



Marta Sofia Morgado Simões

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Coimbra, outubro de 2023



Marta Sofia Morgado Simões

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Trabalho de projeto submetido ao Instituto Superior de Contabilidade e Administração de Coimbra para cumprimento dos requisitos necessários à obtenção do grau de **Mestre em Análise de Dados e Sistemas de Apoio à Decisão**, realizado sob a orientação da Professora Doutora Isabel Pedrosa, coorientação da Professora Doutora Joana Jorge de Queiroz Leite e supervisão de Engenheiro António Manuel de Brito Paulino.

Coimbra, outubro de 2023

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

TERMO DE RESPONSABILIDADE

Declaro ser a autora deste projeto, que constitui um trabalho original e inédito, que nunca foi submetido a outra Instituição de ensino superior para obtenção de um grau académico ou outra habilitação. Atesto ainda que todas as citações estão devidamente identificadas e que tenho consciência de que o plágio constitui uma grave falta de ética, que poderá resultar na anulação do presente projeto.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

AGRADECIMENTOS

Agradeço à professora Isabel Pedrosa, professora Joana Leite e professor António Paulino, que me acompanharam ao longo desta caminhada de quase um ano, ajudaram-me a ultrapassar todos os obstáculos que foram surgindo e a encontrar o melhor caminho a seguir para o estudo.

Agradeço a todas as pessoas mais próximas de mim nesta caminhada, amigos e família. Em especial, quero agradecer à minha amiga Eva e ao meu amigo Ibraimo que me deram sempre uma palavra de apoio quando eu achava que não conseguia avançar. Quero agradecer também ao Alberto, pela paciência, pelo apoio, pelos conselhos e por nunca me deixar desistir. Em último, mas não menos importante, quero agradecer aos meus pais que mesmo à distância me conseguiram apoiar e celebraram cada etapa alcançada mais do que eu, o que me deu força para fazer o meu melhor.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

RESUMO

O Abandono Escolar é um problema global que afeta todos os países. No Ensino Superior, este tema é um problema particularmente complexo, na medida em que poderá originar a várias consequências negativas para os estudantes, as famílias e até mesmo para as entidades de ensino. O objetivo geral deste projeto de mestrado é desenvolver um modelo de previsão de Abandono Escolar no Instituto Politécnico de Coimbra. Dentro deste destaca-se a identificação dos fatores que podem levar ao Abandono no IPC, preparação dos dados para integração no modelo capaz de identificar a problemática. Ao longo do projeto é tido em conta a metodologia CRISP-DM para o desenvolvimento do modelo de deteção de Abandono Escolar. Esta metodologia é variante de *data mining* amplamente utilizada em projetos similares. O projeto foi desenvolvido com recurso a dados concretos da instituição na qual se pretende aplicar o modelo, com auxílio a diversos dados dos alunos como média, notas das UC, curso, nota de ingresso, entre outras. Para o modelo final, todas os dados foram submetidos ao algoritmo KNN, atualmente em vigor para combate a outras vertentes do tema. Sendo que, o modelo final com melhores resultados apresenta uma precisão de 76%.

Palavras-chave: Abandono escolar, Crisp-dm, KNN, *Machine learning*, *Python*

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

ABSTRACT

School dropout is a global problem that affects all countries. In Higher Education, this is a particularly complex problem, as it can have several negative consequences for students, families, and educational institutions, as well as several other negative externalities to societies at large. The overall aim of this master's project is to develop a dropout prevention model for the Polytechnic Institute of Coimbra. This includes identifying the factors that can lead to withdrawal from IPC and preparing the data for integration into a model capable of identifying the problem. Throughout the project, the CRISP-DM methodology is considered in the development of the dropout detection model. This methodology is a variant of data mining widely used in similar projects. The project was developed using specific data from the institution in which the model is to be applied, with the help of various student data such as grade point average, course grades and entry grades, among others. For the final model, all the data was submitted to the KNN algorithm, which is currently used to tackle other aspects of the subject. The final model with the best results has an accuracy of 76%.

Keywords: Dropout, Crisp-dm, KNN, Machine learning, Python

ÍNDICE GERAL

INTRODUÇÃO.....	1
1 Revisão da literatura	7
1.1 Posicionamento do AE no ES	7
1.2 <i>Dropout</i> , a diferença entre os termos <i>optout</i> e <i>stopout</i>	8
1.3 Fatores que podem levar ao AE do ES	9
1.4 Dimensões associadas com o AE.....	13
1.5 Quadro resumo dos fatores	14
2 Entendimento do problema da instituição	17
3 Entendimento e preparação dos dados	19
3.1 Conjunto de dados inicial	19
3.2 Análise descritiva dos dados do conjunto de dados inicial	21
3.3 Conjunto de dados modificado	27
3.3.1 1ª iteração.....	27
3.3.2 2ª iteração.....	34
3.4 Análise descritiva do conjunto de dados modificado	45
3.5 <i>Feature scaling</i>	48
4 Modelação	52
4.1 KNN	52
4.2 Árvore de Decisão	53
4.3 Matriz de confusão, accuracy, recall, precision, f1-score	54
4.4 Aplicação do algoritmo KNN.....	57
4.4.1 1º iteração.....	57

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

4.4.2	2º iteração.....	58
5	Resultados.....	62
5.1	KNN	62
5.2	Árvore de decisão	75
6	Implementação.....	77
	CONCLUSÃO.....	79
	REFERÊNCIAS BIBLIOGRÁFICAS	81
	APÊNDICES	84
	Apendice I – Variáveis do conjunto de dados final.....	84
	Apendice II – Variáveis categóricas, numéricas e de saída.....	85
	Apendice III – Acesso ao código python e ficheiros.....	85

ÍNDICE DE TABELAS

Figura 1 - Esquema das etapas da metodologia CRISP-DM.	3
Figura 2 - Análise descritiva das variáveis MATR_ID, média e nota de ingresso.....	21
Figura 3 - Gráfico de barras com distribuição da amostra pelos estados civis.....	22
Figura 4 - Gráfico de barras com a distribuição do número de alunos por total de inscrições no curso	23
Figura 5 - Gráfico circular da distribuição da variável "bolseiro"	23
Figura 6 - Distribuição dos alunos pelos estatutos	24
Figura 7 - Distribuição dos alunos pelos com mais de 100 alunos	25
Figura 8 – Gráfico de barras com a distribuição dos alunos pelas opções de candidatura..	25
Figura 9 – Gráfico de barras com as variáveis “estado da matrícula” e “renovou ano seguinte”	26
Figura 10 – Gráfico circular com a distribuição dos alunos na variável IFRE_Modo	26
Figura 11 - Código em python para construção da variável "Total sup 10"	29
Figura 12 - Descrição e gráfico circular respetivos à variável "regimes de ingresso"	31
Figura 13 - Código em python da divisão da coluna de "estatutos"	32
Figura 14 - Descrição e gráfico circular respetivos à variável "estatutos"	32
Figura 15 - Descrição dos estatutos pertencentes à variável "estatuto_falta de pagamento de propinas e emolumentos"	33
Figura 16 - Código python com o tratamento de dados para a criação da variável "abandono"	34
Figura 17 - Definição do conjunto de dados "df_local"	38
Figura 18 - Estrutura do conjunto de dados "df_local"	39
Figura 19 - Coordenadas de cada uma das escolas pertencentes ao estudo	40
Figura 20 – Código em python do cálculo da distância de cada uma das coordenadas dos alunos às coordenadas das escolas	40
Figura 21 - Código em python para cesso a base de dados com coordenadas	41
Figura 22 - Exemplo do conjunto de dados das variáveis das instituições.....	42
Figura 23 - Exemplo do conjunto de dados com instituição, curso e número de alunos	42
Figura 24 - Exemplo do conjunto de dados com as variáveis derivadas do estado civil.....	43

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Figura 25 - Exemplo do conjunto de dados com variáveis derivadas de "disciplinas" e "ECTS"	44
Figura 26 - Código em python da transformação que originou a variável "abandono"	45
Figura 27 - Gráfico de barras com a distribuição do número de disciplinas a que um aluno está inscrito no ano letivo 2020/2021	46
Figura 28 - Gráfico de barras com a distribuição do número de disciplinas realizadas com aprovação por aluno, no ano letivo 2020/2021.....	47
Figura 29 - Gráfico circular da variável "proximidade"	47
Figura 30 - Gráfico circular com descrição de como se divide a variável "abandono"	48
Figura 31 - Representação da distribuição da variável "rácio médias"	48
Figura 32 - Ilustração do resultado dos diferentes tipos de feature scaling.....	49
Figura 33 - Exemplo gráfico de como se agrupam os dados após aplicação de normalização e padronização	51
Figura 34 - Código em python para aplicação da normalização na 1ª iteração	57
Figura 35 - Código em python para o algoritmo KNN.....	58
Figura 36 - Código em python que identifica e separa as variáveis categóricas e numéricas automaticamente	59
Figura 37 - Exemplo do resultado da importância das variáveis na 2ª iteração	60
Figura 38 - Representação gráfica do grau de importância das variáveis categóricas na 2ª iteração.....	60
Figura 39 - Representação gráfica do grau de importância das variáveis numéricas na 2ª iteração.....	61
Figura 40 - Resultados do algoritmo da 1ª iteração.....	62
Figura 41 - Código em python para a execução do algoritmo KNN na 2ª iteração	65
Figura 42 - Resultado do algoritmo KNN da 2ª iteração - opção 1.....	66
Figura 43 - Resultado de variáveis do algoritmo KNN da 2ª iteração - opção 1.....	66
Figura 44 - Resultados do algoritmo KNN na partição de teste da 2ª iteração - opção 1 ...	67
Figura 45 - Resultado do algoritmo KNN da 2ª iteração - opção 2.....	68
Figura 46 - Resultado de variáveis do algoritmo KNN da 2ª iteração - opção 2.....	69
Figura 47 - Resultados do modelo KNN na 2ª iteração – opção 3	69

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Figura 48 - Resultado de variáveis do algoritmo KNN da 2ª iteração - opção 3.....	70
Figura 49 - Resultado do modelo no teste opção 3.....	71
Figura 50 - Resultados do algoritmo KNN para o curso de Biotecnologia na 2ª iteração ..	72
Figura 51 - Resultados do algoritmo KNN para o curso de Comunicação Organizacional na 2ª iteração	72
Figura 52 - Resultados do algoritmo KNN para o curso de Fisioterapia na 2ª iteração	73
Figura 53 - Resultados do algoritmo KNN para os cursos de Engenharia Eletrotécnica, Gestão de Empresas e Engenharia Informática na 2ª iteração.....	73
Figura 54 - Resultados do modelo para cada instituição na 2ª iteração	74
Figura 55 - Esquema do modelo.....	75
Figura 56 - Resultados do modelo com o algoritmo Árvores de Decisão.....	76
Figura 57 - Representação gráfica da implementação do algoritmo KNN.....	78
Figura 58 - Conjunto final de variáveis - 2ª iteração.....	85
Figura 59 - Descrição de variáveis categóricas, numéricas e saída - 2ª iteração.....	85

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

ÍNDICE DE FIGURAS

Tabela 1 - Fatores da dimensão organizacional.....	14
Tabela 2 - Fatores da dimensão da gestão de vida pessoal.....	15
Tabela 3 - Fatores da dimensão profissional e relacional/social	16
Tabela 4 - Listagem de variáveis do conjunto de dados inicial.....	19
Tabela 5 - Listagem das variáveis resultantes da 1ª iteração.....	28
Tabela 6 - Listagem das variáveis resultantes da 2ª iteração.....	35
Tabela 7 - Matriz de confusão	55

Lista de abreviaturas, acrónimos e siglas

AE – Abandono Escolar

BI – *Business Intelligence*

DGEEC – Direção Geral de Estatísticas da Educação e Ciência

ECTS - Sistema Europeu de Transferência e Acumulação de Créditos

ES – Ensino Superior

IPC – Instituto Politécnico de Coimbra

KNN - *K-nearest neighbors*

ML – *Machine Learning*

NEE – Necessidades Especiais de Ensino

UC – Unidades Curriculares

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

INTRODUÇÃO

Enquadramento e motivação

A cada dia que passa, são cada vez mais os jovens que associam o sucesso escolar e pessoal com o ingresso e conclusão do Ensino Superior (ES). O que se pode verificar quando apesar do ES não ser o ensino obrigatório, em 2022 foram mais de 446 mil estudantes que pertenciam ao mesmo.

Ainda que, seguir para o ES seja “o novo normal”, não deixa de ser uma decisão determinante para o futuro dos jovens que estão prestes a terminar o ensino obrigatório. Com esta determinação vem também a indecisão, ansiedade e dúvida para o seu futuro. E se a grande maioria consegue ultrapassar estas pequenas barreiras no ES, alguns dos estudantes não encontram as forças necessárias ou não têm apoios suficientes e tomam a decisão de partir para outra aventura que não a conclusão do ES.

Segundo vários estudos fidedignos publicados e entidades como a DGEEC – Direção Geral de Estatísticas da Educação e Ciência, a percentagem de alunos que abandonaram o ES um ano após terem iniciado o seu percurso académico tem aumentado de ano para ano. Os dados obtidos dizem respeito ao ano letivo de 2020/2021 e contam com 10.8% de abandono no grau de Licenciatura. Este valor mais recente representa um aumento de quase dois pontos percentuais face aos últimos dados disponibilizados de 9.1% do ano letivo anterior (2019/2020).

Ao sermos confrontados com valores tão altos, as perguntas que se impõem são qual a verdadeira origem da elevada taxa de Abandono Escolar (AE) e o que poderemos implementar para que a mesma diminua. Assim sendo, o primeiro passo será a tentativa de identificação dos motivos que levam a este fenómeno preocupante. Ao abordar este tema serão sempre muitas as possíveis razões que podemos considerar, como as áreas dos cursos que estão a frequentar, a empregabilidade, o método de ensino, dificuldades financeiras, situação profissional, entre outros. As efetivas causas não são fáceis de encontrar, no entanto as condições financeiras destacam-se, sendo, muitas vezes, apontada como a razão principal para este fenómeno, que possivelmente terá sido

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

agravado pelo corte nas bolsas aos jovens universitários. Ou mesmo o aumento de propinas mensais que temos acompanhado ao longo dos anos.

Tendo em conta que o AE está a aumentar de dia para dia e que as razões poderão ser tão variáveis é necessário que seja um tema a que se deva dar realmente importância. Neste projeto o tema do AE será abordado como urgente para mudar de rumo. Ou seja, é fundamental que se identifiquem quais os fatores que estão a contribuir para esta causa, quais são os principais motivos que levam estudantes a desistir e quais os sinais transmitidos que nos possam levar à diminuição da taxa de abandono escolar no ES.

Objetivos

Neste seguimento, definem-se como objetivos deste projeto apurar quais os fatores que podem ajudar a prever o AE no ES especificamente aplicado a um conjunto de unidades orgânicas de um instituto de ES. Deste modo, o objetivo final passa por criar uma ferramenta que possa ser utilizada a longo prazo para monitorizar e diminuir as demais desistências do ES. Será aplicado a um caso de estudo, concretamente o Instituto Politécnico de Coimbra (IPC).

O IPC é uma das dez maiores instituições de ES portuguesas, e é constituído pelas seguintes escolas:

- ESAC – Escola Superior Agrária
- ESEC – Escola Superior de Educação
- ESTGOH – Escola Superior de Tecnologia e Gestão
- ESTeSC – Escola Superior de Tecnologia e Saúde
- ISCAC – Instituto Superior de Contabilidade e Administração
- ISEC – Instituto Superior de Engenharia

Para a execução foi necessário recorrer aos dados dos alunos na base do NÓNIO, que posteriormente foram a base da criação de um *Data Lake* que agrega as informações dos alunos de diversas bases de dados que foi necessário para a análise de dados. Além disso, foram implementados algoritmos de inteligência artificial, especificamente o *K-nearest neighbors* (KNN) e introdução às árvores de decisão, para detetar padrões nos dados relacionados com o AE e que podem indicar aos docentes os alunos que poderão estar em risco de abandono, meramente como sinal de alerta. Para estes algoritmos executarem da

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

melhor maneira e mostrarem os resultados mais próximos da realidade, visto que se trata de dados que transmitem probabilidades e não certezas, é necessário que passem pelo tratamento de dados adequado. Assim, qual o tratamento mais indicado para os dados em estudo foi também parte dos grandes objetivos do projeto. Todos estes processos mencionados foram auxiliares para criar mecanismos de análise e visualização de dados em tempo real, bem como notificar possíveis casos de AE.

Metodologia

No que diz respeito à metodologia será aplicada a Crisp-dm. Esta é composta por seis etapas: entendimento do negócio, entendimento dos dados, preparação dos dados, modelação, avaliação e implementação. Em que, cada uma destas etapas tem um papel fundamental para o sucesso do projeto e contribuirá para alcançar o objetivo proposto, no entanto não existe a necessidade de ser feita apenas um tipo de sequência. A sua sequência não é rígida, ou seja, é possível que se volte atrás nas fases sempre que necessário. As setas apresentadas apenas indicam as dependências mais importantes e frequentes, como podemos ver na Figura 1.

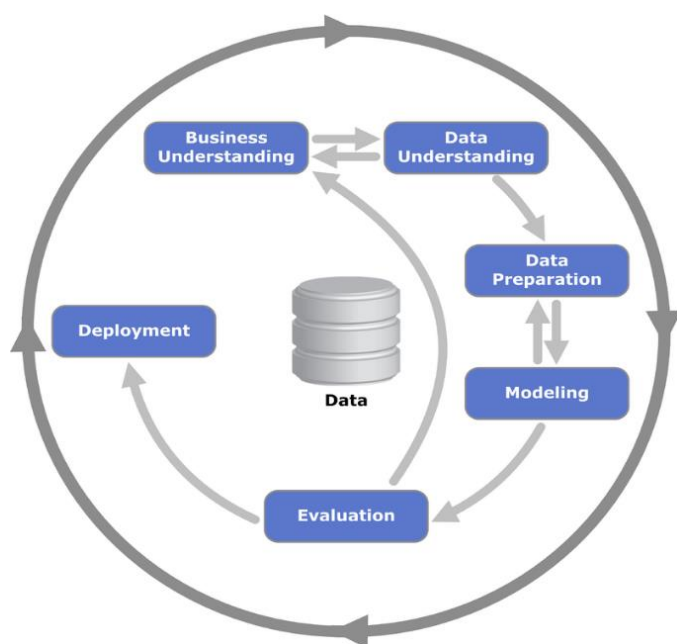


Figura 1 - Esquema das etapas da metodologia CRISP-DM.

Fonte: <https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/>

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

A primeira etapa, entendimento do negócio, irá estabelecer os objetivos do projeto, tendo inerente a recolha de toda a informação respetiva ao negócio, a descrição clara dos objetivos e quais serão os critérios de sucesso do negócio. Após esta definição, será pertinente a descrição da situação que se está perante, tendo uma descrição dos recursos disponíveis, quais os requisitos, pressupostos e até mesmo limitações, descrição dos riscos e ainda apresentação de custos e benefícios. Por fim, todas as informações anteriores irão ajudar na produção do plano de projeto.

Na segunda etapa, entendimento dos dados, dá-se a recolha e exploração os dados disponibilizados pela instituição em questão. Será ainda a altura para identificar as variáveis que podem influenciar o AE e perceber como é que elas se relacionam. Com estes dados, será possível produzir um relatório de exploração dos dados e verificar a qualidade dos mesmos.

A etapa de preparação dos dados irá incluir os processos de limpeza e transformação dos dados para que estes possam seguir para a análise. Será o momento em que se obtém o conjunto de dados final, dados que serão carregados na ferramenta de modelação pretendida. Esta tarefa poderá ser efetuada diversas vezes, em qualquer altura do projeto de modo que se obtenham os melhores resultados.

Após os dados estarem preparados para serem trabalhados, serão selecionados os modelos que vão permitir atingir o objetivo do projeto, constituindo assim a etapa da modelação. Ainda nesta etapa, após a seleção da técnica de modelação é criada uma conceção de teste, em que se criam procedimentos e mecanismos para testar a qualidade do modelo escolhido. E, por fim a construção e revisão do modelo.

Para serem modelos com maior precisão necessitam de ser submetidos à etapa de avaliação, ou seja, onde será feita a avaliação dos resultados, e se os mesmos estão de acordo com os objetivos pré-definidos, sendo assim feita a aprovação ou não do modelo escolhido. Neste ponto é importante fazer uma revisão de todo o processo, o que poderá originar que os passos seguintes a efetuar sejam relativos a tarefas anteriores, para assim se conseguir obter os melhores resultados possíveis.

Por último, na sexta etapa, a implementação será considerada a etapa onde se apresentam

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

as conclusões do projeto, como será efetuada a implementação do mesmo e algumas considerações finais. Com isto, será esperado que haja uma melhor compreensão dos fatores que levam ao AE no ES e ajudar a desenvolver estratégias para reduzir esse problema. Especificamente, será efetuada uma primeira visualização da implementação do projeto no IPC, sendo necessário recorrer a comunicações com os responsáveis e futuros utilizadores do modelo. Com isto, é esperado que sejam acrescentadas ao projeto melhorias, e ainda, futuramente haja uma maior adesão do projeto e maior assertividade de resultados.

Estrutura do documento

No presente documento, a estrutura é meteticulosamente delineada com a finalidade de proporcionar clareza e organização. Inicialmente, no capítulo 1, será apresentada a revisão da literatura, abordando a posição atual do AE e o seu desenvolvimento ao longo do tempo. Este capítulo tem por objetivo estabelecer um enquadramento temporal do tema. Na revisão de literatura serão apresentadas as diferentes iterações do tema, assim como os fatores que são abordados como os mais prováveis de estar na origem do AE.

Em seguida, o capítulo 2 **Error! Reference source not found.** apresenta o entendimento do problema da instituição, ou seja, esclarece mais pormenorizadamente os objetivos do projeto, como se procede a recolha de dados e quais os recursos utilizados ao longo do estudo.

O entendimento e preparação dos dados serão discutidos no capítulo 3, no qual será apresentada a análise exploratória dos dados, bem como as diversas iterações realizadas ao longo do processo e a descrição das novas variáveis criadas.

Prosseguindo com a modelação, o capítulo 4 contemplará a aplicação de algoritmos descritos, gerando diversos resultados que serão minuciosamente analisados e comparados entre si, conforme descrito no capítulo 5.

Finalizando a parte prática do projeto, no capítulo 6, é apresentada a etapa da implementação. A elaboração deste capítulo tem como objetivo esclarecer o funcionamento prático do modelo, para esclarecer os utilizadores e ajudar com a

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

implementação deste novo recurso de prevenção de AE. Uma vez que os utilizadores finais consigam perceber pormenorizadamente como funciona o modelo, é mais fácil que o coloquem em prática visando o objetivo do projeto. Assim, encontra-se explicado como funciona o algoritmo KNN no caso concreto do AE, bem como as limitações e próximos passos da implementação.

Por fim, a conclusão do projeto será apresentada, encerrando o documento de forma a consolidar as principais conclusões obtidas ao longo deste estudo.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

1 Revisão da literatura

O capítulo da revisão de literatura tem por objetivo enquadrar todos os temas abordados anteriormente. Desta forma será mais fácil perceber a evolução de cada ponto e como estes estão inseridos atualmente no mercado, para se conseguir trabalhar neles adaptados à atualidade.

No primeiro ponto, irei abordar como se posiciona o tema do AE no ES, seja na atualidade de Portugal e também a nível de estudos que têm sido elaborados no mesmo âmbito.

Considerando que este tema não é linear e que existem diferentes teorias e abordagens, é também importante que sejam esclarecidos alguns termos para clarificar qual a linha de pensamento adotada ao longo do estudo realizado.

Um dos pontos centrais do estudo foi a identificação dos fatores que levam ao AE para posteriormente serem implementadas medidas para a mitigação do mesmo. Assim, um dos pontos que irá ser abordado mais à frente será o levantamento dos fatores mais frequentemente apontados nos projetos relativos aos problemas de desistência no ES.

1.1 Posicionamento do AE no ES

Inicialmente, é importante clarificar o que efetivamente se irá considerar como AE do ES. A definição de AE pode parecer simples, porém existem muitas variáveis a ter em conta para que tal se possa afirmar como abandono. Vários autores debatem sobre este tema, como é exemplo do artigo de Casanova et al. (2018), que afirma que o AE se dá quando o aluno pede transferência para outro curso ou universidade, complementando o caso com maior alarmística, o aluno estar há mais de dois anos sem matrícula no ES sem ter acabado o curso previamente inscrito. Aplicado a este projeto, serão considerados como abandonos todos os alunos que ainda não concluíram os estudos e não se inscrevem no ano seguinte, assim como os que submetem previamente uma interrupção.

Num estudo feito em 2022 sobre como se posiciona o tema na atualidade mundial por Barroso et al. , foram analisados diversos artigos que abordam o tema até 2018. Dá realce que os anos mais concorridos de estudos do tema são 2014 e 2016 com mais de metade

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

dos artigos publicados. Surge deste modo, a preocupação de o tópico voltar a ser estudado com alguma urgência dado os resultados obtidos nos últimos anos letivos.

Nessa linha, desde cedo se debate este problema que se considera tão atual. No que diz respeito às ações políticas, em 1999 o Ministério da Ciência, Tecnologia e ES emitiu o despacho nº6659/99 que comprova a relevância da questão do abandono.

Posteriormente, em março de 2013 o Governo português emitiu a Resolução da Assembleia da República nº 60/2013, em que é recomendado a elaboração de um relatório anual sobre o AE no ES. Em 2014, foi emitido o documento “Orientações estratégicas para o ES” onde são definidas as medidas de ação, nomeadamente a criação do programa Retomar, que é direcionado para o aumento do sucesso escolar e diminuição do AE.

Mais recentemente, no ano de 2022 contamos com o projeto “Programa de promoção de sucesso e redução de AE no ES” apresentado pela Ministra da Ciência, Tecnologia e ES, Elvira Fortunato. Fundamentalmente o projeto tem o objetivo de estimular o desenvolvimento de instrumentos que apoiem na integração de novos estudantes nas nossas Instituições de ES.

1.2 Dropout, a diferença entre os termos *optout* e *stopout*

Os termos em inglês, *dropout*, *optout* e *stopout* são comumente utilizados neste meio. Na medida em que o termo *dropout* é equivalente ao usado no presente projeto como AE, surge a necessidade de distinguir os restantes dois termos que são variações do primeiro, nomeadamente:

- *Stopout*;
- *Optout*

Para tal, o estudo de Casanova et al. (2018), é crucial para ajudar na distinção. Assim, podemos considerar que um aluno está a passar por um processo de *stopout* quando o mesmo realiza uma pausa de um ano, mas continua com claras intenções de retomar os estudos. São vários os autores que estudam este tema e segundo Stratton et al. (2008) é considerável o número de estudantes que optam por esta pausa. Por norma, a interrupção

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

tende a ser breve e dura apenas um ano, frequentemente com maior adesão a alunos que frequentam o primeiro ano de ES. Por outro lado, temos a expressão de *optout* que diz respeito aos alunos que por diversas razões não se identificam com o curso/área/instituição de ensino atual e procedem a essa mudança de curso ou instituição, aproveitando por sua vez as possibilidades de mobilidade que são permitidas pelo sistema de ES.

1.3 Fatores que podem levar ao AE do ES

“O abandono no ES tem vindo a ganhar maior visibilidade política e social e tem, igualmente, constituído objeto de reflexão por parte da comunidade científica e das instituições do ES” (Ferreira & Fernandes, 2015, p1). Será com base nesta citação que surge a urgência de se desenvolver o tema.

Portugal insere-se numa comunidade cada vez mais inclusiva, em que as diferentes áreas de estudo estão cada vez mais constituídas com maior diversidade de estudantes, seja de género, raça, cultura, etc. Tendo isto em conta, é cada vez mais difícil acompanhar os estudantes da melhor forma para conseguir que todos se encontrem satisfeitos ou perceber alguns fatores que possam antever o AE. Assim, é necessário rever toda a estrutura neste âmbito e perceber quais poderão ser os fatores que estão por trás destas preocupações.

Percebendo e analisando o universo de estudos feitos anteriormente sobre o AE, conseguimos levantar algumas das que foram identificadas como causas ou comportamentos de risco para a desistência de um aluno.

Um dos fatores de risco que mais vezes é percebido e também que mais vezes é mostrado por parte dos alunos, é a ausência frequente das aulas (principalmente as de maior ligação com a área que frequentam), seguida de ausência das avaliações periódicas. Este fator estará muitas vezes ligado com os estudantes de “risco”, que estão previamente inseridos num ambiente pessoal, financeira e familiar instável, (Ferreira et al., 2019).

Neste seguimento, muitos são os estudos que realçam as dificuldades financeiras como sendo uma causa frequente para o AE. Segundo a DGEEC, em 2022, em Portugal encontravam-se matriculados mais de 446mil alunos portugueses e estrangeiros. Deste

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

número, cerca de 29.1%, ou seja, mais de 120 mil alunos, estavam deslocados das suas residências originais. Assim, tendo em conta estes dados, é necessário ponderar a implicações financeiras que foram necessárias para que os mais de 120 mil alunos consigam continuar o seu percurso académico. Para os que mudam de residência, irão ter custos com deslocações para as suas casas natais, numa nova habitação na cidade onde se insere a instituição de ensino e propinas. Ainda que existam ajudas para os mais necessitados das mesmas, muitas das vezes poderá não ser o suficiente.

Com as dificuldades financeiras, surge em alguns dos casos a necessidade de se adquirir o estatuto de trabalhador-estudante. Esta condição, reúne conciliar os estudos e compromisso com a entidade patronal e é igualmente uma situação difícil. É apontada também em alguns estudos como uma das razões para o AE, ou seja, a situação profissional do aluno.

Sejam as dificuldades financeiras que vão surgindo ou a situação profissional em que o aluno se encontra, ambas estão interligadas com a vida pessoal do estudante. Esta dimensão pessoal terá um grande peso no desfecho que terá a sua frequência no ES. Deste modo, entramos em mais um fator que continua a ser mencionado, razões familiares. Dentro das razões familiares podemos lidar com o facto de estudantes terem uma família, encontrarem-se casados, com filhos e que de alguma forma será complicado a conciliação dos dois mundos. Por outro lado, sendo o caso de não ter ainda a constituição da sua família, poderão estar a atravessar problemas familiares e instabilidade familiar que tenham a ver principalmente com parentes diretos. Muitos dos estudos associados afirmam que uma das maiores razões de abandono escolar foram razões familiares (Ferreira & Fernandes, 2015).

Estes fatores são apresentados com realce nos artigos de Sivakumar et al. (2016) e Khanam & Quraishi (2016).

Segundo o estudo feito com estudantes de universidades alemãs em 2020 por Mouton et al, uma das quatro razões apontadas como significativas no abandono precoce “está relacionada com as condições de estudo, nomeadamente superlotação do espaço de estudo”. Assim, é relevante que as instituições proporcionem condições face aos alunos

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

que permanecem na instituição fora do horário de aulas. É também importante que a área desses espaços disponibilizados seja proporcional ao número de alunos que a estrutura recebe. Assim sendo, é apontado como um fator relevante a possíveis faltas de condições da instituição seja, por exemplo, a qualidade de conexão com a internet ou falta de espaços acolhedores para que os estudantes se sintam confortáveis em ter um espaço destinado a estudo após as aulas.

Com tantas mudanças repentinas na vida de um estudante deslocado, muitos deles, segundo o estudo de Albuquerque (2008), demonstram receio face ao ambiente social e revelam o medo de “Não se relacionar bem com os colegas”. Estamos perante a situação em que muitos mudam de local de residência e terão de se adaptar a um novo ritmo de estudo, gerir a relação com colegas (muitas vezes com quem partilham residência), professores e toda a envolvência da instituição de ensino. Ainda, relacionado com o ambiente académico, no mesmo estudo revela “o medo de não se adaptarem ao ambiente”, “não se conseguirem inserir na vida académica”, “a pouca disponibilidade para as atividades extracurriculares” e “a falta de espírito de união entre os cursos da mesma escola”. São vários exemplos de dificuldades que os estudantes enfrentam quando ingressam para o ES e que muitas vezes aliados com outros fatores levam à desistência. Complementando este fator, o estudo de Sivakumar et al. (2016) reforça mais uma vez que uma das grandes razões pode ser o ambiente institucional, por não ir de encontro às expectativas dos alunos ou de não sentirem o devido acompanhamento que necessitam num momento de maior fragilidade.

Por outro lado, existem estudos que avaliaram de que forma os alunos sentem a falta de disponibilidade e apoio dos docentes como razão de desistência e estes mostraram um valor bastante baixo. Segundo Mouton et al. (2020), na página 25, “apenas uma classe, representando 2% da amostra, distinguiu a falta de apoio dos docentes como motivo de abandono”. No que diz respeito aos docentes, poderá estar igualmente relacionado o método de ensino aplicado nas Unidades Curriculares (UC) não ser o mais adequado. Os métodos de ensino poderão ter uma relação com o AE, incluindo os métodos de avaliação adotados (Ferreira & Fernandes, 2015).

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Como possível consequência, e não só, dos fatores apresentados acima os alunos podem não obter os resultados esperados. O que poderá levar a revolta, frustração, e em casos extremos ser mais uma ajuda para o AE.

Neste ambiente de mudança, um dos grandes pilares será sempre a família. Deste modo, alguns estudos avançaram na descoberta da relação entre o AE e o nível de escolaridade dos pais. No estudo de Ferreira et al. (2018), na página 4, incidiram sobre a relação entre as habilitações literárias da mãe do estudante. Neste estudo, chegaram à conclusão de que “os estudantes cujas mães possuem o ES revelam valores médios mais elevados em quase todas as dimensões do AE”, abordadas mais à frente, “excetuando a dimensão profissão/carreira, onde sobressaem os estudantes cujas mães têm menores habilitações académicas, no entanto, não se registam diferenças estatisticamente significativas”. Ainda no que diz respeito à mãe, mostrou-se também relevante a sua situação profissional. Esta variável, no mesmo estudo mostrou-se significativa estatisticamente. Ou seja, os alunos cujas mães estavam ativas profissionalmente obtiveram melhores resultados. No entanto, como no mesmo projeto se tem em consideração diferentes dimensões: por exemplo, “dimensão profissão/carreira” e “dimensão gestão de vida”, a variável anterior mostrou que estudantes cuja mãe está profissionalmente ativa se destacam mais pela positiva na primeira dimensão, e por outro lado, quando as mães não se encontram com atividade profissional os estudantes destacam-se mais na segunda dimensão de gestão de vida. Tal, pode sugerir a menor disponibilidade dos pais em geral para apoiar e comunicar eficazmente com os filhos. Nesta fase, que se mostra exigente a nível de transição da vida escolar poderá impactar e refletir-se negativamente na adaptação ao ES.

Uma particularidade que é muito frequente quando os alunos ingressam no ES é o facto de não entrarem na sua primeira opção (Albuquerque, 2008), ou seja no curso que mais desejariam. Se existem inúmeras dificuldades associadas ao início de uma nova jornada será claramente mais desafiante ingressar num curso não preferencial. Este problema poderá dar-se tanto pela média do curso ser demasiada alta face à média do estudante ou também pelo facto de ano para ano os ingressos ao ES estarem a aumentar. Assim sendo, os alunos aproveitam o seu primeiro ano para “marcar passo e conseguirem transferência de curso”. Associando estes fatores às diversas dificuldades é mais provável que seja

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

gerado “desinteresse, desmotivação, dificuldade de enquadramento dos alunos na escola ou ainda por dificuldade em gerirem a sua aprendizagem e os seus métodos de estudo”. Associado ao desinteresse também está a não identificação com a área de saída, nem com o conteúdo que é lecionado ao longo do curso ou a área aplicada num futuro.

Em 2022, registaram-se valores mais baixos do desemprego entre recém-licenciados. Segundo dados da DGEEC, diminuiu de 5% para 4% num ano. No entanto, a empregabilidade associada à área de estudo continua a ser uma preocupação bastante atual. É necessário ter em conta cursos de baixa empregabilidade, com pouca escolha e ainda com a não identificação com as opções escritas para emprego em cada curso.

1.4 Dimensões associadas com o AE

Existem várias dimensões que podem ser consideradas e que foram tidas em conta na grande maioria dos artigos analisados e supracitados. Foram consideradas 4 dimensões principais dado a complexidade e diversidade de fatores a tidos em conta. Dimensão organizacional, gestão da vida pessoal, profissional e relacional.

- A **dimensão organizacional** está relacionada com as instalações fornecidas pela instituição e tudo o que está diretamente relacionado com o interior das mesmas. Os fatores suprarreferidos que pertencem a esta dimensão são os seguintes: Método de ensino das UC, notas não correspondentes com a expectativa, falta de condições da instituição, falta de disponibilidade e acompanhamento dos docentes, ausência frequente das UC e avaliações importantes.
- **Dimensão da gestão de vida pessoal:** Situação profissional do estudante, dificuldades financeiras, razões familiares, nível de escolaridade e situação profissional dos pais.
- **Dimensão profissional:** Empregabilidade, não entrarem na sua primeira opção e a não identificação com as opções de saída.
- **Dimensão relacional/social:** Receio face ao ambiente social seja dentro da instituição ou com quem irão dividir o espaço de residência, relação com os professores e com os colegas.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

1.5 Quadro resumo dos fatores

Dimensão organizacional

Tabela 1 - Fatores da dimensão organizacional

Artigos /fatores	Método de ensino das UC	Notas não correspondentes com a expectativa	Falta de condições da instituição	Falta de disponibilidade e acompanhamento dos docentes	Ausência frequente das UC e avaliações importantes
(Realinho et al., 2022)		x			
(Ferreira et al., 2018)			x	x	
(Ouatik et al., 2022)		x			
(Ferreira & Fernandes, 2015)			x	x	
(Ferreira et al., 2019)	x		x		
(Barroso et al., 2022)				x	
(Khanam & Quraishi, 2016)					
(Sivakumar et al., 2016)				x	x
(Stratton et al., 2008)	x	x			
(Casanova et al., 2018)	x	x		x	

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Dimensão da gestão de vida pessoal

Tabela 2 - Fatores da dimensão da gestão de vida pessoal

Artigos /fatores	Situação profissional do estudante	Dificuldades financeiras	Razões familiares / problemas pessoais	Nível de escolaridade e situação profissional dos pais
(Realinho et al., 2022)				x
(Ferreira et al., 2018)	x			x
(Ouatik et al., 2022)				
(Ferreira & Fernandes, 2015)		x	x	
(Ferreira et al., 2019)	x	x	x	
(Mouton et al., 2020)	x	x		
(Barroso et al., 2022)	x	x	x	x
(Khanam & Quraishi, 2016)		x	x	
(Albuquerque, 2008)		x	x	
(Sivakumar et al., 2016)		x	x	
(Stratton et al., 2008)		x	x	
(Casanova, 2018)	x	x	x	x

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Dimensão profissional e relacional/social

Tabela 3 - Fatores da dimensão profissional e relacional/social

Artigos /fatores	Empregabilidade	Não entrarem na sua primeira opção	Não identificação com as opções de saída	Receio face ao ambiente social	Relação com os professores e com os colegas
(Realinho et al., 2022)	x				
(Ferreira et al., 2018)			x		
(Ouatik et al., 2022)				x	x
(Ferreira & Fernandes, 2015)					
(Ferreira et al., 2019)				x	
(Mouton et al., 2020)			x		x
(Barroso et al., 2022)				x	x
(Stratton et al., 2008)				x	x
(Casanova, 2018)			x		

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

2 Entendimento do problema da instituição

No capítulo do entendimento do negócio é importante reforçar mais uma vez que o projeto se centra no estudo dos fatores que poderão estar por trás do AE dos alunos do IPC, e deste modo proceder a medidas preventivas para mitigar estes acontecimentos.

Esta previsão é baseada na análise de dados detalhados sobre alunos. Ao longo dos anos, as diferentes escolas do IPC foram recolhendo informações importantes dos seus alunos. Deste modo, o projeto teve como base de dados, os dados concretos de alunos que frequentaram as instituições em que se quer atuar, incluindo dados de registo, histórico académico, desempenho nas disciplinas entre outras informações relevantes para a análise do AE. Assim, serão utilizadas as bases de dados do sistema de gestão académica do IPC, concretamente os sistemas atualmente em uso. Para garantir a confidencialidade e privacidade dos dados dos alunos, foram utilizados procedimentos adequados de anonimização e segurança de dados. Foram tomadas todas as precauções necessárias para garantir que a recolha de dados fosse realizada de forma ética e responsável, em conformidade com as normas e regulamentações aplicáveis.

O IPC, ao longo dos anos passou por vários processos de informatização e de disponibilização dos dados importantes para os seus docentes e alunos. No ano letivo 2019/2020 celebrou-se mais uma dessas etapas. Foi nesse ano que a plataforma NONIO ficou disponível para todos os envolventes do instituto, agregando assim toda a informação necessária para a docência, assim como também todos os campos necessários para os alunos realizarem um melhor acompanhamento do seu percurso académico.

Assim com recurso a essa nova plataforma, foram efetuadas algumas extrações. Foram tidos em conta os dados dos anos letivos 2020/2021 e 2021/2022 no modelo que irá ser posteriormente apresentado, com dados à data de agosto de 2021 e agosto de 2022, respetivamente. No entanto, como é necessário conter dados sobre a variável de saída (AE), os dois campos auxiliares na mesma foram retirados em outubro de 2021 e outubro de 2022 depois de todas as inscrições no ano letivo seguinte efetuadas. Todos os dados vêm com detalhe por aluno, recorrendo apenas a um número fictício e meramente distintivo, sem qualquer valor real, para preservar a identidade dos alunos incluídos no

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

estudo.

Por outro lado, além da acessibilidade da sua utilização, esta plataforma foi também crucial para funcionar como um agregador de todos os dados sobre as unidades orgânicas. Deste modo, sendo que tudo passou a convergir para uma só plataforma, para o projeto em questão é visto como uma mais-valia, ou seja, toda a informação que está disponível para ser incluída no estudo pode ser extraída da base de uma só vez.

Assim, os critérios de sucesso estão relacionados sua capacidade de fornecer informações úteis para a prevenção do AE. Os principais critérios de sucesso incluem:

- Desenvolver um modelo preditivo eficaz que seja capaz de identificar os fatores que contribuem para o abandono escolar.
- Fornecer recomendações práticas com base nos resultados da análise de dados, permitindo a implementação de estratégias de prevenção.
- Garantir que a recolha e análise de dados sejam conduzidas de maneira ética e em conformidade com as normas e regulamentações aplicáveis.

Para a realização deste projeto, foram disponibilizados recursos tecnológicos e de análise de dados. Foram utilizadas as seguintes ferramentas:

- *Python*: A linguagem de programação *Python* foi escolhida para análise de dados e tratamento estatístico. A sua versatilidade e disponibilidade de bibliotecas e módulos específicos para análise de dados tornaram-na a escolha ideal.
- *Power BI Desktop*: O *Power BI Desktop*, é uma ferramenta da Microsoft, foi utilizada para a estatística descritiva dos dados e para visualização de informações. Essa ferramenta desempenhou um papel crucial na tomada de decisões com base em Business Intelligence (BI).

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

3 Entendimento e preparação dos dados

No âmbito desta pesquisa, a etapa de entendimento e preparação dos dados desempenha um papel fundamental, quer seja na obtenção de novos dados como também na consecução dos objetivos traçados. Assim, neste capítulo será apresentado o conjunto de dados iniciais, bem como as transformações significativas que foram aplicadas a esses dados. Além disso, será realizada uma análise de estatística descritiva relevante para destacar alguns aspetos importante e características dos dados, servindo como base para as análises posteriores.

3.1 Conjunto de dados inicial

Entende-se por conjunto de dados inicial, aquele que é extraído diretamente da base para ser submetido aos vários processos apresentados ao longo do projeto. Este apresenta informação em cada linha de cada disciplina para o mesmo aluno, o que significa que existem várias linhas por cada aluno, correspondentes ao número de disciplinas que o mesmo se inscreveu no início do ano letivo. Desta forma, não é possível ter, em cada linha, a visão geral do percurso do aluno no curso em que está inscrito.

Assim sendo, inicialmente são extraídas 25 variáveis descritas e sintetizadas na Tabela 4 que se apresenta seguidamente, contando com 8404 alunos distintos e 93842 linhas.

Tabela 4 - Listagem de variáveis do conjunto de dados inicial

Variável	Descrição	Tipo de dados
MATR_ID	Identificador anonimizado do aluno	<i>Integer</i>
Estado civil	Estado civil do aluno	<i>String</i>
Curso	Nome do curso em que o aluno está inscrito	<i>String</i>
Código da disciplina	Código identificativo da disciplina em que o aluno está inscrito	<i>Integer</i>
Disciplina	Designação da disciplina	<i>String</i>

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Variável	Descrição	Tipo de dados
Disc_ects	Número de ECTS da disciplina	<i>Float</i>
Nota	Nota obtida a cada disciplina	<i>Float</i>
Média	Média atual do aluno: Consideram-se apenas as notas das disciplinas em que houve aprovação (ou seja, superiores a 10 valores)	<i>Float</i>
ECTS aprovados matrícula	Número de ECTS que foram aprovados aquando da matrícula do aluno	<i>Integer</i>
Total de inscrições	Número de inscrições do aluno no curso atual, ou seja, quantidade de anos letivos que o aluno frequentou o curso em que está matriculado	<i>Integer</i>
Bolseiro	Variável binária para identificar se o aluno tem acesso a bolsa ou não.	<i>String</i>
Estatutos	Estatuto do aluno	<i>String</i>
Regime de ingresso	Regime de ingresso do aluno	<i>String</i>
Opção de candidatura	Intervalo de opção de candidatura de 0 a 6, sendo o valor 0 correspondente aos regimes de ingresso que não têm opção de candidatura e de 1 a 6 o regime de ingresso normal. Essa variável é importante para verificar a relação com o curso e área	<i>Integer</i>
Nota de ingresso	Nota com que o aluno se candidatou ao curso. No caso de alguns regimes de ingresso, essa informação pode estar ausente	<i>Float</i>
Localidade	Concelho da morada fiscal do aluno	<i>String</i>
Código postal	Código postal da morada fiscal do aluno	<i>Integer</i>
Cpostal	Primeiros quatro dígitos do código postal da morada fiscal do aluno	<i>Integer</i>

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Variável	Descrição	Tipo de dados
Cpprefixo	Últimos 3 dígitos do código postal da morada fiscal do aluno	Integer
Estado da matrícula	Estado em que se encontra a matrícula do aluno: ativa/inativa	String
Renovou no ano seguinte	Descrição da inscrição do aluno no ano letivo seguinte	String
Entrada_vigor_pe	Ano letivo em que o plano curricular no curso entrou em vigor	String
IFRE_Modo	Variável binária que indica se o aluno está a frequentar o curso de forma integral ou parcial	String
ECTS inscritos	Número de ECTS inscritos no início do ano letivo	Integer
ECTS aprovados inscrição	Número de ECTS aprovados no início do ano letivo	Integer

3.2 Análise descritiva dos dados do conjunto de dados inicial

MATR_ID/Número de alunos, média, nota de ingresso: A análise através do número de MATR_ID conseguiu transmitir quantos alunos tratava o conjunto de dados inicial. Estamos perante 8404 alunos correspondentes ao ano letivo 2020/2021. A média dos alunos é em média 13.17 valores e a média da nota de ingresso é ligeiramente superior com 13.79 valores. Tendo em conta a variável de saída é possível perceber que cerca de 14.7% dos alunos praticaram AE. Estes valores encontram-se resumidos na Figura 2.

DADOS INICIAIS			
Alunos	Média	Média Ingresso	Abandonos
8404	13.17	13.79	1238

Figura 2 - Análise descritiva das variáveis MATR_ID, média e nota de ingresso

Estado civil: No conjunto de dados para treino, o estado civil distribuía-se da seguinte forma apresentada na Figura 3.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

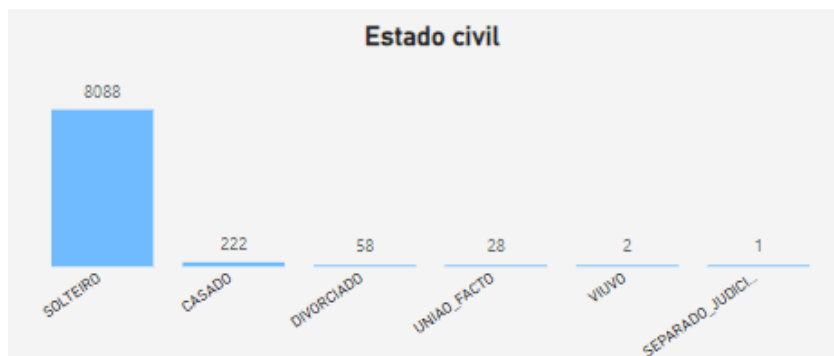


Figura 3 - Gráfico de barras com distribuição da amostra pelos estados civis

Dos 8404 estudantes a serem estudados, apenas 316 apresentavam um estado civil diferente de solteiro.

Curso: A variável curso engloba todos os cursos de licenciatura lecionados no ano letivo 2020/2021 em todas as unidades orgânicas do IPC. Assim, estão abrangidos 61 cursos diferentes. O curso de Engenharia Informática é o curso que possui mais alunos, com um total de 845. Por outro lado, o curso em Biodiversidade e Conservação da Natureza é o que apresenta menos alunos, 6.

Disciplina: Ao nível do número de disciplinas diferentes, existem 1966. Estas disciplinas estão distribuídas por todos os cursos, de todas as escolas e todos os anos.

Disc_ects: Esta variável apresenta o número de ECTS que cada disciplina. Para 668 disciplinas é atribuído o número mais comum de 6 ECTS. No entanto, no máximo uma disciplina pode ter 55 ECTS, correspondente a disciplinas de estágio.

Total de inscrições: A variável de total de inscrições apresenta o número de vezes que um aluno se encontra matriculado no curso atual. Assim sendo, consegue-se perceber com a ajuda da Figura 4, que a maioria da população em estudo se encontra no primeiro ano inscrito no curso. Por outro lado, é constituído por 15 alunos que têm 10 inscrições ou mais, com um máximo de 14 inscrições.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

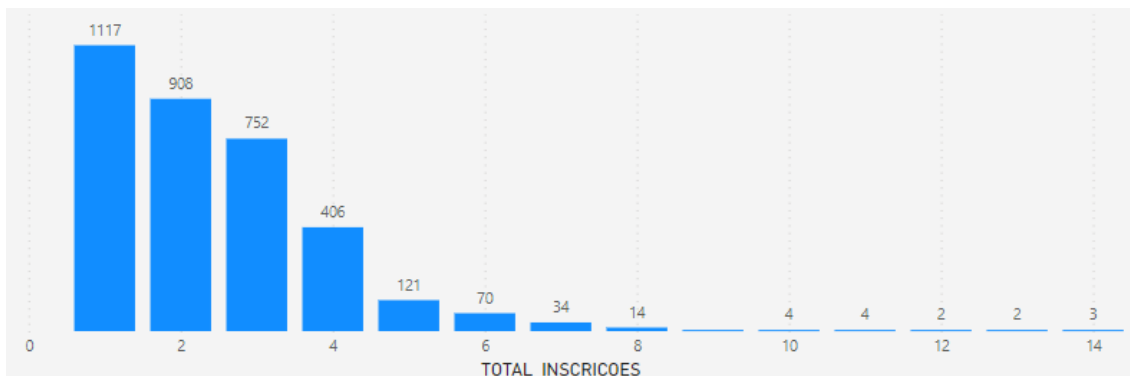


Figura 4 - Gráfico de barras com a distribuição do número de alunos por total de inscrições no curso

Bolseiro: A variável apresenta-se com a distribuição descrita na Figura 5. Cerca de 28% dos alunos é abrangido pela bolsa, ou seja, 2335 alunos.



Figura 5 - Gráfico circular da distribuição da variável "bolseiro"

Estatutos: Os dados apresentam 16 tipos de estatutos diferentes, sendo que cada aluno pode ter mais do que um estatuto associado ou não ter nenhum. De acordo com a Figura 6, é possível perceber que a 7545 alunos não estão associados a qualquer tipo de estatuto. Seguido de 636 alunos com o estatuto de trabalhador-estudante.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar



Figura 6 - Distribuição dos alunos pelos estatutos

Regime de ingresso: Para os regimes de ingresso estão considerados 31 regimes diferentes, estando na Figura 7 apresentada a distribuição dos regimes por alunos, apenas os que contém mais de 100 alunos. Como esperado, o regime com mais alunos é o regime geral de ingresso com 5537 alunos.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar



Figura 7 - Distribuição dos alunos pelos com mais de 100 alunos

Opção de candidatura: As opções disponíveis para a variável opção de candidatura são estar dentro do intervalo de 1 a 6 ou não conter essa informação pelo facto dos alunos estarem associados a regimes de ingressos especiais. Assim, na Figura 8 é possível perceber que apenas menos de metade dos alunos entraram no curso atual como sendo a sua primeira opção.

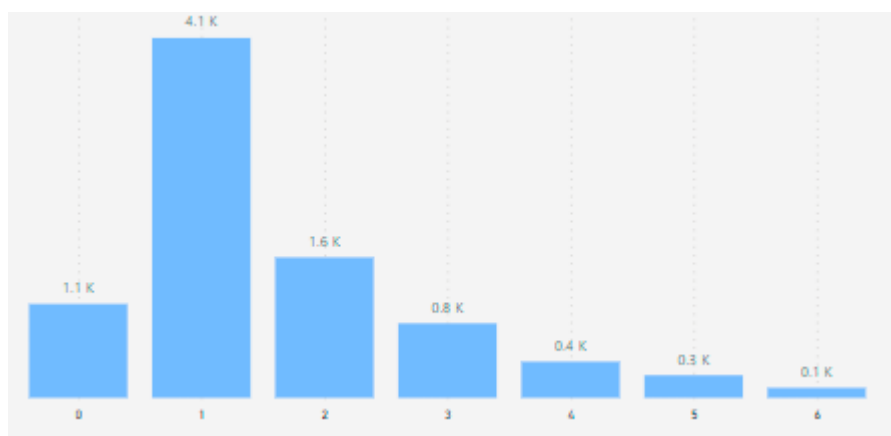


Figura 8 – Gráfico de barras com a distribuição dos alunos pelas opções de candidatura

Estado da matrícula /Renovou no ano seguinte: O estado da matrícula e renovou no ano seguinte são as duas variáveis que deram origem à variável de saída, como está

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

explicado no ponto seguinte. Assim, é possível ver que cada uma destas variáveis é constituída por 2 opções, tendo as opções de “ativo” e “concluído” para o estado da matrícula e “sim” e “não” para o renovou no ano seguinte.

7225 alunos encontravam-se com o estado da matrícula ativo e 1179 como concluído. Dos 8404, 6013 renovaram a matrícula no ano seguinte e os restantes não. Estes dados podem ser comprovados na Figura 9.

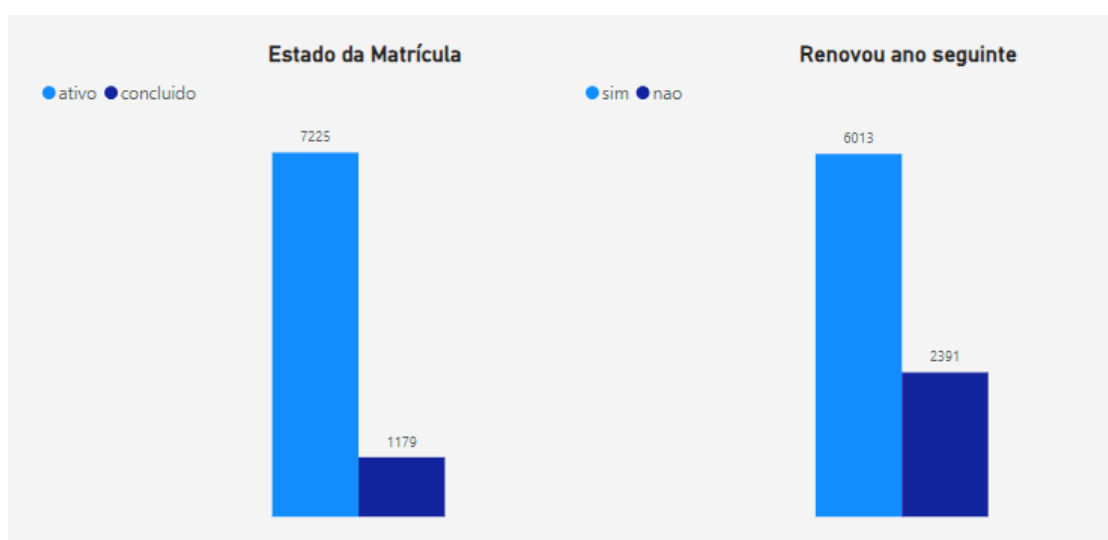


Figura 9 – Gráfico de barras com as variáveis “estado da matrícula” e “renovou ano seguinte”

IFRE_Modo: A variável IFRE_Modo descreve em que regime de estudos o aluno se encontra, tendo em conta que cerca de 88% dos alunos frequenta o curso de forma integral, como se pode comprovar na Figura 10.

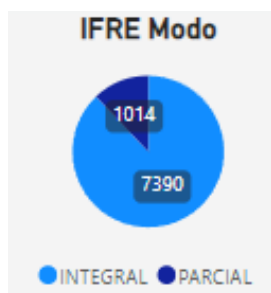


Figura 10 – Gráfico circular com a distribuição dos alunos na variável IFRE_Mod

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

3.3 Conjunto de dados modificado

O conjunto de dados inicial apenas contém a informação em bruto utilizada como ponto de partida deste projeto. Tendo em conta tudo o que foi mencionado no capítulo 1.3 foi necessário fazer alguns ajustes para que os dados conseguissem traduzir o que era esperado. Assim, seguidamente são apresentadas as iterações que tiveram por base os mesmos dados iniciais, mas abordagens diferentes. Com as iterações discriminadas será possível perceber como se comportam as variáveis com os diferentes tratamentos de dados.

3.3.1 1ª iteração

Numa primeira iteração, o objetivo foi relacionar as variáveis iniciais com os fatores apontados como mais relevantes que levam ao AE. Esta primeira iteração contém um total de 8404 alunos distintos e 19 variáveis.

Assim, tendo por base o estudo apresentado na revisão de literatura e as variáveis disponíveis pela plataforma da instituição as seguintes mantiveram a sua forma original:

- **MATR_ID**
- **Estado Civil**
- **Curso**
- **Média**
- **Total de Inscrições**
- **Bolseiro**
- **Opção de Candidatura**
- **Nota de Ingresso**
- **IFRE_Modo**

Tendo por base as restantes foram criadas outras com maior relevância para a forma como os dados foram modelados. Na Tabela 5 são apresentadas e sintetizadas as 19 variáveis

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

adicionais que foram criadas e, consideradas nesta 1ª iteração, como relevantes.

Tabela 5 - Listagem das variáveis resultantes da 1ª iteração

Variável	Descrição	Tipo de dados
Total Disciplinas	Número total de disciplinas que o aluno se inscreveu no ano corrente	<i>Integer</i>
Total Sup10	Número de disciplinas em que o aluno obteve classificação superior a 10 até ao momento, no ano corrente	<i>Integer</i>
Disciplinas em Falta	Número de disciplinas em que o aluno estava inscrito no início do ano, mas ainda não obteve uma classificação superior a 10	<i>Integer</i>
Média das Notas Maiores que 10	Para complementar a média geral, considera-se a média apenas das disciplinas do ano corrente em que o aluno obteve classificação superior a 10 valores	<i>Float</i>
Total de ECTS	Contabilização do número total de ECTS que o aluno se inscreveu no início do ano corrente	<i>Integer</i>
ECTS Feitos	Contabilização dos ECTS correspondentes às disciplinas em que o aluno já obteve classificação superior a 10 valores, no ano corrente	<i>Integer</i>
Proximidade	Variável classificada com base em códigos postais, ou seja, se o aluno pertence ao concelho de Coimbra é considerado "Perto", se é fora do concelho de Coimbra é "Longe", para avaliar a necessidade de alojamento pago e outras despesas	<i>String</i>
Regimes de Ingresso	Variáveis binárias que indicam o tipo de regime de ingresso do aluno, incluindo: • Regime Geral de Acesso • Preferência Regional • Mudança de Par Instituição/Curso • Maiores de 23 anos • Reingresso • Estudante Internacional	<i>Integer</i>

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Variável	Descrição	Tipo de dados
Estatutos	Variáveis binárias que indicam o estatuto do aluno, incluindo: - Falta de pagamento de propinas ou emolumento - Trabalhador-Estudante - Aluno com Necessidades Educativas Especiais (NEE)	Integer
Abandono	Variável de saída para saber se o aluno abandonou ou não o ES	Integer

Originalmente, o conjunto de dados estava organizado com a correspondência de uma linha para cada inscrição numa unidade curricular que por sua vez estaria associado a cada aluno individualmente. No entanto, dado a diversidade de número de linhas que foram geradas por cada aluno, sendo diferentes entre eles, a opção agregadora incidiu sobre a variável MATR_ID que é o número identificativo do aluno. Ou seja, para cada MATR_ID foi apenas impressa uma linha com todos os dados necessários.

Total de disciplinas: Nesta variável realizou-se uma análise distinta do número de disciplinas inscritas por cada aluno, com o objetivo de obter uma compreensão abrangente das UC que os alunos tinham inicialmente no início do ano letivo.

Total superior a 10: É uma variável que tem por base duas variáveis originais, como se pode verificar na Figura 11 abaixo.

```
df_filtrado = df[df['NOTA']>=10]
totalsup10 = df_filtrado.groupby('MATR_ID')['CODIGO_DISCIPLINA'].nunique()
df['totalsup10'] = df['MATR_ID'].map(totalsup10).fillna(0)
```

Figura 11 - Código em python para construção da variável "Total sup 10"

Inicialmente, é criado um conjunto de dados auxiliar para só apresentar as linhas correspondentes às notas superiores a 10. Posteriormente, é efetuada a operação para cada MATR_ID, do número único do código da disciplina, tendo em conta o filtro anterior. Assim, gera-se a nova coluna no conjunto de dados final. Os valores ausentes são

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

substituídos por 0, pois verificaram-se alguns casos em que os alunos não obtiveram nenhum código da disciplina com valores superiores a 10.

Media de notas superiores a 10: Na elaboração desta variável, é seguida uma abordagem semelhante à variável anterior. Foi tido em conta o conjunto de dados com as notas superiores a 10, e foi agrupado pela média, pois só a média de notas superiores a 10 foi considerada para a média do aluno. É interessante ter uma distinção entre a média geral que o aluno obteve em todas as suas aprovações no curso, e a média no ano letivo em análise, pois este campo também poderá ser indicativo de alerta. Com esta variável consegue-se perceber que a diferença, destas duas médias dos alunos do estudo é, no ano letivo 2020/2021, -0.06. No total dos dados, a média é de 13.17 valores, tendo em conta todas as disciplinas que o aluno já obteve classificação superior a 10 valores. Por outro lado, a média é 13.12 valores, tendo em conta apenas as disciplinas que o aluno obteve no ano em estudo a classificação superior a 10 valores.

ECTS: Seguindo a mesma lógica das disciplinas, foram criadas variáveis relativas aos ECTS, o número total de ECTS que cada aluno se inscreveu no início do ano letivo e o número de ECTS respetivos às disciplinas que ficaram concluídas, ou seja, que obtiveram uma classificação superior a 10 valores. Deste modo, é possível perceber e comparar alunos que realizam o mesmo número de disciplinas, no entanto, com ECTS díspares, o que leva a carga horária e esforço diferentes também. Assim, dá origem à coluna de **ECTS feitos**.

Código postal: A variável do código postal inclui uma coluna auxiliar que regista a correspondência dos primeiros 4 dígitos. Assim, tendo em conta a maioria dos estudos de caso de todas as instituições envolvidas se localizam no concelho de Coimbra e o objetivo com esta variável é perceber a distância que os alunos se encontram das suas moradas fiscais, foi definido um intervalo de códigos postais que seriam denominados de “Perto” e os restantes seriam de “Longe”. Com esta distinção das duas saídas da variável é possível efetuar uma estimativa dos alunos que vão enfrentar problemas de 2ª habitação e tudo o que isso agrega, como foi referido anteriormente na revisão da literatura. O intervalo para a consideração de “perto” é: 3000 a 3049, os restantes valores são

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

considerados como “longe”. Além desta regra, e como temos estudantes internacionais com códigos postais que não são do tipo numérico, procedeu-se à ação de transformar todos os códigos postais que não eram numéricos, no número 0, pois este valor de 0 não será incluído no intervalo acima apresentado, e os estudantes estarão associados à etiqueta de “longe”. Apesar de ter esta variável em consideração, a mesma apresenta uma limitação, pois uma das instituições localiza-se na cidade de Oliveira do Hospital. No entanto, a amostra mostra-se reduzida.

Após a transformação, a variável corresponde à denominação de “**Proximidade**”.

Regimes de ingresso: os dados sobre o regime estão originalmente com a disposição de uma lista, o que não daria para ter uma perceção de como se distribuem pelos alunos em estudo. Posto isto, com código *python* foi criada uma coluna para cada opção de regime de ingresso no conjunto de dados. Inicialmente é criada uma lista com todas as opções, para que com uma função seja percorrida cada opção da lista como cabeçalho de uma nova coluna. Após terem sido criadas as colunas é inserido o valor de 1 na nova coluna se a opção de regime de ingresso atual estiver presente na coluna original e um valor de 0 caso contrário.

Deste modo, é possível visualizar na Figura 12 com recurso à soma quais são os regimes que mais vezes são utilizados na aplicação à universidade, e por essa ordem optar por quais regimes manter como variáveis que integram o algoritmo final.

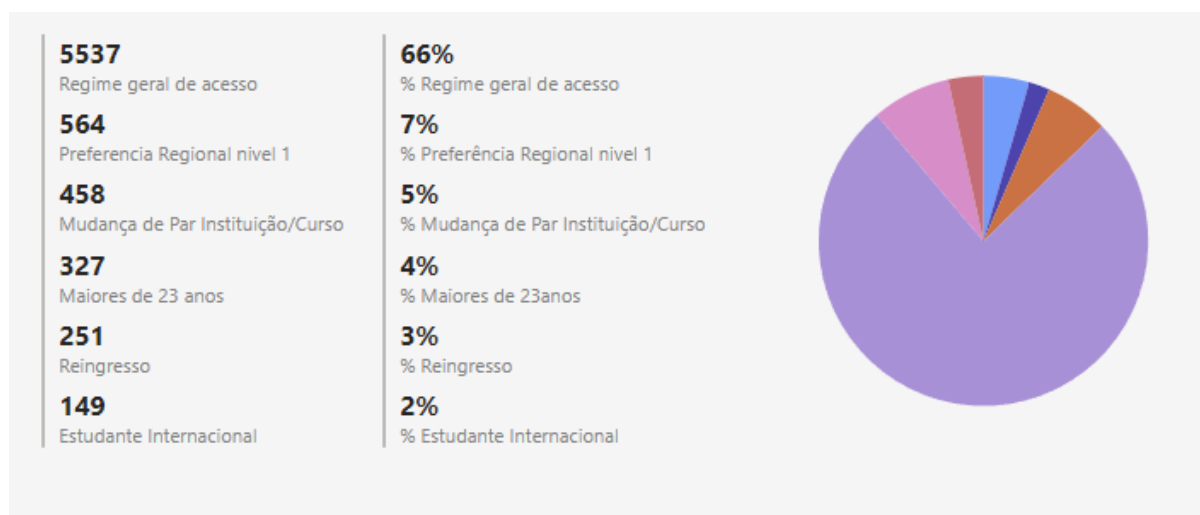


Figura 12 - Descrição e gráfico circular respetivos à variável "regimes de ingresso"

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Tendo em conta a visão final da distribuição dos regimes, conseguimos perceber que a grande maioria recorre ao regime geral de acesso, agregando 66%, o que equivale a 5537 alunos. Nos restantes regimes que agregam mais alunos, representam 21% da fatia, o que equivale a 1749 dos alunos. Os restantes alunos e % total corresponde a regime que não se mostraram relevantes para o estudo.

Estatutos: a coluna dos estatutos, originalmente encontrava-se na mesma disposição do que a anterior, dos regimes. Portanto, primeiramente, pela mesma razão supracitada, foi criada uma lista com todas as possibilidades de estatutos. Por outro lado, na descrição dos estatutos, encontravam-se alguns espaços em branco e caracteres especiais, o que levou a ser necessário fazer a seguinte limpeza representada na Figura 13.

```
split_colunas = df['ESTATUTOS'].str.split('|', expand=True).applymap(lambda x: x.strip() if isinstance(x, str) else x)
```

Figura 13 - Código em python da divisão da coluna de "estatutos"

Após esta transformação recorri ao mesmo método dos regimes. Com a qual foram considerados os estatutos como mais relevantes para o estudo, os apresentados seguidamente na Figura 14.

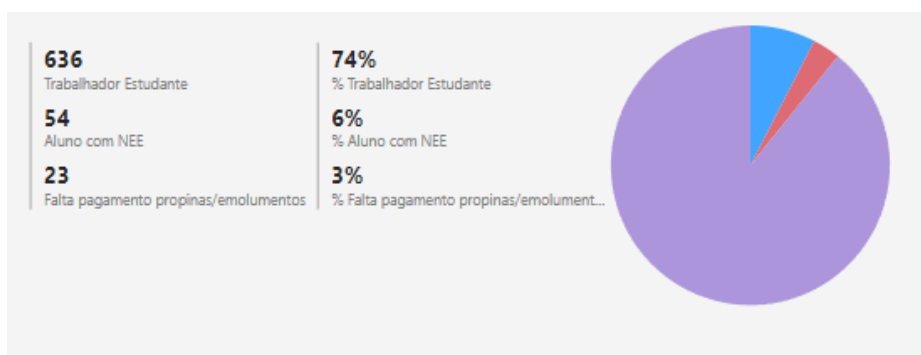


Figura 14 - Descrição e gráfico circular respetivos à variável "estatutos"

Na Figura 14 é tido em conta um universo de 859 estudantes que têm algum tipo de estatuto associado. Dentro desses 859 alunos, 636 deles têm acesso a todos os benefícios associados ao estatuto trabalhador-estudante, o que representa cerca de 74% do universo. Dado o número tão elevado e sendo, na revisão de literatura referido como uma possível

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

causa associada ao AE, é considerado para integrar o algoritmo. Seguidamente, os alunos com necessidades especiais são o segundo maior grupo dentro dos alunos do IPC deste ano letivo, correspondendo a 6% do universo, pelo que também se considera relevante para o estudo. Por fim, e tendo em conta o grande tema das dificuldades financeiras dos alunos, o tema da falta de pagamento de propinas e emolumentos será também incluída como uma variável. Para esta variável considerou-se um conjunto de estatutos que se resumiram a “Falta de pagamento de propinas e emolumentos” como se pode verificar na Figura 15.

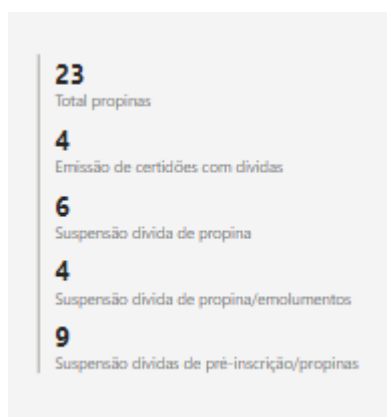


Figura 15 - Descrição dos estatutos pertencentes à variável "estatuto_falta de pagamento de propinas e emolumentos"

Nota de ingresso: As notas de ingresso, correspondem à nota com a qual o aluno se candidatou ao curso, apresentam valores entre 95 e 200, pois são os valores possíveis para um aluno se candidatar. No entanto, esta variável quando é extraída poderá ter, erradamente, o formato de 0 a 20. Para uniformizar a variável, foi criada uma condição em que quando a variável apresentava valores inferiores a 94, considerar o valor multiplicado por 10, para assim ficarem todas as notas de ingresso na mesma escala.

Abandono: Por fim, a variável de saída é o abandono. No entanto, não existe nenhum campo registado que mostre se se trata de um aluno que abandonou o ES ou não. Assim, houve a necessidade de conjugar duas variáveis: Estado da matrícula e renovou no ano seguinte. O estado da matrícula é recolhido aquando das restantes variáveis, por outro

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

lado a variável renovou no ano seguinte, é posteriormente recolhida quando terminam os prazos de inscrição no novo ano letivo.

```
condicoes = [
    ((df['ESTADO_MATRICULA'] == 'Ativo') & (df['RENOVOU_ANO_SEGUINTE'] == 'Não')),
    ((df['ESTADO_MATRICULA'] == 'Interrompido') & (df['RENOVOU_ANO_SEGUINTE'] == 'Não')),
    ((df['ESTADO_MATRICULA'] == 'Ativo') & (df['RENOVOU_ANO_SEGUINTE'] == 'Sim')),
    ((df['ESTADO_MATRICULA'] == 'Concluído') & (df['RENOVOU_ANO_SEGUINTE'] == 'Não')),
    ((df['ESTADO_MATRICULA'] == 'Interrompido') & (df['RENOVOU_ANO_SEGUINTE'] == 'Sim')),
]

valores = ['Abandono', 'Abandono', 'Não abandono', 'Não abandono', 'Não abandono']
df['Abandono'] = np.select(condicoes, valores, default='Outro')
```

Figura 16 - Código python com o tratamento de dados para a criação da variável "abandono"

Tendo estas opções apresentadas em conta e ainda acrescentando a opção de “interrompido” na variável de estado da matrícula, pois poderá ser também uma opção, construíram-se as seguintes opções apresentadas na Figura 16 para definir como Abandono ou Não abandono.

Terminando assim as alterações efetuadas às variáveis para estas conseguirem integrar corretamente e devolver os melhores resultados possíveis.

3.3.2 2ª iteração

Avaliando os resultados da descrição do ponto anterior, concluiu-se que a abordagem anterior não se tornou a mais eficaz. É relevante que se tenha em conta todas as transformações anteriores, pois todas as variáveis definidas estão diretamente relacionadas com estudos sobre o tema.

Seguindo as mesmas bases de tratamento de dados anteriores, foi necessário verificar quais as variáveis que incluem valores nulos e quais os tipos de dados que estavam a ser considerados. Assim, os valores nulos, dependendo das variáveis, foram substituídos por “NA” ou “0” e os tipos de dados foram convertidos para o que se pretendia. A síntese das variáveis obtidas nesta 2ª iteração apresenta-se na

Tabela 6.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Tabela 6 - Listagem das variáveis resultantes da 2ª iteração

Variável	Descrição	Tipo de dados
MATR_ID	Identificador anonimizado e utilizado como variável agregadora	<i>Integer</i>
Estado civil	Estado civil do aluno dividida nas seguintes variáveis: • Casado • Divorciado • Separado judicialmente • Solteiro • União de facto • Viúvo	<i>Integer</i>
Média	Média atual do aluno ao longo do curso: Consideram-se apenas as notas das disciplinas em que houve aprovação (ou seja, superiores a 10 valores)	<i>Float</i>
ECTS aprovados matrícula	Número de ECTS foram aprovados aquando da matrícula do aluno	<i>Integer</i>
Total de inscrições	Número de inscrições do aluno no curso atual, ou seja, quantidade de anos letivos que o aluno frequentou o curso em que está matriculado	<i>Integer</i>
Bolseiro	Variável binária para identificar se o aluno tem acesso a bolsa ou não	<i>Integer</i>
Nota ingresso	Nota com que o aluno se candidatou ao curso. No caso de alguns regimes de ingresso, essa informação pode estar ausente	<i>Float</i>

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Variável	Descrição	Tipo de dados
Distância instituição	Distância em km entre a morada fiscal do aluno e a escola que frequenta.	<i>Float</i>
Alentejo, Algarve, Açores, Centro, Madeira, Norte e Outro	Variáveis binárias com identificação a qual região do país o aluno pertence. Alunos estrangeiros considerados na coluna de outros.	<i>Integer</i>
Opção de matrícula	Os alunos foram distinguidos pelo facto de terem escolhido o respetivo curso como primeira opção (atribui 1), caso contrário são associados a 0	<i>Integer</i>
Modo de frequência	Aos alunos que frequentem de forma integral foram identificados com a opção 1, e com 0 caso seja de forma parcial	<i>Integer</i>
Estatutos	Cada estatuto diferente deu origem a uma variável. Essas variáveis apresentam-se binárias, sendo a atribuição de 1 para casos positivos e 0 para casos negativos.	<i>Integer</i>
Regimes de ingresso:	Cada regime de ingresso diferente deu origem a uma variável. Essas variáveis apresentam-se binárias, sendo a atribuição de 1 para casos positivos e 0 para casos negativos.	<i>Integer</i>
Curso	Na diversidade de cursos existente no IPC, estes dividem-se maioritariamente nas seguintes áreas: Engenharia, Saúde e Gestão. Assim, foram criadas variáveis para cada uma destas áreas e os restantes foram agregados na variável de outros cursos, tendo em consideração que darão origem a variáveis binárias, sendo que o valor 1 é para alunos que se insiram nesses grupos e 0 para os que não se inserem	<i>Integer</i>

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Variável	Descrição	Tipo de dados
Instituição	Cada aluno tem associado a que instituição pertence, dando assim origem a uma coluna por cada escola como variável binária.	<i>Integer</i>
Disciplina_aprov	Número de disciplinas que obtiveram classificação superior a 10 valores	<i>Integer</i>
Disciplina_reprov	Número de disciplinas que obtiveram classificação inferior a 10 valores	<i>Integer</i>
Disciplina_desist	Número de disciplinas que não obtiveram classificação	<i>Integer</i>
ECTS_aprov	Total de ECTS que obtiveram classificação superior a 10 valores	<i>Integer</i>
ECTS_reprov	Total de ECTS que obtiveram classificação inferior a 10 valores	<i>Integer</i>
ECTS_desist	Total de ECTS que não obtiveram classificação	<i>Integer</i>
Número de alunos	Número de alunos que o curso respetivo contém no total da amostra	<i>Integer</i>
Médio rácio	Diferença que existe entre a média geral do aluno com a média do ano atual.	<i>Float</i>
Ects inscritos	Número total de ECTS que o aluno no início do ano se inscreveu	<i>Integer</i>
ECTS por inscrição	Número de ECTS total por inscrição	<i>Integer</i>

Distância da Instituição de ensino: Dificuldades financeiras, dada a atualidade e maioria dos estudos relacionados com AE ou até mesmo apenas o tema do ES, é considerado um tema urgente de se avaliar. Sendo que, anteriormente apenas se definiu como “perto” ou

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

“longe” a distância da morada do aluno para o concelho de Coimbra e a mesma não mostrou relevância, nesta abordagem optou-se por uma análise diferente, dado que as despesas monetárias têm tendência a crescer conforme o crescimento da distância da escola. Portanto, ao contrário do que se verificou anteriormente para definir a distância da escola agora tem-se em conta a distância em km. Para isso, recorri à ferramenta bastante conhecida de *web scraping* através de bibliotecas de python.

Web scraping é definida com uma técnica de extração de dados da web com o objetivo de os tornar disponíveis numa base de forma estruturada. Para isso, com recurso a bibliotecas de *python* é possível ter acesso ao código HTML das páginas web e assim conseguir aceder ao conteúdo pretendido. Neste âmbito, é necessário ter em consideração as questões éticas e legais da informação que se está a aceder, na medida em que é importante respeitar a privacidade dos utilizadores e evitar a violação de direitos de autor.

Para o estudo, o pedido de *web scraping* foi utilizado para recolher e armazenar as coordenadas das moradas dos alunos, com recurso aos seguintes *websites*:

- <https://www.codigo-postal.pt> , pois é considerado o *website* de moradas mais confiável em Portugal.
- <https://www.google.com/maps> , para colmatar falhas apresentadas pelo *website* anterior. Ou seja, quando as moradas após a recolha do *website* anterior apresentarem as coordenadas com os valores NAN ou Desconhecido, a função que faz o pedido ao Google Maps tenta corrigir.
- E, ainda <https://worldpostalcode.com> que tem como objetivo complementar todas as informações já recolhidas.

Para aplicar esta técnica, foi criado um conjunto denominado `df_local` que apenas inclui as variáveis necessárias do conjunto de dados inicial. O `df_local` agrega as variáveis representadas seguidamente na Figura 17.

```
df_local = df[['MATR_ID', 'LOCALIDADE', 'CODIGO_POSTAL', 'CPOSTAL', 'CPPREFIXO']].drop_duplicates()
```

Figura 17 - Definição do conjunto de dados "df_local"

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Df_local apresenta a estrutura exemplificada na Figura 18.

	MATR_ID	LOCALIDADE	CODIGO_POSTAL	CPOSTAL	CPPREFIXO
0	1019	Formoselha	SANTO VARÃO	3140	363
16	1026	Golães	GOLÃES	4820	451
27	1027	Casal Fernando	TENTÚGAL	3140	435
42	1028	Travasso	VACARIÇA	3050	510
57	1029	Agualva	AGUALVA	9760	18

Figura 18 - Estrutura do conjunto de dados "df_local"

Inicialmente, no *web scraping* ao website <https://www.codigo-postal.pt>, foi criada uma função "postal_code ()" que retorna as coordenadas de latitude e longitude do código postal de cada linha do conjunto de dados criado e se o site não encontrar as coordenadas do código postal, a função retorna o valor "NAN". Todas as coordenadas recolhidas são armazenadas na coluna nova criada "Coordenadas". Para esta técnica foi usada a biblioteca "Requests" para fazer solicitações HTTP ao site e a biblioteca "BeautifulSoup" para aceder ao HTML.

Após esta primeira abordagem, alguns dados ficaram sem correspondência de coordenadas, o que poderá ser corrigido com o mesmo *request* ao Google Maps.

Por fim, como estão incluídos alunos internacionais e por vezes as duas alternativas anteriores poderão não responder corretamente, foi criada uma outra função "worldpostalcode" que acede ao site: <https://worldpostalcode.com>.

Após as três alternativas para preencher a coluna "coordenadas" corretamente, foi necessário criar uma coluna que traduz a distância, das coordenadas de cada aluno para as escolas, em km. Para isso, na função "distance" é utilizada a fórmula de *Haversine*. Esta é uma equação matemática que ajuda a calcular a distância entre 2 pontos na Terra, usando as coordenadas de latitude e longitude. É abordada da seguinte forma apresentada em baixo, na Figura 19 e Figura 20, para as diferentes unidades orgânicas.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

```
instituicao_locdata = [
    ['ESTESC', '40.198180381753296,-8.461037203538314'],
    ['ISEC', '40.19265136700732,-8.41158990353859'],
    ['ESTGOH', '40.36097647207093,-7.8612131612014755'],
    ['ESEC', '40.205292759481296,-8.40595711888099'],
    ['ESAC', '40.21107213150218,-8.452502590457854'],
    ['ISCAC', '40.209633394362136,-8.452438222538632']
]

df_inst_loc = pd.DataFrame(instituicao_locdata, columns = ['Instituição', 'Instituição Coordenadas'])
```

Figura 19 - Coordenadas de cada uma das escolas pertencentes ao estudo

```
def distance(x):
    coord_morada = x['COORDENADAS']
    coord_instit = x['Instituição Coordenadas']

    coord_morada = coord_morada.split(',')
    coord_instit = coord_instit.split(',')

    lon1 = radians(float(coord_morada[1].strip()))
    lon2 = radians(float(coord_instit[1].strip()))
    lat1 = radians(float(coord_morada[0].strip()))
    lat2 = radians(float(coord_instit[0].strip()))

    dlon = lon2 - lon1
    dlat = lat2 - lat1
    a = sin(dlat / 2)**2 + cos(lat1) * cos(lat2) * sin(dlon / 2)**2

    c = 2 * asin(sqrt(a))
    r = 6371
    return abs(round(c * r, 3))

df_local['DIST_INSTITUICAO'] = df_local.apply(distance, axis = 1)
```

Figura 20 – Código em python do cálculo da distância de cada uma das coordenadas dos alunos às coordenadas das escolas

Sendo que r = raio equatorial da Terra, necessário para efetuar a conversão da distância para km.

Uma vez que os dados sobre as localidades dos alunos já foram transformados, é necessário guardá-los sem ter a necessidade de executar novamente as funções de *web scrapping*, a não ser que se trate de novos dados de alunos que ainda não foram submetidos a estes processos, isto porque, os processos descritos acima dependem de tempo e recursos. Com vista a melhorar o processo, recorreu-se a uma conexão de SQL da biblioteca *sqlite3* do *python* para guardar os dados numa base de dados. Uma vez

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

guardados os dados, é possível aceder aos mesmo através de uma *query*, como no exemplo da Figura 21.

```
df_local = pd.read_sql('SELECT * FROM t_local', conn)
```

Figura 21 - Código em python para cesso a base de dados com coordenadas

A variável que será considerada para os modelos será a que representa a distância em km: “dist_instituição”.

Localidade / Região: A localidade / região de origem do aluno foi tida em conta como uma variável adicional ao modelo, na medida em que tal como as coordenadas e distância medida em km, esta poderá estar relacionada com o AE.

Para o conjunto de dados ser distribuído corretamente pelas regiões foi utilizado um ficheiro csv auxiliar. Este ficheiro contém a correspondência entre os concelhos, cidade e regiões com os primeiros 2 dígitos do código postal, para assim poder ser feita a ligação com os dados dos alunos.

Para esta etapa, foi necessário criar um conjunto de dados temporário que só incluía as variáveis correlacionadas: Localidade e código postal. Assim, é possível que seja criada uma variável para a relação, que apenas incluía os primeiros 2 dígitos do código postal. Deste modo, é possível que seja associada a que região pertence cada aluno. Por fim, são criadas variáveis para cada região: Alentejo, Algarve, Açores, Centro, Madeira, Norte e Outro (os restantes que não pertencem a nenhuma das anteriores). Cada variável, terá o valor binário com o valor 1 a representar se o aluno pertence a essa região e o valor 0 caso contrário.

Cursos e instituições: Para uma melhor divisão por instituições de ensino da amostra de alunos que se está a trabalhar, foi utilizado mais um documento em excel que agrega todos os cursos ativos no momento da extração com as respetivas escolas. Assim, conseguimos ter mais uma variável para o modelo, tal como proceder a alguns testes específicos apresentados mais à frente. Cada instituição deu origem a uma variável, ficando o conjunto de dados das instituições com o apeto da Figura 22.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

	MATR_ID	ESAC	ESEC	ESTESC	ESTGOH	ISCAC	ISEC
0	1019	0	0	1	0	0	0
1	1026	0	0	1	0	0	0
2	1027	0	0	1	0	0	0
3	1028	0	0	1	0	0	0
4	1029	0	0	1	0	0	0

Figura 22 - Exemplo do conjunto de dados das variáveis das instituições

De seguida, para se ter uma ideia da população de cada curso e perceber quais os cursos com maior expressividade ou não na mesma escola, acrescentou-se uma coluna que faz a soma de alunos no curso que o aluno frequenta. Assim, em resumo consegue-se obter um conjunto de dados com a associação de cada curso à instituição respetiva com o número total de alunos respetivos, como mostra na seguinte Figura 23.

	Instituição		CURSO	NUM_ALUNOS
0	ESAC	Licenciatura em Agricultura Biológica		29
1	ESAC	Licenciatura em Biodiversidade e Conservação d...		6
2	ESAC	Licenciatura em Biotecnologia		169
3	ESAC	Licenciatura em Ciências Florestais e Recursos...		53
4	ESAC	Licenciatura em Enfermagem Veterinária		37

Figura 23 - Exemplo do conjunto de dados com instituição, curso e número de alunos

Rácio de média: Tal como na iteração anterior, nesta também se mostrou relevante ter uma variável com a média do ano atual que o aluno está a frequentar, mas teve-se em conta uma abordagem diferente. A média do ano atual também considera as UC em que o aluno ainda não obteve classificação, que ficam com a representatividade de 0. Assim, consegue-se ter uma perceção de que se o aluno tem muitas UC para concluir e se nas

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

restantes tem valores baixos (incluindo valores abaixo de 10 valores) poderá ser caso de alarme.

Em suma, com esta análise é possível perceber se a diferença para a sua média em geral é demasiado grande para a média apenas do ano letivo em questão. Considerando esta diferença como uma nova variável é possível perceber se existe realmente relação com o que se quer estudar ou não.

Regime de ingresso e estatutos: Os regimes de ingresso e estatutos, inicialmente encontravam-se todos na mesma coluna. No entanto, para a análise se tornar eficaz houve necessidade de fazer uma separação, como referido na 1ª iteração.

Estado civil: Originalmente no conjunto de dados existiam 6 opções diferentes de estado civil dos alunos. Entre estes, alguns como “Casado” ou “união de facto” que poderão estar diretamente ligados com os fatores referidos na revisão de literatura, como apoio a família. Assim, na mesma linha das variáveis anteriores, o conjunto de dados do estado civil ficou com o seguinte aspeto da Figura 24.

	MATR_ID	CASADO	DIVORCIADO	SEPARADO_JUDICIALMENTE	SOLTEIRO	UNIAO_FACTO	UNKNOWN	VIUVO
0	1019	0	0	0	1	0	0	0
1	1026	0	0	0	1	0	0	0
2	1027	0	0	0	1	0	0	0
3	1028	0	0	0	1	0	0	0
4	1029	0	0	0	1	0	0	0

Figura 24 - Exemplo do conjunto de dados com as variáveis derivadas do estado civil

Tendo em conta que a coluna “UNKNOWN” corresponde aos alunos que não têm associado a si qualquer opção no estado civil.

UC: As variáveis associadas às disciplinas e às suas respetivas notas são tratadas separadamente: Código da disciplina, disciplina, ECTS da disciplina e nota.

Dentro da variável nota poder-se-ão ver as seguintes variações:

- **Aprovado:** quando a variável nota é ≥ 10 ;
- **Reprovado:** quando a variável nota é < 10 ;

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

- **Desistência:** valores em branco.

Cada variação, dará origem a uma nova variável que conta cada caso para cada aluno individualmente.

Com recurso à criação de uma tabela dinâmica com os dados acima mencionados foi possível obter a seguinte estrutura: por cada aluno é mostrado o número de disciplinas aprovadas, reprovadas e desistidas assim como os respetivos ECTS para cada uma das 3 variações, como se pode verificar na Figura 25.

	MATR_ID	DISCIPLINA Aprov	DISCIPLINA Desist	DISCIPLINA Reprov	ECTS Aprov	ECTS Desist	ECTS Reprov
0	1019	8.0	5.0	3.0	48.0	22.0	12.0
1	1026	11.0	0.0	0.0	60.0	0.0	0.0

Figura 25 - Exemplo do conjunto de dados com variáveis derivadas de "disciplinas" e "ECTS"

Finalização das variáveis: Todas as variáveis que continham texto na composição foram modificadas, conforme descrito anteriormente. Isso deve-se ao fato de que os algoritmos não conseguem interpretar essas variáveis de outra forma, o que tornou necessário transformar essas variáveis categóricas. Após o pré-processamento de dados mencionado, surgiu a necessidade de melhorar os resultados, levando à utilização da biblioteca `"get_dummies()"` para a codificação. Como o nome sugere, essa função converte variáveis categóricas em variáveis fictícias. Isso significa que, por exemplo, a variável de estado civil deixa de ser uma única variável com descrições de cada estado civil e é transformada num conjunto de variáveis correspondentes ao número de estados diferentes presentes nos dados. Essas novas variáveis são preenchidas com valores 0 e 1, em que 0 indica a ausência e 1 a presença da categoria.

Portanto, é importante considerar a possibilidade de multicolinearidade entre as variáveis. Isso ocorre quando as variáveis independentes possuem relações lineares exatas ou muito próximas disso, ou seja, quando o valor de uma coluna pode ser deduzido automaticamente com base nas outras. No exemplo, se houver 6 tipos de estados civis diferentes, e correspondendo a isso, 6 variáveis fictícias, a última variável pode ser derivada das outras 5, o que cria multicolinearidade.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Para minimizar os erros associados ao tratamento de dados, optou-se por utilizar o argumento “*drop_first*” como verdadeiro na função “*get_dummies()*”, o que implica a remoção da primeira variável fictícia criada.

Variável de saída: Abandono: A variável de saída neste modelo segue as mesmas diretrizes do modelo apresentado anteriormente, como se pode comprovar na Figura 26.

```
df_target = df[['MATR_ID', 'ESTADO_MATRICULA', 'RENOVOU_ANO_SEGUINTE', 'TOTAL_INSCRICoes']].drop_duplicates()

df_target = pd.DataFrame((df_target.pivot_table(index = ['MATR_ID', 'RENOVOU_ANO_SEGUINTE'],
                                                columns = 'ESTADO_MATRICULA', values = 'TOTAL_INSCRICoes',
                                                aggfunc = 'count')).to_records())

df_target['RENOVOU_ANO_SEGUINTE'] = df_target['RENOVOU_ANO_SEGUINTE'].apply(lambda x: 1 if x == 'Sim' else 0)
df_target.fillna(0, inplace=True)
df_target = df_target.astype('int')

def target(row):
    ra = row['RENOVOU_ANO_SEGUINTE']
    act = row['Ativo']
    end = row['Concluído']
    inr = -1
    if 'Interrompido' in row.keys():
        inr = row['Interrompido']

    if (act == 1) & (ra == 0):
        return 1
    elif (inr == 1) & (ra == 0):
        return 1
    elif (act == 1) & (ra == 1):
        return 0
    elif (end == 1) & (ra == 0):
        return 0
    else:
        return -1

df_target['Abandono'] = df_target.apply(target, axis = 1)
df_target[['MATR_ID', 'Abandono']].drop_duplicates()
```

Figura 26 - Código em python da transformação que originou a variável "abandono"

Para o passo seguinte, é necessário juntar todos os subgrupos criados e descritos nos pontos acima. Assim, o conjunto de dados final apresenta a formatação apresentada no Apêndice I.

3.4 Análise descritiva do conjunto de dados modificado

No seguimento da análise descritiva apresentada no ponto 3.2 é apresentada agora a análise descritiva complementar referente às novas variáveis geradas para o novo

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

conjunto de dados modificados. Assim, neste ponto apenas serão apresentadas as estatísticas consideradas relevantes para o estudo e que sejam adicionais às já descritas.

De seguida, serão apresentadas as variáveis originadas na 1ª iteração.

Total de disciplinas: Dentro do conjunto de dados disponível, o número máximo de disciplinas em que um único aluno está inscrito é de 21.

No entanto, é importante notar que a situação mais frequente é que um aluno esteja inscrito em 12 disciplinas, englobando uma expressiva representatividade de 2092 alunos com essa característica específica.

Podemos ver a distribuição de número de disciplinas a que um aluno está inscrito por ano letivo através da Figura 27.

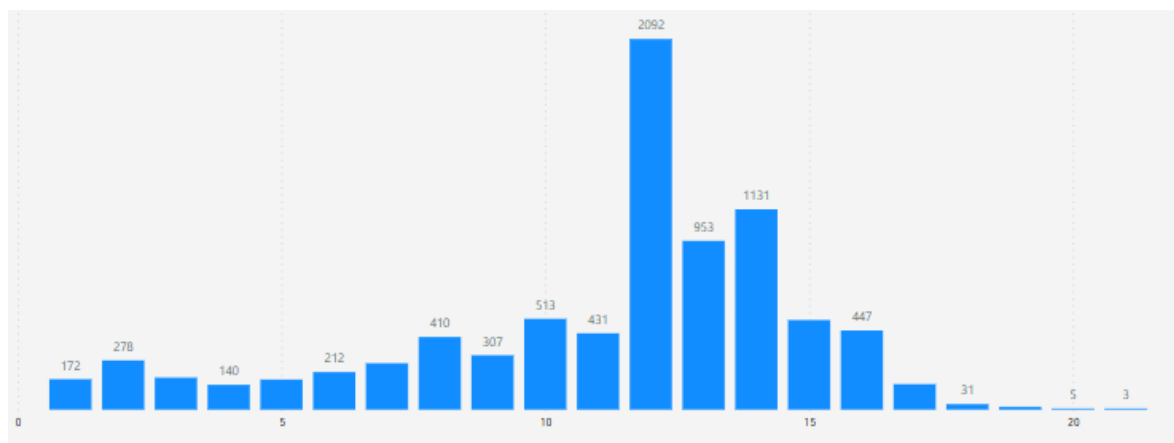


Figura 27 - Gráfico de barras com a distribuição do número de disciplinas a que um aluno está inscrito no ano letivo 2020/2021

Total superior a 10: Dadas as discrepâncias entre alunos, pode-se considerar que em média os alunos em estudo completaram 8.2 disciplinas ao longo do ano. Apenas 3 alunos completaram 18 disciplinas e 226 não completaram nenhuma. Ainda, à semelhança da variável anterior o mais frequente será os alunos completarem 12 disciplinas, com uma representatividade de 926 alunos, como podemos ver em maior detalhe na Figura 28.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

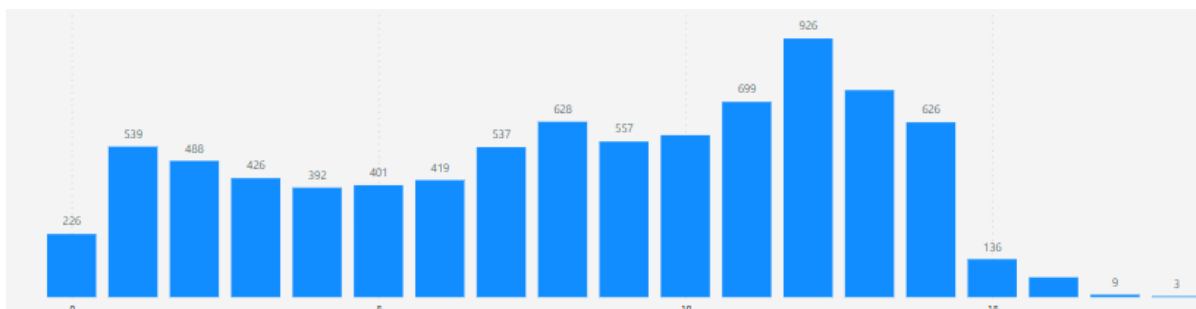


Figura 28 - Gráfico de barras com a distribuição do número de disciplinas realizadas com aprovação por aluno, no ano letivo 2020/2021

Proximidade: A distribuição dos alunos entre “perto e “longe”, está em 28% e 72%, respetivamente, como está apresentado na Figura 29.



Figura 29 - Gráfico circular da variável "proximidade"

Abandono: Na sua forma final, o conjunto de dados encontra-se com 1238 alunos que abandonaram o ES, como se pode verificar na Figura 30.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar



Figura 30 - Gráfico circular com descrição de como se divide a variável "abandono"

Nesta segunda parte, serão apresentadas as estatísticas descritivas das novas variáveis geradas na 2ª iteração.

Rácio de média: Dos dados disponíveis a distribuição da variável de rácio de médias apresenta-se como representado na Figura 31, tendo em conta que se excluíram o aluno que não apresentam variação.



Figura 31 - Representação da distribuição da variável "rácio médias"

3.5 Feature scaling

Para concluir a preparação de dados foi também efetuada um *feature scaling*. Todas as variáveis sofreram as modificações necessárias para integrarem no algoritmo. No entanto, as demais encontram-se em diferentes escalas. Na medida em que, se está a considerar a integração efetiva no algoritmo de variáveis com escalas de 0 a 20 (notas dos alunos), variáveis binárias, entre outras será crucial aplicar regras para normalizar os dados e que estes não tenham mais peso em relação aos outros apenas pela sua dimensão de escala, a este processo de normalização é comumente chamado de *feature scaling*.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

feature scaling é um passo importante em inúmeros processos de ML, pois diferentes variáveis podem ter diferentes intervalos de valores. Esta ação pode ocorrer através de dois recursos: normalização e padronização. Por um lado, a normalização transforma todas as variáveis de modo a estas estarem limitadas no intervalo $[0,1]$ ou $[-1,1]$. Por outro lado, a padronização é mais radical e transforma todos os dados para não possuírem unidade, de modo que tenham média 0 e variação 1. Na Figura 32 é apresentado graficamente como se comportam cada uma das opções de *feature scaling*.

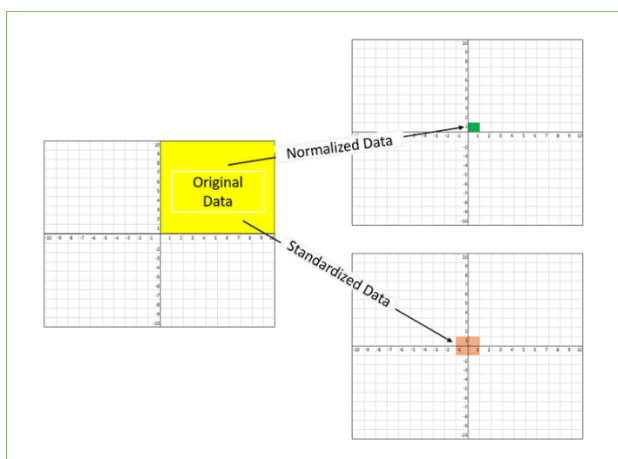


Figura 32 - Ilustração do resultado dos diferentes tipos de *feature scaling*

Fonte: <https://www.turing.com/kb/how-and-where-to-apply-feature-scaling-in-python>

Aplicando as diferentes escalas ao algoritmo KNN, consegue-se medir o seu impacto. Na medida em que, irá ser considerada a distância entre dois pontos (pela distância euclidiana). Portanto, se uma variável tiver um intervalo maior de valores, irá ser utilizada como a principal, o que pode não corresponder à verdade dos dados.

Este processo, na sua grande maioria é aplicado como uma etapa de pré-processamento, o que além de benefícios que poderá trazer para o modelo, como foi referido anteriormente, pode também estar suscetível a alguns problemas.

Uma das normalizações que podem ser usadas é a normalização *min-max*, que consiste num método simples e que tem como objetivo redimensionar o intervalo das variáveis

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

para que variem entre [0,1] ou [-1,1], em que este intervalo é baseado nos dados e qual fará mais sentido. É seguida pela seguinte fórmula:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Onde x é um valor original, x' é o valor normalizado.

Para redimensionar um intervalo entre um conjunto arbitrário de valores [a, b], a fórmula seria seguinte:

$$x' = a + \frac{(x - \min(x))(b - a)}{\max(x) - \min(x)}$$

Onde a, b são os valores mínimo-máximo.

Por outro lado, além da normalização *min-max* também pode ser utilizada a padronização, que ao contrário do exemplo anterior faz com que os valores de cada variável tenham média zero (ao subtrair a média no numerador) e a distância unitária. A fórmula para calcular a média seria:

$$x' = \frac{x - \bar{x}}{\sigma}$$

Onde x é o vetor de variável original, ($\bar{x}$) é a média de x e (σ) é o seu desvio padrão.

Este método de padronização é ideal para ser aplicado quando os dados apresentam uma distribuição normal.

Com isto, podemos perceber as diferenças entre os dois métodos apresentados em cima através da representação gráfica de Figura 33.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

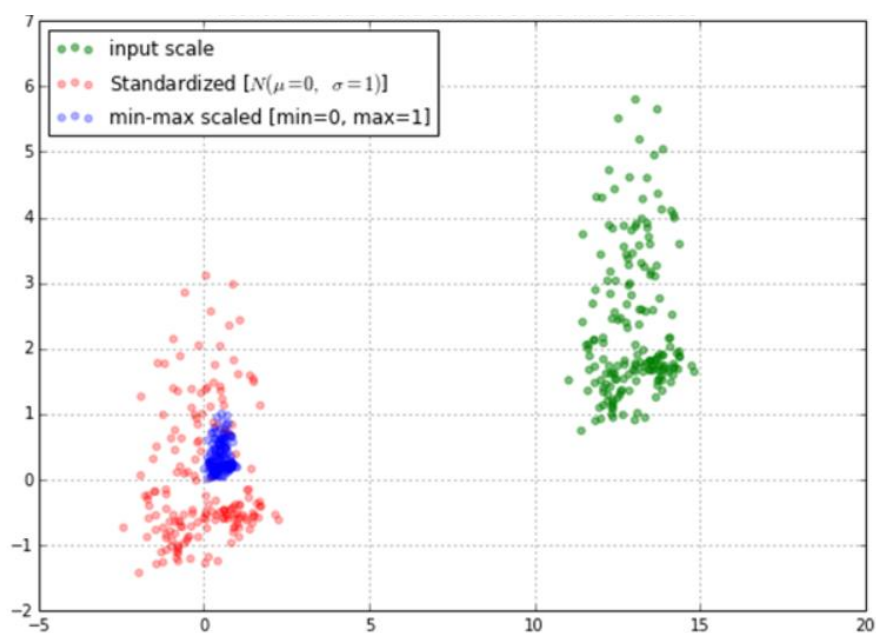


Figura 33 - Exemplo gráfico de como se agrupam os dados após aplicação de normalização e padronização

Fonte <https://atoti.io/articles/when-to-perform-a-feature-scaling/>

4 Modelação

Aprendizagem automática (ML, do inglês *machine learning*) é uma área da inteligência artificial que permite aos dispositivos informáticos que aprendam a partir dos dados sem serem especificamente programados. Cada vez mais vemos este tema em crescimento, seja na área científica, ou no nosso dia a dia, mesmo que não nos demos conta.

Netflix ou recomendações do Facebook são alguns exemplos em que podemos ver esta aprendizagem em ação.

No tema de AE, ML é útil na sua utilização para identificar os alunos que poderão estar em risco de AE. Ao analisar os dados recolhidos da plataforma sobre os alunos, como o desempenho académico e comportamento nas diversas áreas de recolha, os computadores podem identificar padrões que posteriormente identificam os alunos em risco. Estes mesmos dados e informações geradas poderão posteriormente ser utilizadas para desenvolver programas de intervenção personalizados para ajudar os alunos a permanecer as instituições de ensino.

4.1 KNN

Um dos algoritmos mais conhecidos neste meio é o KNN. Ele é comumente utilizado para resolver problemas de classificação e regressão. É considerado um dos algoritmos mais simples e fáceis de compreender, pois funciona medindo a distância entre um novo ponto de dados e todos os pontos de dados no conjunto de treino. O número de pontos considerados na execução do algoritmo (representado por k) é definido pelo utilizador. Este valor corresponde ao número de pontos mais próximos que são utilizados para determinar a classificação do novo ponto de dados.

Apesar da simplicidade deste algoritmo, existem diversas vantagens em ser utilizado, nomeadamente:

- **Algoritmo não paramétrico**, isto é, não efetua suposições explícitas sobre a distribuição dos dados. Ou seja, torna-se um algoritmo flexível e adequado para uma maior diversidade de temas, incluindo a previsão do AE.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

- **Boa adaptação à mudança constante de dados**, uma vez que não ocorre um treino formal. Desta forma, o algoritmo adapta-se perfeitamente a um conjunto de dados que vai sendo atualizado ao longo do tempo, como é o caso dos dados dos alunos. Assim, as variáveis como média, número de ECTS aprovados, etc, poderão ser atualizados sem se ter a necessidade de treinar novamente o modelo.
- **Robustez a Outliers**, pois como apenas são considerados os vizinhos mais próximos, consegue facilmente não ser afetado por valores díspares.
- **Não requer um treino formal nem exigente**, pois vai armazenando no momento os dados de treino.

Com todas estas vantagens, é um algoritmo bastante utilizado e com bons resultados em diversos temas.

4.2 Árvore de Decisão

As árvores de decisão, à semelhança do KNN, são também um algoritmo de ML que tanto pode ser usado para classificação como para regressão. As árvores de decisão de classificação têm o objetivo de qualificar os registos, associá-los com a classe determinada. No caso do AE terá uma classificação binária. À medida que se forma a árvore de decisão, a mesma vai formando uma simples de representação / relação existente num conjunto de dados. Vai criando assim, subconjuntos cada vez menores, com base num conjunto de regras simples. Inicialmente, é utilizado o conjunto de dados inicial, que é dividido em duas partes. Estas partes estão relacionadas entre si tendo por base um atributo específico selecionado. Este processo, é repetido ao longo da árvore até que cada subconjunto tenha dados apenas de uma classe. Assim, resulta uma árvore de decisão com regras que podem ser utilizadas para prever a variável definida como alvo.

Apesar de estarmos perante um algoritmo simples, existem outras vantagens neste modelo que são valorizadas:

- **Facilidade de compreensão.**

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

- **Boa performance tanto para classificação como regressão.**
- **Versatilidade.**

Por outro lado, existem algumas desvantagens que precisam de ser tidas em conta, como por exemplo:

- Algoritmo submetido a certos dados, poderá tornar-se **complexo** e de difícil compreensão.
- **Overfitting:** pode dar-se *overfitting* quando o modelo se ajusta excessivamente aos dados de treino, perdendo capacidade para aplicar o modelo a dados novos.

4.3 Matriz de confusão, accuracy, recall, precision, f1-score

A matriz de confusão considera-se uma tabela que compara, em números absolutos, as classes reais com as classes preditas. Na diagonal da matriz estão os valores que indicam classificações corretas. Esta é composta por quatro quadrantes, dos quais:

- **Verdadeiro positivo (TP):** Dados que foram classificados corretamente como positivos.
- **Falso positivo (FP):** Dados que foram classificados incorretamente como positivos.
- **Verdadeiro negativo (TN):** Dados que foram classificados corretamente como negativos.
- **Falso negativo (FN):** Dados que foram classificados incorretamente como negativos.

Através desta matriz, conseguimos ainda calcular todas as métricas apresentadas anteriormente. É considerada como uma ferramenta versátil que pode ser usada para avaliar o desempenho de algoritmos de classificação, conseguindo mostrar quais são os pontos positivos e quais o modelo necessita de ser melhorado. Na Tabela 7 é apresentado como se apresenta a matriz para o problema em causa.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

		Valor predito	
		Não Abandono	Abandono
R E A L	Não Abandono	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Abandono	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Tabela 7 - Matriz de confusão

Todas as métricas apresentadas seguidamente podem ser calculadas para problemas de classificação binária e multiclasse.

A métrica de avaliação *accuracy* é a mais utilizada quando falamos de medir a viabilidade de um algoritmo. Esta medida mede a proporção de pontos de dados que foram classificados corretamente. A mesma é calculada como a razão entre o número de pontos de dados classificados corretamente e o número total de pontos de dados.

A métrica *recall*, é semelhante à anterior, na medida em que também se baseia na medida de proporção dos pontos de dados. No entanto, esta medida foca-se nos dados que são realmente positivos e que foram classificados como positivos. É uma métrica bastante relevante para o estudo em causa, pois queremos o maior número de alunos que têm comportamentos preditivos de AE a serem identificados atempadamente, como medida preventiva.

Assim, esta medida tem a seguinte sintaxe:

$$\text{recall} = TP / (TP + FN)$$

onde:

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

- **TP:** número de pontos de dados positivos que foram classificados corretamente.
- **FN:** número de pontos de dados positivos que foram classificados incorretamente.

Considera-se, também uma métrica complementar à precisão, que irá ser apresentada em seguida. Em conclusão, *recall* é uma métrica importante de avaliação de algoritmos de aprendizagem supervisionada que é útil para problemas de classificação onde é importante evitar o falso negativo.

No mesmo alinhamento da métrica anterior, enquadra-se a precisão que mede a proporção de pontos de dados que foram classificados corretamente. É calculada como a razão entre o número de pontos de dados classificados corretamente e o número total de pontos de dados. Trata-se de uma métrica importante para problemas onde é crucial evitar o falso positivo. Especificamente, a precisão aplicada ao algoritmo KNN depende dos seguintes fatores:

- O número de vizinhos K: Controla o número de pontos de dados considerados para fazer a previsão. Um valor maior de K tende a aumentar o valor da precisão, por outro lado poderá afetar a sensibilidade, uma vez que tem tendência de considerar pontos muito distantes e possivelmente foram da região desejada.
- A distância: a distância utilizada para medir a proximidade entre os pontos de dados (distância Euclidiana, *Manhatan* ou *Minkowski*).

Tendo em consideração a combinação das métricas *recall* e precisão, foi criada a nova métrica *f1-score*, sendo calculada pela média das duas métricas referidas anteriormente. O *f1-score* poderá devolver valores entre 0 e 1, sendo o 1 quando é aplicado a um modelo perfeito e 0 quando aplicado a um modelo que não é capaz de distinguir as classes.

É calculado tendo em conta a seguinte fórmula:

$$F1 = 2 * (\text{precisão} * \text{recall}) / (\text{precisão} + \text{recall})$$

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

onde:

- **precision:** proporção de pontos de dados classificados como positivos e que realmente são positivos.
- **recall:** proporção de pontos de dados que são realmente positivos e que foram classificados como positivos.

O *f1-score* é uma métrica complementar à precisão e ao *recall*. A precisão e o *recall* são métricas de deteção e precisão. Um alto valor de *f1-score* indica que o modelo é bom em ambas as tarefas.

4.4 Aplicação do algoritmo KNN

4.4.1 1ª iteração

Nesta primeira fase de implementação o algoritmo preferencial será o KNN, com possibilidades de ajustes e ponderação de adição de complexidade, seja no próprio algoritmo ou implementação de outro.

Tendo em conta o tratamento de dados realizado em 3.3, deu o resultado das seguintes colunas: Média, Total de disciplinas, Total superior a 10, Total ECTS, ECTS feitos, Totais inscrições, Regime maiores de 23 anos, Estatuto trabalhador-estudante, Estatuto falta de pagamento propinas ou emolumentos, Estatuto de aluno com NEE e Abandono

Para normalizar todas as colunas, tendo em conta a sua discrepância de escalas, utilizou-se o método “*MinMaxScaler*” como apresenta a Figura 34.

```
min_max_scaler = MinMaxScaler()
data_min_max = min_max_scaler.fit_transform(df_knn)
```

Figura 34 - Código em python para aplicação da normalização na 1ª iteração

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Com isto, as escalas dos dados irão estar entre [0,1] e assim devolvem melhores resultados.

Após isto, o conjunto de dados foi dividido em *features* e *target* e aplicou-se a divisão dos dados com 80% dos dados para treino e 20% para teste. Para terminar, aplicou-se o algoritmo, tendo em consideração nove vizinhos e a métrica euclidiana, como se pode ver na Figura 35.

```
clf = KNeighborsClassifier(n_neighbors=9, p=2, metric='euclidean')
```

Figura 35 - Código em python para o algoritmo KNN

4.4.2 2ª iteração

Para melhor interpretação e tratamentos dos dados, as variáveis são armazenadas em diferentes conjuntos de dados consoante a sua tipologia, categóricas ou numéricas. Assim, as variáveis categóricas ficam agregadas no conjunto de dados “cat_cols” e as variáveis numéricas constituem o conjunto de dados “val_cols”. Para as análises que se seguem será também necessário guardar a variável de saída: abandono, como *target*.

As variáveis categóricas, numéricas e de saída consideradas para integrar o modelo nesta 2ª iteração estão discriminadas no Apêndice II – Variáveis categóricas, numéricas e de saída. Nesta lista, tiveram-se em consideração os resultados de vários testes realizados. O que levou à exclusão das variáveis associadas às regiões e aos cursos a que cada aluno pertencia.

Ao escolher manualmente as variáveis como representado acima, poderia levar a alguma intervenção humana num futuro, quando o modelo fosse submetido a diferentes dados de diferentes anos. Isto porque, em cima estão descritas as variáveis, por exemplo, dos estatutos que foram obtidas por uma divisão de todas as opções apresentadas na coluna. Uma vez que, essa coluna, noutras circunstâncias, poderia incluir outros exemplos, será

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

vantajoso que as variáveis sejam agregadas automaticamente. Para isso, recorreu-se ao código apresentado na Figura 36.

```
cat_cols_list = [[1,0],[0,1],[False,True],[True,False]]

cat_cols = []
val_cols = []
target = 'Abandono'

for c in df_modelo.columns:
    if c != target and c != 'tipo_dados':
        resp = list(df_modelo[c].unique())
        if (len(resp) == 2) and (resp in cat_cols_list):
            cat_cols.append(c)
        else:
            val_cols.append(c)
```

Figura 36 - Código em python que identifica e separa as variáveis categóricas e numéricas automaticamente

Com estes automatismos, será desnecessária qualquer intervenção humana para identificação das variáveis.

Importância das variáveis

Com o objetivo de avaliar as variáveis através do seu grau de importância, para as variáveis categóricas foi utilizado o “*ordinalencoder()*” que é um método pertencente ao *scikit learn* que codifica as variáveis categóricas ordinais em valores numéricos. Este método atribui um valor inteiro a cada categoria, com base na ordem natural das categorias. Por outro lado, para a variável de saída foi submetida ao método “*LabelEncoder()*” que codifica a variável entre 0 e 1. Este método atribui um valor inteiro a cada categoria de forma arbitrária, ou seja, não há necessariamente uma ordem entre as categorias.

Para as variáveis numéricas, uma vez que já se encontram no formato pretendido e ideal para integrar o algoritmo não foi aplicada mais nenhuma alteração.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Tendo em conta quais as variáveis finais a integrarem o algoritmo é relevante perceber de que maneira se distribuem as suas importâncias. Para calcular a importância das variáveis, foi tida em conta a sua divisão, entre variáveis categóricas e variáveis numéricas.

Por um lado, para as variáveis categóricas foi usada o método “*SelectKBest*” que será novamente abordado e explicado no próximo ponto. Com isto foi possível ter acesso à atribuição das pontuações de importância de cada variável, tem-se como exemplo, as quatro variáveis categóricas apresentadas na Figura 37.

```
BOLSEIRO_COD: 7.845857
Regime__(ISEC) Regimes Especiais: 2.541663
Regime__(ISEC) Transitou de Plano de Estudos: 1.209322
Regime__Acores Continente nivel 1: 0.985685
```

Figura 37 - Exemplo do resultado da importância das variáveis na 2ª iteração

Ou em formato de gráfico, na Figura 38, para se conseguir perceber a ordem pela qual elas se encontram.

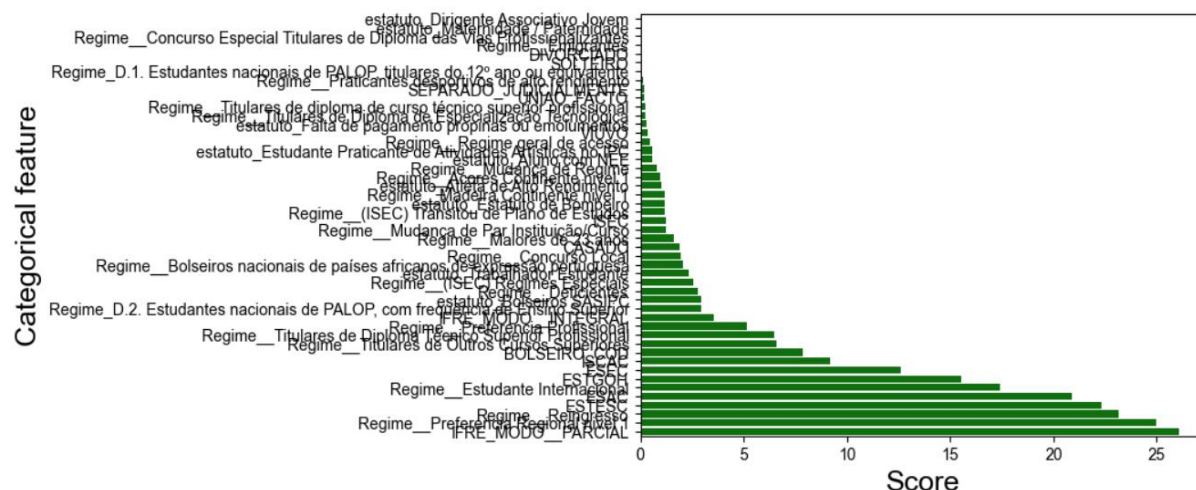


Figura 38 - Representação gráfica do grau de importância das variáveis categóricas na 2ª iteração

Por outro lado, para a importância das variáveis numéricas foi utilizada o método “*ExtraTreesClassifier*” que é indicada para fornecer uma lista de valores representando

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

o grau de importância de cada característica no processo de classificação. O gráfico representativo na Figura 39 para estas variáveis.

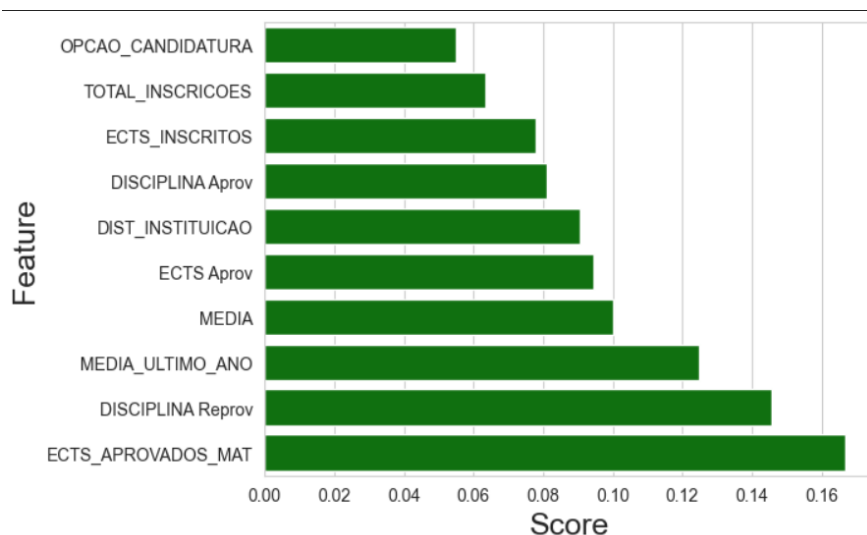


Figura 39 - Representação gráfica do grau de importância das variáveis numéricas na 2ª iteração

Todas as modificações efetuadas aos dados em *python* e aplicação dos modelos estão disponibilizados no Google Drive. Em Apêndice III – Acesso ao código python e ficheiros é apresentado o link que dá acesso a todos os ficheiros.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

5 Resultados

5.1 KNN

No que diz respeito à 1ª iteração apresentada em cima, os resultados não foram os melhores, o que fez com que surgisse a necessidade de criar um modelo mais robusto apresentado na 2ª iteração. Antes de se avançar em maior pormenor para os resultados que se mostraram mais aliciantes, importa ter em atenção os resultados da 1ª iteração, apresentados na Figura 40.

	precision	recall	f1-score	support
0.0	0.89	0.97	0.93	1431
1.0	0.67	0.33	0.44	250
accuracy			0.88	1681
macro avg	0.78	0.65	0.69	1681
weighted avg	0.86	0.88	0.86	1681

Figura 40 - Resultados do algoritmo da 1ª iteração

Tendo em consideração que a classe 0 é não abandono e 1 abandono, conseguem-se tirar as seguintes conclusões:

- Obteve-se um *recall* para a previsão do AE abaixo de 0.5, o que pode ser considerado baixo e tornar o modelo não confiável. Neste resultado apenas 33% dos alunos foram identificados com a probabilidade de praticar AE e realmente se verificou verdadeira a previsão.
- O valor de precisão obtido para o AE foi de 0.67. Tendo em conta estes resultados, obtém-se 0.44 de *f1-score* que fica abaixo do que seria esperado.

Avaliando os resultados obtidos, houve necessidade de avançar para uma 2ª iteração tendo em conta alguns pontos de melhorias nesta 1ª iteração, nomeadamente:

- Ter em consideração outras variáveis.
- Fazer um tratamento de dados mais criterioso.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

- Nesta 1ª iteração foram atribuídos valores para cada opção de variáveis, o que pode enviesar a interpretação do algoritmo e pode não estar 100% correto. Assim, na iteração seguinte será um ponto a melhorar e apenas considerar dois tipos de variáveis: categóricas variando apenas entre 0 e 1 e numéricas.

Referente à 2ª iteração, nesta última parte antes da integração dos dados para o algoritmo tiveram-se em conta vários pontos: a separação das variáveis categóricas e numéricas, variável de saída, divisão dos dados em treino e teste (mesmo que se refira a um teste dentro da totalidade do treino, usualmente também chamada de validação), normalização, *loop* para escolher as variáveis, função para escolher o *k*, aplicação do KNN.

Primeiramente, foi possível, como anteriormente referido, nomear o conjunto de variáveis categóricas e numéricas separadamente. Assim como, nomear a variável de saída. Em seguida, foi aplicado a divisão treino e teste. Neste caso, utilizou-se a divisão de 80%, 20% para o conjunto de treino e teste, respetivamente.

Após a divisão, aplicou-se a função de normalização. Tendo em conta a divisão feita anteriormente entre variáveis categórica e numéricas, consegue-se perceber que as escalas de ambas as partições diferem entre si. O que releva a necessidade de aplicar a normalização referida no ponto 3.5.

- Numa primeira fase aplicou-se a normalização antes da divisão dos dados treino e teste. Esta normalização apenas se aplicou nas variáveis numéricas, para que as mesmas passem a estar entre o intervalo [0,1], na medida em que as restantes variáveis categóricas já se encontram neste formato. Neste caso para a normalização foi utilizada a biblioteca “*MinMaxScaler()*”. Sendo utilizada antes da divisão poderá traduzir em alguns erros, o que levou a aplicação da mesma biblioteca depois da divisão em treino e teste.
- Com a modificação para a normalização após a divisão treino e teste, notou-se a presença de alguns outliers, nesta segunda versão optou-se pela padronização para normalizar os dados, com o que se obteve resultados melhores.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Concluída a normalização, efetuou-se a escolha automática das variáveis a integrar no modelo. Uma das falhas apresentadas anteriormente seria a seleção das variáveis a ter em conta no modelo, para que o mesmo obtivesse os melhores resultados. Deste modo, nesta segunda iteração do modelo optou-se por fazer esta seleção automática que poderá ajustar-se aos dados a que o modelo é submetido. Ou seja, sendo que o modelo será utilizado ao longo dos anos e com dados de entrada diferentes, o modelo está com a capacidade de escolher as melhores variáveis, assim como o número ótimo de k como se poderá ver a descrição posteriormente. Para a escolha das variáveis foi utilizado o método “*SelectKBest*”. Este é um algoritmo de seleção de variáveis que seleciona as k variáveis com maior importância para a variável dependente. Nesse contexto, foi utilizado o algoritmo “*mutual_info_classif*,” que aplica a Estatística de Informação Mútua para avaliar as relações entre as variáveis independentes e a variável dependente. Essa abordagem é robusta, pois consegue detetar associações mesmo em conjuntos de dados que contenham variáveis categóricas e numéricas.

A ideia é encontrar uma combinação que resulte no menor erro (RMSE) ao ajustar o modelo KNN aos dados de treino.

Para a escolha do melhor número de vizinhos apresentam-se duas opções. Na medida em que, inicialmente no desenvolvimento do código python foi desenvolvida uma função que testava vários valores de “ k ” e selecionava o valor em que obtivesse o menor erro médio quadrático (RMSE). Os números possíveis de k são guardados num *array* com auxílio da biblioteca “*NumPy*”, em que constam no intervalo de 1 a 26, exclusive. Este número foi definido consoante o que é considerado o mais eficaz para o algoritmo KNN.

No entanto, apesar dos bons resultados obtidos pela função anterior o modelo ficava com uma maior carga para ser executado o que faria com que o tempo e recursos utilizados fossem superiores. Assim, houve a necessidade de pesquisar uma nova versão. A segunda versão, consiste na utilização de um método diferente para a procura de melhor número de vizinhos no modelo (“ k ”): *GridSearchCV()*.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Este método da biblioteca scikit-learn percorre todos os parâmetros que estão inseridos na tabela de parâmetros e produz a melhor combinação numa métrica escolhida pelo utilizador. Neste caso, a métrica escolhida foi a *accuracy*. Um dos campos adicionais desta biblioteca é o número de *folds* para a validação cruzada, que nesta situação foi escolhido o número 10. E, por fim, é possível definir o “*n-jobs*”, que no caso de ser escolhido o valor -1 é sinónimo de usar todos os núcleos disponíveis da máquina que está a executar o código. Em suma, o objetivo será sempre encontrar o melhor valor para os parâmetros de vizinhos do classificador do KNN.

Todos os passos descritos acima foram os necessários e utilizados para a execução do algoritmo. Em *python*, a seguinte Figura 41 representa o código utilizado para a execução do mesmo.

```
report, l1, l2, acc = knn_result(k, X_train_fs, y_train, X_test_fs, y_test)
```

Figura 41 - Código em python para a execução do algoritmo KNN na 2ª iteração

Em que:

- K é definido pelo método *GridSearchCV()*.
- Conjunto de treino e teste apenas contém as variáveis definidas automaticamente.

Ao longo desta pesquisa, foram realizados diversos testes com diferentes configurações e técnicas de modelação com o objetivo de aprimorar a precisão das previsões.

É importante realçar que, neste estudo, observam-se variações nos resultados obtidos durante esses testes. Embora algumas técnicas tenham apresentado desempenho superior em certos casos, a variabilidade nos resultados não deve ser interpretada como uma razão para descartar outras abordagens. Em vez disso, percebe-se que os resultados encontrados podem estar diretamente relacionados às características exclusivas dos dados usados.

Assim, as abordagens utilizadas são as seguintes:

- Opção 1: Dados do ano letivo 2020/2021 utilizados como treino. Estes dados ao longo da modelação do algoritmo são divididos em treino e teste, correspondendo esta última divisão a uma validação do treino efetuado. O teste efetivo será

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

aplicado nos novos dados em que se pretende testar o modelo em tempo real, quando o mesmo estiver implementado no IPC.

- Opção 2: Conjugação dos dois anos letivos disponíveis, 2020/2021 e 2021/2022, dividindo como treino e teste, servindo o segundo ano letivo como teste.
- Opção 3: E, por fim, apenas considerar um ano letivo para todo o modelo. Isto, com o ano letivo 2020/2021 efetuaram as divisões em treino e teste, sendo que mais uma vez o treino está dividido em treino e teste novamente, correspondendo esta segunda divisão de teste a uma validação.

Os resultados da **opção 1** revelam-se os seguintes apresentados na Figura 42 e as variáveis consideradas mais importantes e vantajosas para o modelo, apresentadas na Figura 43.

k	prediction	precision	recall	f1-score	support	features	accuracy
441	9	1	0.76	0.52	0.62	250.0	BOLSEIRO_COD; Regime__(ISEC) Transitou de Plan... 0.903629

Figura 42 - Resultado do algoritmo KNN da 2ª iteração - opção 1

```
[ 'BOLSEIRO_COD',
  'Regime__(ISEC) Transitou de Plano de Estudos',
  'Regime__Bolseiros nacionais de países africanos de expressão portuguesa',
  'Regime__D.2. Estudantes nacionais de PALOP, com frequência de Ensino Superior',
  'Regime__Deficientes',
  'Regime__Emigrantes',
  'Regime__Estudante Internacional',
  'Regime__Preferencia Profissional',
  'Regime__Regime geral de acesso',
  'Regime__Reingresso',
  'Regime__Titulares de Diploma Técnico Superior Profissional',
  'Regime__Titulares de Diploma de Especialização Tecnológica',
  'Regime__Titulares de Outros Cursos Superiores',
  'estatuto_Bolseiros SASIPC',
  'estatuto_Maternidade / Paternidade',
  'ESAC',
  'ESEC',
  'ESTESC',
  'ISCAC',
  'ISEC',
  'SEPARADO_JUDICIALMENTE',
  'IFRE_MODALIDADE_INTEGRAL',
  'IFRE_MODALIDADE_PARCIAL',
  'ECTS_APROVADOS_MAT']
```

Figura 43 - Resultado de variáveis do algoritmo KNN da 2ª iteração - opção 1

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Com os resultados desta opção podemos concluir que o modelo apresenta uma alta precisão (0.76). Esta precisão indica a capacidade do modelo de minimizar os falsos positivos, ou seja, identificar corretamente a maioria dos alunos que estão realmente em risco de AE. Por outro lado, na métrica de *recall* o valor apresenta-se mais baixo, mas com um valor considerado bom. O que significa que o modelo tem uma capacidade de aproximadamente 52% de identificar todos os alunos que estão em risco de AE. Poderão estar a existir alguns alunos que apresentam risco de abandonar, mas o modelo não está a identificar. Este valor ligeiramente abaixo da primeira métrica apresentada poderá estar justificado pelo universo dos dados de alunos que efetivamente praticaram AE. Com isto, dificulta a capacidade do modelo de identificar corretamente os alunos que efetivamente abandonaram. Deste modo, a baixa taxa de *recall* indica que o modelo deixa espaço para melhorias na identificação de todos os alunos em risco. O que pode levar a melhorias futuras do modelo, identificando outras variáveis que ajudem a maximizar estes resultados.

Conjugando as duas métricas apresentadas, precisão e *recall*, obteve-se um valor de *f1-score* de 62%, o que se revela um valor bom.

Por fim, apresenta um valor de *accuracy* de 0.90 que sugere que o modelo é eficaz na classificação geral dos alunos.

Tendo em consideração o número de *k* e as variáveis escolhidas no modelo de treino, aplicou-se o modelo nos dados de teste. Os resultados estão descritos na Figura 44.

	precision	recall	f1-score	support
0	0.91	0.97	0.94	1431
1	0.73	0.46	0.56	250
accuracy			0.89	1681
macro avg	0.82	0.71	0.75	1681
weighted avg	0.88	0.89	0.88	1681

Figura 44 - Resultados do algoritmo KNN na partição de teste da 2ª iteração - opção 1

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Os resultados apresentam-se inferiores, o que poderá ser justificado pela menor quantidade de dados presente. No entanto, apesar de apresentar valores de *recall* inferiores, apenas apresenta um valor de *f1-score* seis pontos abaixo.

Os resultados da **opção 2** revelam-se os seguintes apresentados na Figura 45 e as variáveis consideradas mais importantes e vantajosas para o modelo, apresentadas na Figura 46. É necessário ter em atenção que, ao contrário das restantes opções, nesta opção não se mostrou relevante retirar do modelo as variáveis respetivas a cursos dos alunos e portanto foram consideradas.

	k	prediction	precision	recall	f1-score	support	features	accuracy
305	7	1	0.51	0.44	0.48	1375.0	CURSO_ENGENHARIA_COD; CURSO_GESTAO_CONTAB_COD;...	0.848214

Figura 45 - Resultado do algoritmo KNN da 2ª iteração - opção 2

```
[ 'CURSO_ENGENHARIA_COD',
  'CURSO_GESTAO_CONTAB_COD',
  'BOLSEIRO_Sim',
  'Regime__(ISEC) Mudança Interna de Curso',
  'Regime__(ISEC) Regimes Especiais',
  'Regime__(ISEC) Transitou de Plano de Estudos',
  'Regime_B Cidadãos Portugueses Bolseiros ou Equiparados',
  'Regime_Bolseiros nacionais de países africanos de expressão portuguesa',
  'Regime_Concurso Local',
  'Regime_Deficientes',
  'Regime_Estudante Internacional',
  'Regime_Mudança de Regime',
  'Regime_Preferencia Profissional',
  'Regime_Preferencia Regional nivel 1',
  'Regime_Regime geral de acesso',
  'Regime_Reingresso',
  'Regime_Titulares de Diploma Técnico Superior Profissional',
  'Regime_Titulares de Diploma de Especialização Tecnológica',
  'Regime_Titulares de Outros Cursos Superiores',
  'Regime_Titulares de Outros Cursos Superiores Lista Única',
  'Regime_Titulares de diploma de curso técnico superior profissional',
  'Regime_Vagas Autónomas',
  'estatuto_Estatuto de Bombeiro',
  'estatuto_Bolseiros SASIPC',
  'estatuto_Dirigente Associativo Jovem',
  'estatuto_Maternidade / Paternidade',
  'estatuto_Estudante Praticante de Atividades Artísticas no IPC',
  'estatuto_Aluno com NEE',
  'estatuto_Atleta de Alto Rendimento',
  'estatuto_Pagamento por Entidade Terceira Pagadora',
```

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

```
'estatuto_Suspensão de dívida de propina e de emolumentos de estudantes concluídos',
'estatuto_Praticante de Confissões Religiosas',
'estatuto_Justificação de faltas a exames (Exceção prevista nos regulamentos)',
'estatuto_Processo em cobrança pela Autoridade Tributária',
'estatuto_Estudante Atleta do IPC',
'estatuto_Artigo 8º do Regulamento Prescrições (05)',
'estatuto_Isenção de Validação de Dívidas de PréInscrição de Propina',
'estatuto_Falta de pagamento propinas ou emolumentos',
'DIVORCIADO',
'SEPARADO_JUDICIALMENTE',
'SOLTEIRO',
'UNIAO_FACTO',
'VIUVO',
'IFRE_MODALIDADE_PARCIAL',
'ECTS_APROVADOS_MAT',
'DISCIPLINA_Reprov',
'ECTS_Aprov',
'MEDIA_ULTIMO_ANO']
```

Figura 46 - Resultado de variáveis do algoritmo KNN da 2ª iteração - opção 2

Os resultados da opção 2 revelaram-se inferiores em todas as métricas relativamente à opção 1 apresentada. Ou seja, em vez de 0.62 da métrica f1-score, apresenta 0.48, o que mostra ser um valor menos confiável. Na precisão apenas apresenta 0.51, o que aumenta a existência de falsos positivos. Neste caso, tem-se 1375 casos de abandono para obter este resultado como “support”, o que se pode considerar um valor alto, no entanto comparam-se anos completamente diferentes a nível de funcionamento das escolas e que o treino feito com o ano letivo 2020/2021 poderá não ser um bom treino para os dados do ano seguinte.

Os resultados da **opção 3** revelam-se os seguintes apresentados na Figura 47 e um dos conjuntos de variáveis consideradas mais importantes e vantajosas para o modelo, apresentadas na Figura 48.

k	prediction	precision	recall	f1-score	support	features	accuracy
187	9	1	0.69	0.52	0.6	182.0 Regime_Acores Continente nivel 1; Regime_Con...	0.904089

Figura 47 - Resultados do modelo KNN na 2ª iteração – opção 3

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

```
'Regime_Maiores de 23 anos',
'Regime_Mudança de Par Instituição/Curso',
'Regime_Mudança de Regime',
'Regime_Preferencia Profissional',
'Regime_Preferencia Regional nível 1',
'Regime_Regime geral de acesso',
'Regime_Reingresso',
'Regime_Titulares de Diploma Técnico Superior Profissional',
'Regime_Titulares de Diploma de Especialização Tecnológica',
'Regime_Titulares de Outros Cursos Superiores',
'Regime_Titulares de diploma de curso técnico superior profissional',
'estatuto_Trabalhador Estudante',
'estatuto_Estatuto de Bombeiro',
'estatuto_Bolseiros SASIPC',
'estatuto_Dirigente Associativo Jovem',
'estatuto_Estudante Praticante de Atividades Artísticas no IPC',
'estatuto_Aluno com NEE',
'ESEC',
'ESTESC',
'ESTGOH',
'ISCAC',
'ISEC',
'DIVORCIADO',
'SEPARADO_JUDICIALMENTE',
'SOLTEIRO',
'UNIAO_FACTO',
'VIUVO',
'IFRE_MODALIDADE_PARCIAL',
'ECTS_APROVADOS_MAT',
'DISCIPLINA_Reprov']
```

Figura 48 - Resultado de variáveis do algoritmo KNN da 2ª iteração - opção 3

Por fim, na opção 3 são apresentados os melhores resultados que a opção 2, mas inferiores aos da opção 1, ainda que semelhantes. Apenas são considerados 182 casos de abandono o que pode justificar os valores não conseguirem atingir patamares acima. Deste modo, considera-se uma precisão alta de 0.69 e um *recall* de 0.52. Assim, resulta um *f1-score* de 0.6.

Na medida que esta 3ª opção divide duas vezes os dados, o mesmo modelo também foi aplicado para a partição de 20% de teste. Uma vez escolhidos automaticamente o número ideal de vizinhos e quais as variáveis ideais, aplicou-se o modelo tendo isso em conta. Os resultados constam na Figura 49.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

	precision	recall	f1-score	support
0	0.90	0.98	0.94	1431
1	0.76	0.41	0.53	250
accuracy			0.89	1681
macro avg	0.83	0.69	0.73	1681
weighted avg	0.88	0.89	0.88	1681

Figura 49 - Resultado do modelo no teste opção 3

Os resultados de *precision* e *accuracy* apresentam valores bons. No entanto, face ao teste efetuado no conjunto de treino a métrica *recall* apresenta uma variação de onze pontos. Esta variação poderá estar explicada pela quantidade da amostra ser mais reduzida.

Avaliando as 3 opções, para este exemplo os melhores resultados constatarem-se na opção 1. Estes resultados poderão estar influenciados pelos dados característicos desse ano letivo, que se mostrou particular. Assim sendo, quando comparado com outros dados de outro ano, os mesmos não têm o mesmo comportamento.

Deste modo, nos resultados da opção 1 que se mostraram os mais conseguidos o número de vizinhos, “k”, escolhido foi de 9. Tendo associados os resultados das métricas à variável de saída “Abandono”.

Tendo em consideração a opção anterior com melhores resultados, optou-se por aplicar o mesmo algoritmo aos alunos dos maiores cursos de cada instituição. Ou seja, para cada curso que tenha um maior número de alunos dentro de cada instituição, aplica-se o algoritmo apenas tendo em conta alunos do mesmo curso. Os resultados obtidos foram diversos, mas em grande parte deve-se à pouca população de estudantes que se está a ter em conta.

Para o ESAC, a licenciatura escolhida é “Biotecnologia” que contém uma população de 169 alunos e obteve os resultados mostrados na Figura 50.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

ESAC: Licenciatura em Biotecnologia, número de alunos: 169				
	precision	recall	f1-score	support
0	0.87	0.97	0.92	133
1	0.81	0.47	0.60	36
accuracy			0.86	169
macro avg	0.84	0.72	0.76	169
weighted avg	0.86	0.86	0.85	169

Figura 50 - Resultados do algoritmo KNN para o curso de Biotecnologia na 2ª iteração

Para este curso, obtiveram-se bons resultados. Principalmente na precisão, pois apresenta um valor superior ao conjunto total de dados, com 0.81. O *recall* também é positivo, ainda que abaixo de 0.5.

Em seguida, para a escola ESEC e ESTESC, o curso com o maior número de alunos é “Comunicação Organizacional” com 167 alunos e “Fisioterapia”. No entanto, ao contrário do curso anterior, os resultados não se mostraram bons. Os valores de precisão são muito baixo para serem considerados resultados confiáveis como mostram as figuras: Figura 51, Figura 52.

ESEC: Licenciatura em Comunicação Organizacional, número de alunos: 167				
	precision	recall	f1-score	support
0	0.97	0.94	0.96	160
1	0.23	0.43	0.30	7
accuracy			0.92	167
macro avg	0.60	0.68	0.63	167
weighted avg	0.94	0.92	0.93	167

Figura 51 - Resultados do algoritmo KNN para o curso de Comunicação Organizacional na 2ª iteração

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

ESTESC: Licenciatura em Fisioterapia, número de alunos: 191				
	precision	recall	f1-score	support
0	0.98	0.98	0.98	186
1	0.20	0.20	0.20	5
accuracy			0.96	191
macro avg	0.59	0.59	0.59	191
weighted avg	0.96	0.96	0.96	191

Figura 52 - Resultados do algoritmo KNN para o curso de Fisioterapia na 2ª iteração

Os seguintes cursos apresentados têm um aumento exponencial de alunos face aos anteriores. Mas, ainda assim não conseguem apresentar valores semelhantes aos conseguidos no conjunto total de dados. Assim pode-se inferir que, dado o número diminuído de desistências no conjunto total (apenas aproximadamente 15%), ao fazer a repartição apenas para os maiores cursos de cada unidade orgânica, o número de abandonos ainda é menor. E, com um número tão reduzido, o modelo não atua da melhor forma. Como se pode verificar na Figura 53.

ESTGOH: Licenciatura em Engenharia Electrotécnica, número de alunos: 211				
	precision	recall	f1-score	support
0	0.82	0.82	0.82	179
1	0.21	0.23	0.22	40
accuracy			0.71	219
macro avg	0.52	0.52	0.52	219
weighted avg	0.71	0.71	0.71	219
ISCAC: Licenciatura em Gestão de Empresas, número de alunos: 513				
	precision	recall	f1-score	support
0	0.89	0.84	0.87	451
1	0.20	0.27	0.23	64
accuracy			0.77	515
macro avg	0.54	0.56	0.55	515
weighted avg	0.80	0.77	0.79	515
ISEC: Licenciatura em Engenharia Informática, número de alunos: 845				
	precision	recall	f1-score	support
0	0.88	0.83	0.86	741
1	0.16	0.22	0.18	105
accuracy			0.76	846
macro avg	0.52	0.53	0.52	846
weighted avg	0.79	0.76	0.77	846

Figura 53 - Resultados do algoritmo KNN para os cursos de Engenharia Eletrotécnica, Gestão de Empresas e Engenharia Informática na 2ª iteração

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

O curso de “Engenharia eletrotécnica” da ESTGOH apresenta um *recall* de 23% para a variável de saída Abandono e uma *accuracy* de 71%. O curso de “Gestão de empresas” do ISCAC que apresenta 513 alunos, obteve o resultado de *recall* para Abandono de 27% e tem uma *accuracy* de 77%. Por fim, para o maior curso apresentado no conjunto de dados, “Engenharia Informática” do ISEC com 845 alunos, os resultados foram de 22% de *recall* e 76% de *accuracy*.

Complementando o estudo com testes de aplicação do modelo mais abrangente, é crucial debater sobre os resultados aplicados a cada instituição, ao invés de cada curso com maior população como apresentado acima. Assim os resultados são apresentados na Figura 54.

ESTESC					ESEC				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.92	0.67	0.77	1162	0	0.92	0.97	0.95	1435
1	0.13	0.45	0.21	130	1	0.62	0.38	0.47	188
accuracy			0.65	1292	accuracy			0.90	1623
macro avg	0.53	0.56	0.49	1292	macro avg	0.77	0.67	0.71	1623
weighted avg	0.84	0.65	0.72	1292	weighted avg	0.89	0.90	0.89	1623
ISEC					ESAC				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.89	0.83	0.86	1678	0	0.86	0.95	0.90	601
1	0.25	0.36	0.30	269	1	0.66	0.38	0.48	155
accuracy			0.76	1947	accuracy			0.83	756
macro avg	0.57	0.59	0.58	1947	macro avg	0.76	0.66	0.69	756
weighted avg	0.80	0.76	0.78	1947	weighted avg	0.81	0.83	0.81	756
ESTGOH					ISCAC				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.82	0.85	0.84	453	0	0.86	0.91	0.89	1865
1	0.32	0.27	0.29	115	1	0.40	0.29	0.34	381
accuracy			0.74	568	accuracy			0.80	2246
macro avg	0.57	0.56	0.57	568	macro avg	0.63	0.60	0.61	2246
weighted avg	0.72	0.74	0.73	568	weighted avg	0.78	0.80	0.79	2246

Figura 54 - Resultados do modelo para cada instituição na 2ª iteração

Após os resultados, podemos concluir que a precisão apenas se mostra com bons resultados nas unidades orgânica ESEC e ESAC, tendo também estas unidades apresentado os melhores valores em *f1-score*. Por outro lado, a ESTESC apresentou o melhor resultado em *recall*, no entanto não se mostraram valores bons, pois não foram suportados pela precisão. Todos os valores que não se mostraram coerentes com os mostrados na opção 1 poderão dever-se à pouca amostra de desistências que existe em cada uma das escolas.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Para uma melhor compreensão do modelo KNN foi criado o esquema apresentado seguidamente na Figura 55.

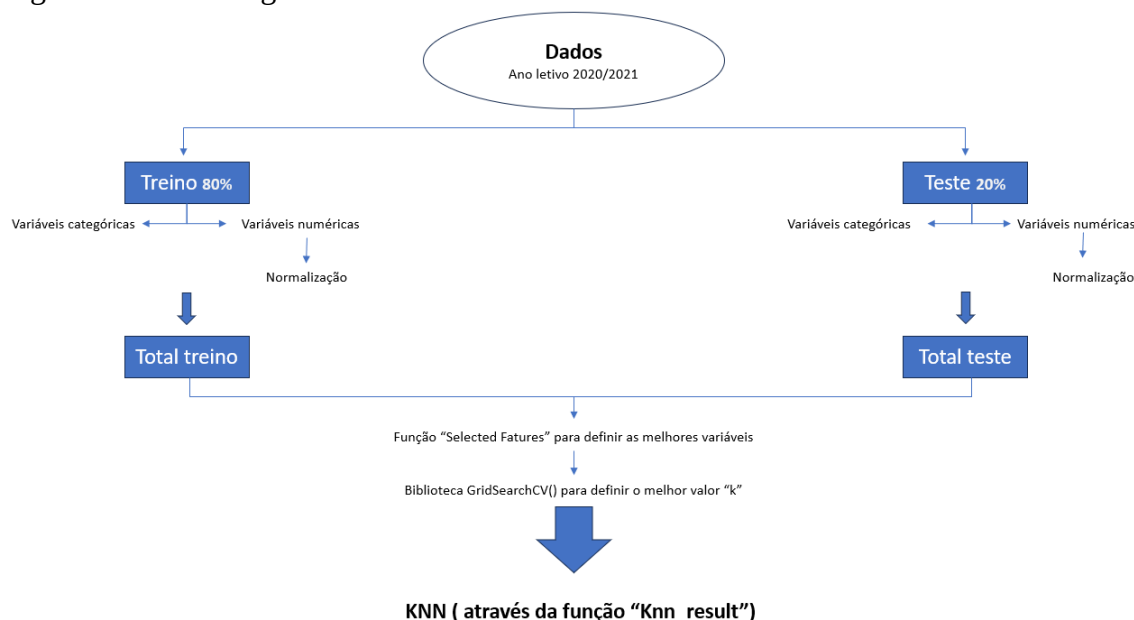


Figura 55 - Esquema do modelo

5.2 Árvore de decisão

O algoritmo KNN será tido em conta como o modelo implementado para o problema exposto neste projeto. No entanto, é interessante conseguir comparar o algoritmo mais trabalhado com outro igualmente eficaz para a questão. Esta aplicação de um algoritmo diferente poderá ser útil apenas para comparar resultados ou como hipótese de próximos passos na implementação da prevenção de AE nas instituições do IPC.

Deste modo, de uma maneira sucinta estes foram os passos para a aplicação do modelo de árvores de decisão:

- É efetuada a escolha das variáveis pretendidas, retirando aqueles que neste modelo não fazem sentido.
- Divisão dos dados em treino e teste como anteriormente, com a divisão de 80% / 20%, respetivamente.
- São escolhidos os seguintes parâmetros a serem tidos em conta no modelo: *"criterion"*, *"max_depth"*, *"min_samples_split"*, *"min_samples_leaf"*,

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

“*max_leaf_nodes*”. Com estes parâmetros é possível que consoante os dados seja escolhido o melhor método de divisão, variando entre a entropia e índice de *Gini*. De seguida, é apresentada qual a profundidade máxima das árvores, o número mínimo de amostras necessárias para dividir um nó, o número mínimo de amostras necessárias para ser uma folha e o número máximo de folhas na árvore.

- Recorre-se à biblioteca “*GridSearchCV*” que tem como objetivo testar todas as combinações possíveis dos parâmetros mencionados no ponto anterior.
- Por fim, é dada a instrução para devolver os melhores parâmetros encontrados para o modelo e que combinados devolvem o melhor resultado.

Os resultados do modelo para a árvore de decisão aplicando a opção 1 mostraram-se os representados na Figura 56.

```
accuracy score: 0.90
precision score: 0.72
recall score: 0.48
f1 score: 0.58
```

Figura 56 - Resultados do modelo com o algoritmo Árvores de Decisão

Tanto o modelo de KNN com as árvores de decisão obtiveram resultados semelhantes, sendo que os melhores resultados de KNN obteve quatro pontos a mais na métrica de *recall*. Ou seja, para as árvores de decisão obteve-se uma precisão de 0.72, *recall* de 0.48, *f1-score* de 0.58 e *accuracy* de 0.90.

Todo o tratamento de dados foi ajustado para se conseguir obter o melhor resultado possível para o algoritmo KNN, e por essa razão poderá ter implicado nos resultados deste segundo algoritmo. No entanto, mesmo sem qualquer ajuste e uma aplicação “básica” mostrou-se igualmente eficaz. Neste sentido, será um algoritmo bom para os dados em questão, que ajustado da melhor forma se conseguirá obter bons resultados.

6 Implementação

Nesta etapa pretende-se descrever alguns dos pontos que serão relevantes para a implementação do modelo. A implementação no IPC ainda não se realizou, o que faz com que não se tenham quaisquer tipos de conclusões sobre isso. No entanto, considerou-se importante descrever neste capítulo como funciona, ao pormenor, o modelo desenvolvido. Com esta descrição é possível que os utilizadores finais compreendam como se vai processar o projeto e assim ajudá-los para que obtenham os melhores resultados.

Explicação do modelo

O algoritmo KNN é uma técnica de ML que ajuda a fazer previsões com base na proximidade dos dados. Assim, segue os seguintes passos:

1. Escolhe o ponto que irá classificar – esse ponto corresponde aos valores das variáveis de um aluno. Assim, para fazer a previsão para esse estudante, primeiramente é escolhido qual será submetido a isso.
2. Seguidamente, foram encontrados os vizinhos mais próximos. Ou seja, graficamente será como se os dados de todos os restantes alunos estivessem dispostos e o KNN encontra quais são os alunos que possuem dados mais parecidos com o que está a ser analisado
3. Ao encontrar os vizinhos terá em consideração o número de vizinhos definido no código de python que dá a instrução. Assim, se o número de k for igual a 5, encontrará quais são os 5 vizinhos (alunos) que mais se assemelham com o aluno que se está a querer ter a previsão de Abandono ou não.
4. Após escolher os vizinhos e comparar os valores com os mesmos, o algoritmo vai avaliar qual a variável de saída desses vizinhos. Isto é, avalia se cada um dos alunos vizinhos praticou ou não AE. Assim, por exemplo, se a maioria tiver o resultado de AE, a previsão será de abandono também para o aluno que se está a estudar.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Na Figura 57 estão representados graficamente os pontos apresentados anteriormente. No exemplo apresentado na figura também se considera um k igual a 5. Sendo que os vizinhos são representados pelos pontos coloridos, vermelho e azul, e o aluno do qual quer obter a previsão está representado pelo ponto de interrogação.

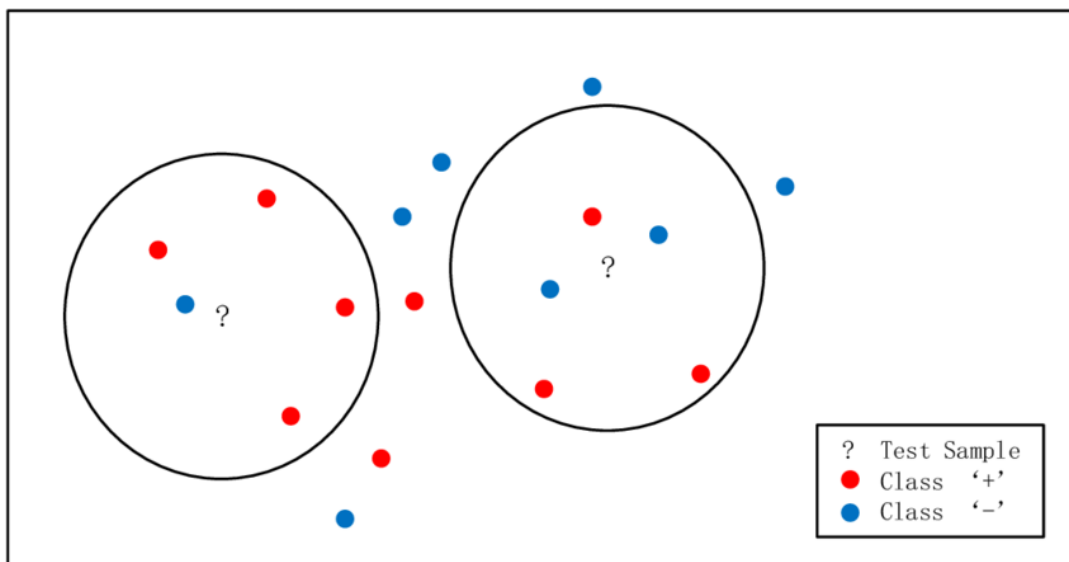


Figura 57 - Representação gráfica da implementação do algoritmo KNN

Fonte: <https://akriti-gupta2015.medium.com/knn-k-nearest-neighbours-f57265377f95>

Próximas implementações

Sendo um modelo que ainda tem algumas limitações os próximos passos serão os seguintes:

1. Tendo em conta, como já for referido, os dados de anos atípicos, é necessário efetuar ajustes no modelo para que se possam colmatar algumas falhas inerentes a esses dados.
2. Após estarem ajustados a anos letivos com maior “normalidade”, o objetivo será em conjunto com os utilizadores finais definir quais variáveis poderão fazer sentido adicionar.
3. Uma vez aceites pela população que irá utilizar, será benéfico ir acrescentando complexidade ao modelo, assim como reavaliar qual o algoritmo que melhor se adequa aos dados em vigor.

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

CONCLUSÃO

O AE no ES continua a ser um desafio significativo e preocupante que afeta não só a qualidade da educação, mas também o futuro dos estudantes. Neste âmbito foi criado este modelo preditivo baseado no algoritmo KNN que pode ser adaptado e implementado nas seis escolas referidas ao longo do projeto. Com este projeto pretende-se, contribuir com uma ferramenta eficaz para a identificação precoce de estudantes em risco de AE, permitindo que as instituições desenvolvam estratégias preventivas personalizadas. No que diz respeito ao IPC, como um todo irá estar associado a esta causa de grande relevância o que contribuirá para o reforço da reputação da instituição no que diz respeito à qualidade da educação e ao apoio do sucesso dos alunos. Com esta implementação o IPC poderá ter o resultado de aumento da sua taxa de retenção de estudantes. Outro dos contributos importantes será alocado também aos diretores de cada curso que desempenham um papel fundamental na jornada educacional dos alunos. Este projeto irá fornecer indicadores relevantes de alerta sobre os alunos de cada curso o que levará a uma ação de orientação personalizada e melhoria da qualidade do ensino oferecido. Em suma, o projeto mostra potencial para revolucionar a abordagem do IPC na prevenção do AE no ES.

Limitações

Uma das limitações do projeto reside no facto de se incluir no conjunto de treino dados do ano em que aconteceu a pandemia de COVID-19. Sendo considerado por todos um ano atípico poderá ter influenciado em alguns resultados imprevisíveis. No futuro, tendo mais anos letivos em que não se encontrem no registo de pandemia seria benéfico para o modelo desconsiderarem-se os dados deste ano e do seguinte (2021/2022). Tendo isto em conta, caso os dados de treino pudessem ter sido referentes a outro ano letivo não relacionado com este tema, os dados seriam mais reais e obter-se-iam melhores resultados. Outra limitação foi identificada na escolha do algoritmo. O algoritmo escolhido como principal foi o KNN pelo mesmo estar já implementado na instituição. Por um lado, agregou a ele inúmeras vantagens como foi descrito anteriormente. Por outro

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

lado, veio trazer algumas dificuldades na obtenção de bons resultados com os dados que estavam a ser utilizados.

Trabalhos futuros

No âmbito do estudo apresentado, é fundamental considerar a continuação de estudos e futuros trabalhos na área de AE. Uma das áreas que ao longo deste desenvolvimento se mostrou possível de bons resultados foi a Engenharia das variáveis. Na medida em que, com esta melhoria pudesse melhorar a qualidade da modelação preditiva e compreender melhor os fatores que influenciam o desempenho dos alunos, aperfeiçoar as variáveis existentes ou adicionar novas.

Uma das abordagens mais imediatas seria a inclusão de novas variáveis, tais como a assiduidade dos alunos nas aulas e nos momentos de avaliação. Os dados existentes acerca da assiduidade tanto nas aulas como nas avaliações, à data deste projeto não estão disponível, no entanto em breve será uma das funcionalidades a ter em conta o que poderá originar uma melhoria adicionar estas variáveis.

Além disto, é de igual forma importante que num futuro sejam testados diferentes tipos de algoritmos. Com os testes numa maior variedade de algoritmos poderá levar a um melhor resultado. Assim como, elevar a robustez no algoritmo de árvores de decisão que neste estudo apenas foi testada a sua forma mais simples. Por fim, como indicação de trabalho futuro, seria de realçar uma maior janela temporal de dados para treinar o modelo e não apenas os dados de um ano. Com um maior volume de dados, os modelos podem ser treinados de maneira mais sólida, levando a previsões mais confiáveis e precisas.

REFERÊNCIAS BIBLIOGRÁFICAS

- Albuquerque, Teresa (2008). Do abandono à permanência num curso de ensino superior. *Revista de ciências da educação* nº7.
- Barroso, P. C. F., Oliveira, Í. M., Noronha-Sousa, D., Noronha, A., Mateus, C. C., Vázquez-Justo, E., & Costa-Lobo, C. (2022). Dropout factors in higher education: a literature review. *Psicologia Escolar e Educacional*, 26.
<https://doi.org/10.1590/2175-35392022228736>
- Casanova, J. R. (2018). Abandono no ensino superior: modelos teóricos, evidências empíricas e medidas de intervenção. *Educação: Teoria e Prática*, 28(57), 05–22.
<https://doi.org/10.18675/1981-8106.vol28.n57.p05-22>
- Ferreira, F., & Fernandes, P. (2015). Factors influencing the school dropout in higher education and initiatives towards its prevention: The students' perspective do Porto (FPCEUP) (Porto/Portugal). (Vol. 45).
- Ferreira, M., Duarte, J., Abrantes, J. L., Cabral, L., Guiné, R., Campos, S., & Cardoso, A. P. (2019). Adaptation and validation to Portuguese of the reasons for higher education dropout scale. *Revista de Enfermagem Referencia*, 2019(21), 35–44.
<https://doi.org/10.12707/RIV18090>
- Ferreira, M., Duarte, J., Cardoso, A. P., Cabral, L., Guiné, R., Campos, S., & Alves, C. (2018). Competências emocionais e prevenção do abandono nos estudantes do ensino superior politécnico. *Revista Portuguesa de Enfermagem de Saúde Mental*, Especial 6. <https://doi.org/10.19131/rpesm.0208>
- Khanam, A., & Quraishi, U. (2016a). A Study of Reasons and Implications of the Dropout Phenomenon in Women at University.
<https://www.researchgate.net/publication/326723714>
- Ouatik, F., Erritali, M., Ouatik, F., & Jourhmane, M. (2022). Predicting Student Success Using Big Data and Machine Learning Algorithms. *International Journal of Emerging Technologies in Learning*, 17(12), 236–251.
<https://doi.org/10.3991/ijet.v17i12.30259>
- Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting Student Dropout and Academic Success. *Data*, 7(11). <https://doi.org/10.3390/data7110146>

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

- Sivakumar, S., Venkataraman, S., & Selvaraj, R. (2016). Predictive modeling of student dropout indicators in educational data mining using improved decision tree. *Indian Journal of Science and Technology*, 9(4), 1–5.
<https://doi.org/10.17485/ijst/2016/v9i4/87032>
- Stratton, L. S., O’Toole, D. M., & Wetzels, J. N. (2008). A multinomial logit model of college stopout and dropout behavior. *Economics of Education Review*, 27(3), 319–331. <https://doi.org/10.1016/j.econedurev.2007.04.003>
- Mounton, Divan, Zhang, Hui, Ertl, Bernhard (2020). German university student’s reasons for dropout. Identifying latent classes. *Journal for educational research online* 12. <https://doi.org/10.25656/01:20977>
- Casanova, R., Joana, Cervero, Antonio, Núñez, Carlos, José, Almeida, S., Leandro, Bernardo, Ana, (2018). Factors that determine the persistence and dropout of university students. <https://doi.org/10.7334/psicothema2018.155>
- Assembleia da República. *Pela prevenção e combate efetivo ao abandono escolar precoce*. DetalheIniciativa.
<https://www.parlamento.pt/ActividadeParlamentar/Paginas/DetalheIniciativa.aspx?BID=110512>
- Pandey, P. (2022, junho 07). *Beware of the dummy variable trap in Pandas*. Built In. <https://builtin.com/machine-learning/dummy-variable-trap-pandas>
- Carlos, J. (2020, dezembro 30). *Construindo o k-nearest neighbors (KNN) com Linguagem python*. LinkedIn. <https://www.linkedin.com/pulse/construindo-o-k-nearest-neighbors-knn-jos%C3%A9-carlos/?originalSubdomain=pt>
- Team, T.A. (2022, julho 05). *Which feature scaling technique to use- standardization vs Normalization*. Towards AI. <https://towardsai.net/p/l/which-feature-scaling-technique-to-use-standardization-vs-normalization>.
- Gupta, A. (2020, novembro 19). *KNN(K nearest neighbours)*. Medium. <https://akriti-gupta2015.medium.com/knn-k-nearest-neighbours-f57265377f95>
- Silva, E. (2023, outubro 09). *O que É multicolinearidade?*. Medium.
<https://medium.com/@lainetnr/o-que-%C3%A9-multicolinearidade-e4adff76345a>

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

- Okamura, S. (2020, dezembro 28). *GridSearchCV for Beginners*. Toward DataScience. <https://towardsdatascience.com/gridsearchcv-for-beginners-db48a90114ee>
- Ross, B.C. (2014). *Mutual information between discrete and continuous data sets*. PloS one. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3929353/>
- Santos, G. (2021, janeiro 10). *Seleção de Variáveis (parte 2)*. Medium. <https://medium.com/data-hackers/sele%C3%A7%C3%A3o-de-vari%C3%A1veis-parte-2-aa5fc824323c>
- Scikit-learn developers (2023). *Sklearn.preprocessing.LabelEncoder*. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html>
- Rueda, J.F.V. (2019). *CRISP-DM: Una Metodología para Minería de Datos en salud, healthdataminer*. <https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud>
- Turing (2022). *How and where to apply feature scaling in python*. <https://www.turing.com/kb/how-and-where-to-apply-feature-scaling-in-python>
- Python tutorials. *Real Python*. <https://realpython.com/>
- Vashisht, R. (2021, janeiro 06). *Machine learning: When to perform a feature scaling?*. Atoti Community. <https://atoti.io/articles/when-to-perform-a-feature-scaling>

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

APÊNDICES

Appendice I – Variáveis do conjunto de dados final

MATR_ID	MEDIA	ECTS_INSCRITOS	ECTS_APROVADOS_INSC	BOLSEIRO_COD	PRIM_OPCAO_CANDIDATURA	IFRE_MODO_COD	CURSO_ENGENHARIA_COD																
0	1019	12.610	82.0	48.0	1	1	1	0															
1	1026	14.789	60.0	60.0	0	0	1	0															
CURSO_SAUDE_COD	CURSO_GESTAO_CONTAB_COD	CURSO_OUTROS_COD	DIST_INSTITUICAO	DISCIPLINA_Aprov	DISCIPLINA_Desist	DISCIPLINA_Reprov	ECTS_Desist	ECTS_Reprov															
1		0	0	12.980	8.0	5.0	3.0	22.0	12.0														
1		0	0	141.535	11.0	0.0	0.0	0.0	0.0														
ECTS_POR_INSCRICAO	ECTS_REPROVADOS_INSCRICAO	ECTS_RACIO_REPROV_INSC	Regime_Praticantes desportivos de alto rendimento	Regime_Regime geral de acesso	Regime_Maiores de 23 anos	Regime_Mudança de Par Instituição/Curso																	
48.666667		34.0	0.414634	1	0	0	0																
60.000000		0.0	0.000000	0	1	0	0																
Regime_Bolseiros nacionais de países africanos de expressão portuguesa	Regime_Titulares de Outros Cursos Superiores	Regime_Reingresso	Regime_Estudante Internacional	Regime_Titulares de Diploma de Especialização Tecnológica	Regime_(ISEC) Transitou de Plano de Estudos	Regime_Titulares de diploma de curso técnico superior profissional	Regime_(ISEC) ESTGOH (Despacho 55/2011-P - 4 de Outubro 2011)																
0	0	0	0	0	0	0	0																
0	0	0	0	0	0	0	0																
Regime_(ISEC) Mudança Interna de Curso	Regime_(ISEC) Regimes Especiais	Regime_G - Naturais e filhos de naturais do Território de Timor Leste	Regime_D.1. Estudantes nacionais de PALOP, titulares do 12º ano ou equivalente	Regime_D.2. Estudantes nacionais de PALOP, com frequência de Ensino Superior	Regime_B: Cidadãos Portugueses Bolseiros ou Equiparados	Regime_(ISCAC) Alteração LBCG p/ LBIQ, Port 466-E/2000 de 21/07	Regime_(ISCAC) Integração lic biet, artº 31º Port 413-A/98, 17/07	Regime_Preferencia Regional nivel 1															
0	0	0	0	0	0	0	0	0															
0	0	0	0	0	0	0	0	0															
Regime_Preferencia Regional nivel 1	Regime_Preferencia Profissional	Regime_Acores Continente nivel 1	Regime_Emigrantes	Regime_Concurso Local	Regime_Deficientes	Regime_Madeira Continente nivel 1																	
0	0	0	0	0	0	0	0																
0	0	0	0	0	0	0	0																
Regime_Titulares de Diploma Técnico Superior Profissional	Regime_Mudança de Regime	Regime_Vagas Autônomas	Regime_Concurso Especial Titulares de Diploma das Vias Profissionalizantes	estatuto_NA	estatuto_Trabalhador Estudante	estatuto_Estatuto de Bombeiro	estatuto_Bolseiros SASIPC																
0	0	0	0	1	0	0	0																
0	0	0	0	1	0	0	0																
estatuto_Permite emissão de certidões com dívidas	estatuto_Suspensão de dívida de propina	estatuto_Suspensão de dívidas de pré-inscrição e propinas	estatuto_Suspensão de dívida de propina e de emolumentos	estatuto_Dirigente Associativo Jovem	estatuto_Maternidade / Paternidade																		
0	0	0	0	0	0	0	0																
0	0	0	0	0	0	0	0																
estatuto_Estudante Praticante de Atividades Artísticas no IPC	estatuto_Aluno com NEE	estatuto_Atleta de Alto Rendimento	estatuto_Pagamento por Entidade Terceira Pagadora	estatuto_Suspensão de dívida de propina e de emolumentos de estudantes concluídos	estatuto_Praticante de Confissões Religiosas	estatuto_Justificação de faltas a exames (Exceção prevista nos regulamentos)																	
0	0	0	0	0	0	0	0																
0	0	0	0	0	0	0	0																
estatuto_Processo em cobrança pela Autoridade Tributária	estatuto_Estudante Atleta do IPC	estatuto_Falta de pagamento propinas ou emolumentos	NUM_ALUNOS	ESAC	ESEC	ESTESC	ESTGOH	ISCAC	ISEC	CASADO	DIVORCIADO												
0	0	0	191	0	0	1	0	0	0	0	0	0											
0	0	0	191	0	0	1	0	0	0	0	0	0											
SEPARADO_JUDICIALMENTE														SOLTEIRO		UNIAO_FACTO		VIUVO		MEDIA_RACIO_MEDIAS		Abandono	
0	1	0	0	0.584853	0																		
0	1	0	0	1.026562	0																		

Fatores preditivos para o Abandono Escolar dos estudantes do IPC: Mitigação do Abandono escolar

Figura 58 - Conjunto final de variáveis - 2ª iteração

Apêndice II – Variáveis categóricas, numéricas e de saída

```
cat_cols = ['BOLSEIRO_COD', 'PRIM_OPCAO_CANDIDATURA', 'IFRE_MODAL_COD', 'CURSO_ENGENHARIA_COD',
            'CURSO_SAUDE_COD', 'CURSO_GESTAO_CONTAB_COD', 'CURSO_OUTROS_COD',
            'Regime_Praticantes desportivos de alto rendimento',
            'Regime_Regime geral de acesso', 'Regime_Maiores de 23 anos', 'Regime_Mudanca de Par Instituição/Curso',
            'Regime_Bolseiros nacionais de países africanos de expressão portuguesa', 'Regime_Titulares de Outros Cursos Superiores', 'Regime_Reingresso',
            'Regime_Estudante Internacional', 'Regime_Titulares de Diploma de Especialização Tecnológica',
            'Regime_(ISEC) Transitou de Plano de Estudos', 'Regime_Titulares de diploma de curso técnico superior profissional',
            'Regime_(ISEC) ESTGOH (Despacho 55/2011-P - 4 de Outubro 2011)', 'Regime_(ISEC) Mudança Interna de Curso',
            'Regime_(ISEC) Regimes Especiais', 'Regime_G - Naturais e filhos de naturais do Território de Timor Leste',
            'Regime_D.1. Estudantes nacionais de PALOP, titulares do 12º ano ou equivalente',
            'Regime_D.2. Estudantes nacionais de PALOP, com frequência de Ensino Superior',
            'Regime_B: Cidadãos Portugueses Bolseiros ou Equiparados',
            'Regime_(ISCAC) Alteração LBCG p/ LBIG, Port 466-E/2000 de 21/07',
            'Regime_(ISCAC) Integração lic biet, artº 31º Port 413-A/98, 17/07',
            'Regime_Preferencia Regional nível 1', 'Regime_Preferencia Profissional', 'Regime_Acores Continente nível 1',
            'Regime_Emigrantes', 'Regime_Concurso Local', 'Regime_Deficientes', 'Regime_Madeira Continente nível 1',
            'Regime_Titulares de Diploma Técnico Superior Profissional', 'Regime_Mudanca de Regime', 'Regime_Vagas Autónomas',
            'Regime_Concurso Especial Titulares de Diploma das Vias Profissionalizantes',
            'estatuto_NA', 'estatuto_Trabalhador Estudante', 'estatuto_Estatuto de Bombeiro', 'estatuto_Bolseiros SASIPC',
            'estatuto_Permite emissão de certidões com dívidas', 'estatuto_Suspensão de dívida de propina',
            'estatuto_Suspensão de dívidas de pré-inscrição e propinas', 'estatuto_Suspensão de dívida de propina e de emolumentos',
            'estatuto_Dirigente Associativo Jovem', 'estatuto_Maternidade / Paternidade',
            'estatuto_Estudante Praticante de Atividades Artísticas no IPC', 'estatuto_Aluno com NEE', 'estatuto_Atleta de Alto Rendimento',
            'estatuto_Pagamento por Entidade Terceira Pagadora', 'estatuto_Suspensão de dívida de propina e de emolumentos de estudantes concluídos',
            'estatuto_Praticante de Confissões Religiosas', 'estatuto_Justificação de faltas a exames (Exceção prevista nos regulamentos)',
            'estatuto_Processo em cobrança pela Autoridade Tributária',
            'estatuto_Estudante Atleta do IPC', 'estatuto_Falta de pagamento propinas ou emolumentos',
            'DIVORCIADO', 'SEPARADO_JUDICIALMENTE', 'SOLTEIRO', 'UNIAO_FACTO',
            'VIUVO']

val_cols = ['DIST_INSTITUICAO', 'DISCIPLINA Aprov', 'DISCIPLINA Desist', 'DISCIPLINA Reprov', 'ECTS Desist',
            'ECTS Reprov', 'ECTS_POR_INSCRICAO', 'NUM_ALUNOS', 'ECTS_INSCRITOS', 'ECTS_APROVADOS_INSC', 'MEDIA',
            'ECTS_REPROVADOS_INSCRICAO', 'MEDIA_RACIO_MEDIAS', 'ECTS_RACIO_REPROV_INSC']

target = 'Abandono'
```

Figura 59 - Descrição de variáveis categóricas, numéricas e saída - 2ª iteração

Apêndice III – Acesso ao código python e ficheiros

<https://drive.google.com/drive/folders/1ZuPixQO2QkdoK4LisanDILKO9qaYTvEE?usp=sharing>