalhado é descrito no Capítulo 4. Foi igualmente identificado um desequilíbrio significativo na distribuição da variável-alvo, com uma proporção aproximada de três alunos com conclusão para cada aluno com não conclusão, o que determinou a necessidade de estratégias de balanceamento na fase seguinte. Por fim, a compreensão dos dados confirmou a ausência de registos de interação com plataformas digitais de aprendizagem, delimitando o espaço de variáveis disponíveis a dados administrativos, condição que

constitui simultaneamente uma limitação e uma característica que torna este estudo representativo da realidade da maioria das instituições de ensino profissional.

Fase 3 — Preparação dos Dados (*Data Preparation*): A fase de preparação dos dados compreendeu o conjunto de transformações necessárias para tornar os dados adequados à modelação. As decisões mais relevantes desta fase dizem respeito à construção da variável-alvo, ao tratamento de valores em falta, à codificação de variáveis categóricas e ao balanceamento das classes.

A construção da variável-alvo seguiu uma lógica de identificação do estado final observado de cada aluno. Os casos com outcome definitivo (conclusão ou não conclusão) foram separados dos casos em estado transitório, que foram excluídos da amostra de modelação. Esta distinção é metodologicamente necessária para garantir que o modelo aprende exclusivamente a partir de percursos completos, evitando a contaminação com informação incompleta.

O desequilíbrio de classes identificado na fase anterior foi tratado através da técnica SMOTE-ENN (*Synthetic Minority Over-sampling Technique with Edited Nearest Neighbours*), que combina a geração de exemplos sintéticos da classe minoritária com a eliminação de exemplos ambíguos na fronteira de decisão (Batista et al., 2004). Em comparação com abordagens mais simples, como o *oversampling* aleatório ou a ponderação de classes, a técnica distingue-se pela capacidade de produzir uma fronteira de decisão mais limpa, reduzindo o ruído introduzido por exemplos de difícil classificação. A divisão dos dados em conjuntos de treino e teste foi efetuada com estratificação, preservando a proporção das classes originais e garantindo a representatividade dos resultados na avaliação.

Fase 4 — Modelação (*Modelling*): A fase de modelação integra quatro componentes analíticas complementares, cada uma respondendo a uma questão de investigação distinta: a caracterização de perfis de risco através de *clustering*, a predição supervisionada da não conclusão em dois horizontes temporais, a explicabilidade das previsões e a análise temporal do abandono escolar.

***Clustering*:** A análise de *clustering* foi conduzida previamente à modelação supervisionada para identificar perfis de alunos sem recorrer à variável-alvo de não conclusão como critério de agrupamento. A sequência é intencional: os *clusters* caracterizam a população estudantil de forma independente, e o seu perfil não é influenciado pelos resultados dos modelos preditivos, constituindo um contributo analítico autónomo.

Foram comparados três algoritmos de *clustering*, K-Means (MacQueen, 1967), Hierárquico Ward (Ward, 1963) e Gaussian Mixture Models (GMM) (McLachlan & Peel, 2000), com base em três métricas de validação interna, coeficiente de Silhouette (Rousseeuw, 1987), índice Davies-Bouldin (Davies & Bouldin, 1979) e índice Calinski-Harabasz (Caliński & Harabasz, 1974). O K-Means com $k=3$ foi selecionado como solução principal por apresentar os melhores resultados nas três métricas simultaneamente e por satisfazer um requisito operacional fundamental: a capacidade de atribuir novos alunos a um *cluster* sem necessidade de retreinar o modelo, o que é indispensável para a aplicação prospectiva do *framework* a *cohorts* futuras e para a eventual expansão a outros polos da instituição. Os três *clusters* identificados correspondem a perfis operacionalmente interpretáveis, designados como Sucesso, Abandono e Risco Intervencionável, conforme descrito no Capítulo 4. Uma análise complementar com $k=4$ é igualmente apresentada, revelando uma segmentação adicional dentro do perfil de Risco Intervencionável com implicações para a diferenciação das estratégias de intervenção.

Modelos preditivos: A previsão baseada em análise supervisionada da não conclusão do curso, foi estruturada em dois modelos complementares, diferenciados pelo horizonte temporal a que se destinam. O Modelo M1b é um instrumento de avaliação prospectiva: utiliza exclusivamente variáveis disponíveis no momento de matrícula do aluno, permitindo classificar o risco antes de qualquer interação com a instituição. O Modelo M1a é um instrumento de avaliação retrospectiva: utiliza o conjunto completo de variáveis comportamentais e académicas registadas ao longo do percurso, produzindo uma avaliação mais precisa, mas dependente de informação que só existe após o início do ano letivo. A arquitetura de dois modelos é deliberada, refletindo dois momentos distintos de decisão institucional: a admissão e o acompanhamento contínuo.

Para ambos os modelos, foram comparados três algoritmos de aprendizagem supervisionada, XGBoost, Random Forest e Regressão Logística, através de validação cruzada estratificada e do teste de McNemar para comparação estatística de classificadores. Os resultados indicaram equivalência estatística entre os três algoritmos, tendo o XGBoost sido selecionado como algoritmo final por consistência metodológica entre os dois modelos (Chen & Guestrin, 2016). Face ao desequilíbrio de classes, o M1b recorreu adicionalmente à técnica SMOTE-ENN descrita na fase de preparação de dados. A calibração isotónica das probabilidades foi avaliada, mas rejeitada após verificação

empírica de que deteriorava a qualidade probabilística do modelo, quantificada pelo Brier Score.

A definição do limiar de classificação de risco, que determina a fronteira entre as classificações semafóricas de vermelho, amarelo e verde, seguiu uma lógica de minimização dos falsos negativos, reconhecendo que, em contextos educacionais, o custo de não identificar um aluno em risco é sistematicamente superior ao custo de um falso alarme (Torres-Cruz et al., 2025; Mduma, 2023). O limiar foi estabelecido empiricamente com base na análise da curva Precision-Recall, privilegiando a sensibilidade na detecção da classe de risco.

Explicabilidade: A interpretabilidade das previsões dos modelos foi assegurada através da técnica SHAP (SHapley Additive exPlanations), que produz valores de contribuição individuais para cada variável em cada previsão, com base na teoria dos jogos cooperativos (Lundberg & Lee, 2017). A adoção de um método de explicabilidade algorítmica justifica-se pelo contexto de utilização do *framework*: os destinatários das previsões são gestores e coordenadores pedagógicos sem formação técnica em *Machine Learning*, para quem a transparência das recomendações é condição necessária para a confiança e a adoção. O SHAP pertence à fase de Modelação, e não à Avaliação, porque o seu propósito é interpretar o funcionamento do modelo e não medir o seu desempenho.

Análise de sobrevivência: A análise de sobrevivência através do estimador de Kaplan-Meier (Kaplan & Meier, 1958), foi incorporada para responder a uma questão que a classificação binária não consegue abordar: em que momento do percurso académico ocorre tipicamente o abandono e se essa dinâmica temporal é consistente com os perfis de risco identificados pelo *clustering*. O estimador de Kaplan-Meier permite estimar a probabilidade de permanência ao longo do tempo, considerando o tempo até à ocorrência do evento de abandono. Os casos de conclusão do curso foram tratados como observações censuradas, ou seja, situações em que o evento de abandono não ocorreu durante o período de observação. Aplicado separadamente por *cluster*, o estimador permite verificar se os perfis identificados apresentam curvas de sobrevivência estatisticamente distintas. No âmbito do CRISP-DM, esta análise integra a fase de Avaliação, onde serve como validação da coerência temporal dos perfis de risco.

Fase 5 — Avaliação (*Evaluation*): A fase de avaliação do CRISP-DM incide sobre a qualidade técnica dos modelos e dos perfis produzidos na fase de Modelação. A avaliação dos modelos preditivos recorre a métricas de classificação estabelecidas, AUC-ROC (Fawcett, 2006), F1-score e Brier Score (Brier, 1950), complementadas por testes

de significância estatística para comparação entre classificadores. A avaliação dos clusters utiliza métricas de validação interna, como o coeficiente de Silhouette (Rousseeuw, 1987), o índice Davies-Bouldin (Davies & Bouldin, 1979) e o índice Calinski-Harabasz (Caliński & Harabasz, 1974), sendo complementada pela análise de Kaplan-Meier descrita na secção anterior, que valida a coerência temporal dos perfis identificados. Os resultados desta avaliação técnica são apresentados em detalhe no Capítulo 4. A avaliação da utilidade operacional do artefacto junto dos responsáveis institucionais constitui a componente de avaliação sumativa da DSR, documentada na secção 3.2.2.

Fase 6 — Implementação (*Deployment*): A fase de implementação operacionalizou o *framework* desenvolvido em duas componentes. Os modelos treinados foram aplicados ao universo de alunos ativos no ano letivo 2024/25, produzindo classificações de risco com base em variáveis de matrícula, o que constitui a demonstração operacional do *framework* em condições reais de utilização. Os resultados também foram integrados num sistema de *Business Intelligence* usando a ferramenta Power BI, estruturado segundo um modelo multidimensional star schema (Kimball & Ross, 2013), com uma camada analítica de *Machine Learning*, que serve de interface entre os outputs dos modelos e os utilizadores institucionais. Com esta fase fecha-se o ciclo iterativo do CRISP-DM e completam-se as fases de Demonstração e Avaliação da DSR, com os resultados técnicos integrados num artefacto pronto para utilização institucional.

3.4 Considerações Éticas, Privacidade e Proteção de Dados

O desenvolvimento desta investigação envolveu o tratamento de dados pessoais de alunos, o que implica a adoção de um conjunto de medidas para garantir o respeito pelos princípios de proteção de dados consagrados no Regulamento Geral sobre a Proteção de Dados (RGPD, Regulamento UE 2016/679). O tratamento dos dados foi realizado ao abrigo de autorização da Profitecla, instituição detentora e responsável pelos dados, e circunscrito exclusivamente para os fins desta investigação académica, não tendo os dados sido partilhados com terceiros em nenhum momento do processo.

Os dados fornecidos pela instituição parceira foram previamente pseudonimizados, cada aluno é identificado por um código numérico interno, sem qualquer correspondência direta à informação nominativa. A chave de associação entre os identificadores numéricos e as identidades reais dos alunos é detida exclusivamente

pela Profitecla e não foi, em nenhum momento, partilhada com o investigador. Esta separação garante que o tratamento analítico efetuado nesta dissertação não permite a identificação individual dos alunos. O conjunto de dados inclui variáveis de natureza potencialmente sensível, designadamente indicadores socioeconómicos relativos ao escalão de apoio social dos alunos, cujo tratamento foi sujeito a particular cuidado no que respeita à sua utilização analítica e à forma como os resultados são apresentados, de modo a evitar qualquer forma de estigmatização de grupos ou indivíduos.

Uma dimensão ética central deste *framework* diz respeito ao papel que os modelos preditivos desempenham no processo de decisão institucional. Os outputs dos modelos (classificações de risco, perfis de *clustering* e valores de importância de variáveis) constituem instrumentos de informação e de priorização para uso dos responsáveis pedagógicos, sem substituir o julgamento humano sobre situações individuais. A decisão final cabe sempre aos coordenadores pedagógicos e responsáveis institucionais, que devem interpretar as recomendações do sistema, considerar o contexto específico de cada aluno e determinar a natureza e a oportunidade de qualquer ação. Este posicionamento coloca o *framework* fora do âmbito do artigo 22.º do RGPD relativo a decisões automatizadas com efeitos jurídicos significativos, e alinha-se com os princípios de human-in-the-loop que a literatura de *Learning Analytics* identifica como condição necessária para a adoção ética de sistemas de apoio à decisão em contextos educacionais (Khosravi et al., 2022).

Esta escolha de design tem implicações concretas na arquitetura do *framework*. A incorporação de explicabilidade algorítmica através do SHAP serve um propósito que ultrapassa a técnica: para que um responsável institucional possa exercer julgamento informado sobre a situação de um aluno, precisa de compreender as razões da classificação de risco atribuída pelo modelo. Um sistema que classifica sem explicar impede a verificação humana e reduz o aluno a um resultado numérico, o que é incompatível com os princípios de transparência e responsabilização que devem orientar a utilização de IA (*Machine Learning*) em educação. A transparência das previsões satisfaz, assim, uma exigência simultaneamente técnica e ética.

3.5 Ferramentas e Ambiente Tecnológico

A implementação do *framework* recorreu a um conjunto de ferramentas selecionadas com base em três critérios: adequação técnica, reprodutibilidade dos processos e alinhamento com o perfil dos utilizadores finais do artefacto de *Business*

Intelligence. As ferramentas adotadas são igualmente congruentes com o perfil técnico do Mestrado em *Business Intelligence and Analytics* do ISCAP, no âmbito do qual foram desenvolvidas as competências necessárias à sua aplicação, garantindo o domínio técnico adequado à condução rigorosa da investigação.

Todo o processo analítico, desde a preparação dos dados até ao *deployment* dos modelos, foi desenvolvido em Python 3.11, linguagem de programação *open source* que se tornou o padrão de facto em projetos de *Data Science* e *Machine Learning* aplicado. A escolha do Python justifica-se, para além da sua adoção generalizada, pela disponibilidade de bibliotecas especializadas que cobrem integralmente as necessidades técnicas desta investigação sem recurso a ferramentas externas: o ecossistema científico do Python oferece soluções nativas para manipulação de dados tabulares, algoritmos de *Machine Learning*, balanceamento de classes, explicabilidade algorítmica e análise de sobrevivência. O desenvolvimento foi conduzido em ambiente Jupyter Notebook integrado no Visual Studio Code, numa configuração que favorece a iteratividade característica do processo CRISP-DM, permitindo documentar, executar e rever cada fase analítica de forma incremental e rastreável.

O sistema de *Business Intelligence* foi desenvolvido recorrendo ao Microsoft Power BI Desktop. A plataforma suporta a integração direta de outputs de modelos externos, permite a construção de modelos multidimensionais complexos e disponibiliza uma linguagem de expressões (DAX) adequada às medidas analíticas necessárias. A escolha é reforçada pelo facto de a Profitecla já utilizar internamente o Power BI, o que elimina barreiras de adoção tecnológica e permite integrar o artefacto nos fluxos de trabalho existentes sem necessidade de infraestrutura adicional. A Tabela 4 apresenta o inventário completo das ferramentas e bibliotecas utilizadas.

Tabela 4 — Ferramentas e Bibliotecas utilizadas

Ferramenta / Biblioteca	Versão	Finalidade
Python	3.11.15	Linguagem de desenvolvimento
pandas	2.3.3	Manipulação e transformação de dados tabulares
numpy	2.4.4	Computação numérica e operações matriciais
scikit-learn	1.8.0	Pré-processamento, <i>clustering</i> , validação cruzada e métricas de avaliação

XGBoost	3.2.0	Modelação preditiva supervisionada (M1a e M1b)
imbalanced-learn	0.14.1	Balanceamento de classes (SMOTE-ENN)
statsmodels	0.14.6	Testes de significância estatística (McNemar)
SHAP	0.51.0	Explicabilidade algorítmica
lifelines	0.30.3	Análise de sobrevivência (Kaplan-Meier)
matplotlib	3.10.8	Produção de figuras analíticas e visualizações de desempenho dos modelos
seaborn	0.13.2	Visualizações estatísticas exploratórias e análise de distribuições
Jupyter Notebook + VS Code	—	Ambiente de desenvolvimento iterativo
Microsoft Power BI Desktop	—	<i>Dashboards</i> e artefacto de apoio à decisão

4 Implementação

Este capítulo executa as fases de *Data Understanding* a *Deployment* do CRISP-DM, cujos fundamentos e mapeamento ao projeto foram estabelecidos na Secção 3.3. A fase de *Business Understanding*, já documentada nessa secção, não é aqui repetida. A apresentação segue a ordem do processo: descrição e preparação dos dados (4.1 e 4.2), aprendizagem não supervisionada e supervisionada, incluindo a interpretação dos modelos com Inteligência Artificial Explicável (4.3, 4.4 e 4.5), análise de sobrevivência como instrumento de avaliação (4.6) e operacionalização num sistema de apoio à decisão (4.7).

4.1 Breve Descrição dos Dados

Os dados utilizados nesta investigação foram disponibilizados pela Profitecla e dizem respeito exclusivamente ao polo do Porto, cobrindo os anos letivos de 2018/19 a 2023/24. O ficheiro original em formato EXCEL contém informação pseudonimizada, organizada em três tabelas independentes, cada uma com uma finalidade distinta.

A **Fich1** descreve o perfil geral do aluno, incluindo dados demográficos, socioeconómicos e informação sobre a escola de origem. A **Fich2** regista o desempenho académico anual de cada aluno: assiduidade, módulos em atraso e situação curricular. A **Fich3** contém dados de empregabilidade recolhidos após a conclusão do curso, disponíveis apenas para os alunos que concluíram e que foram contactados. Esta tabela foi incluída no ficheiro original, mas não foi utilizada em nenhuma componente de modelação desta investigação, dado que a sua cobertura parcial a tornaria incompatível com os objetivos preditivos definidos; fica identificada como fonte potencial para investigação futura sobre trajetórias pós-formação.

As tabelas Fich1 e Fich2 partilham a mesma granularidade, uma linha por aluno por ano letivo. O conjunto totaliza 2 922 registos, correspondentes a 1 086 alunos únicos, que permite a definição da variável-alvo, (Secção 4.2.1). Esta organização reproduz a taxonomia de variáveis preditoras consolidada na literatura de *Educational Data Mining*, que distingue dados demográficos e socioeconómicos, histórico académico anterior e desempenho durante o percurso (Alturki et al., 2022; Costa-Mendes et al., 2022).

O dicionário completo das variáveis das duas tabelas utilizadas é apresentado nas Tabelas 5 e 6.

Tabela 5 — Dicionário de variáveis: Fich1 (Perfil do aluno)

Variável	Tipo	Descrição	Grupo
ID	Categórica	Identificador único do aluno (pseudonimizado)	Identificador
Curso	Categórica	Código do curso	Académico
Curso_Designação	Categórica	Designação completa do curso	Académico
Ano Letivo	Categórica	Ano letivo do registo	Temporal
Ano Letivo Entrada	Categórica	Ano letivo de primeira matrícula	Temporal
Ano Plano Curricular	Ordinal (1–3)	Ano do plano curricular do registo	Académico
Data Nascimento	Data	Data de nascimento	Demográfico
Sexo	Binária	Género (Feminino / Masculino)	Demográfico
Nacionalidade	Categórica	Nacionalidade do aluno	Demográfico
Imigrante	Binária	Indica se o aluno é imigrante	Socioeconómico
Escola Origem	Categórica	Nome da escola de origem (716 designações distintas, não normalizadas)	Origem
Escola Origem Tipo Ensino	Categórica	Tipo de estabelecimento de origem (Regular, Profissional)	Origem
Ano Conclusão 9.º Ano	Ordinal	Ano de conclusão do 9.º ano	Histórico académico
Escola Anterior Ano Frequentado	Ordinal	Último ano frequentado na escola anterior	Histórico académico
Escalão Abono	Ordinal	Escalão de abono de família (valores altos = menor rendimento)	Socioeconómico
Aluno Situação	Categórica	Estado do aluno no final do percurso	Outcome

Tabela 6 — Dicionário de variáveis: Fich2 (Desempenho académico anual)

Variável	Tipo	Descrição	Grupo
ID	Categórica	Identificador único do aluno	Identificador
Curso / Curso_Designação	Categórica	Curso do registo	Académico
Ano Letivo / Entrada	Categórica	Anos letivos do registo e de entrada	Temporal
Ano Plano Curricular	Ordinal (1–3)	Ano do plano curricular	Académico
Módulos Atraso	Numérica	Módulos curriculares em atraso no final do ano	Desempenho
Média Final	Numérica	Nota final (apenas para quem concluiu)	Desempenho

Aulas Ano	Numérica	Carga horária programada (constante por curso/ano)	Excluída
Falta Ano	Numérica	Total de faltas em valor absoluto	Excluída
%Faltas Ano	Numérica	Taxa de absentismo anual (faltas / aulas programadas)	Desempenho
Aluno Situação	Categórica	Estado do aluno no ano letivo	Outcome

4.1.1 Problemas de Qualidade dos Dados

A qualidade dos dados é uma condição prévia para a fiabilidade de qualquer modelo de aprendizagem automática. Nesta fase foram identificados quatro tipos de problemas, com natureza e impacto distintos.

Integração de tabelas com colunas partilhadas. As tabelas Fich1 e Fich2 foram disponibilizadas de forma independente, sem chave de junção pré-definida. Ambas contêm colunas com o mesmo nome (Curso, Curso_Designação, Ano Letivo, Ano Letivo Entrada, Ano Plano Curricular e Aluno Situação), o que obrigou a definir explicitamente qual das tabelas seria a fonte autoritativa de cada coluna e qual a chave primária que permitiria cruzá-las sem perda nem duplicação de registos.

Inconsistências em variáveis invariantes. A variável Ano de Conclusão do 9.º Ano é um facto biográfico fixo que não deveria variar entre registos do mesmo aluno. Foram detetados casos com valores contraditórios entre anos letivos, indicando erros de introdução de dados na fonte.

Variáveis com semântica enganosa. A variável Aulas Ano não regista a presença individual do aluno, mas a carga horária total programada para o curso, apresentando variância nula dentro do mesmo curso e ano letivo. A variável Falta Ano, em valor absoluto, não é comparável entre cursos com cargas horárias distintas. Ambas foram identificadas como inadequadas para modelação (Secção 4.2.2).

Nomes de escolas de origem não normalizados. O campo Escola Origem apresentava 716 designações distintas, com variações ortográficas e abreviaturas heterogéneas. Este problema de alta cardinalidade é reconhecido na literatura como um obstáculo relevante em dados educativos reais. Fernández-García et al. (2021), num estudo de previsão de abandono escolar, identificaram o mesmo problema na variável *Origin Educational Institution*, 137 valores distintos com 45,4% de registos omissos, concluíram que a fragmentação excessiva das classes reduz o poder preditivo e introduz esparsidade no espaço de variáveis. O impacto na presente investigação é, contudo,

assimétrico, os modelos preditivos utilizam a variável Escola Origem Tipo Ensino, o tipo de estabelecimento, de baixa cardinalidade, pelo que a validade da modelação não é comprometida. A consequência manifesta-se apenas no eixo de captação (Secção 4.7.4), onde é necessária granularidade ao nível da escola individual, e está documentada como limitação.

4.1.2 Estratégias de Correção e Validação Aplicadas

Cada problema identificado foi tratado com base num critério explícito, privilegiando a transparência sobre a correção especulativa.

Integração das tabelas. Adotou-se o par composto [ID, Ano Letivo] como chave primária, que identifica de forma única cada observação anual. A Fich1 foi estabelecida como tabela-base, em todas as colunas partilhadas, os valores provêm da Fich1, eliminando ambiguidade. A correção foi validada pela verificação de unicidade após a junção, que não retornou registos duplicados.

Inconsistências no Ano de Conclusão do 9.º Ano. Aplicou-se imputação pela moda por aluno, cada ID recebeu o valor mais frequente entre os seus registos. Esta abordagem é consistente com práticas adotadas na literatura de predição de retenção estudantil para variáveis categóricas (Matz et al., 2023) e parte do pressuposto de que o valor modal reflete o facto biográfico correto e de que o erro é de natureza aleatória e pontual, condição não verificável sem fonte externa, mas metodologicamente mais sólida do que eliminar os registos afetados.

Alunos com múltiplos outcomes. Sete alunos registaram mudança de curso durante o percurso, apresentando mais do que um estado final. Uma vez que o objetivo dos modelos é prever a probabilidade de conclusão do percurso formativo, e não a conclusão de um curso específico, atribuiu-se $\text{target} = 1$ a qualquer aluno que tenha concluído pelo menos um dos cursos frequentados.

Nomes de estabelecimentos de ensino não normalizados. O campo relativo à escola de origem apresentou 716 designações distintas para um universo estimado substancialmente inferior, evidenciando a ausência de controlo de qualidade no momento de recolha. A heterogeneidade observada resultou de múltiplos fatores cumulativos: variações ortográficas e de acentuação, uso inconsistente de abreviaturas, e referência ora pelo nome da escola individual ora pela unidade orgânica de agrupamento. Situação análoga é descrita por Fernández-García et al. (2021), que, perante a inviabilidade de desambiguação sistemática, optaram pela exclusão total da variável.

Nesta investigação, identificado o problema numa fase posterior à preparação de dados, foi desenvolvida uma estratégia de normalização em duas etapas, normalização determinística por transformações sem ambiguidade, seguida de correspondência aproximada de *strings* validada de acordo com o registo oficial de estabelecimentos da DGEEC. O processo permitiu reduzir as 716 designações originais para 495 entidades únicas, com 80,7% dos registos enriquecidos com informação de agrupamento, concelho e natureza institucional. Ainda assim, a implementação completa ficou condicionada à validação junto da Profitecla, nomeadamente à adoção de um campo de entrada pré-definido no sistema de registo, ancorado no mesmo registo, de forma a garantir consistência em ciclos de dados futuros.

Concluída esta fase, foram consolidados dois conjuntos de dados, um agregado ao nível do aluno, destinado à aprendizagem supervisionada e não supervisionada, e um longitudinal, com estrutura anual, necessário para a análise de sobrevivência (Secção 4.6).

4.2 Preparação dos Dados

4.2.1 Construção da Variável-Alvo

A definição da variável-alvo é uma decisão teórica antes de ser técnica pois exige clarificar o que se entende por sucesso formativo no contexto da Profitecla. Nesta investigação, o sucesso é operacionalizado como a conclusão do percurso formativo, correspondendo ao estado “Concluiu” no campo Aluno Situação.

Como os dados têm granularidade anual, o processo começou pela identificação, para cada aluno, do registo com o valor máximo de Ano Plano Curricular, o ponto mais avançado do percurso com informação disponível. A partir desse registo, foi atribuída à variável binária *target_concluiu*, o valor 1 corresponde a “Concluiu” e o valor 0 corresponde a qualquer estado de saída não conclusivo (Abandonou, Anulou Matrícula, Excluído por Faltas, entre outros).

Dois grupos foram excluídos do universo de aprendizagem. Os alunos com estado “Transitou” encontravam-se ainda a frequentar o curso, com o outcome indeterminado, pelo que a sua inclusão introduziria censura não tratada na variável-alvo. Os alunos com o estado “CEF — Certificado Escolar” correspondem a percursos de dupla certificação (Cursos de Educação e Formação) com estrutura curricular, duração e critérios de conclusão distintos dos cursos de nível 3 do Quadro Europeu de Qualificações; o outcome associado não é comparável à conclusão dos cursos modelados, pelo que a sua inclusão enviesaria a definição do alvo.

O conjunto resultante contém 1 086 alunos, 814 concluíram (75,0%) e 272 não concluíram (25,0%), correspondendo a um rácio de desequilíbrio de 2,99:1. Este nível de desbalanceamento não é negligenciável. Modelos treinados sem tratamento tendem a maximizar a taxa de exatidão global prevendo sistematicamente a classe maioritária, suprimindo a deteção da classe minoritária, precisamente os alunos em risco. As estratégias de mitigação são descritas na Secção 4.4.3.

4.2.2 Seleção e Tratamento de Variáveis

A seleção seguiu o critério de que apenas variáveis com informação discriminante ao nível do aluno individual deveriam integrar os conjuntos de *features*. A variável Aulas Ano, sendo uma constante do programa (variância nula dentro do mesmo curso e ano), e a variável Falta Ano em valor absoluto, não comparável entre cursos com durações distintas, foram excluídas (Fernández-García et al., 2021).

A variável Média Final exigiu tratamento especial, já que está disponível exclusivamente para alunos que concluíram. A sua utilização direta no Modelo Retrospectivo (M1a) introduziria *data leakage*, o modelo aprenderia a prever conclusão a partir de uma variável que só existe para quem concluiu. A solução foi criar uma variável binária auxiliar *tem_nota_final* (1 = nota disponível; 0 = ausente), que preserva o sinal informativo da presença da nota sem expor o seu valor.

As variáveis %Faltas Ano e Módulos Atraso foram mantidas, a primeira normaliza as faltas pela carga horária, tornando-as comparáveis entre cursos; a segunda reflete diretamente o ritmo de progressão curricular, um dos preditores mais consistentes de abandono escolar (Fernández-García et al., 2021; Costa-Mendes et al., 2022).

Uma decisão transversal foi a distinção entre os conjuntos de *features* do M1a e do M1b. O **Modelo Retrospectivo (M1a)** tem acesso ao percurso completo, inclui desempenho acumulado. O **Modelo Prospetivo (M1b)** está restrito a variáveis disponíveis no momento da matrícula, simulando uma previsão feita antes do início do curso. O diferencial de desempenho entre ambos é interpretado não como fraqueza do M1b, mas como a quantificação do valor informativo do comportamento académico ainda indisponível à matrícula (Secção 4.4.4).

4.2.3 Criação de Novas *Features*

A mesma taxa de absentismo tem significados distintos consoante o curso, 15% de faltas num curso onde a média é 5% representa uma posição muito diferente de 15% num curso onde a média é 20%. Para contextualizar o desempenho individual face à

norma do programa, foram criadas *features* de desvio relativo ao curso: *desvio_faltas_curso*: diferença entre a taxa de absentismo do aluno e a média do seu curso no mesmo ano; *desvio_modulos_curso*: diferença entre os módulos em atraso do aluno e a média do seu curso; *ranking_faltas_curso*: posição percentual do aluno na distribuição de absentismo do seu curso.

Esta abordagem de contextualização intra-grupo é consistente com práticas de *feature engineering* em *Educational Data Mining* que reconhecem a heterogeneidade entre programas como fonte de enviesamento se não for explicitamente controlada (Alvarez-Garcia et al., 2024).

4.2.4 Tratamento de Valores Omissos, *Outliers* e Codificação

O único campo com omissão estrutural relevante foi Média Final, tratado pela *flag* binária descrita na Secção 4.2.2. As variáveis numéricas com omissões residuais foram imputadas pela mediana.

Relativamente aos *outliers*, não foi aplicada nenhuma remoção. Em previsão de abandono escolar, os valores extremos, um aluno com 80% de absentismo ou dez módulos em atraso, não são erros de medição, mas os casos de maior risco que o modelo tem de aprender a identificar. A sua remoção eliminaria informação relevante para a deteção precoce (Matz et al., 2023).

A codificação foi adaptada à finalidade de cada módulo. Para os modelos preditivos (M1a e M1b), as variáveis categóricas foram codificadas com *OrdinalEncoder*, adequado para algoritmos de árvore como *Random Forest* e *XGBoost*, que não pressupõem distâncias métricas entre categorias: os valores inteiros atribuídos a cada categoria servem apenas como pontos de corte (identificadores de *split*) nas regras de decisão da árvore, sem que a sua ordem ou magnitude seja interpretada como uma relação métrica real entre categorias (Chen & Guestrin, 2016). Para o *clustering*, adotou-se *Label Encoding*, com 2 a 3 categorias por variável; a expansão dimensional do *one-hot encoding*, isto é, a criação de uma coluna binária adicional por categoria, não acrescenta informação discriminante ao *K-Means* e introduziria esparsidade desnecessária. A variável Escalão Abono foi tratada como variável ordinal numérica em ambos os contextos, preservando a ordenação natural dos escalões (valores crescentes correspondem a maior carência económica).

4.2.5 Divisão Treino/Teste e Validação Cruzada

O conjunto de 1 086 alunos foi dividido em 80% para treino (868 alunos) e 20% para teste (218 alunos), com estratificação pela variável-alvo, preservando em cada subconjunto a mesma proporção de concluintes e não concluintes do conjunto total, o que evita que uma divisão aleatória simples produza, por acaso, um conjunto de teste com uma taxa de abandono distinta da real e assim enviesasse a avaliação do modelo (Kohavi, 1995), garantindo que o rácio de 2,99:1 se mantém em ambos os conjuntos. Os dois modelos partilham exatamente o mesmo *split*, assegurando que as mesmas pessoas estão no mesmo conjunto para ambos.

A divisão foi aleatória e não temporal. A separação por ano letivo seria metodologicamente mais rigorosa para avaliar a capacidade do modelo em prever resultados de grupos de alunos futuros, mas foi preterida por duas razões, com cerca de 180 alunos com *outcome* por ano, reduziria o treino a um *n* insuficiente para a estabilidade dos modelos; e o objetivo é demonstrar a capacidade preditiva das *features*, não simular a capacidade do modelo em prever resultados de grupos de alunos futuros, tarefa abordada separadamente no *deployment* sobre os alunos ativos (Secção 4.4.4). Esta limitação é reconhecida no Capítulo VI.

Para a seleção de hiperparâmetros e comparação entre algoritmos, foi aplicada validação cruzada estratificada de 5 *folds* exclusivamente sobre os 80% de treino. Este conjunto é sucessivamente particionado em 5 subconjuntos, treinando o modelo em 4 e validando no restante, repetindo o processo 5 vezes até que cada instância tenha sido usada uma vez para validação (Kohavi, 1995). O procedimento é distinto e complementar à divisão treino/teste inicial. A validação cruzada orienta a seleção de hiperparâmetros e a comparação entre algoritmos usando apenas dados de treino, enquanto os 20% de teste, nunca vistos durante este processo, ficam reservados para a avaliação final e não enviesada do modelo já selecionado. A escolha de $k = 5$ equilibra o viés e a variância da estimativa com o custo computacional (Refaeilzadeh et al., 2009), mantendo o rácio de desequilíbrio em cada partição.

4.3 Clustering de Perfis de Alunos

A análise de *clustering* identifica grupos naturais de alunos com base no seu comportamento e perfil de entrada, sem recorrer à variável-alvo. Esta abordagem de aprendizagem não supervisionada é complementar à aprendizagem supervisionada / análise preditiva, enquanto o M1a e o M1b estimam a probabilidade individual de

conclusão, o *clustering* produz perfis coletivos que caracterizam tipologias de risco e orientam intervenções diferenciadas. Os *clusters* não foram construídos a partir do *outcome*, a taxa de conclusão de cada grupo foi calculada posteriormente, como validação da sua relevância operacional.

4.3.1 Pipeline de Pré-Processamento e Seleção do Algoritmo

Foram selecionadas 14 variáveis (11 numéricas e 3 categóricas), combinando comportamento durante o percurso e perfil à entrada, com exclusão explícita de variáveis de *outcome* (média final, flag de conclusão, ano máximo do plano curricular e a própria variável-alvo). As categóricas foram codificadas com *Label Encoding* e as 14 variáveis foram standardizadas com *StandardScaler* antes do *clustering*, condição recomendada para algoritmos baseados em distância euclidiana como o K-Means, de forma a evitar que variáveis de escala maior dominem o cálculo da distância e distorçam a formação dos grupos (Milligan & Cooper, 1988). A Análise de Componentes Principais (PCA) foi aplicada apenas para visualização bidimensional, não influenciando a formação dos grupos, realizada no espaço completo das 14 *features* (Pecuchova & Drlik, 2024).

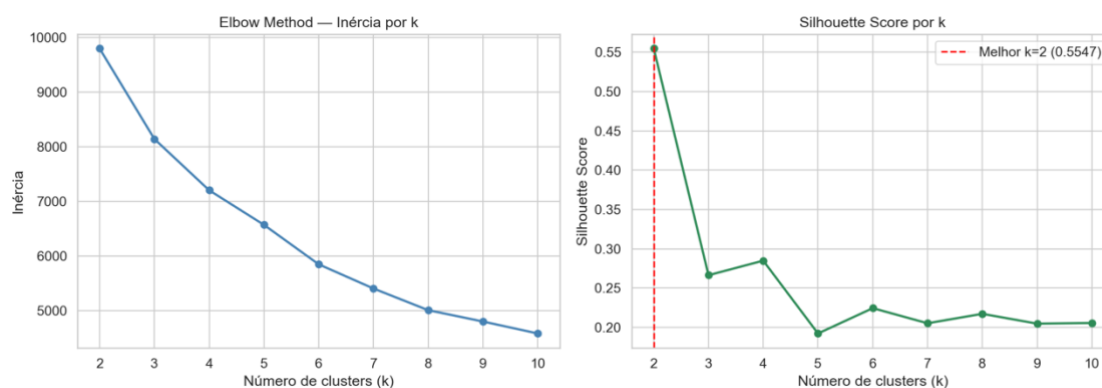
Tabela 7 — Comparação de algoritmos de clustering ($k = 3$)

Método	Silhouette ↑	Davies-Bouldin ↓	Calinski-Harabasz ↑
K-Means	0,2664	1,4214	469,8
Hierarchical (Ward)	0,2127	1,5383	384,9
GMM	0,1367	2,5894	182,9

4.3.2 Determinação do Número de Clusters

A determinação de k recorreu ao método do cotovelo (variação da inércia intra-*cluster*) e à análise do coeficiente de Silhouette para k entre 2 e 10, conforme apresentado na Figura 3. A solução $k = 2$ foi rejeitada por trivialidade, pois reproduz, essencialmente, a dicotomia da variável-alvo (taxas de 82,6% e 0%), não acrescentando informação além da classificação supervisionada. A solução $k = 3$ foi adotada como principal, pela combinação de cotovelo visível, Silhouette mais elevado entre as soluções com $k \geq 3$ e interpretabilidade operacional dos grupos. A solução $k = 4$ é discutida como análise complementar na Secção 4.3.4.

Figura 3 — Elbow Method e Silhouette Score



4.3.3 Perfis Identificados

A aplicação do K-Means com $k = 3$ produziu três *clusters* com características e taxas de conclusão distintas (Tabela 8 e Figura 4).

Figura 4 — Síntese visual dos três perfis de alunos identificados por K-Means ($k = 3$)

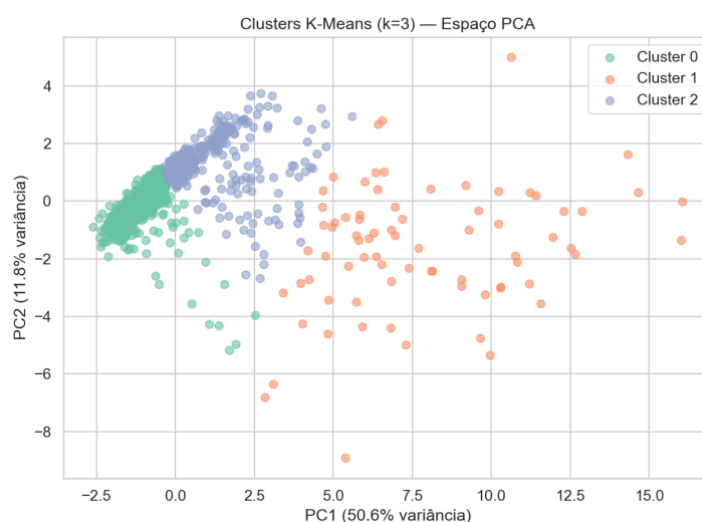


Tabela 8 — Caracterização dos clusters ($k = 3$)

Cluster	Designação	n	% total	Taxa conclusão	%Faltas (méd.)	Mód. atraso (méd.)
0	Sucesso	737	67,9%	86,0%	0,03	0,36
1	Abandono	76	7,0%	0,0%	0,38	24,72
2	Risco Intervencionável	273	25,1%	65,9%	0,18	2,19

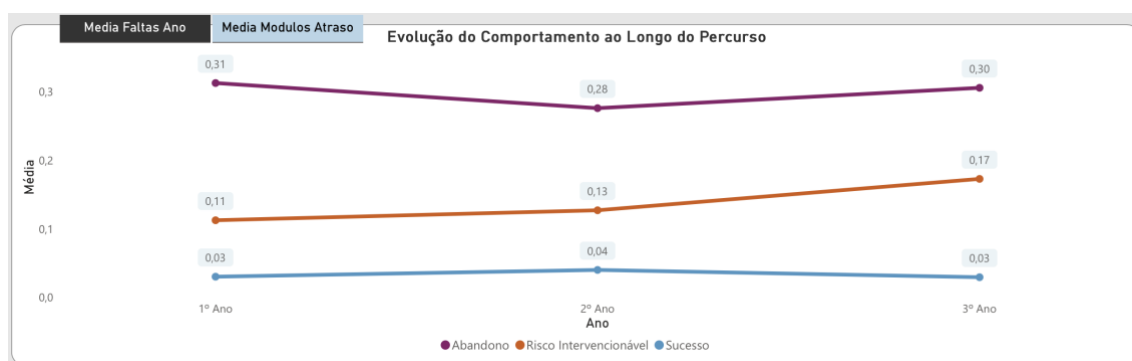
Cluster 0 — Sucesso (n = 737, 86,0%): Concentra a maioria dos alunos; caracteriza-se por baixa taxa de absentismo (0,03 em média), número reduzido de

módulos em atraso (0,36) e desvios negativos face à média do curso. A percentagem residual de insucesso (14%) merece atenção preventiva.

Cluster 1 — Abandono (n = 76, 0,0%): Apesar de representar apenas 7% da amostra, nenhum dos 76 alunos concluiu. O perfil é marcado por absentismo elevado (0,38) e acumulação acentuada de módulos em atraso (24,72 em média), correspondendo a situações de rutura já instalada. A concentração do abandono no primeiro ano (Secção 4.6) reforça a necessidade de alerta precoce.

Cluster 2 — Risco Intervencionável (n = 273, 65,9%): Grupo de maior relevância estratégica. Com absentismo moderado (0,18) e atrasos controláveis (2,19), apresenta uma taxa de conclusão acima do *Cluster 1*, mas abaixo da média global. A designação reflete o pressuposto de que a intervenção atempada, através de medidas como o acompanhamento tutorial, o apoio socioeconómico e as aulas de recuperação, pode alterar a trajetória académica dos estudantes, evidenciando uma probabilidade razoável de sucesso.

Figura 5 — Evolução das médias de absentismo ao longo dos três anos curriculares, por cluster



4.3.4 Refinamento dos Perfis de Risco

A solução $k = 4$ foi analisada para verificar se o *Cluster 2* encobre sub-perfis com necessidades de intervenção distintas. Dois dos quatro clusters resultantes sobrepõem-se maioritariamente a este perfil, distribuindo-se por dois grupos diferenciáveis, um sub-perfil ($n \approx 262$, 65,6%) dominado por absentismo moderado, indicando intervenção centrada na assiduidade; e um sub-perfil emergente ($n \approx 42$, 45,2%) com menor absentismo, mas mais módulos em atraso e escalão de abono mais elevado, indicando intervenção combinada de apoio académico e social.

Esta distinção tem valor operacional, permite diferenciar o tipo de resposta para grupos que, em $k = 3$, receberiam o mesmo protocolo. Não foi adotada como solução

principal por duas razões, o sub-perfil de $n \approx 42$ é estatisticamente frágil para uma instituição de polo único, com centróides potencialmente instáveis entre *cohorts*; e a distinção, embora real, é subtil e exige validação com os diretores de curso antes de ser formalizada. A solução $k = 4$ torna-se robusta com a expansão a outros polos, onde a dimensão amostral aumentaria a estabilidade do grupo.

4.4 Modelos de Previsão de Conclusão

O desenvolvimento dos modelos seguiu uma lógica de instrumentação temporal dupla, dois modelos treinados sobre conjuntos de *features* diferentes, que respondem a duas questões operacionais distintas. O Modelo Retrospectivo (M1a) opera sobre o percurso completo e serve de instrumento de caracterização histórica e de validação da relevância das variáveis comportamentais. O Modelo Prospetivo (M1b) opera apenas sobre informação disponível aquando do processo de matrícula e constitui o instrumento de intervenção precoce, sendo o modelo que alimenta o semáforo de risco. O diferencial entre ambos não é uma fraqueza, mas a quantificação do valor informativo dos dados comportamentais (Secção 4.4.4).

4.4.1 Seleção dos Algoritmos e Configuração do *Pipeline*

Para cada modelo foram comparados três algoritmos, Regressão Logística (Cox, 1958), *Random Forest* (Breiman, 2001) e *XGBoost* (Chen & Guestrin, 2016), estes dois últimos métodos de *ensemble* baseados em árvores, amplamente validados na literatura de previsão de abandono escolar (Mduma, 2023; Fernández-García et al., 2021; Torres-Cruz et al., 2025). O *pipeline* aplica imputação pela mediana às variáveis numéricas, substituindo valores omissos pelo valor central da distribuição, menos sensível a outliers do que a média e, às categóricas, imputação pela moda (substituição pelo valor mais frequente) seguida de codificação *OrdinalEncoder*, que converte cada categoria num valor inteiro distinto sem impor uma relação de ordem interpretável entre elas. O desequilíbrio de classes foi inicialmente compensado por `scale_pos_weight` (*XGBoost*) e `class_weight='balanced'` (restantes); uma estratégia de reamostragem mais robusta foi adotada para o M1b (Secção 4.4.3).

A avaliação comparativa dos algoritmos usou métricas distintas consoante o propósito de cada modelo, para o M1a, o AUC-ROC, por captar a capacidade discriminativa global, adequada à caracterização histórica de uma *cohort* já concluída; para o M1b, o F1 da Classe 0 (não conclusão), por priorizar a deteção de alunos em risco,

num contexto em que o custo de não sinalizar um aluno que virá a não concluir (falso negativo) supera o de sinalizar indevidamente um aluno que concluiria (falso positivo) (Torres-Cruz et al., 2025; Mduma, 2023).

Em M1a, o *XGBoost* obteve o valor médio de AUC-ROC mais elevado nos 5 *fold*s da validação cruzada ($0,9812 \pm 0,0110$), seguido do *Random Forest* ($0,9783 \pm 0,0070$) e da Regressão Logística ($0,9456 \pm 0,0155$). Em M1b, o F1 da Classe 0 apresentou uma ordenação distinta, com a Regressão Logística à frente ($0,5193 \pm 0,0387$), seguida do *Random Forest* ($0,5032 \pm 0,0695$) e do *XGBoost* ($0,4935 \pm 0,0644$). Em ambos os modelos, a diferença entre a média mais alta e a mais baixa é inferior ao desvio-padrão dos algoritmos envolvidos, indicando sobreposição considerável nas distribuições de desempenho entre *fold*s. A própria inversão da ordenação entre os dois modelos reforça esta leitura: nenhum algoritmo mantém vantagem consistente, o que sugere, numa primeira análise descritiva, equivalência entre os três.

Para verificar formalmente esta equivalência, aplicou-se o Teste de McNemar (McNemar, 1947) a cada par de algoritmos. Em todos os pares e em ambos os modelos, os valores-p foram superiores a 0,05, não permitindo rejeitar a hipótese de que os algoritmos têm desempenho equivalente, resultado consistente com a sobreposição já observada nos desvios-padrão da validação cruzada. Optou-se pelo Teste de McNemar, em detrimento de um teste t emparelhado sobre os *scores* dos 5 *fold*s, por este último assumir independência entre as observações, pressuposto violado quando os *fold*s partilham parcialmente os mesmos dados de treino, o que infla artificialmente a taxa de erro Tipo I e pode identificar diferenças estatisticamente significativas inexistentes (Dietterich, 1998). O Teste de McNemar, aplicado sobre as previsões do conjunto de teste independente, opera sob pressupostos compatíveis com este cenário e mantém um controlo adequado do erro Tipo I (Dietterich, 1998; Rainio et al., 2024). Deste modo, o desvio-padrão da validação cruzada funciona como evidência descritiva da estabilidade de cada algoritmo entre partições, enquanto o McNemar fornece a evidência inferencial sobre a significância das diferenças no conjunto de teste.

Estabelecida a equivalência estatística entre os três algoritmos, a seleção do *XGBoost* para ambos o modelo assentou num critério de consistência metodológica, a utilização do mesmo algoritmo em M1a e M1b permite a comparação direta dos valores SHAP entre os dois modelos, sem confundir diferenças de importância das variáveis com diferenças de algoritmo (Secção 4.5). Na ausência de vantagem estatisticamente

significativa de qualquer alternativa, esta comparabilidade constitui um critério de desempate legítimo e alinhado com o objetivo de explicabilidade do estudo.

4.4.2 M1a — Modelo Retrospectivo

O M1a utiliza 17 *features* (12 numéricas e 5 categóricas), combinando perfil de entrada e comportamento acumulado (taxas de absentismo médias e máximas, módulos em atraso, desvios relativos ao curso e *ranking* de absentismo intra-curso). O XGBoost foi selecionado pelo critério de AUC-ROC. O desempenho no conjunto de teste consta da Tabela 9; as Figuras 6 e 7 apresentam a comparação de métricas e a matriz de confusão.

Os modelos M1a e M1b foram comparados para quantificar o ganho preditivo associado à incorporação de variáveis comportamentais acumuladas ao longo do percurso escolar. Dado que o M1a utiliza informação retrospectiva indisponível no momento da matrícula, a comparação não visa identificar o melhor modelo em termos absolutos, mas sim avaliar o contributo incremental dessas variáveis para a capacidade preditiva.

Tabela 9 — Métricas de desempenho no conjunto de teste (M1a e M1b)

Métrica	M1a — Retrospectivo	M1b — Prospetivo
AUC-ROC	0,9867	0,7347
Accuracy	0,9679	0,7110
F1 Classe 0 (não conclusão)	0,9333	0,4878
F1 Classe 1 (conclusão)	0,9789	0,7987
Recall Classe 0	0,8909	0,5455
Precision Classe 0	0,9800	0,4412
Brier Score	0,0289	0,1888

Figura 6 — Comparação das principais métricas entre o Modelo Retrospectivo e o Modelo Prospetivo

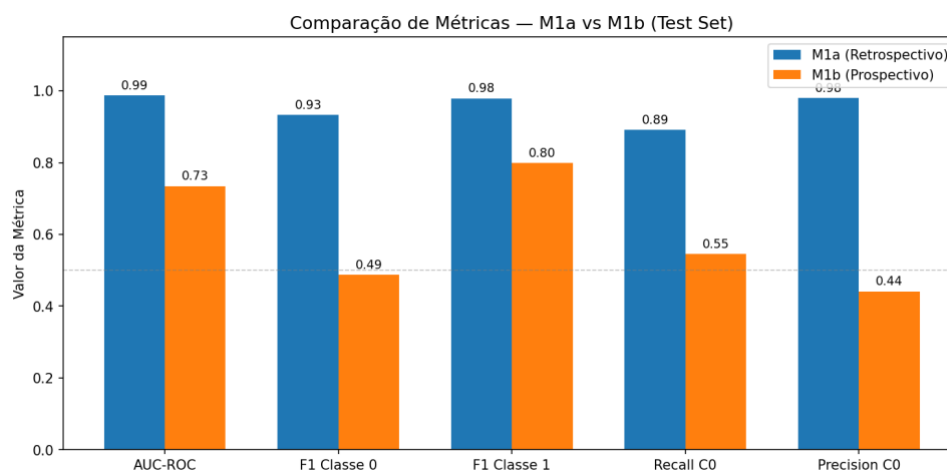
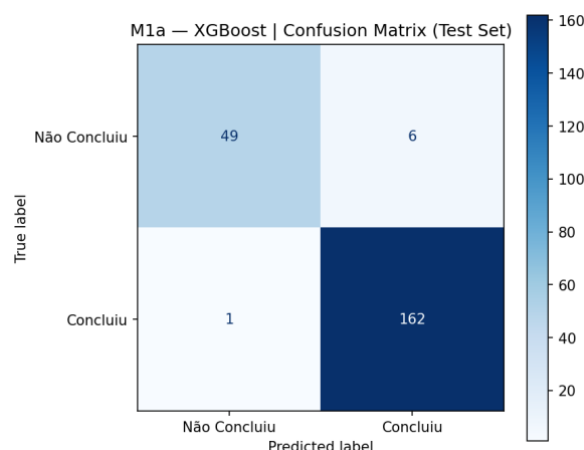


Figura 7 — Matriz de confusão do Modelo Retrospectivo (M1a) no conjunto de teste



O AUC-ROC do M1a (0,987) aproxima-se do limite teórico, refletindo a elevada capacidade discriminativa quando o percurso comportamental completo está disponível, em linha com estudos análogos em ensino profissional (Fernández-García et al., 2021). A sua limitação é simétrica à vantagem, só é aplicável a alunos com percurso já registado, não podendo ser usada em tempo real para novos alunos.

4.4.3 M1b — Modelo Prospetivo

O alvo do M1b é a Classe 0 (não conclusão), que constitui a minoria dos casos. Sem sinais comportamentais, o classificador tende a ajustar-se à classe maioritária (conclusão) e a apresentar baixa sensibilidade sobre os alunos em risco, precisamente os que o modelo existe para sinalizar. Por esse motivo, as estratégias de reamostragem foram avaliadas pelo F1 da Classe 0 em validação cruzada estratificada, métrica que sintetiza a *precision* e o *recall* dessa classe e traduz de forma direta a utilidade do modelo para a deteção precoce. A reamostragem foi aplicada apenas dentro das partições de treino de cada *fold*, através de um pipeline, para não contaminar as partições de avaliação. O M1b utiliza 8 *features* (3 numéricas e 5 categóricas), informações exclusivamente disponíveis no momento da matrícula, idade de entrada, ano frequentado na escola anterior, escalão de abono, curso, sexo, nacionalidade, estatuto de imigrante e tipo de escola de origem.

Tabela 10 — Comparação de estratégias de reamostragem (M1b, validação cruzada)

Condição	F1 Classe 0 (CV)
Sem reamostragem	0,4116
SMOTE	0,4171
SMOTE-ENN	0,4968

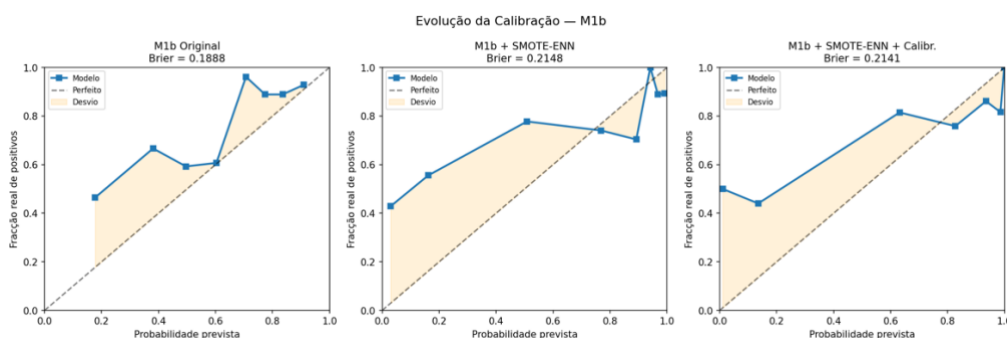
O SMOTE gera exemplos sintéticos da classe minoritária por interpolação entre vizinhos, aumentando a sua representação, mas podendo introduzir pontos em zonas de sobreposição entre classes (Chawla et al., 2002). O SMOTE-ENN acrescenta a esse passo o *Edited Nearest Neighbor*, que elimina instâncias de ambas as classes classificadas incorretamente pelos respectivos vizinhos mais próximos, limpando a fronteira de decisão (Batista et al., 2004; Mduma, 2023). No conjunto de treino, este procedimento reduziu as 868 amostras iniciais (651 de conclusão, 217 de não conclusão) para 732 (317 e 415), invertendo a proporção original a favor da classe de interesse. O SMOTE-ENN foi selecionado por maximizar o F1 da Classe 0 (0,4968, contra 0,4171 do SMOTE e 0,4116 sem reamostragem — Tabela 10).

Não foi utilizado o parâmetro `scale_pos_weight` do *XGBoost*. Este parâmetro corrige o desequilíbrio ao nível do algoritmo, penalizando mais os erros na classe minoritária, e desempenha a mesma função que a reamostragem ao nível dos dados. Aplicar ambos em simultâneo corrigiria o desequilíbrio duas vezes, com risco de sobre-sinalização. Optou-se por o resolver uma única vez, na preparação dos dados, mantendo o classificador com a sua função de custo padrão.

As probabilidades do M1b alimentam o semáforo de risco, o que suscita a questão de saber se devem ser calibradas, ou seja, ajustadas para que um valor de 0,7 corresponda a 70% de casos efetivamente positivos. A qualidade da calibração é medida pelo Brier Score, o erro quadrático médio entre a probabilidade prevista e o resultado observado, em que valores menores indicam melhor calibração. A reamostragem por SMOTE-ENN piorou o Brier Score, de 0,1888 para 0,2148, resultado esperado, porque a inversão da proporção de classes afasta as probabilidades das frequências reais. A calibração isotónica foi testada sobre esse modelo para tentar recuperar essa correspondência, mas alterou o Brier Score apenas de 0,2148 para 0,2141, variação sem significado prático, e é ajustada sobre dados já rebalanceados, cuja distribuição não reflete a população real.

Não havendo ganho, e sendo os limiares do semáforo derivados empiricamente da própria distribuição de probabilidades e não da sua leitura como frequências absolutas, a calibração não acrescenta valor à finalidade do modelo. O modelo final utiliza as probabilidades diretas do *XGBoost* com SMOTE-ENN, sem calibração (Figura 8).

Figura 8 — Curvas de fiabilidade e Brier Score nas três configurações



Definição do semáforo de risco: O semáforo de risco traduz a probabilidade de conclusão estimada pelo modelo prospetivo (M1b) numa classificação de três níveis, concebida para orientar a priorização do acompanhamento pedagógico por utilizadores sem formação estatística. A cada aluno é atribuída uma cor consoante a probabilidade prevista: vermelho para alto risco, amarelo para risco moderado e verde para trajetória estável (Tabela 11). A definição das categorias exige limiares de corte sobre a probabilidade contínua. O limiar inicial de alto risco, fixado intuitivamente em 0,40, foi revisto através de uma análise de sensibilidade que fez variar o ponto de corte entre 0,10 e 0,90 (Figura 9). O valor de 0,34 foi adotado por maximizar o F1 da Classe 0 (não conclusão) e coincidir com o ponto de equilíbrio empírico em que a Precision e o Recall dessa classe se igualam em 0,527 na curva de sensibilidade ao limiar (Figura 9), o que traduz um compromisso simétrico entre falsos positivos e falsos negativos. A fronteira de baixo risco foi fixada em 0,75, valor acima do qual a probabilidade de conclusão é suficientemente elevada para dispensar monitorização reforçada. A distribuição resultante nos 1086 alunos históricos é de 300 em vermelho (28%), 176 em amarelo (16%) e 610 em verde (56%).

Figura 9 — Evolução de F1, Recall e Precision da Classe 0 em função do threshold, e curva ROC do M1b

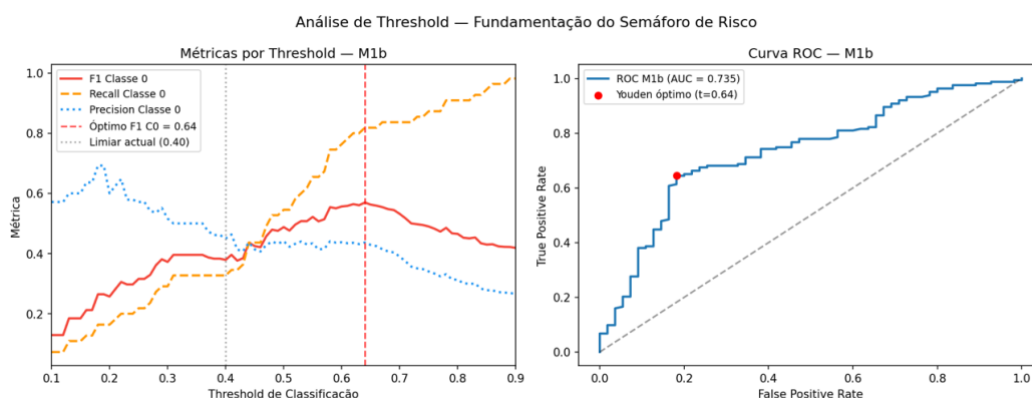


Tabela 11 — Sistema de semáforo de risco (M1b)

Nível	Critério	Leitura operacional	Resposta sugerida
Vermelho	$P < 0,34$	Alto risco — sinal débil à matrícula	Onboarding diferenciado; acompanhamento tutorial desde a 1. ^a semana
Amarelo	$0,34 \leq P < 0,75$	Risco moderado — monitorização ativa	Check-ins periódicos; suporte ao 1. ^o sinal comportamental
Verde	$P \geq 0,75$	Baixo risco — trajetória estável	Acompanhamento padrão; reavaliação por M1a

4.4.4 Análise Comparativa e *Deployment*

A aplicação dos modelos aos alunos ativos constitui a operacionalização da fase de *Deployment* do CRISP-DM (Secção 3.3). O diferencial de AUC entre M1a e M1b não é uma fraqueza do modelo prospetivo, mas a medida direta da informação que os dados comportamentais acrescentam ao perfil de entrada, o risco de abandono é em parte determinado por fatores observáveis à matrícula e em parte por comportamentos que emergem durante o curso. O M1b define quem observar com mais atenção desde o início; o M1a determina quem já está em trajeto de risco e necessita de intervenção imediata.

Os dois modelos foram aplicados aos 396 alunos ativos do ano letivo 2024/25 (estado “Transitou”). Os resultados do semáforo constam da Tabela 12; a Figura 10 apresenta a deterioração do risco ao longo do percurso.

Tabela 12 — Semáforo de risco para os 396 alunos ativos (M1b vs. M1a)

Nível	M1b — Na matrícula	M1a — Avaliação atual
Alto risco	109 (27,5%)	236 (59,6%)
Risco moderado	76 (19,2%)	105 (26,5%)
Baixo risco	211 (53,3%)	55 (13,9%)

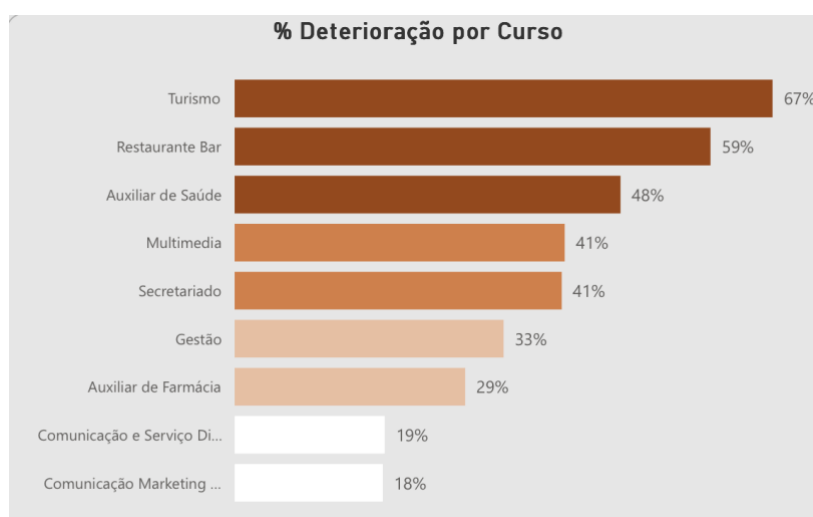
A divergência entre as distribuições não é uma inconsistência, é o sistema a funcionar como desenhado. Dos 211 alunos classificados como baixo risco à matrícula, apenas 35 mantêm essa classificação; 102 transitaram para alto risco e 74 para risco moderado. Estes alunos representam o output operacionalmente mais valioso do sistema, seriam invisíveis a qualquer triagem baseada apenas no perfil de entrada, mas o seu comportamento posterior ativou um alerta claro. A matriz de transição completa consta da Tabela 13.

Tabela 13 — Matriz de transição M1b → M1a

Matrícula (M1b) ↓ / Atual (M1a) →	Alto risco	Mod. risco	Baixo risco	Total
Baixo risco (verde)	102	74	35	211
Risco moderado (amarelo)	53	11	12	76
Alto risco (vermelho)	81	20	8	109
Total	236	105	55	396

A taxa de deterioração varia substancialmente por curso, Turismo (67%) e Restaurante e Bar (59%) apresentam as concentrações mais elevadas de alunos cujo perfil de risco piorou após a matrícula, conforme pode ser observado na Figura 10, dessa constituindo as prioridades de intervenção imediata para a gestão da Profitecla.

Figura 10 — Comparação entre o risco à matrícula (M1b) e o risco atual (M1a) por curso



4.5 Interpretação dos Modelos com Inteligência Artificial Explicável

A capacidade preditiva elevada tende a implicar uma perda de legibilidade interna do modelo. Como referido na secção 2.6, esta opacidade não é exclusiva dos algoritmos de ensemble baseados em árvores; caracteriza igualmente as redes neuronais e outras famílias de modelos de maior complexidade. No caso dos métodos baseados em árvores, uma árvore de decisão isolada é diretamente interpretável, dado que a previsão corresponde a um percurso legível de regras. É a agregação de centenas de árvores num modelo de *gradient boosting*, como o utilizado neste estudo, que eleva a capacidade preditiva à custa dessa legibilidade e converte o modelo em caixa-preta. Em decisão educativa esta opacidade é um problema, porque um diretor de curso não pode agir sobre

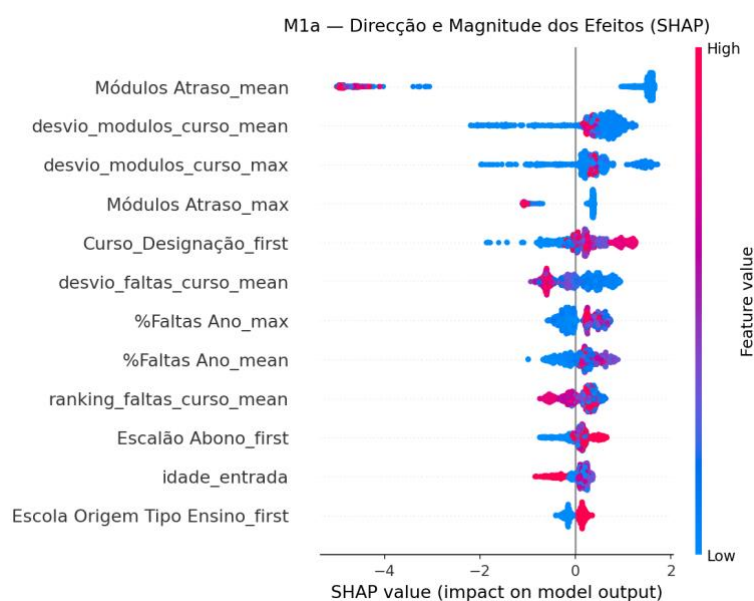
uma previsão de risco sem compreender os fatores que a determinam, e a confiança institucional depende da transparência (Alvarez-Garcia et al., 2024).

Nesse sentido, foi adotado o SHAP (SHapley Additive exPlanations), que constitui um método baseado na teoria de jogos cooperativos que atribui a cada *feature* uma contribuição quantificada para cada previsão (Lundberg & Lee, 2017). A escolha sobre o LIME (outro algoritmo de XAI) justifica-se pelas garantias de consistência e aditividade que o SHAP oferece. Para modelos de árvore, o TreeExplainer calcula estas contribuições de forma exata e eficiente. O uso do mesmo algoritmo em M1a e M1b permite comparação direta das importâncias entre modelos.

4.5.1 Interpretação do Modelo Retrospectivo (M1a)

No M1a, as *features* dominantes são variáveis de comportamento acumulado, os módulos em atraso (médios e máximos) e os desvios relativos ao curso surgem como os preditores de maior magnitude, com a importância máxima a corresponder a Módulos Atraso (média). Valores elevados empurram a previsão para a não conclusão, conforme apresenta a Figura 11. Esta dominância confirma a relação, amplamente documentada, entre ritmo de progressão curricular e sucesso formativo (Fernández-García et al., 2021; Costa-Mendes et al., 2022) e valida mutuamente o *clustering*, o perfil do *Cluster 1* corresponde exatamente às *features* que o SHAP identifica como mais determinantes da não conclusão.

Figura 11 — Direção e magnitude dos efeitos das features no M1a



4.5.2 Interpretação do Modelo Prospetivo (M1b)

No M1b, restrito à informação aquando do processo da matrícula, a *feature* dominante é o tipo de escola de origem ($|\text{SHAP}| = 0,64$), seguida do curso designação (0,47), da idade de entrada (0,35) e do escalão de abono (0,34)). O efeito médio por tipo de escola consta da Tabela 14.

Figura 12 — Importância global das features do M1b

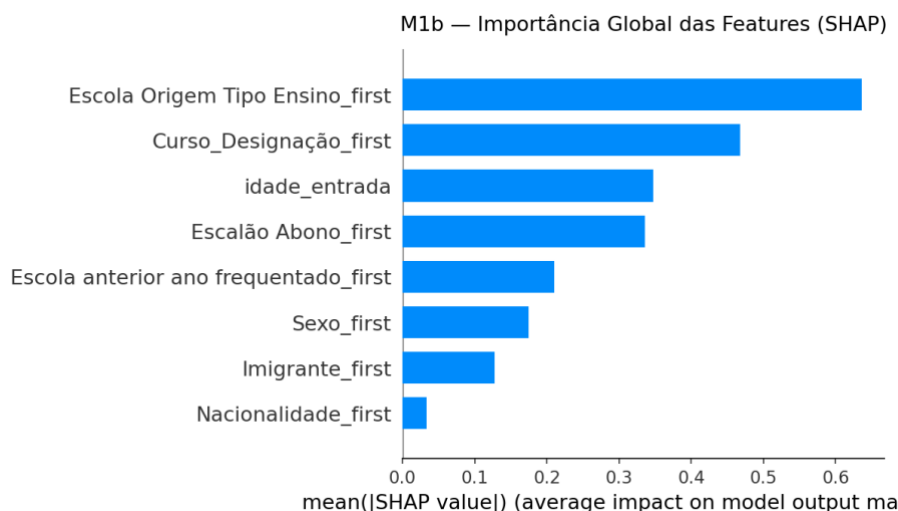


Tabela 14 — Taxa de conclusão por tipo de escola de origem (M1b)

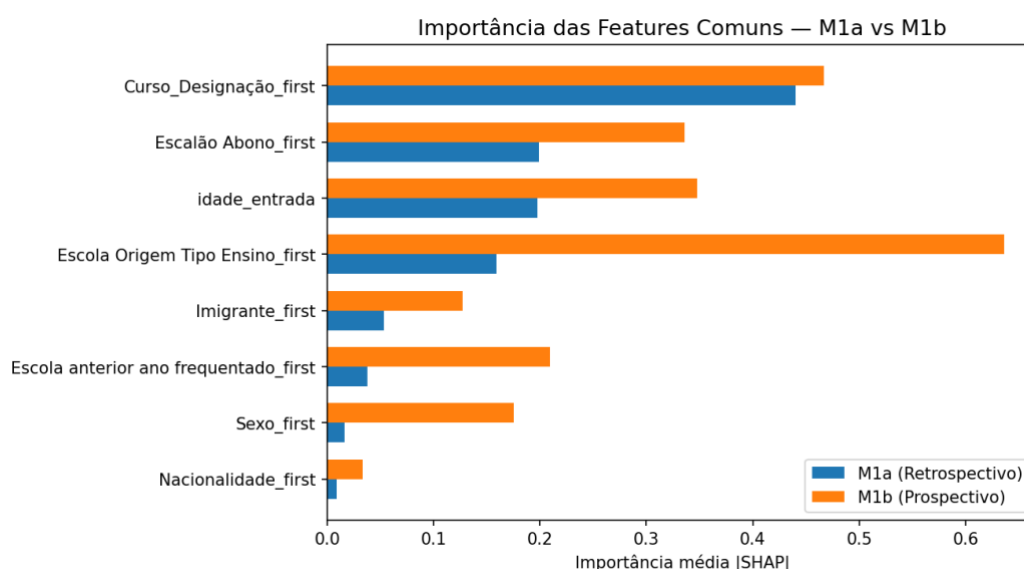
Tipo de escola de origem	Taxa de conclusão
Regular	83%
Profissional	63%
Desconhecido	63%

A importância elevada do tipo de escola de origem é uma associação estatística, não uma relação causal. O modelo identifica que alunos de determinados tipos de estabelecimento têm, em média, maior ou menor probabilidade de conclusão, mas não estabelece que a escola de origem cause o abandono. Fatores não observados e correlacionados (motivação, apoio familiar, preparação prévia) podem explicar a relação. Por isso, a estratégia de captação (Secção 4.7.4) não deve interpretar o tipo de escola como critério determinístico de seleção, mas como um sinal a combinar com apoio académico após a matrícula. Tratar uma associação correlacional como base para decisões prescritivas é uma das fragilidades mais comuns na aplicação de *Machine Learning* à educação (Matz et al., 2023).

4.5.3 Comparação entre Modelos e Limitações da XAI

A comparação das *features* (Figura 13) comuns quantifica o argumento da Secção 4.4.4, variáveis de perfil de entrada (escalão de abono, escola anterior) têm peso relativo maior no M1b, onde são quase a única informação disponível, e menor no M1a, onde são ofuscadas pelas variáveis comportamentais de maior poder discriminante.

Figura 13 — Importância média ($|SHAP|$) das *features* comuns aos dois modelos



As limitações da XAI devem ser explicitadas. O SHAP explica o comportamento do modelo, não a realidade; se o modelo aprendeu uma associação enviesada, o SHAP irá explicá-la fielmente. Os valores SHAP são contribuições marginais médias e não capturam intuitivamente interações complexas. A interpretabilidade não elimina o risco ético de decisões automatizadas; explicar uma previsão não a torna justa nem apropriada para fundamentar uma intervenção. Estas limitações reforçam que o sistema deve funcionar como apoio à decisão humana, não como seu substituto.

4.6 Análise de Sobrevivência: Tempo até ao Abandono

Enquadrada na fase de *Avaliação* do CRISP-DM, a análise de sobrevivência valida a coerência temporal dos perfis identificados na fase de Modelação. Responde a uma questão que a classificação não aborda, quando o aluno abandona.

4.6.1 Fundamentação

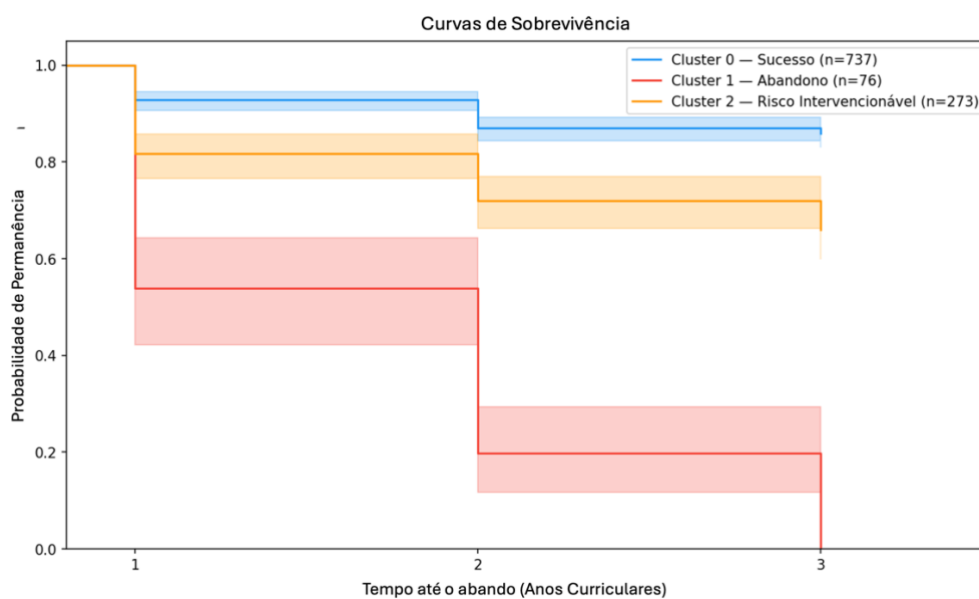
A estimativa de Kaplan-Meier modela a probabilidade de permanência ao longo do tempo, tratando adequadamente os dados censurados, alunos ainda em percurso (estado “Transitou”), cujo *outcome* é desconhecido, constituem censura natural à direita e são incorporados na estimativa em vez de descartados (Kaplan & Meier, 1958). A

duração foi definida como o ano máximo do plano curricular atingido, e o evento como a conclusão. A análise foi estratificada pelos três *clusters* da Secção 4.3.

4.6.2 Curvas de Sobrevivência e Validação Estatística

As curvas de Kaplan-Meier (Figura 14), com intervalos de confiança de 95%, revelam dinâmicas distintas de permanência entre os clusters. O eixo vertical representa a probabilidade de permanência ao longo dos anos curriculares, enquanto o eixo horizontal corresponde ao tempo até ao abandono. O resultado com maior relevância prática observa-se no Cluster 1, em que se estima que cerca de 46% dos alunos abandonem antes do final do primeiro ano. Esta concentração precoce do abandono sugere que a janela de intervenção mais relevante para os alunos de maior risco se situa no período de integração, durante as primeiras semanas do percurso académico. O Cluster 2 apresenta um decréscimo gradual da probabilidade de permanência ao longo dos três anos, justificando uma monitorização contínua. Em contraste, o Cluster 0 mantém uma elevada e estável probabilidade de permanência ao longo do período de observação.

Figura 14 — Curvas de Kaplan-Meier por cluster, com intervalos de confiança de 95%



A diferença entre as curvas foi validada com o teste de log-rank, que compara as funções de sobrevivência sob a hipótese nula de igualdade. A comparação global entre os três *clusters* é altamente significativa ($p < 0,0001$), confirmando que os perfis têm dinâmicas de sobrevivência estatisticamente distintas, o que reforça o seu valor não só na taxa final de conclusão, mas também na calendarização das intervenções.

4.7 Sistema de Apoio à Decisão: Arquitetura e *Dashboards*

A fase de *Deployment* materializa os modelos num sistema de *Business Intelligence* acessível aos decisores da Profitecla. Esta secção apresenta a arquitetura de dados e os três *dashboards* desenvolvidos; a sua validação empírica com os *stakeholders* é objeto do Capítulo V.

4.7.1 Arquitetura de Dados

Os *notebooks* Python exportam os resultados de cada fase analítica para ficheiros CSV intermédios, que o Power BI consome via Power Query. Esta separação física entre o ambiente de modelação e o ambiente de visualização tem uma consequência prática direta, os modelos podem ser re-treinados e os CSV regerados sem qualquer alteração à estrutura dos *dashboards*.

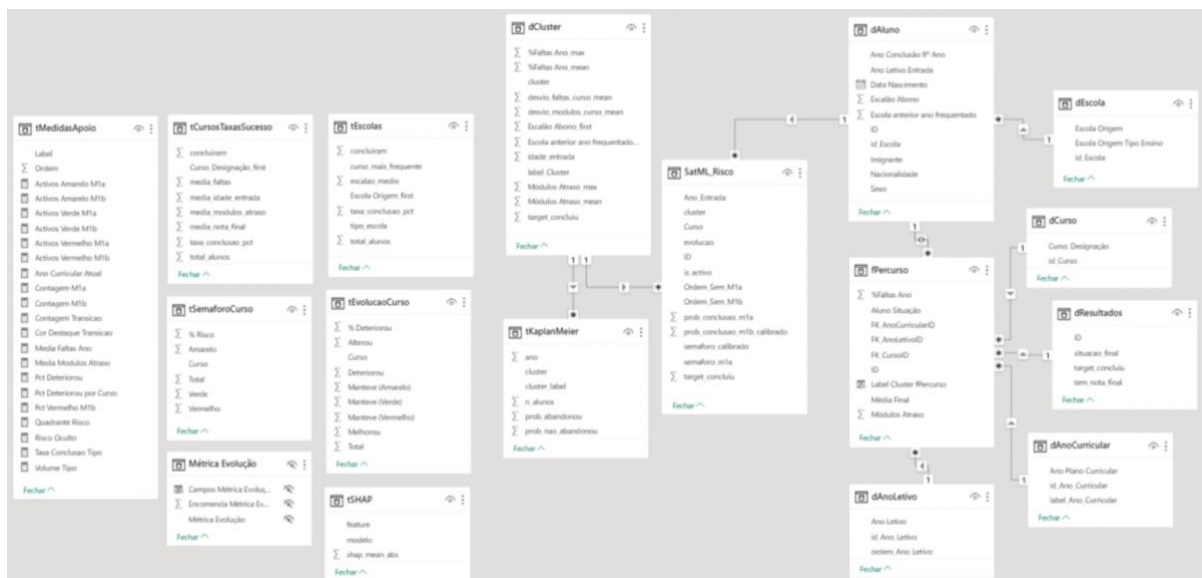
O modelo de dados subjacente, conforme apresentado na Figura 15, combina duas lógicas de organização com propósitos distintos. O núcleo segue o esquema em estrela de Kimball & Ross (2013), onde a tabela de factos fPercurso regista os indicadores do percurso académico de cada aluno por ano curricular, taxa de absentismo, módulos em atraso, média final, situação do aluno, incluindo as respetivas chaves estrangeiras para as cinco dimensões mencionadas: dAluno, dCurso, dAnoCurricular, dAnoLetivo e dResultados. A opção pela desnormalização característica deste paradigma é deliberada, os modelos dimensionais priorizam a velocidade de leitura sobre a eficiência de escrita, o que se traduz em consultas DAX mais simples e melhor desempenho nas agregações (Kimball & Ross, 2013). A dimensão dEscola foi mantida separada de dAluno pelo peso preditivo do tipo de escola de origem no M1b (SHAP = 0,64), tornando-a um eixo de análise com identidade própria no *dashboard* de captação, e não um atributo acessório do aluno.

Os outputs dos modelos de *Machine Learning* ocupam uma posição diferente no modelo. As probabilidades individuais de conclusão (M1a e M1b), os semáforos, o *cluster* K-Means e a evolução de risco têm granularidade de uma linha por aluno, são calculados externamente ao Power BI e podem ser regerados de forma independente do percurso histórico. Integrá-los na tabela de factos misturaria medições de negócio com atributos calculados de natureza distinta; integrá-los numa dimensão existente introduziria dependências artificiais entre a camada analítica e o núcleo dimensional. A solução foi materializar estes dados na tabela SatML_Risco, ligada a dAluno por uma relação n:1. A estrutura remete para o conceito de *satellite table* do modelo Data Vault

(Linstedt & Olschimke, 2015), os satélites estendem uma entidade central com atributos contextuais ou derivados, preservando a integridade do núcleo e permitindo que a camada de *Machine Learning* evolua sem impacto na estrutura de base. O modelo resultante é uma arquitetura híbrida, o núcleo Kimball assegura a performance e a intuitividade operacional; o satélite acomoda a natureza incremental e re-computável dos resultados preditivos.

Uma camada adicional de tabelas com prefixo t — tSHAP, tKaplanMeier, tEvolucaoCurso, tCursosTaxasSucesso, tEscolas, tSemaforoCurso, materializa transformações pré-calculadas específicas a cada *dashboard*. Estas tabelas não mantêm relações com o núcleo dimensional; são artefactos analíticos autónomos, carregados diretamente dos CSV de output dos notebooks, que evitam lógica DAX complexa no lado da visualização. A arquitetura de dados do sistema pode ser visualizada na Figura 15, assim composta por um núcleo em estrela (fPercurso e seis dimensões), tabela satélite SatML_Risco e camada analítica complementar (tabelas t-).

Figura 15 — Arquitetura de Dados do Sistema



A principal limitação da arquitetura atual é a dependência de extração e carregamento manuais, os dados saem do ERP institucional por exportação manual, passam pelos notebooks e chegam ao Power BI como ficheiros estáticos. A integração automatizada por *pipeline* com atualização incremental diretamente do ERP constitui o requisito de evolução para uma eventual transição do sistema para produção, e é identificada como trabalho futuro.

4.7.2 Ecrã Inicial e *Dashboard* de Monitorização de Risco

O ecrã inicial sintetiza o âmbito do sistema (1 086 alunos analisados, 75% de taxa de conclusão, três perfis de risco, seis anos de dados históricos) e encaminha para os três *dashboards* temáticos, cada um dirigido a um público distinto (Figura 16).

Figura 16 — Ecrã inicial do sistema, com indicadores globais e acesso aos dashboards de monitorização, perfis e captação



O *dashboard* de Monitorização de Risco (Figura 17), dirigido aos coordenadores de curso, é o instrumento operacional central. Apresenta os 396 alunos ativos com o seu semáforo aquando do processo de matrícula (M1b) e atual (M1a), a matriz de transição entre estados, a distribuição de risco por curso e uma lista acionável dos alunos em alto risco atual, com filtros por curso, ano curricular e evolução. O seu valor reside em traduzir as probabilidades dos modelos numa hierarquia de prioridades de intervenção.

Figura 17 — Dashboard operacional de monitorização de risco

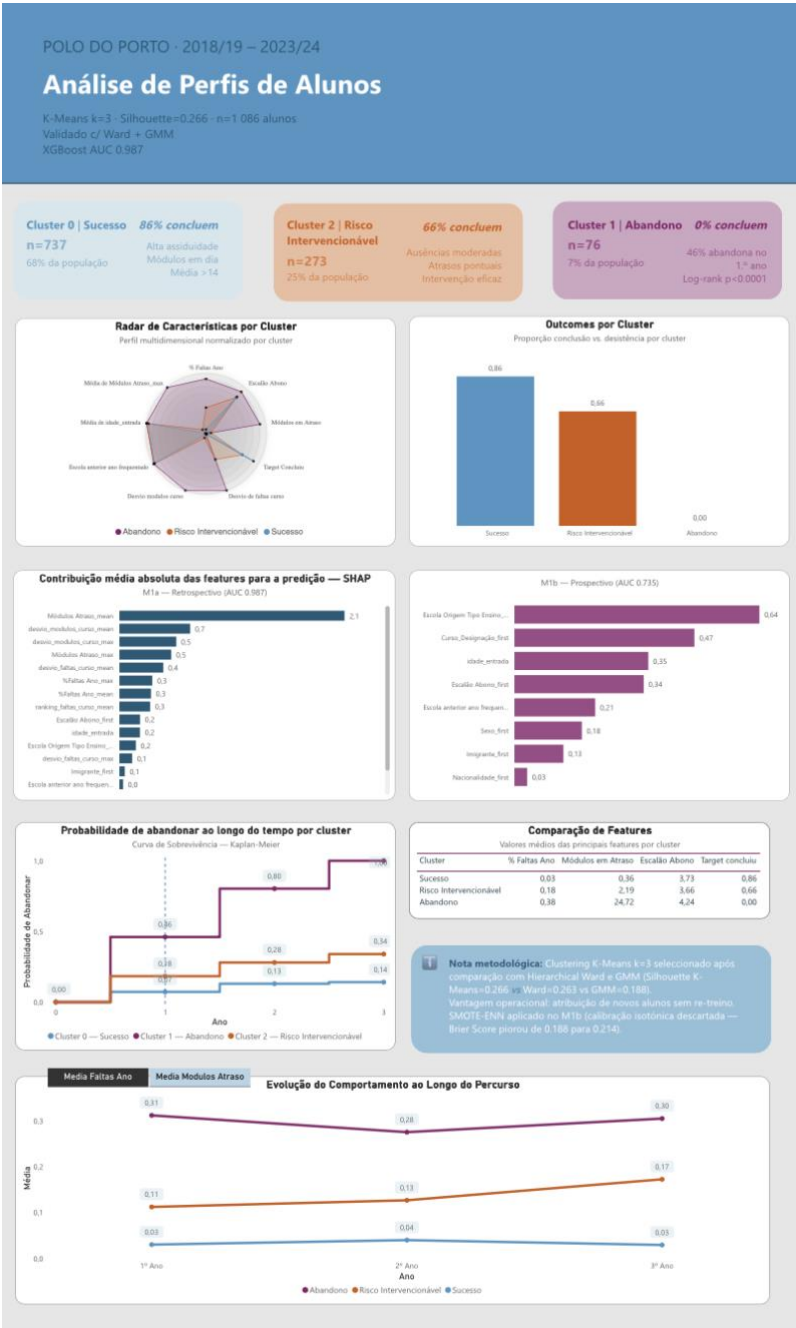


4.7.3 Dashboard de Perfis de Alunos

Dirigido à gestão académica, o *dashboard* de Perfis de Alunos (Figura 18) apresenta os três *clusters* com as suas características multidimensionais (radar normalizado), a proporção de conclusão por *cluster*, a evolução comportamental ao longo do percurso, a importância global SHAP dos dois modelos e as curvas de Kaplan-Meier. Integra, num único painel, os resultados das fases de aprendizagem não supervisionada e

supervisionada, interpretação e sobrevivência, servindo de instrumento de leitura estratégica dos padrões de sucesso e abandono escolar.

Figura 18 — Dashboard de perfil de alunos

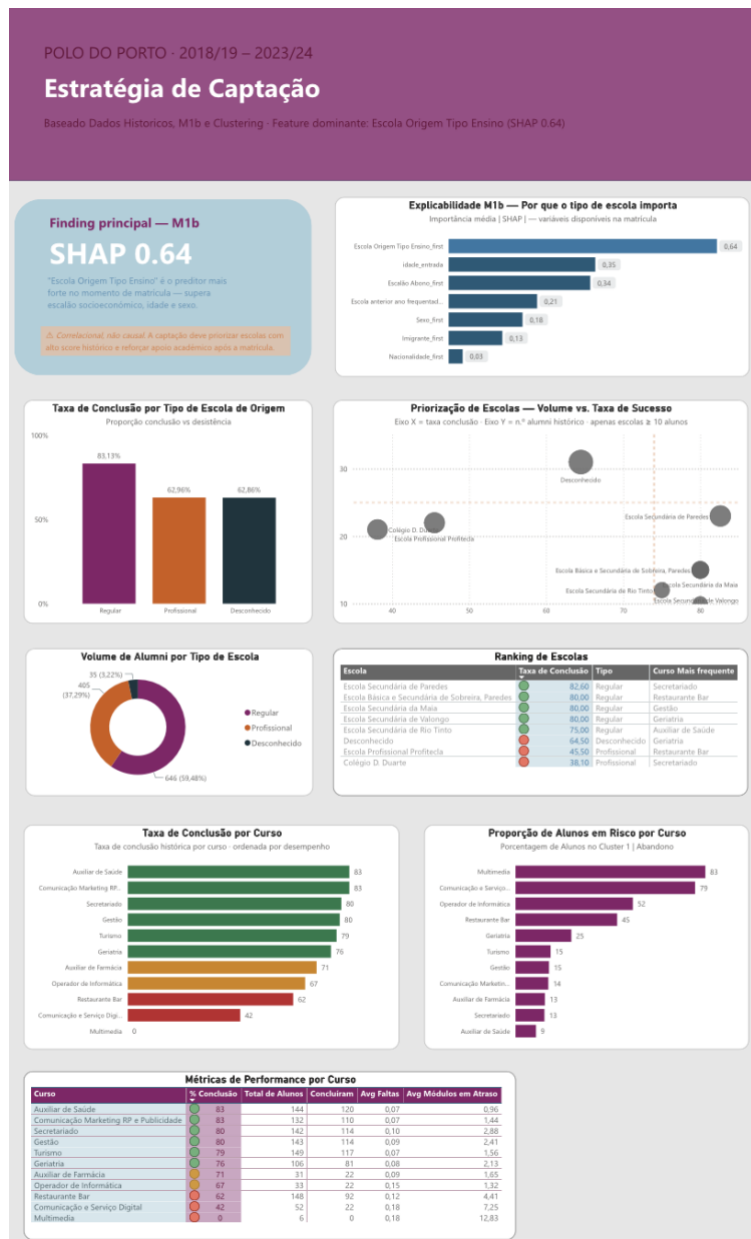


4.7.4 Dashboard de Captação e Recrutamento

Dirigido às áreas de marketing e à direção da escola, o *dashboard* de Captação operacionaliza o segundo eixo da investigação (Figura 19). Cruza a taxa de conclusão histórica por tipo de escola de origem e por escola individual (limitada às dez escolas com maior volume, dada a limitação de normalização documentada na Secção 4.1.2) com a

explicabilidade SHAP do M1b, identificando o perfil de origem associado a maior probabilidade de sucesso.

Figura 19 — Dashboard de captação e recrutamento



Os dados confirmam um diferencial relevante: as escolas de tipo Regular apresentam uma taxa de conclusão histórica de 83,13%, face a 62,96% das escolas Profissionais. O *ranking* individual destaca estabelecimentos com maior taxa de sucesso e volume relevante. Reitera-se a advertência da Secção 4.5.3, esta informação é correlacional e deve orientar a priorização do esforço de captação combinada com reforço de apoio académico após a matrícula, nunca como critério determinístico de seleção de candidatos.

Importa ainda assinalar uma fragilidade amostral, cursos com reduzido número de alunos (como Multimédia, com seis alunos no histórico) apresentam taxas extremas, neste caso, 0% de conclusão, que não devem ser lidas como padrão estável, mas como sinal a confirmar com dados de *cohorts* adicionais.

CAPÍTULO V – RESULTADOS E DISCUSSÃO

5 Resultados e Discussão

Este capítulo apresenta e discute a avaliação do artefacto desenvolvido junto dos seus destinatários institucionais. Enquanto o Capítulo 4 documentou o desempenho técnico interno dos modelos e a construção do sistema de apoio à decisão, o presente capítulo responde em que medida o artefacto é considerado útil, claro e adotável pelos profissionais que o utilizariam na prática. Esta avaliação constitui a componente sumativa da fase de Avaliação no quadro da *Design Science Research* (Venable et al., 2016), e a sua função é estabelecer a validade do artefacto não apenas pelo seu rigor estatístico, mas pela sua relevância operacional em contexto real.

A avaliação combina duas fontes de evidência, uma componente quantitativa, recolhida através de um questionário estruturado de utilidade percebida, e uma componente qualitativa, resultante da discussão estruturada conduzida durante a sessão de demonstração. A integração das duas componentes permite triangular a perceção dos participantes e identificar não apenas o grau de aceitação do artefacto, mas também as condições e reservas que qualificam essa aceitação.

5.1 Caracterização da Sessão de Avaliação

A sessão de avaliação foi realizada a 9 de junho de 2026, de forma online, com a participação de quatro responsáveis da instituição com funções de gestão pedagógica e operacional. A sessão teve a duração aproximada de uma hora e seguiu o formato estruturado de apresentação do processo de desenvolvimento e dos resultados, demonstração ao vivo dos *dashboards* desenvolvidos em Power BI, discussão aberta orientada por questões, e aplicação do questionário de avaliação. A estrutura da sessão e do instrumento da recolha encontram-se mais bem detalhados no Apêndice I.

A dimensão da amostra ($N=4$) é adequada ao propósito desta avaliação. No contexto da *Design Science Research*, a avaliação de artefactos por um número reduzido de especialistas do domínio constitui uma prática estabelecida, sobretudo quando os participantes detêm conhecimento aprofundado do contexto operacional em que o artefacto será utilizado (Venable et al., 2016). Os resultados quantitativos são, por conseguinte, interpretados como indicadores descritivos da perceção dos participantes, e não como base para inferência estatística, uma distinção que orienta toda a análise subsequente.

5.2 Resultados Quantitativos

O questionário aplicado, descrito em detalhe na secção 3.2.1, organizou-se em quatro dimensões adaptadas do *Technology Acceptance Model* (Davis, 1989), utilidade percebida, facilidade de utilização percebida, relevância e transferibilidade, e uma avaliação global. A Tabela 15 apresenta as estatísticas descritivas para cada item.

Tabela 15 — Resultados do questionário de avaliação (N=4, escala 1–5)

Item	Dimensão	Mediana	Média	Desvio-padrão
A1 — Informação útil para apoiar a intervenção	Utilidade	5,0	4,75	0,50
A2 — Melhoria da deteção atempada de risco	Utilidade	5,0	4,75	0,50
A3 — Perfis refletem situações reais	Utilidade	4,0	4,00	0,00
A4 — Relevância da informação para captação	Utilidade	4,5	4,50	0,58
B1 — Clareza das visualizações	Usabilidade	5,0	4,75	0,50
B2 — Intuitividade do sistema semafórico	Usabilidade	4,5	4,50	0,58
B3 — Facilidade de aprendizagem autónoma	Usabilidade	4,0	4,25	0,50
C1 — Congruência com o conhecimento dos alunos	Relevância	4,0	4,00	0,82
C2 — Pertinência da expansão a outros polos	Transferibilidade	5,0	4,75	0,50
C3 — Intenção de utilização regular	Adoção	5,0	4,75	0,50
D1 — Avaliação global da utilidade	Global	4,5	4,50	0,58

A avaliação global do artefacto foi consistentemente positiva, com uma média de 4,50 no item de síntese (D1) e medianas iguais ou superiores a 4,0 em todos os itens. Os valores mais elevados concentraram-se nas dimensões de utilidade percebida (A1 e A2, ambos com mediana 5,0) e de adoção e transferibilidade (C2 e C3, ambos com mediana

5,0), indicando que os participantes reconheceram valor prático no artefacto e manifestaram intenção clara de o utilizar e expandir.

O desvio-padrão revela um padrão informativo. A concordância foi total no item A3 (perfis refletem situações reais), onde os quatro participantes atribuíram a classificação 4, um consenso que sustenta a aderência dos perfis à realidade institucional. A maior dispersão registou-se no item C1 (congruência com o conhecimento dos alunos), com um desvio-padrão de 0,82 e a única classificação 3 do questionário. Esta dispersão não constitui uma fragilidade do artefacto, pelo contrário, é coerente com o feedback qualitativo discutido na secção seguinte, que qualifica temporalmente o conceito de risco e ajuda a explicar por que razão a congruência percebida é mais matizada do que a utilidade percebida.

5.3 Resultados Qualitativos

A discussão estruturada durante a sessão produziu um conjunto de observações que se organizam em quatro temas. A análise temática que se segue prioriza os contributos com maior relevância para a interpretação dos resultados e para a fundamentação das escolhas metodológicas da investigação.

Tema 1 — Adesão dos perfis e dos resultados à prática institucional: Os participantes confirmaram que a lógica dos três perfis de risco corresponde à forma como a instituição já acompanha informalmente os seus alunos, com base em indicadores de faltas, módulos em atraso e capacidade de conclusão atempada. Esta convergência entre a segmentação produzida estatisticamente e a prática operacional é particularmente relevante porque sustenta a validade dos *clusters*, os perfis identificados não são meros artefactos estatísticos, mas correspondem a categorias reconhecíveis pelos profissionais da instituição. De igual modo, os participantes confirmaram explicitamente que a concentração do abandono no primeiro ano, evidenciada pela análise de sobrevivência, coincide com a perceção interna da escola, reforçando a interpretação do primeiro ano como janela crítica de retenção.

Tema 2 — Qualificação temporal do risco intervencionável: O contributo qualitativo mais significativo da sessão diz respeito aos limites práticos do perfil de Risco Intervencionável. Os participantes observaram que, embora o risco seja efetivamente intervencionável, a capacidade de atuação da instituição diminui acentuadamente quando o aluno já acumulou um número elevado de faltas e módulos em atraso. Um dos responsáveis descreveu este processo como um efeito de "bola de neve", em que as faltas

geram atrasos, os atrasos geram desmotivação, e a desmotivação culmina no abandono. Esta observação não enfraquece o conceito de risco intervencionável, qualifica-o temporalmente. O valor da intervenção depende criticamente do momento em que o sinal de risco é identificado, tendo um maior impacto quanto mais cedo ocorrer. Este feedback fundamenta diretamente a opção metodológica central da investigação, a arquitetura de dois modelos preditivos complementares. O modelo prospetivo, aplicável no momento da matrícula, permite a sinalização precoce antes da acumulação dos sinais de deterioração; o modelo retrospectivo, alimentado pelo comportamento observado, permite monitorizar a evolução do risco ao longo do percurso. A perceção dos *stakeholders* converge, assim, com a justificação técnica da dualidade dos modelos apresentada no Capítulo 4.

Tema 3 — Necessidade de interpretação contextual dos *dashboards*. Os participantes consideraram as visualizações úteis e claras, em consonância com os resultados quantitativos das dimensões de usabilidade, mas introduziram uma leitura crítica pertinente. Manifestaram surpresa com o grau de deterioração observado entre o risco inicial e o risco ao longo do percurso, e um dos participantes alertou para o facto de que, em determinados cursos, uma taxa de deterioração reduzida não significa necessariamente uma situação favorável, pode antes indicar que o risco já era elevado desde a entrada. Esta observação revela que as visualizações, ainda que claras, requerem interpretação contextual para evitar leituras simplificadas e aponta para o valor de refinar os indicadores visuais de modo a incorporar o nível de risco inicial como referência de leitura. Um dos participantes sugeriu, na resposta aberta, a apresentação de valores absolutos em complemento das percentagens, recomendação diretamente acionável para o aperfeiçoamento dos *dashboards*. Dadas as recomendações, aprimoramentos foram incorporados ainda no decurso da investigação. O *dashboard* de monitorização foi ajustado para apresentar, em complemento das taxas percentuais, a evolução da taxa de deterioração por curso, bem como mudanças nas cores de alguns visuais, melhorando a legibilidade contextual. Este refinamento ilustra o carácter iterativo do processo de *Design Science Research*, em que a avaliação do artefacto retroalimenta diretamente a sua melhoria.

Tema 4 — Condições de adoção e limitações reconhecidas: Os participantes demonstraram recetividade concreta à implementação do sistema, reconhecendo que a maioria das variáveis necessárias já existe no sistema de informação da instituição. Identificaram, contudo, que a extração manual recorrente dos dados constitui um

obstáculo à sustentabilidade da atualização, sugerindo uma periodicidade trimestral como ponto de partida mais realista do que a atualização mensal. No plano das limitações, os participantes confirmaram, do lado dos utilizadores, os problemas de qualidade dos dados de escola de origem, informação em falta e inconsistência nos nomes das escolas, que haviam sido identificados na fase de preparação de dados. Esta confirmação é metodologicamente valiosa, porque valida que as limitações apontadas na investigação são limitações reais do sistema de informação institucional, e não apenas uma preocupação técnica do investigador. Quanto à estratégia de captação, os participantes reconheceram o seu valor exploratório, relativizando-a simultaneamente porque nem sempre a instituição pode atuar livremente em todas as escolas de origem, e os padrões históricos podem alterar-se entre anos. Os resultados de captação devem, por conseguinte, ser interpretados como instrumentos de apoio estratégico e não como relações causais ou permanentes.

5.4 Triangulação e Validade da Avaliação

A integração das componentes quantitativa e qualitativa revela um padrão coerente. As dimensões com avaliação quantitativa mais elevada, utilidade percebida e intenção de adoção, são consistentes com a recetividade prática à implementação manifestada na discussão. Simultaneamente, a maior dispersão registada na congruência percebida (item C1) encontra explicação direta no feedback qualitativo: a perceção de que o risco intervencionável depende do momento da sua identificação introduz uma nuance que se reflete numa avaliação mais matizada da congruência entre os resultados do modelo e o conhecimento experiencial dos alunos. A convergência entre as duas fontes de evidência reforça a fiabilidade interpretativa da avaliação.

A avaliação confirma, no essencial, a utilidade e a aceitabilidade do artefacto, enquanto produz contributos que enriquecem a sua interpretação. O mais relevante é a qualificação temporal do conceito de risco, que valida empiricamente, a partir da experiência dos profissionais, a opção pela arquitetura dual de modelos. Este resultado ilustra o valor da avaliação por especialistas no quadro da *Design Science Research*, que não se limita a confirmar o artefacto, mas devolve conhecimento que aprofunda a compreensão do problema e fundamenta as decisões de design.

5.5 Discussão sobre a Transferibilidade e a Implementação

A avaliação revelou um interesse expressivo na expansão do artefacto aos restantes polos da Profitecla, traduzido na elevada classificação do item C2. A discussão sobre as modalidades dessa expansão identificou três estratégias possíveis, com implicações distintas. A primeira consiste em replicar o *framework* metodológico em cada polo, retreinando os modelos com os dados locais de cada contexto, opção que preserva a especificidade de cada polo e é coerente com o princípio, defendido nesta investigação, de que o que é transferível é o *framework* e não os parâmetros calibrados localmente. A segunda consiste em utilizar os perfis e padrões identificados no Polo do Porto para informar a estratégia de outros polos, abordagem mais rápida, mas que comporta o risco de importar padrões não generalizáveis. A terceira consiste em desenvolver um modelo único e globalizado para toda a rede, simplificando a manutenção, mas sacrificando a sensibilidade às particularidades locais.

Os próprios participantes reconheceram que a utilização do histórico do Porto como base para outros polos pode comprometer a fiabilidade, devido a diferenças contextuais e geográficas. Esta reserva sustenta a primeira estratégia como a mais defensável: a transferibilidade do artefacto reside na replicabilidade do seu processo de construção, e não na aplicação direta dos seus resultados a contextos distintos. A generalização exige validação local e, idealmente, a adaptação dos modelos a cada polo, uma orientação que se inscreve naturalmente no trabalho futuro.

6 Conclusões

6.1 Síntese da Investigação

A presente dissertação partiu de um problema concreto e recorrente nas instituições de ensino profissional, a ausência de mecanismos sistemáticos de antecipação do abandono e de apoio à decisão baseado em evidência, num contexto onde os recursos são escassos, os perfis dos alunos são vulneráveis e os dados disponíveis se restringem a registos administrativos. O ensino profissional é, paradoxalmente, o setor com maior necessidade de intervenção precoce e o menos representado na investigação em *Learning Analytics* e *Educational Data Mining*, áreas dominadas por estudos conduzidos em ambientes de dados ricos do ensino superior.

O *framework* desenvolvido respondeu a este desafio articulando quatro componentes analíticas num artefacto integrado: definição de perfis de risco por *clustering*, previsão de não conclusão em dois horizontes temporais, interpretabilidade por SHAP e operacionalização em *Business Intelligence*. Construído sobre seis anos letivos de dados administrativos do polo do Porto da Profitecla (2018/19–2023/24), correspondentes a 1 086 alunos, o artefacto foi desenvolvido segundo o paradigma de *Design Science Research* e o processo CRISP-DM, e validado junto dos responsáveis institucionais numa sessão estruturada de avaliação. Os resultados foram parcialmente disseminados no artigo científico *A Machine Learning Approach to Dropout Prediction in Vocational Secondary Education*, aceite para apresentação na conferência ICITED 2026 (Springer LNNS).

6.2 Resposta às Questões de Investigação

Q1 — Quais os fatores associados à conclusão do curso no ensino profissional? A análise de *clustering* K-Means com $k=3$ identificou três perfis operacionalmente distintos e estatisticamente coerentes. O *Cluster 0* (Sucesso, $n=737$, 86,0% de conclusão) caracteriza-se por absentismo reduzido (0,03) e progressão curricular estável (0,36 módulos em atraso). O *Cluster 1* (Abandono, $n=76$, 0,0% de conclusão) apresenta valores extremos de absentismo (0,38) e de módulos em atraso (24,72 em média), correspondendo a situações de rutura instalada. O *Cluster 2* (Risco Intervencionável, $n=273$, 65,9% de conclusão) ocupa uma posição de relevância estratégica: com absentismo moderado e atrasos controláveis, é o grupo onde a intervenção atempada tem maior potencial de alterar a trajetória.

A análise de sobrevivência de Kaplan-Meier complementou estes perfis ao revelar a dimensão temporal do abandono, cerca de 46% dos alunos do *Cluster* 1 abandonam antes do final do primeiro ano curricular, delimitando a janela crítica de intervenção ao período de onboarding. O teste de log-rank confirmou que as curvas de sobrevivência dos três *clusters* são estatisticamente distintas ($p < 0,0001$), validando a coerência temporal dos perfis. O feedback dos *stakeholders* na sessão de avaliação corroborou esta concentração do abandono no primeiro ano, conferindo validade ecológica ao resultado estatístico.

A análise SHAP mostrou que, na ausência de dados comportamentais (M1b), o tipo de escola de origem ($|\text{SHAP}|=0,64$), o curso, a idade de entrada e o escalão de abono são os preditores mais determinantes. No M1a, os módulos em atraso acumulados assumem primazia absoluta, confirmando que o comportamento académico é o melhor preditor da não conclusão quando disponível, e que as variáveis socioeconómicas funcionam como proxy da escassez de informação no início do percurso, não como determinantes causais do resultado, distinção com implicações éticas diretas para o desenho das intervenções.

Q2 — É possível prever, com desempenho aceitável, o risco de não conclusão ao longo do percurso? A resposta é afirmativa para os dois horizontes temporais, com desempenho diferenciado e interpretação complementar. O Modelo Retrospectivo (M1a), alimentado pela trajetória comportamental completa, alcança uma AUC-ROC de 0,987 e um F1 de 0,933 para a classe de não conclusão, valores que o posicionam entre os resultados de referência da literatura de EDM para contextos de dados administrativos restritos. O Modelo Prospetivo (M1b), restrito a variáveis disponíveis na matrícula e tratado com SMOTE-ENN, obtém uma AUC-ROC de 0,735 e F1 de 0,488, desempenho modesto em termos absolutos, mas coerente com a limitação de informação disponível nesse momento e com estudos análogos em contextos comparáveis.

O diferencial de AUC entre os dois modelos ($\Delta=0,25$) quantifica o valor informativo do comportamento académico acumulado, e não uma fragilidade do M1b. Os dois modelos operam em momentos distintos do ciclo de decisão institucional: o M1b define quem observar com atenção desde o primeiro dia; o M1a determina quem já está em trajetória de risco e necessita de intervenção imediata. A aplicação dos dois modelos aos 396 alunos ativos em 2024/25 revelou uma divergência operacionalmente decisiva: dos 211 alunos classificados como baixo risco à matrícula pelo M1b, apenas 35 mantiveram essa classificação no M1a, com 102 a transitarem para alto risco. Estes alunos seriam invisíveis a qualquer triagem baseada unicamente no perfil de entrada, mas o

comportamento posterior ativou alertas claros. A divergência é o sistema a funcionar como desenhado.

A equivalência estatística entre os três algoritmos comparados, XGBoost, *Random Forest* e Regressão Logística, confirmada pelo teste de McNemar em ambos os modelos ($p > 0,05$), sustenta a escolha do XGBoost por consistência metodológica, e não por superioridade empírica. Esta distinção é relevante para a reprodutibilidade do *framework* em outros contextos institucionais.

Q3 — Como traduzir os resultados dos modelos em *dashboards* prescritivos úteis para a gestão? Os três *dashboards* desenvolvidos em Power BI operacionalizam a resposta a esta questão em três registos distintos. O *Dashboard* de Monitorização de Risco, destinado aos coordenadores de curso, traduz as probabilidades dos modelos numa hierarquia de prioridades de intervenção assente num semáforo de três níveis, com filtros por curso, ano curricular e evolução temporal. O *Dashboard* de Perfis de Alunos, dirigido à gestão académica, integra *clustering*, SHAP e curvas de Kaplan-Meier num painel de leitura estratégica dos padrões de sucesso e abandono. O *Dashboard* de Captação e Recrutamento operacionaliza o segundo eixo da investigação ao cruzar a escola de origem com a taxa de conclusão histórica e com os valores SHAP do M1b.

A avaliação junto de quatro responsáveis da Profitecla produziu uma pontuação média global de 4,50/5,0 (escala Likert), com medianas de 5,0 nas dimensões de utilidade percebida, intenção de adoção e transferibilidade. O feedback qualitativo acrescentou uma qualificação temporal do conceito de risco intervencionável que não emerge dos dados: a capacidade de atuação da instituição diminui quando o aluno já acumulou faltas e módulos em atraso, descrito pelos participantes como um efeito de "bola de neve". Esta observação valida empiricamente, a partir da experiência dos profissionais, a arquitetura dual de modelos e reforça a prioridade do modelo prospetivo para a intervenção preventiva.

6.3 Contributos da Investigação

Do ponto de vista académico, a dissertação apresenta três contributos principais. O primeiro é a demonstração da viabilidade de um *framework* preditivo e prescritivo em contextos de dados restritos, sem LMS e sem métricas de admissão padronizadas, estabelecendo um patamar de desempenho conservador e transferível para as instituições do setor, contribuindo diretamente para suprir a sub-representação do ensino profissional na literatura de *Learning Analytics* identificada por Fernández-García et al. (2021) e

Gedrimiene et al. (2019). O segundo é a documentação da evolução temporal da importância dos preditores através da comparação SHAP entre dois modelos de horizonte diferente, sustentando a interpretação de que as variáveis socioeconómicas funcionam como *proxies* da falta de informação na matrícula, e não como determinantes causais do resultado da distinção com implicações éticas diretas. O terceiro é a integração, num único artefacto operacional, da modelação não supervisionada, da previsão supervisionada dual-horizonte, da explicabilidade algorítmica e da camada prescritiva de *Business Intelligence*, respondendo à lacuna de integração identificada na revisão e avançando o estado da arte na transição da analítica preditiva para a prescritiva.

A nível aplicado, o artefacto entrega à Profitecla quatro instrumentos: um semáforo de risco para os alunos ativos, um painel de monitorização operacional, um mapa estratégico de captação e uma análise de perfis históricos. Estes instrumentos representam uma passagem da gestão intuitiva para a gestão orientada por dados, numa instituição que, até esta investigação, não dispunha de qualquer sistema formal de análise preditiva do percurso dos seus alunos. O valor destes artefactos realiza-se quando os profissionais que acompanham os alunos os incorporam nas suas práticas, desde a sinalização precoce até à conversa individualizada ou ao plano de recuperação.

6.4 Limitações

Esta investigação está sujeita a limitações que condicionam a generalização dos seus resultados. A principal limitação metodológica reside na divisão aleatória do conjunto de dados, em detrimento de uma divisão temporal que permitisse avaliar a capacidade do modelo em prever resultados de grupos de alunos futuros, simulando de forma mais rigorosa o desempenho em produção com novas entradas. Esta opção foi motivada pela dimensão amostral (1 086 alunos), insuficiente para garantir estabilidade dos modelos com uma única coorte de teste, e constitui uma prioridade de refinamento à medida que a base histórica cresce.

O âmbito geográfico restrito ao polo do Porto implica que os parâmetros dos modelos refletem os padrões específicos deste contexto. A aplicação direta a outros polos ou instituições com populações distintas requer retreino local, a transferibilidade do artefacto reside no processo de construção, não nos parâmetros estimados. A dependência da extração manual dos dados constitui um obstáculo à sustentabilidade e frequência de atualização, sendo a integração por *pipeline* automatizado diretamente com o ERP institucional o pré-requisito para a transição do sistema para produção. A impossibilidade

de normalizar completamente o campo de escola de origem (716 designações distintas) restringe o *dashboard* de captação às dez escolas de maior volume, embora a validade dos modelos preditivos não seja comprometida, dado que utilizam o tipo de escola de origem, de baixa cardinalidade. Por último, o horizonte de avaliação externa foi limitado a uma única sessão com quatro participantes de um único polo, o que restringe a inferência sobre a percepção de outros contextos institucionais.

6.5 Direções para Investigação Futura

Os resultados desta investigação abrem várias direções para trabalho futuro. A extensão do *framework* aos restantes polos da Profitecla permitirá validar a estabilidade dos perfis de *clustering* com amostras maiores, avaliar a necessidade de calibração local dos modelos e explorar a solução $k=4$, cuja estabilidade estatística requer dimensão amostral superior à disponível no polo do Porto. A implementação de um *pipeline* de atualização automática, com integração direta ao ERP escolar, é a condição necessária para a transição do sistema para utilização institucional continuada com a periodicidade trimestral sugerida pelos *stakeholders* na sessão de avaliação.

Do ponto de vista metodológico, a transição para uma divisão temporal do conjunto de dados, à medida que a série histórica se alarga, permitirá uma estimativa mais conservadora e rigorosa do desempenho preditivo em produção. A incorporação de técnicas de inferência causal, como modelos de variáveis instrumentais ou *frameworks* de *potential outcomes*, permitiria superar a limitação correlacional dos modelos atuais e avançar para recomendações de intervenção com garantia de efeito, respondendo diretamente à lacuna de conhecimento causal identificada na revisão de literatura. A exploração de abordagens de *federated learning*, que permitiriam o treino partilhado entre polos sem centralização de dados, representa uma direção relevante para contextos multi-institucionais com restrições de privacidade e proteção de dados. Finalmente, a análise das trajetórias pós-formação, recorrendo aos dados de empregabilidade da Fich3 excluídos desta investigação, permitiria estender o *framework* do sucesso formativo ao sucesso profissional, completando o ciclo de valor que o ensino profissional se propõe oferecer.

6.6 Reflexão Final

O ensino profissional serve os jovens com maior necessidade de suporte e menor margem de erro. Um abandono que ocorre no primeiro ano não representa apenas um

insucesso administrativo, representa, muitas vezes, o término de uma trajetória formativa que poderia ter sido diferente se tivesse tido uma intervenção atempada. Esta dissertação demonstrou que é possível construir um sistema de apoio à decisão útil, interpretável e eticamente fundamentado a partir de dados de uma instituição de ensino profissional. O valor deste trabalho não reside nos algoritmos em si, mas na sua articulação com as práticas dos profissionais que acompanham os alunos e na sua capacidade de tornar visível o que, sem análise de dados e aplicação de uma solução de aprendizagem supervisionada e não supervisionada, permaneceria invisível até ser tarde demais para intervir.

A concretização plena do potencial deste sistema extravasa o âmbito técnico desta dissertação e depende de decisões organizacionais que os artefactos aqui apresentados visam informar: a capacitação das equipas em literacia de dados, a eventual afetação de recursos ao acompanhamento individualizado e o desenho de protocolos de intervenção diferenciados. A tecnologia, por si só, não retém nenhum aluno; quem retém são os professores, os coordenadores e as equipas de apoio que, com esta ferramenta, podem fazer chegar o acompanhamento a quem mais precisa e fazê-lo de forma atempada

REFERÊNCIAS BIBLIOGRÁFICAS

Ancheta, R. F., Daniel, D., & Ahmad, R. (2021). EFFECT OF CLASS ATTENDANCE ON ACADEMIC PERFORMANCE. *European Journal of Education Studies*. <https://doi.org/10.46827/ejes.v8i9.3887>

Ahmed, W., Azhar, M., Asim, M., Sadiq, S., & Ahmed, M. (2025). Machine learning-based academic performance prediction with explainability for enhanced decision-making. *Scientific Reports*, 15, Article 26879. <https://doi.org/10.1038/s41598-025-12353-4>

Alalawi, K., Athauda, R., & Chiong, R. (2025). Student performance prediction and action (SPPA): A framework to pilot learning analytics interventions. *International Journal of Artificial Intelligence in Education*, 35, 1239–1287. <https://doi.org/10.1007/s40593-024-00429-7>

Ali, I., Cakici, K., Danish, M., Saleh, A., & Zhu, X. (2023). Using the AIRR framework for equitable early alert system development: A case study. *Journal of Learning Analytics*, 10(1), 125–139. <https://doi.org/10.18608/jla.2023.7827>

Alturki, S., Cohausz, L., & Stuckenschmidt, H. (2022). Predicting Master's students' academic performance: an empirical study in Germany. *Smart Learning Environments*, 9. <https://doi.org/10.1186/s40561-022-00220-y>

Alvarez-Garcia, M., Arenas-Parra, M., & Ibar-Alonso, R. (2024). Educational data mining and machine learning for student profiling: An explainable cluster analysis approach. *Computers & Education*, 223, Article 105166. <https://doi.org/10.1016/j.compedu.2024.105166>

Aremu, R., Okfalisa, O., Gbolagade, K. A., Misra, S., Damasevicius, R., & Maskeliunas, R. (2022). A fuzzy-based decision support model for prediction of academic performance of students in higher education. *Sensors*, 22(16), Article 6137. <https://doi.org/10.3390/s22166137>

Asha, S., & Joshith, V. P. (2025). Learning analytics in higher education: A comprehensive bibliometric analysis using the SPAR-4-SLR approach (2013–2023). *Journal of Educational Technology Systems*, 53(4), 337–363. <https://doi.org/10.1177/00472395241300060>

Aulck, L. S., Velagapudi, N., Blumenstock, J., & West, J. D. (2016). Predicting Student Dropout in Higher Education. *ArXiv*, abs/1606.06364. <https://doi.org/10.48550/arxiv.1606.06364>

Baker, R. S., & Hawn, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32, 1041–1066. <https://doi.org/10.1007/s40593-021-00285-9>

Bañeres, D., Rodríguez-González, M. E., Guerrero-Roldán, A.-E., & Karadeniz, I. (2023). An early warning system to identify and intervene online dropout learners. *International Journal of Educational Technology in Higher Education*, 20, Article 3. <https://doi.org/10.1186/s41239-022-00368-y>

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20–29. <https://doi.org/10.1145/1007730.1007735>

Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H.-Y., & Hussain, A. (2022). Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, 28, 905–971. <https://doi.org/10.1007/s10639-022-11152-y>

Bean, J. P., & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research*, 55(4), 485–540. <https://doi.org/10.3102/00346543055004485>

Brahim, G. B., Mabrouk, A. B., & Al-Obeidat, F. (2022). A multi-layer perception approach for predicting student's performance. *International Journal of Information and Education Technology*, 12(6), 503–509. <https://doi.org/10.18178/ijiet.2022.12.6.1651>

Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/a:1010933404324>

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3. [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2)

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3(1), 1–27. <https://doi.org/10.1080/03610927408827101>

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc.

Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *ArXiv*, abs/1106.1813. <https://doi.org/10.1613/jair.953>

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge*

Discovery and Data Mining (pp. 785–794). ACM.
<https://doi.org/10.1145/2939672.2939785>

Choi, S. I. (2024). A study on the development of teaching and learning methods using AI clustering analysis. *International Journal of Internet, Broadcasting and Communication*, 16(1), 113–118. <https://doi.org/10.7236/IJIBC.2024.16.1.113>

Costa-Mendes, R., Cruz-Jesus, F., Oliveira, T., & Castelli, M. (2022). Deep learning in predicting high school grades: A quantum space of representation. *Emerging Science Journal*, 6(Edição Especial), 1–20. <https://doi.org/10.28991/ESJ-2022-SPER-01>

Cox, D. (1958). The Regression Analysis of Binary Sequences. *Journal of the royal statistical society series b-methodological*, 20, 215-232. <https://doi.org/10.1111/j.2517-6161.1958.tb00292.x>

Danaher, M., Waycott, J., & Danaher, P. A. (2023). Student and diversity and inclusion leaders'/administrators' perspectives on bias and equity in the uses of learner data. *Journal of Learning Analytics*, 10(1), 22–38. <https://doi.org/10.18608/jla.2023.7833>

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2), 224–227. <https://doi.org/10.1109/TPAMI.1979.4766909>

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340. <https://doi.org/10.2307/249008>

Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, 10, 1895-1923. <https://doi.org/10.1162/089976698300017197>

Direção-Geral de Estatísticas da Educação e Ciência. (2024). Estatísticas da educação 2022/23 (EDUSTAT). <https://www.dgeec.mec.pt/np4/EstatEducacao/>

Durachman, Y. (2025). Clustering Student Behavioral Patterns: A Data Mining Approach Using K-Means for Analyzing Study Hours, Attendance, and Tutoring Sessions in Educational Achievement. *Artificial Intelligence in Learning*. <https://doi.org/10.63913/ail.v1i1.5>

Etemadpour, R., Zhu, Y., Zhao, Q., Hu, Y., Chen, B., Sharier, M. A., Zheng, S., & Paiva, J. G. (2020). Role of absence in academic success: an analysis using visualization tools. *Smart Learning Environments*, 7. <https://doi.org/10.1186/s40561-019-0112-3>

Eurostat. (2024). Early leavers from education and training — Statistics explained. European Commission. https://ec.europa.eu/eurostat/statistics-explained/index.php/Early_leavers_from_education_and_training

Fahd, K., & Miah, S. J. (2023). Design, development, and evaluation of an AI-based big data analytics system (BDAS) for higher education. *Education and Information Technologies*, 28, 13865–13889. <https://doi.org/10.1007/s10639-022-11426-3>

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>

Fernández-García, A. J., Preciado, J. C., Melchor, F., Rodríguez-Echeverría, R., Conejero, J. M., & Sánchez-Figueroa, F. (2021). A real-life machine learning experience for predicting school dropout at different stages. *IEEE Access*, 9, 136060–136072. <https://doi.org/10.1109/ACCESS.2021.3115851>

Fischer, C., Pardos, Z., Baker, R., Williams, J., Smyth, P., Yu, R., Slater, S., Baker, R. B., & Warschauer, M. (2020). Mining Big Data in Education: Affordances and Challenges. *Review of Research in Education*, 44, 130 - 160. <https://doi.org/10.3102/0091732x20903304>

Gaftandzhieva, S., Talukder, A., Hussain, S., & Doneva, R. (2023). Exploring the possibilities of using learning analytics in higher education. *Sustainability*, 15(1), Article 232. <https://doi.org/10.3390/su15010232>

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends*, 59(1), 64–71. <https://doi.org/10.1007/s11528-014-0822-x>

Gedrimiene, E., Silvola, A., Pursiainen, J., Rusanen, J., & Muukkonen, H. (2019). Learning Analytics in Education: Literature Review and Case Examples From Vocational Education. *Scandinavian Journal of Educational Research*, 64, 1105 - 1119. <https://doi.org/10.1080/00313831.2019.1649718>

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105. <https://doi.org/10.2307/25148625>

Hora, M. T., Bouwma-Gearhart, J., & Park, H. J. (2017). Data use in higher education: Understanding the barriers and possibilities. *Journal of Higher Education*, 88(6), 901–931. <https://doi.org/10.1080/00221546.2017.1291259>

Hung, H.-C., Liu, I., Liang, C.-T., & Su, Y.-S. (2020). Applying Educational Data Mining to Explore Students' Learning Patterns in the Flipped Learning Approach for Coding Education. *Symmetry*, 12, 213. <https://doi.org/10.3390/sym12020213>

Jayaprakash, S. M., Moody, E. W., Lauría, E. J. M., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, 1(1), 6–47. <https://doi.org/10.18608/jla.2014.11.3>

Jones, K. M. L., Goben, A., Perry, M. R., Regalado, M., Salo, D., Asher, A. D., Smale, M., & Briney, K. A. (2023). Transparency and Consent: Student Perspectives on Educational Data Analytics Scenarios. *portal: Libraries and the Academy*, 23, 485 - 515. <https://doi.org/10.1353/pla.2023.a901565>

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457–481. <https://doi.org/10.1080/01621459.1958.10501452>

Karunaratne, T. (2021). For Learning Analytics to Be Sustainable under GDPR—Consequences and Way Forward. *Sustainability*. <https://doi.org/10.3390/su132011524>

Kassarnig, V., Bjerre-Nielsen, A., Mones, E., Lehmann, S., & Lassen, D. (2017). Class attendance, peer similarity, and academic performance in a large field study. *PLoS ONE*, 12. <https://doi.org/10.1371/journal.pone.0187078>

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 3, Article 100074. <https://doi.org/10.1016/j.caeai.2022.100074>

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling* (3.^a ed.). John Wiley & Sons.

Klenke, J., Massing, T., & Hanck, C. (2022). How directed acyclic graphs can help us to draw causal claims in learning analytics. *Journal of Learning Analytics*, 9(3), 85–104. <https://doi.org/10.18608/jla.2022.7575>

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Em Proceedings of the 14th International Joint Conference on Artificial Intelligence* (pp. 1137–1145). Morgan Kaufmann.

Linstedt, D., & Olschimke, M. (2015). *Building a scalable data warehouse with Data Vault 2.0*. Morgan Kaufmann.

Liu, Q., & Khalil, M. (2023). Understanding privacy and data protection issues in learning analytics using a systematic review. *Br. J. Educ. Technol.*, 54, 1715-1747. <https://doi.org/10.1111/bjet.13388>

Liu, Z., Zhou, X., & Liu, Y. (2025). Student Dropout Prediction Using Ensemble Learning with SHAP-Based Explainable AI Analysis. *Journal of Social Systems and Policy Analysis*. <https://doi.org/10.62762/jsspa.2025.321501>

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. 4765-4774. <https://doi.org/10.4230/oasics.dx.2024.27>

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. Em L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). University of California Press.

Martins, P. S., Nunes, L. C., Reis, P., & Thomas, T. (2025). A expansão dos cursos profissionais em Portugal: Que impacto na educação, no emprego e no empreendedorismo? (Policy Papers da Fundação, #04). Fundação Francisco Manuel dos Santos. <https://ffms.pt/pt-pt/estudos/policy-papers/expansao-dos-cursos-profissionais-em-portugal-que-impacto-na-educacao-no-emprego-e-no>

Matz, S. C., Bukow, C. S., Peters, H., Deacons, C., Dinu, A., & Stachl, C. (2023). Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. *Scientific Reports*, 13, Article 5705. <https://doi.org/10.1038/s41598-023-32484-w>

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. John Wiley & Sons.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157. <https://doi.org/10.1007/BF02295996>

Mduma, N. (2023). Data balancing techniques for predicting student dropout using machine learning. *Data*, 8(3), Article 49. <https://doi.org/10.3390/data8030049>

Milligan, G. W., & Cooper, M. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, 5, 181-204. <https://doi.org/10.1007/bf01897163>

Nafuri, A. F. M., Sani, N., Zainudin, N. F. A., Rahman, A. H. A., & Aliff, M. (2022). Clustering Analysis for Classifying Student Academic Performance in Higher Education. *Applied Sciences*. <https://doi.org/10.3390/app12199467>

Ndukwe, I. G., & Daniel, B. K. (2020). Teaching analytics, value and tools for teacher data literacy: A systematic and tripartite approach. *International Journal of Educational Technology in Higher Education*, 17, Article 22. <https://doi.org/10.1186/s41239-020-00201-6>

Newman-Ford, L., Fitzgibbon, K., Lloyd, S., & Thomas, S. (2008). A large-scale investigation into the relationship between attendance and attainment: a study using an innovative, electronic attendance monitoring system. *Studies in Higher Education*, 33, 699 - 717. <https://doi.org/10.1080/03075070802457066>

Omar, T., Alzahrani, A. M., & Zohdy, M. (2020). Clustering Approach for Analyzing the Student's Efficiency and Performance Based on Data. *Journal of Data Analysis and Information Processing*. <https://doi.org/10.4236/jdaip.2020.83010>

Palacios, C. A., Reyes-Suarez, J., Bearzotti, L. A., Leiva, V., & Marchant, C. (2021). Knowledge Discovery for Higher Education Student Retention Based on Data Mining: Machine Learning Algorithms and Case Study in Chile. *Entropy*, 23. <https://doi.org/10.3390/e23040485>

Palacios, C., Reyes-Suarez, J., Lazo-Salas, P., Chamba-Eras, L., & García, M. V. (2021). Machine learning model for the prediction of academic performance of students in a university. Em K. Arai (Ed.), *Advances in information and communication* (Vol. 1363, pp. 513–525). Springer. https://doi.org/10.1007/978-3-030-73103-8_37

Pamungkas, L., Dewi, N. A., & Putri, N. A. (2024). Classification of Student Grade Data Using the K-Means Clustering Method. *Jurnal Sisfokom (Sistem Informasi dan Komputer)*. <https://doi.org/10.32736/sisfokom.v13i1.1983>

Pecuchova, J., & Drlik, M. (2024). Enhancing the early student dropout prediction model through clustering analysis of students' digital traces. *IEEE Access*, 12, 56218–56236. <https://doi.org/10.1109/ACCESS.2024.3388070>

Peffer, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>

Petrović, A., Jovanović, L., Stanković, M., Živković, M., & Bacanin, N. (2024). Identifying and understanding student dropouts using metaheuristic-optimized machine learning classifiers and explainable artificial intelligence techniques. *IEEE Access*, 12, 49669–49692. <https://doi.org/10.1109/ACCESS.2024.3385178>

Quille, K., & Bergin, S. (2022). PreSS: Predicting student success early in CS1: A pilot international replication and generalization study. Em *Proceedings of the 2022*

ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE '22) (pp. 41–47). Association for Computing Machinery. <https://doi.org/10.1145/3502718.3524755>

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14. <https://doi.org/10.1038/s41598-024-56706-x>

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. Em L. Liu & M. T. Özsu (Eds.), *Encyclopedia of database systems* (pp. 532–538). Springer. https://doi.org/10.1007/978-0-387-39940-9_565

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10. <https://doi.org/10.1002/widm.1355>

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)

Santos, G. M., Gonçalves, C. T., Gonçalves, M. J. A., & Silva, A. J. A. (2026). A machine learning approach to dropout prediction in vocational secondary education. Em *Proceedings of ICITED 2026*. Springer Lecture Notes in Networks and Systems.

Schmid, E., & Garrels, V. (2021). Parental involvement and educational success among vulnerable students in vocational education and training. *Educational Research*, 63, 456 - 473. <https://doi.org/10.1080/00131881.2021.1988672>

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, 181, 526–534. <https://doi.org/10.1016/j.procs.2021.01.199>

Staneviciene, E., Gudoniene, D., Punys, V., & Kukstys, A. (2024). A Case Study on the Data Mining-Based Prediction of Students' Performance for Effective and Sustainable E-Learning. *Sustainability*. <https://doi.org/10.3390/su162310442>

Sun, J. C. (2023). Gaps, guesswork, and ghosts lurking in technology integration: Laws and policies applicable to student privacy. *Br. J. Educ. Technol.*, 54, 1604-1618. <https://doi.org/10.1111/bjet.13379>

Syakur, M., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering*, 336. <https://doi.org/10.1088/1757-899x/336/1/012017>

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1), 89–125. <https://doi.org/10.3102/00346543045001089>

Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition* (2.^a ed.). University of Chicago Press.

Torres-Cruz, F., Pari-Condori, E. Y., Tumi-Figueroa, E. N., Coyla-Idme, L., Tito-Lipa, J., Gonzalez, L. A., & Tumi-Figueroa, A. (2025). Prediction of university dropouts through random forest-based models. *Journal of Advanced Pharmacy Education & Research*, 15(1), 78–83.

Umargono, E., Suseno, J. E., & Gunawan, S. K. V. (2020). K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula. *Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019)*. <https://doi.org/10.5220/0009908402340240>

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A framework for evaluation in design science research. *European Journal of Information Systems*, 25(1), 77–89. <https://doi.org/10.1057/ejis.2014.36>

Villegas-Ch, W., Palacios-Pacheco, X., & Luján-Mora, S. (2020). Application of a business intelligence method for the analysis of learning outcomes in a higher education institution. *Sustainability*, 12(14), Article 5745. <https://doi.org/10.3390/su12145745>

Waheed, H. H. (2020). The effect of using a deep learning-based framework on students' engagement in online learning environment. *International Journal of Distance Education Technologies*, 18(4), 56–74. <https://doi.org/10.4018/IJDET.2020100104>

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236–244. <https://doi.org/10.1080/01621459.1963.10500845>

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9. <https://doi.org/10.1186/s40561-022-00192-z>

Apêndice I – Instrumento de Avaliação e Estrutura da Sessão

A.1 Enquadramento e Propósito

Este apêndice documenta o processo de avaliação do artefacto desenvolvido, conduzido a 9 de junho de 2026, através de uma videoconferência na plataforma Microsoft Teams. A avaliação foi estruturada para recolher evidência quantitativa e qualitativa sobre a utilidade percebida, a usabilidade e a potencial adoção do sistema de apoio à decisão junto dos seus destinatários finais. O instrumento foi desenvolvido com base nos construtos do Technology Acceptance Model (Davis, 1989) e no enquadramento FEDS de Venable et al. (2016) para avaliação de artefactos em Design Science Research.

A.2 Estrutura da Sessão de Avaliação

A sessão foi realizada com a participação de quatro responsáveis da Profitecla com funções de gestão pedagógica e operacional, e seguiu um formato estruturado em cinco segmentos:

Segmento 1 — Abertura e Contextualização (5 minutos)

O investigador apresentou brevemente o enquadramento da investigação, o processo de desenvolvimento do artefacto ao longo de três meses, e o mapeamento das seis fases da Design Science Research à aplicação prática. Foi enfatizado que a sessão constituía a Fase 5 (Avaliação) do processo de investigação, e que a opinião dos participantes era essencial para validar a utilidade operacional do sistema.

Segmento 2 — Apresentação do Processo Metodológico (8 minutos)

Apresentação visual em HTML das 6 etapas do processo CRISP-DM em linguagem de negócio, sem jargão técnico:

- Compreender (reuniões com a Profitecla)
- Explorar (6 anos de dados, 1.086 alunos)
- Preparar (limpeza e construção de variáveis)
- Modelar (comparação de algoritmos)
- Avaliar (sessão em curso)
- Implementar (sistema ativo com 396 alunos 2024/25)

Incluiu também contexto de mercado, sobre não existir solução equivalente no ensino profissional português; o sistema é calibrado para os dados da Profitecla; o framework é transferível a outros polos.

Segmento 3 — Apresentação dos Resultados (25 minutos)

Demonstração ao vivo dos dashboards desenvolvidos em Power BI, abrangendo:

- Perfis de Alunos (Clustering)
- Apresentação dos três clusters: Sucesso (86% conclusão), Risco Intervencionável (66% conclusão), Abandono (~0% conclusão)
- Explicação do funcionamento e da aplicação prática para priorização de intervenção
- Modelos Preditivos — Dois Horizontes Temporais
- Modelo M1b (na matrícula): com dados limitados de entrada
- Modelo M1a (ao longo do percurso): com dados comportamentais acumulados
- Apresentação em linguagem operacional dos resultados
- Análise da deterioração entre os dois modelos e discussão das implicações para intervenção
- Demonstração do Power BI
- Semáforo de risco por aluno
- Alunos que deterioraram desde a matrícula
- Análise por curso e escola de origem
- Explicabilidade dos resultados (SHAP — importância de variáveis)
- Limitações Identificadas
- Normalização de escolas de origem: 716 variações de nomes → recomendação de campo pré-definido
- Qualidade dos dados históricos
- Capacidade de replicação em outros polos com ressalvas de adaptação local
- Estratégia de Atualização
- Estrutura do loop mensal/trimestral de atualização
- Requisitos operacionais para sustentabilidade

Segmento 4 — Discussão Estruturada (12 minutos)

Discussão guiada por questões abertas orientadas para validar a aderência do artefacto à prática institucional:

- "Os três perfis identificados correspondem à forma como acompanham informalmente os alunos?"
- "A concentração do abandono no primeiro ano coincide com a vossa percepção?"
- "Que limitações vêm na utilização prática do sistema?"
- "Que condições seriam necessárias para uma implementação sustentável?"
- "Há interesse em expandir a outros polos? Se sim, que abordagem vos parece mais realista?"

Segmento 5 — Aplicação do Questionário (10 minutos)

Distribuição do questionário estruturado em Google Forms, com instruções orais breves sobre a escala Likert 1–5 e a expectativa de tempo de resposta (~5 minutos). Os participantes responderam de forma independente.

A.3 Instrumento de Avaliação — Questionário Estruturado

Instruções para Aplicação

Enquadramento para o respondente:

As afirmações que se seguem referem-se ao sistema de apoio à decisão que lhe foi apresentado. Avalie cada uma numa escala de 1 (Discordo totalmente) a 5 (Concordo totalmente). Não existem respostas certas ou erradas, a sua opinião honesta é o que importa. O tempo estimado de resposta é de 5 minutos.

Secção A — Utilidade Percebida

A1. Este sistema fornece informação útil para apoiar decisões de intervenção junto de alunos em risco.

Escala: 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

A2. A utilização deste sistema melhoraria a capacidade da Profitecla de identificar alunos em risco atempadamente.

Escala: 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

A3. Os perfis de alunos identificados (Sucesso / Risco Intervencionável / Abandono) refletem situações reais que reconheço na instituição.

Escala: 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

A4. A informação sobre escola de origem e perfil de entrada é relevante para as decisões de captação de novos alunos.

Escala: 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Secção B — Facilidade de Utilização Percebida

B1. As visualizações dos dashboards são claras e fáceis de interpretar.

Escala: 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

B2. O sistema semafórico (verde / amarelo / vermelho) é intuitivo para utilização operacional.

Escala: 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

B3. Seria fácil aprender a utilizar este sistema de forma autónoma no dia-a-dia.

Escala: 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Secção C — Relevância, Transferibilidade e Adoção

C1. Os resultados apresentados são congruentes com o conhecimento que tenho dos alunos da Profitecla.

Escala: 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

C2. Este sistema deveria ser expandido aos outros polos da Profitecla.

Escala: 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

C3. Utilizaria este sistema regularmente se estivesse disponível e atualizado.

Escala: 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Secção D — Avaliação Global

D1. De forma geral, este sistema constitui uma ferramenta útil para a gestão pedagógica e operacional da Profitecla.

Escala: 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐

Secção E — Questões Abertas

E1. Que informação adicional gostaria de ver neste sistema?

Espaço de resposta:

E2. Que limitações ou preocupações identifica na utilização deste sistema na Profitecla?

Espaço de resposta:

E3. Que sugestão daria para tornar este sistema mais útil no seu trabalho?

Espaço de resposta:

A.4 Legenda e Interpretação da Escala

Escala Likert 1–5:

1 = Discordo totalmente — A afirmação não descreve a sua opinião ou experiência

2 = Discordo — A afirmação descreve parcialmente a sua opinião, mas com reservas significativas

3 = Neutro / Indeciso — Não tem uma opinião clara ou a afirmação é neutro

4 = Concordo — A afirmação descreve amplamente a sua opinião, com pequenas ressalvas

5 = Concordo totalmente — A afirmação descreve completamente a sua opinião ou experiência

A.5 Protocolo de Anonimização

Os resultados apresentados no Capítulo 5 reportam agrupamentos de respostas, não individualizações. As observações qualitativas recolhidas durante a discussão estruturada são referidas por função genérica (ex.: "um dos responsáveis") em vez de identificação nominal, garantindo a confidencialidade dos participantes.

A.6 Notas Operacionais para a Aplicação

Tempo total estimado da sessão: 60 minutos (distribuição aproximada: abertura 5' + processo 8' + resultados 25' + discussão 12' + questionário 10')

Materiais necessários:

Link do formulário de recolha em Google Forms

Apresentação visual (partilha de ecrã)

Acesso ao Power BI Desktop ou ambiente equivalente para demonstração ao vivo

Espaço: Sala de reunião com capacidade para 5–8 pessoas, mesa central, acesso a ecrã ou projetor para visualização de dashboards

Dinâmica de facilitação: O investigador conduziu toda a sessão; os participantes foram incentivados a colocar questões de clarificação durante a apresentação dos resultados e a discussão foi conduzida de forma não-diretiva, evitando influenciar as respostas qualitativas

Recolha de dados: Os questionários foram aplicados através de formulário Google ao final da sessão e posteriormente tabulados para análise quantitativa. As notas da discussão foram registadas através da ferramenta Granola (ferramenta de Inteligência Artificial para registo de reuniões e notas).

A.7 Referências do Instrumento

Davis, F. A. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340.

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A framework for evaluation in design science research. *European Journal of Information Systems*, 25(1), 77–89.

Apêndice II – Publicações

No âmbito desta investigação foram produzidas as seguintes publicações científicas:

Trabalho aceite para publicação

Santos, G. M., Gonçalves, C. T., Gonçalves, M. J. A., & Silva, A. J. A. (no prelo). A machine learning approach to dropout prediction in vocational secondary education. Em Proceedings of the International Conference on Information Technology & Education (ICITED 2026) (Lecture Notes in Networks and Systems). Springer.

Trabalhos em preparação

Santos, G. M., Gonçalves, C. T., Gonçalves, M. J. A. (2026). A business intelligence decision support system for early detection of dropout risk in vocational education: architecture and dashboards. Manuscrito em preparação.