Utilização de técnicas de aprendizagem automática em contexto académico para tipificação do risco de abandono escolar

Nkanga Pedro

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática
(2º ciclo de estudos)

Orientador: Prof. Doutor Paulo André Pais Fazendeiro

junho de 2024

Declaração de Integridade

Eu, Nkanga Pedro, que abaixo assino, estudante com número de inscrição M12367 do 2º Ciclo de Engenharia Informática da Faculdades, declaro ter desenvolvido o presente trabalho e elaborado o presente texto em total consonância com Código de Integridades da Universidade da Beira Interior.
Mais concretamente afirmo não ter incorrido em qualquer das variedades de Fraude Académica, e que aqui declaro conhecer, que em particular atendi à exigida referenciação de frases, extratos, imagens e outras formas de trabalho intelectual, e assumindo assim na íntegra as responsabilidades da autoria.

Universidade da Beira Interior, Covilhã 11/06/2024

Nkanga pedro

Dedicatória

A Deus, pelo dom de vida e saúde que me tem concedido.

Aos meus pais, Lucau Pedro e Matsanga Zola, pela educação que souberam dar-me desde a infância até o homem que hoje sou.

Agradecimentos

Agradeço a todos que contribuíram para a formação final deste projeto da dissertação através de críticas e sugestões. Entre eles destaco:

O meu orientador, Professor Doutor Paulo Fazendeiro, cuja orientação foi crucial para este trabalho de dissertação, no que tange a colocação de questões pertinentes e reorganização do tema, na paciência que teve comigo, na partilha de experiência científica e pela honestidade.

Ao Instituto de Telecomunicações (Delegação da Covilhã) pelo acolhimento durante grande parte do tempo dedicado à realização dos trabalhos de desenvolvimento.

A minha rainha Dikizeco Suzana, pela paciência, motivação e sobretudo pela intercessão.

Aos meus filhos, Abraão Nkanga e Zola Nkanga, a minha bênção tê-los.

Finalmente, gostaria de agradecer à minha família e amigos que ficarão no anonimato, para não correr o risco de esquecer algum.

O meu muito obrigado a todos.

Prefácio

Caro leitor,

É com grande entusiasmo que apresento este trabalho, fruto de meses de pesquisa e dedicação. O tema que aqui abordamos é de extrema relevância no contexto académico: o abandono escolar e a alta taxa de reprovações. Através da aplicação de técnicas de aprendizagem automática, buscamos entender e tipificar os fatores que contribuem para esse fenómeno.

Ao longo das próximas páginas, encontrará uma análise profunda das seguintes questões:

O problema do abandono escolar e a alta taxa de reprovações: exploramos por que alguns alunos interrompem os seus estudos antes da conclusão. Será que fatores socioeconómicos, individuais, familiares ou métodos de avaliação na instituição são determinantes para este problema?

Aprendizagem automática como ferramenta: Investigamos como os algoritmos de aprendizagem automática podem ser aplicados para prever o risco de abandono escolar utilizando os dados da UNIKIVI (Universidade Kimpa Vita) como caso de estudo.

Modelo Preditivo: Detalhamos os modelos que serão construídos e as suas métricas de desempenho. Como podemos identificar precocemente os alunos em risco? Quais variáveis são mais relevantes?

Implicações, soluções e finalmente, discutimos as implicações práticas desta investigação. Como podemos utilizar esses *insights* para desenvolver estratégias preventivas e apoiar os alunos.

Boa leitura!

Nkanga Pedro, Universidade da Beira Interior

Resumo

A evasão de estudantes no ensino superior, é um problema sério e preocupante para as autoridades competentes. Esta questão, não afeta apenas o indivíduo que abandona, mas também a instituição, família e a sociedade em geral. Com o desenvolvimento atual da ciência e da tecnologia, a partir dos dados educacionais registados, a previsão eficiente do abandono dos alunos é atualmente um tema importante de investigação.

Este trabalho centra-se na identificação e prevenção do risco de abandono na UNIKIVI com base no desempenho escolar dos alunos. Propomos um conjunto de dados académicos como atributos preditivos e apresentamos modelos de aprendizagem automática que procedem ao agrupamento de alunos utilizando algumas destas características. Espera-se que o modelo sintetizado seja capaz de tipificar o risco de desistência.

O objetivo geral é criar um modelo de aprendizagem automática capaz de prever problemas como a evasão dos alunos inscritos e o aumento da taxa de reprovações na instituição. A aplicação desse modelo pode ajudar a identificar os alunos que estão em risco de abandonar a escola, permitindo que a instituição tome medidas preventivas para evitar a evasão escolar. Além disso, a análise dos dados académicos pode ajudar a identificar situações iminentes de abandono e propor ações para minimizar a evasão escolar.

Nesse sentido, foram recolhidos dados académicos de cinco cursos diferentes correspondentes a 7 anos letivos(de 2016/2017 a 2022/2023 dos cursos de Contabilidade e gestão, Hidráulica e saneamento de água, Agronomia, Engenharia Informática e Enfermagem). Após a recolha do percurso académico dos alunos, a anonimização da informação e o pré-processamento dos dados, foi conduzido um processo de engenharia e seleção de atributos, construindo assim os conjuntos de dados. Para encontrar padrões de comportamento entre os alunos, foi usado um algoritmo de aprendizagem automática de agrupamento, conhecido como modelo *Kprototypes*. Isso pode ajudar a dividir os alunos em grupos com características semelhantes. Também utilizamos alguns modelos de classificação, como *Random Forest* (RF), *Support Vector Machines* (SVM) e *Decision Tree* (DT), para prever a probabilidade de desistência de cada aluno com base nesses grupos e, por fim, usamos métricas apropriadas, como precisão, *recall* e medida F1, para avaliar a precisão dos modelos. A utilização desta metodologia pode ajudar as instituições do ensino superior a melhorar o desempenho dos alunos e reduzir as taxas de evasão escolar.

Palavras-chave

Abandono escolar, Aprendizagem Automática, Desempenho académico, Classificação e Kprototypes

Abstract

The evasion of students at UNIKIVI (Kimpa Vita University) is a serious and worrying problem for the competent authorities. It affects not only the individual who leaves, but also the institution, the family, and the province in general. With the current development of science and technology. From the recorded educational data, efficient prediction of student abandonment is currently a hot topic of research.

This work focuses on the identification and prevention of the risk of abandonment at UNIKIVI based on the school performance of students. We propose a set of academic information as predictive attributes and present machine learning model that is grouping of students according to some characteristics and the model will present an accuracy at risk of withdrawal.

The overall objective is to create a machine learning model capable of predicting problems such as the evasion of enrolled students and the increase in the rate of rejection in the institution. Applying this technique can help identify students who are at risk of leaving school, allowing the institution to take preventive measures to prevent school leaving. Furthermore, the analysis of academic data can help to identify outstanding abandonment situations and propose actions to minimize school dropout.

In this regard, 7 years of academic data from five different courses were collected (academic years from 2016/2017 to 2022 and courses in Accounting and Management, Hydraulics and Water Sanitation, Agronomy, Computer Engineering and Nursing). After the collection of the students' academic path, the anonymization of the information and the pre-processing of the data, a process of engineering and selection of attributes was carried out, thus constructing the datasets. To find patterns of behavior among students, a grouping machine learning algorithm, known as the K-prototypes model, was used. This can help to divide students into groups with similar characteristics. We also use some classification models, such as Random Forest (RF), Support Vector Machines (SVM) and Decision Tree (DT), to predict the probability of each student's withdrawal based on these groups and, finally, we use appropriate metrics such as precision, recall and F1 points to assess the accuracy of the models. Using this technique can help the institution improve student performance and reduce school dropout rates.

Keywords

School dropout, Machine learning, Academic performance, Grading and K-prototypes

Índice

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

ADASYN	*Adaptive Synthetic Sampling* (Amostragem sintética adaptative)
ANN	*Artificial Neural Network (Rede Neural Artificial)*
CLDA	Análise Discriminante Linear Beseada em Cluster
CNAES	*National Competition for Access to Higher Education*
CUM	*Cost Under Margin* (Custo abaixo da margem)
DT	*Decision Tree* (Árvore de Decisão)
EDM	Mineração de dados educativos
ERIC	*Education Resources Information Center* (Centro de Informação sobre Recursos Educativos)
GPA	*Grade Point Average* (Media de nota)
ICT	*Information and Comunication Technology* (Tecnologia da informação e comunicação)
IES	Instituto de Ensino Superior
MESCTI	Ministério do Ensino Superior, Ciência, Tecnologia e Inovação
ML	*Machine Learning* (Aprendizagem Automática)
UNIKIVI	Universidade Kimpa Vita
UBI	Universidade da Beira Interior
RF	*Random Forest* (floresta aleatória)
SVM	*Support Vector Machines* (Máquina de Vetores de Suporte)
K-Prototypes	*K-prototypes Clustering Algorithm*
SDP	*Dropout Student Prediction System* (Sistema de previsão de abandon escolar
PCA	*Principal Component Analysis* (Análise de componentes principais)
SES	Estado Socioeconómico
QID	*Quasi-identifiers* (Quasi-identificadores)
UCI	*University of California, Irvine* (Universidade da Califórnia, Irvine)
SSCI	*Social Sciences Citation Index* (Índice de Citação de Ciências Sociais)
SMOTE	*Synthetic Minority Over-Sampling Technique* (Técnica de sobre amostragem de minorias sintéticas)
SMOTE-ENN	*Synthetic Minority Over-sampling Technique- Edited Nearest Neighbors* (Técnica de sobreamostragem de minorias sintéticas - vizinhos mais próximos editados)
SMOTETomek	*Synthetic Minority Over-sampling Technique- Tomek Links* (Técnica de sobre amostragem de minorias sintéticas - Tomek Links)
SATDAP	Capacitação da Administração Pública

Capítulo 1

1. Introdução

A busca pela excelência no ensino superior é uma preocupação central para qualquer sociedade que almeja o desenvolvimento e a formação de recursos humanos capacitados. Em Angola, a Universidade Kimpa Vita (UNIKIVI) desempenha um papel fundamental na formação de jovens talentos em diversas áreas. No entanto, uma questão que tem despertado a atenção na província do Uíge, onde está inserida a UNIKIVI e não só, é a taxa de reprovações, desistência de alunos e emigração de alunos para outras instituições de ensino superior maioritariamente privada. Os estudantes que reprovam num determinado ano são mais propensos a desistir do curso, o que pode levar ao aumento do desemprego e desigualdade social.

O presente projeto da dissertação centra-se na identificação do risco de abandono na UNIKIVI com base no desempenho escolar dos alunos com objetivo de criar um modelo de aprendizagem automática capaz de prever problemas como a evasão dos alunos inscritos e o aumento da taxa de reprovações na instituição.

Existem várias explicações possíveis para a elevada taxa de reprovações na UNIKIVI. Uma das explicações, é que a maioria dos alunos não estão preparados para o ensino superior, má escolha do curso, falta de acesso à educação de qualidade, desigualdades sociais (incluindo pobreza) e métodos de avaliação inadequados. Esse é um problema complexo que requer uma solução multifacetada.

A aplicação das técnicas de aprendizagem automática com algoritmos de aglomeração (*clustering*) e implementação de classificadores pode ajudar a identificar o grupo de alunos que estão em risco de abandonar a escola, permitindo que a instituição tome medidas preventivas para evitar a evasão escolar. Além disso, o estudo quantitativo poderá analisar os dados académicos sobre o desempenho dos alunos para identificar os principais fatores de evasão escolar e sugerir medidas para diminuir a evasão escolar.

1.1 Objetivo

No contexto da problemática do abandono escolar no contexto académico para tipificação do risco de abandono com base no desempenho escolar dos alunos, o presente projeto da dissertação tem por objetivo principal analisar o abandono académico no 1º

ciclo do ensino superior na UNIKIVI ao criar um modelo de aprendizagem automática capaz de prever problemas como:

- A evasão dos alunos inscritos e;
- O aumento da taxa de reprovações na instituição.

1.2 Motivação

A academia e a sociedade estão cada vez mais interessadas nesta questão. Isso é mais claro à medida que a preocupação social com este assunto aumenta no contexto atual de evasão escolar e taxas de reprovação, com ecos cada vez mais amplos na comunicação social.

Em termos de abandono no ensino superior, em geral, são cada vez mais comuns as notícias acerca das crescentes necessidades de vária ordem que assolam as famílias. De acordo com uma pesquisa do Jornal de Angola de 2021, a desistência está relacionada com o aumento do número de reprovações na UNIKIVI: Excesso de reprovações na "Kimpa Vita" leva estudantes à desistência" [1]. Além disso, pesquisas realizadas por outros jornais do país mostram preocupação com a fuga de alunos da instituição, como menciona o jornal AngoEmprego [2], também os alunos têm se manifestados contra o número de reprovações, recursos, taxa de emolumentos e alegações de injustiças [3].

Neste sentido, este fenómeno tem suscitado uma grande preocupação na sociedade em geral, uma vez que tem efeitos em vários planos, não só a nível individual e familiar, como a um nível mais macro, para as instituições de ensino superior, para a economia nacional (finanças públicas, produtividade e competitividade global do país) e internacional.[4] Quando se trata da qualidade do ensino superior em Angola o site1 do Ministério do Ensino Superior, Ciência, Tecnologia e Inovação (MESCTI) enfatiza questões como a falta de integração entre ensino, pesquisa e extensão, o recrutamento de pessoal e a proliferação e confusão de instituições de ensino superior que não atendem às condições necessárias. Além disso, fala sobre a falta de programas de formação pós-graduada e investigação científica, a necessidade de avaliações internas e externas e os impedimentos burocráticos para avançar na carreira. Além disso, realça a formação precária dos alunos e a falta de atualização dos cursos para atender às demandas do mercado de trabalho [5].

[1] www.ciência.ao

1.3 Questões de pesquisa

Este projeto da dissertação vai tentar responder-se às seguintes questões, organizadas por dois temas principais:

Taxa de reprovações:

- Quais são os fatores que contribuem para a elevada taxa de reprovações dos alunos na UNIKIVI?

Evasão escolar:

- Quais são as recomendações ou métodos para reduzir a taxa de evasão de alunos?

1.4 Enquadramento

A aprendizagem automática pode ser utilizada para traçar o perfil dos alunos e prever a probabilidade de abandonar a escola, que pode ser replicado em ambiente de produção real utilizando dados reais para permitir intervenções informadas e reduzir desistências [6]. Foi constatado que, apesar dos dados educacionais limitados, vários algoritmos de aprendizagem automática podem prever com precisão a evasão dos alunos.

Implicações para a investigação e prática: procedimentos de aprendizagem automática, como demonstrado no trabalho, oferecem a promessa de permitir que os administradores identifiquem de forma confiável os alunos em risco de abandonar a escola, de modo a fornecer programas direcionados e intensivos com o menor custo possível [6]. As técnicas da área da aprendizagem automática podem reduzir as taxas de abandono escolar e o ajuste de híper-parâmetros contribui para melhorar o desempenho preditivo dos classificadores [7].

1.5 Estrutura do Projeto da dissertação

Este trabalho oferece uma análise abrangente e multifacetada dos fenómenos de abandono escolar e taxas de reprovações. Cada capítulo foi cuidadosamente projetado para abordar várias facetas deste tema complexo, a fim de fornecer uma compreensão completa, o trabalho é composto por quatro capítulos organizados da seguinte forma:

- No primeiro capítulo, intitulado Introdução, procura-se apresentar a contextualização dos objetivos, a motivação, questões de pesquisa e o enquadramento da pesquisa.
- No segundo capítulo, apresenta-se um estudo do estado da arte onde se mostra uma descrição resumida sobre os trabalhos científicos publicados nas áreas de abandono escolar e taxa de reprovações com as suas métricas de avaliação do desempenho.
- No terceiro capítulo, são apresentados os fundamentos de aprendizagem automática são discutidos neste capítulo. Ele aborda tanto métodos não supervisionados (como Kprototypes e as suas métricas de validade) quanto técnicas supervisionadas (como SVM, Árvores de decisão e Florestas aleatórias).
- No quarto capítulo, são discutidos os métodos pelos quais o modelo de aglomeração será usado para tipificar o grupo de alunos em termos de desempenho. Em seguida, os modelos sugeridos neste trabalho serão usados para fazer uma previsão do risco de desistência para cada grupo de alunos.
- No quinto capítulo, são apresentadas as discussões sobre as técnicas e os resultados
- Por último, no quarto capítulo são feitas as considerações finais e apresentação de algumas perspetivas para trabalho nessa linha de pesquisa.

Capítulo 2

2. Estado da Arte

O estado da arte da análise para tipificação e classificação do risco de abandono escolar e elevada taxa de reprovações será discutido neste capítulo. Vários autores tentaram apresentar algumas abordagens para resolver esse problema tão preocupante para a sociedade.

2.1 Introdução

O objetivo deste capítulo é apresentar o estado da arte no uso de técnicas de agrupamento e classificação para analisar dados académicos e identificar padrões e tendências na relação entre desempenho académico, situação socioeconómica e taxa de reprovação para o risco de abandono escolar. Para atingir esse objetivo, serão revistos os principais jornais e artigos da literatura científica sobre o assunto. Serão discutidos os métodos utilizados, os resultados obtidos, as limitações, as contribuições e a relação com pesquisas anteriores durante os últimos cinco anos.

Além disso, o capítulo enfatiza o contexto específico de Angola, particularmente na UNIKIVI, localizada na Província do Uíge, no norte do país. Na UNIKIVI, serão realizados estudos sobre desempenho académico, eficiência da produção educacional e avaliação do ensino superior.

O desempenho académico é uma parte vital da educação e é afetado por muitos fatores, incluindo a situação socioeconómica e a qualidade da instituição de ensino [7]. Este capítulo examinará várias maneiras pelas quais a aprendizagem automática pode ser usada para analisar dados académicos relacionados com o desempenho académico; por exemplo, examinaremos a relação entre o desempenho académico de estudantes universitários e o seu estilo de vida socioeconómico e a qualidade da instituição de ensino.

2.2 Revisão da literatura

a) **Relação entre desempenho académico do aluno e a sua situação socioeconómica**

Existem vários trabalhos e artigos que já abordaram a relação entre o desempenho académico, abandono escolar e a taxa de reprovações nas instituições de ensino. Um exemplo de um artigo relacionado com o desempenho académico é "*Socio-economic status and academic performance in higher education: A systematic review*" [7], que discute a relação entre situação socioeconómica (SES) e desempenho académico no ensino superior, argumentando que essa relação é importante para entender e melhorar o desempenho académico. O artigo utiliza uma metodologia mista, incluindo pesquisa bibliográfica, seleção de literatura, análise crítica e análise de conteúdo para analisar as diferentes medidas de SES e desempenho académico, e determinar o papel dos fatores como mediadores na relação entre SES e desempenho académico. A revisão sistemática deste estudo mostrou que existe uma relação fraca, mas significativa entre SES e desempenho académico no ensino superior, com um tamanho médio de efeito de 0.06. A análise do conteúdo sugeriu que o desempenho académico anterior, a experiência académica e experiência institucional podem ser fatores mediadores nessa relação. Adicionalmente, pesquisas recentes utilizando sistemas preditivos baseados em redes neuronais tem mostrado ser uma abordagem eficiente para estudar a relação entre SES e desempenho académico, e incentivando o uso dessa abordagem para obter uma compreensão mais completa da relação.

Outro artigo relevante é "*Effects of learner-centred education on academic achievement: a meta-analysis*" [8], este estudo utilizou uma metodologia de revisão sistemática para avaliar a eficácia da educação centrada no aluno no desempenho académico dos estudantes. Foi incluída uma revisão de estudos publicados entre 2010 e 2020, que investigaram diferentes modelos de ensino centrado no aluno, como aprendizagem cooperativa, aprendizagem independente, aprendizagem experimental e técnicas de sala de aula invertida. A meta-analise foi realizada utilizando técnicas quantitativas e 42 dos 81 artigos revistos foram selecionados para a análise. Os autores calcularam as diferenças médias padronizadas como tamanhos de efeitos e utilizaram um modelo de efeitos aleatórios para investigar a variação nos tamanhos de efeitos nos estudos. A análise de subgrupo foi usada para identificar moderadores. O resultado geral mostrou que a educação centrada no aluno teve um efeito moderado positivo no desempenho académico dos alunos. As recomendações incluem a normalização da educação centrada no aluno na prática diária de ensino dos professores e a escolha apropriado de estratégias de ensino e forma de aprendizagem.

Tabela 1: Artigos relacionados com o desempenho e estado socioeconómico do aluno

Ref	Problema Tratado	Técnicas Utilizadas	Resultados Conseguidos	Limitações
[10]	A influência de fatores socioeconómico no desempenho académico dos estudantes do ensino superior	Método que combina análises qualitativas (como a análise de conteúdo sumativa) e quantitativas (como a meta-análise).	A revisão encontrou três medidas principais de desempenho académico: desempenho académico, competências e persistência. O GPA, os testes padronizados e a conclusão do grau foram usados para calcular essas medidas.	Limitação na definição, mensuração do SES e desempenho acadêmico, falta de generalização dos resultados, ausência de análise de fatores mediadores, viés de publicação.
[11]	A eficácia da educação centrada no aluno para melhorar desempenho académico e no desenvolvimento de habilidades de pensamento criativo e crítico.	Atividades de aprendizagem autónoma, grupos quantitativos de controle experimental, escolha de estratégias e métodos de aprendizagem e meta-análise	O efeito melhorou significativamente o desempenho académico dos alunos e foi classificado como médio, positivo e significativo (P < 0,001). Além das variáveis independentes, influenciou todas as três abordagens de ensino, com a aprendizagem cooperativa sendo a mais eficaz.	Devido ao tamanho inadequado da amostra, a pesquisa não levou em consideração possíveis diferenças culturais relacionadas ao uso de métodos de aprendizagem cooperativo. Outras variáveis que não foram levadas em consideração

b) A preocupação de elevada taxa de reprovações dos alunos na UNIKIVI

Após uma breve noção do assunto em estudo, procuramos solicitar jornais que pudesse nos fornecer dados de alunos em Angola mais foi sem sucesso então pesquisamos alguns artigos que já havia abordado um assunto semelhante, este estudo analisou os dados académicos da UNIKIVI, que é uma universidade pública angolana situada no norte de país. Como já foi mencionada no capítulo anterior sobre a elevada taxa de reprovações e abandono escolar.

Tabela 2: Fonte relacionado às situações de desempenho e situação socioeconómica do aluno em Angola

Ref.	Problema Tratado	Técnicas Utilizadas	Resultados Conseguidos
[1]	Excesso de reprovações no "Kimpa Vita" leva estudantes à desistência.	Entrevista e Análise de dados	No decorrer dos últimos três anos, uma grande quantidade de estudantes das unidades orgânicas da Universidade Kimpa Vita (UNIKIVI) se mudaram para estudar em outras instituições, principalmente privadas, localizadas na cidade do Uíge.
[2]	Estudantes da Universidade Kimpa Vita, no Uíge, protestam contra decreto presidencial	Entrevista e Análise de dados	Em causa as muitas reprovações, recursos e custos. Estudantes da Universidade Kimpa Vita, na cidade do Uíge, província angolana do mesmo nome, protestam nesta terça-feira, 23, pelo segundo dia consecutivo contra o elevado índice de reprovações, recursos, taxa de emolumentos e alegadas injustiças que dizem ser alvo.

2.3 Utilização de técnica de aprendizagem automática

a) Tipificação de aluno em risco de abandono escolar com abordagem de dados misto

Já com as noções obtidos nos artigos anterior, abordamos agora a questão da desistência de certos estudantes por razões mencionadas nos artigos acima, então exploramos mais os artigos "*Generating descriptive model for student dropout: a review of clustering approach*" [9] publicado em 2017 e "*Hybridization of cluster-based LDA and ANN for student performance prediction and comments evaluation*" [10], discutem o uso da mineração de dados para melhorar a retenção de estudantes e entender o comportamento deles. Ambos os artigos mencionam que a mineração de dados educacionais é um campo de pesquisa em crescimento que usa métodos computacionais para analisar grandes coleções de dados educacionais. Os pesquisadores usam técnicas, como algoritmos de ML e agrupamento, para desenvolver modelos preditivos para identificar estudantes em risco. Este último artigo tem pretende fornecer comentários motivacionais e recomendações de vídeo aos estudantes para escolher o curso certo e melhorar o desempenho e o sucesso acadêmico. A pesquisa se baseia em dados gerados pelo uso de tecnologias de informação e comunicação (ICT) nas universidades e usa dois conjuntos de dados, um disponível publicamente e outro gerado sinteticamente, para avaliar a eficácia da abordagem híbrida proposta. A metodologia consiste em três fases: identificação da importância da abordagem, aplicação de técnicas de mineração de dados e análise dos resultados e implementação das soluções.

Com artigo de "*Machine learning approach for reducing students dropout rates*" [11] publicado em 2019 tenta mostrar o quanto é importante o uso de aprendizagem automática para a resolução de problema de abandono escolar. Este aborda o uso de aprendizagem automática para reduzir as taxas de evasão escolar dos estudantes, visando identificar os fatores que contribuem para evasão e desenvolver modelos preditivos que possam identificar alunos em riscos de abandono escolar. Os autores utilizaram diferentes modelos de aprendizagem automática tal como Regressão Logística, Árvore de decisão, Redes neuronais, SVM e *Random Forest* para ter os resultados desejados. De igual modo, o artigo "*Student Dropout Prediction for University with High Precision and Recall*" [12], aborda a previsão de evasão de alunos numa universidade, utilizando métricas precisão e *recall*. O objectivo é identificar antecipadamente os alunos que estão em risco de abandonar os seus estudos, permitindo que a universidade tome medidas preventivas para ajudá-los a permanecem

matriculados. Este artigo propõe um modelo híbrido para prever os alunos que estão prestes a abandonar a universidade. O modelo tenta aumentar a precisão e a taxa de *recall* na previsão dos abandonos. Em seguida, o conjunto de recursos foi comprimido com PCA e aplicado o agrupamento *K-Means* para analisar a razão do abandono. O sistema mostrou um valor de precisão de 0,963, que é 0,093 maior do que o modelo de maior precisão das obras existentes. O *recall* e os escores F1 do abandono, 0,766 e 0,808, respetivamente, também foram melhores do que os do *boosting* por 0,117 e 0,011, tornando-os os mais altos entre trabalho comparados. Em seguida, foram classificadas as razões do abandono em quatro categorias: "Empregado", "Não Registado", "Questão Pessoal" e "Admitido em Outra Universidade". A precisão do abandono de "Admitido em Outra Universidade" foi a mais alta, em 0,672. Na pós-verificação, o sistema designado *Dropout Student Prediction System* (SDP) aumentou a eficiência do aconselhamento prevendo com precisão os abandonos com alta precisão no grupo "Alto Risco", incluindo mais abandonos no total. Além disso, prevendo as razões dos abandonos e apresentando diretrizes para cada departamento, os alunos podem receber aconselhamento personalizado.

Após analisar alguns fatores e métodos apresentados em artigos anteriores, chegamos a um consenso de que combinaríamos esses alunos com os nossos dados mistos para ser identificados melhor. Em seguida, utilizamos o artigo "*Determining the number of clusters using information entropy for mixed data*" [13], proposta uma nova maneira de descobrir quantos clusters existem em conjuntos de dados mistos com atributos numéricos e categóricos. Este artigo discute a tarefa de agrupar dados mistos, que consistem em atributos numéricos e categóricos. Ele apresenta uma taxonomia para estudar algoritmos de agrupamento de dados mistos, identificando cinco principais temas de pesquisa: *particional*, hierárquico, baseado em modelo, baseado em redes neuronais e outros. O artigo fornece uma revisão do estado da arte dos trabalhos de investigação dentro de cada tema de pesquisa e analisa as forças e fraquezas dos métodos existentes, apontando para possíveis direções futuras de investigação. A abordagem utiliza uma combinação de entropia de Rényi e entropia complementar para caracterizar a entropia dentro do cluster e a entropia entre os clusters, e para identificar o pior cluster num conjunto de dados misto. Além disso, é introduzida uma nova medida de dissimilaridade no algoritmo k-protótipos e é desenvolvido um algoritmo para determinar o número de clusters num conjunto de dados misto. O desempenho do algoritmo proposto foi testado em vários conjuntos de dados sintéticos e reais e foi mostrado que o algoritmo proposto é mais eficaz em detetar o número ótimo de clusters e gera melhores resultados de agrupamento do que outros algoritmos de agrupamento.

Os autores mencionam a falta de comparação entre algoritmos de agrupamento competitivos devido a escolhas diferentes de conjuntos de dados por diferentes investigadores. Alguns conjuntos de dados populares usados para avaliar algoritmos incluem Heart (Cleveland), Heart (Statlog) e Australian Credit data, mas esses conjuntos são pequenos e podem não ser representativos de problemas reais e complexos. A maioria desses pacotes de software está disponível em R. Existem pacotes como o K-prototypes clustering, ClustMD, ClustOfVar, CluMix, KAMILA, mixed data clustering algorithm de Macbar et al., e Ahmad e Dey mixed data clustering algorithm disponíveis para agrupamento de dados mistos [14].

No artigo "*An integrated clustering approach for high dimensional categorical data*" [15] é proposta uma nova abordagem para agrupamento de dados categóricos e tipos mistos de alta dimensão, que integra o algoritmo k-means existente com um algoritmo de ligação aglomerativa. A abordagem visa melhorar a precisão dos resultados de agrupamento e comprovar a propriedade de convergência do processo de agrupamento. O método proposto foi testado em vários conjuntos de dados e apresentou resultados promissores, buscando fornecer resultados precisos e eficientes para utilizadores acessando bases de dados. A abordagem utiliza a técnica de conjunto de agrupamentos para superar problemas com algoritmos de agrupamento tradicionais e fornecer soluções mais robustas e estáveis.

Além desses artigos acima mencionados, temos outros artigos baseados na técnica de classificação para prever o abandono escolar ou a desistência. Demonstraram, por meio dos *datasets* abaixo mencionados, que a sua contribuição cientificamente foi significativa relativamente ao modelo de Árvores de decisão e floresta aleatória predominante nesses estudos [16], [17][18], [19], [20], [21], [22].

A tabela abaixo é usada no nosso estudo para vários objetivos. Para começar, ela permite uma rápida visualização das variáveis e métodos comuns usados em estudos anteriores, o que facilita a identificação de áreas de pesquisa existente.

Esta compilação de artigos serve como base para o desenvolvimento das nossas hipóteses e metodologias. Isso garante que a nossa pesquisa seja informada, contextualizada e compatível com estudos anteriores no campo da educação. Assim, a Tabela 3 não é apenas um resumo dos artigos; Ele é um componente essencial no desenvolvimento de uma pesquisa sólida e informada destinada a abordar os desafios persistentes da reprovação estudantil e do abandono escolar.

Tabela 3: Utilização de ML para identificar antecipadamente abandono de aluno

Ref.	Problema Tratado	Técnicas Utilizadas	Resultados Conseguidos	Pontos fracos
[12]	O artigo discute o problema do abandono das IES, bem como a necessidade de IES mudarem para atender à mudança no ambiente empresarial e às expectativas dos alunos. A perda devido à desistência alunos é cada vez mais crucial à medida que o número de matrículas ultrapassa as 10.000.	Ferramentas de apoio à decisão incluindo técnicas analíticas e de *data mining*, Mineração de dados educativos (EDM) e Abordagem de *clustering*	A abordagem de *clustering* foi eficaz na identificação de grupos de alunos com características semelhantes e na previsão da sua probabilidade de evasão. descobriu que certos fatores, como desempenho académico e tipo de entrada na universidade, são fortes preditores de evasão de estudantes.	Estudo reconhece que pode haver outros fatores além do escopo dos dados analisados que contribuem para a evasão de estudantes, como questões pessoais ou financeiras.
[13]	O artigo discute o desafio de prever o desempenho académico devido ao aumento do número de evasões em muitos países, escolhas incorretas de cursos, clusters sobrepostos e identificação de valores de limite em técnicas de mineração de texto para classificar alunos em abandono; complexidade das tarefas de previsão e baixa precisão na classificação de alunos com notas semelhantes; e a falta de proveito.	- Análise Discriminante Linear Baseada em Cluster (CLDA) - Rede Neural Artificial (ANN)	A precisão da classificação do CLDA (algoritmo proposto) é melhor em comparação com os algoritmos de Naïve Bayes, Rede Neural e Agrupamento Hierárquico, com uma média de 93% de precisão de classificação.	- A previsão do índice de evasão é imprecisa e pode ser melhorada. - A precisão da classificação diminui ao usar conjuntos de dados sintéticos.
[14]	A alta taxa de abandono escolar entre alunos. O objectivo é identificar os alunos em risco, permitindo intervenções precoces e ações preventivas para reduzir as taxas de desistência.	Classificadores como Árvore de Decisão, Regressão Logística, KNN e Regressão Linear.	Os dois classificadores LR e MLP provaram ser superiores a todos os outros classificadores, alcançando o melhor desempenho quando a técnica de sobre amostragem foi utilizada.	Disponibilidade limitada de conjuntos de dados públicos de países em desenvolvimento.
[15]	A queda na taxa de matrícula na Coreia do Sul, e o modelo proposto visa ajudar as instituições a identificar e fornecer suporte personalizado aos alunos que estão em risco de abandonar os estudos.	RandomUnderSampler, SMOTE, ADASYN, SMOTEENN e SMOTETomek. modelos de ensemble, regressão logística, ANN e Gradiente *boosting*.	A taxa média de evasão durante os cinco anos foi de 5,1%. O sistema SDP utilizado melhorou a precisão e o recall das previsões O sistema SDP obteve as melhores pontuações no F1 de	O SDP não é especificado como o sistema equilibra estas duas métricas e se existe um compromisso entre elas.

		Modelo híbrido de predição que aumente a precisão e o índice de recuperação dos alunos em risco de abandonar a instituição.	evasão. A Suas pontuações foram de 0,989 para precisão, 0,819 para *recall* e 0,786 para F1.	
[16]	O problema de identificar o cluster mais inadequado de um conjunto de dados misto, bem como a caracterização uniformemente da entropia dentro do cluster e entre cluster. A avaliação dos resultados de agrupamento para dados mistos é outra preocupação abordada. O artigo apresenta um índice de validade de cluster eficaz para avaliar os resultados do clustering, bem como uma nova medida de dissimilaridade e um algoritmo para calcular o número de clusters num conjunto de dados misto baseado no algoritmo Kprototypes	Uso da entropia de Renyi e entropia complementar para caracterizar uniformemente a entropia dentro do cluster e Kprototypes	Algoritmo proposto é superior aos outros algoritmos na maioria dos conjuntos de dados em termos de CUM e ARI.5.3. comparações da capacidade de detetar o número ótimo de clusters e obter melhores resultados de clustering.	Difícil definir um tamanho apropriado do passo do valor limite da similaridade. Portanto, o valor limite da similaridade varia de 0,01 a 1 com passo 0,01 para todos os conjuntos de dados usados neste experimento.
[17]	As matrizes de similaridade de pares e associação binária de agrupamentos são usadas nos métodos de análise categórica de dados e prejudicar os resultados de agrupamento porque valores e informações de nível geral são desconhecidos. O objetivo do estudo é melhorar o processo de agrupamento K-Means existente, resolver o problema de convergência e aumentar a precisão dos resultados de agrupamento para tipos de dados mistos e categóricos de alta dimensão.	Utiliza duas técnicas principais: o algoritmo K-Means e o algoritmo de conjunto de cluster baseado em links.	A abordagem conseguiu melhorar a precisão dos resultados de agrupamento em comparação com outras abordagens existentes. Num conjunto de dados de teste com 10.000 objetos, a abordagem proposta obteve um índice de Rand ajustado de 0,83, enquanto o K-Means puro obteve um índice de 0,62.	Uma das limitações é que a abordagem pode ser sensível a outliers, que são objetos que são muito diferentes dos outros objetos no conjunto de dados.

2.4 Anonimização de dados

Como os dados pertencem ao governo, precisam ser protegidos. Este artigo nos ajudará a usar a técnica de privacidade "*Quantifying the Vulnerability of Attributes for Effective Privacy Preservation Using Machine Learning*" [23]. Uma abordagem automatizada de aprendizagem automática para quantificar a vulnerabilidade de cada item entre os atributos de utilizadores, visando preservar a privacidade dos utilizadores sem prejudicar a utilidade dos dados. O trabalho aborda quatro problemas técnicos na área de preservação de privacidade, incluindo equilíbrio entre privacidade e utilidade, garantias de privacidade em dados desequilibrados, problemas sobre anonimização e a aplicabilidade de modelos de privacidade anteriores. O método proposto utiliza uma técnica de RF para identificar e classificar *quasi-identifiers* (QID) mais vulneráveis, e realiza anonimização flexível para equilibrar a privacidade e a utilidade dos dados. Os resultados foram validados em dois conjuntos de dados de *benchmark* reais, mostrando-se eficaz na preservação do equilíbrio entre privacidade e utilidade. O artigo sugere que o método proposto pode revigorar os métodos existentes e abrir caminhos para futuras pesquisas e desenvolvimentos na área de privacidade de dados. Além disso, é apresentada uma metodologia de anonimização que leva em conta a vulnerabilidade de cada QID relativamente ao atributo sensível e utiliza técnicas de ML para limitar violações de privacidade.

Tabela 4: Artigo relacionado anonimização de dados

Ref.	Problema Tratado	Técnicas Utilizadas	Resultados Conseguidos	Limitações
[18]	- Preservação da privacidade de dados pessoais - Quantificação da vulnerabilidade de cada item entre os atributos - Solução de quatro problemas técnicos no campo de preservação da privacidade	Random Forest (Floresta Aleatória)	Implementação da técnica de ML para medir o risco de cada item entre os atributos dos dados pessoais e evitar a divulgação explícita da privacidade e a perda de utilidade dos dados publicados	O método proposto é aplicável apenas para dados que possuem desequilíbrio de diversidade de atributos sensíveis. Também não leva em consideração a correlação entre os atributos, o que pode afetar a privacidade do conjunto de dados.

2.5 Análise de artigos e relações com o estudo

Após revisar os artigos, descobrimos que existem quatro áreas de estudo principais. Os artigos "Socioeconomic status and academic achievement: A *systematic review and*

meta-analysis of the literature" de Dika et al. (2020) e "*Effects of learner-centered education on academic achievement: a meta-analysis*" de Li et al. (2021) destacam como a vida socioeconómica e a qualidade da instituição estão ligadas ao desempenho Eles usam metodologias de revisão sistemática e meta-análise para analisar os dados empíricos e chegar a conclusões sobre como esses fatores estão relacionados.

Os artigos "*Generating descriptive model for student dropout: a review of clustering approach*" publicado em 2017 e "*Hybridization of cluster-based LDA and ANN for student performance prediction and comments evaluation*" discutem técnicas de clustering para analisar dados educacionais e identificar padrões e tendências de desistência de estudantes. Eles também apresentam metodologias de cluster misto para lidar com dados incompletos e híbridos, como "*Determining the number of clusters using information entropy for mixed data*" e "*An integrated clustering approach for high dimensional categorical data*". Estes últimos artigos mostram o quanto importante contar com aprendizagem automática em questão de abandono escolar que são "*Machine learning approach for reducing students dropout rates*" e "*Student Dropout Prediction for University with High Precision and Recall*.

Usamos esses artigos como base de estudo para a ideia de criar perfis dos alunos com base na classificação do grau de risco de abandono escolar. Também fornecemos clareza sobre a maneira como todos eles usaram a classificação como modelo principal de previsão de risco, com o modelo de árvore de decisão e floresta aleatória dominando a classificação de evasão ou abandono escolar.

Por fim, "*Quantifying the Vulnerability of Attributes for Effective Privacy Preservation Using Machine Learning*" discute como os métodos de ML podem proteger os dados dos alunos e a privacidade deles.

Em resumo, esses artigos fornecem uma visão geral das dificuldades e obstáculos encontrados na análise de dados educacionais que é nosso caso também visto que estamos a trabalhar com dados educacionais da UNIKIVI. Eles também apresentam vários métodos e resultados relacionados à aplicação de técnicas de *cluster* e classificação na análise de dados educacionais e na identificação de padrões e tendências relacionadas ao desempenho académica, situação socioeconómica e qualidade da instituição de ensino. Além disso, eles contribuem significativamente para a pesquisa em andamento na área.

2.6 Dados público utilizados nos artigos analisados

Ao estudar taxas de abandono escolar e abandono escolar, uma compreensão completa dos padrões e variáveis históricos é essencial para fazer previsões precisas e planos de intervenção eficazes. Atualmente, apresentamos uma tabela resumida, compilada minuciosamente, que destaca os conjuntos de dados utilizados em estudos anteriores neste campo.

Na Tabela 5, não apenas fornece uma base sólida para a análise comparativa, mas também fornece uma base para entender as características utilizadas e fatores que foram identificados como significativos na previsão das taxas de abandono escolar e reprovações nos conjuntos de dados públicos existentes.

Tabela 5: Conjunto de dados público

Tipo de dados	Ref	Características	Tipo de problema que pode ser resolvido
Estudantes Matriculados de 2021 do Instituto Politécnico de Portalegre (Portugal)	[DS1]	Dados reais de 44.200 estudantes, incluindo o seu percurso académico, dados demográficos e factores, SATDAP - Capacitação da Administração Pública em 36 variáveis diferentes.	O conjunto de dados inclui informações conhecidas no momento da matrícula (trajetória académica, demografia e fatores socioeconómicos) e o desempenho académico dos alunos ao final do primeiro e segundo semestres. Os dados são usados para construir **modelos de classificação** para predizer a evasão e o sucesso académico dos alunos.
Higher Education Competency Dataset based on the TEC21 Educational Model of Tecnologico de Monterrey	[DS2]	Dados contem 121,584 registos de agosto-dezembro de 2019 a fevereiro-junho de 2022. Com atributos sociodemográficas, admissão e académica.	O conjunto de dados inclui informações anónimas relacionadas com estudantes de licenciatura que se inscreveram e frequentaram pelo menos um semestre no Tecnológico de Monterrey, no México, Os dados são usados para construir **modelos SVM, k-Nearest Neighbor (KNN), DT,RF, Adaptive Boosting (ADA_Boosting), Extreme Gradient (XG_Boosting), Bayesian Classifier (BC), e LDA**
Conjunto de dados MOOC e dados fictícios de Mockaroo	[DS3][DS4]	Dados de referência (benchmark dataset) e dados sintéticos (synthetic dataset). Com atributos, idade, gênero, notas em disciplinas, entre outro atributo desempenho e demográfico gerado pelo site mockarro	Dados usados agrupar e classificar os alunos em risco de abandonar a escola com os modelos **Cluster-based linear discriminant analysis (CLDA), ANN algorithm e KNN**
Registos estudantis da Universidade Nacional de Gyeongsang do ano de 2016 a 2022.	[DS5]	Dados utilizado no estudo foi constituído por 7718 registos dos alunos, incluiu 6 cursos de matemática e 19 cursos	Dados usados para agrupar os alunos com modelo Kmeans e PCA combinamos os resultados de previsão de abandono do **XGBoost e do CatBoost para produzir alta precisão e recall**

		demográficos e socioeconómicos.	
Rajamangala university of technology thanyaburi dropout dataset (RDD)	[DS6]	Dados inclui 2 137 estudantes de licenciatura de 2013 a 2019 e segue o modelo CRISP-DM, utilizando fontes de dados internas da ARIT.	Permite quantificar a vulnerabilidade de cada atributo num conjunto de dados, a fim de preservar a privacidade dos utilizadores sem comprometer a utilidade dos dados. Utilizaram o modelo **Random Forest, SVM, Cluster**, para analisar as correlações entre os atributos e os dados sensíveis.
The data relating to the National Competition for Access to Higher Education (CNAES)	[DS7]	Dados 13.992 linhas e 398 colunas inclui dados demográficos, socioeconómicos, macroeconómicos e académicos. Extraído do programa "SATDAP - Capacitação da Administração Pública	Tem finalidade de construir modelos de **classificação** para prever a evasão e o sucesso académico dos estudantes de diferentes cursos de graduação de uma instituição de ensino superior (evasão, matriculado e formado) no final da duração normal do curso.
Utilização da extração de dados para prever o desempenho dos alunos do ensino secundário	[DS8]	Dados de 649 instâncias e 33 característica baseado em desempenho dos alunos no ensino secundário de duas escolas portuguesas, incluem notas dos alunos, características demográficas, sociais e relacionadas com a escola) e foram recolhidos através de relatórios escolares e questionários.	o desempenho dos alunos no ensino secundário de duas escolas portuguesas. Os atributos dos dados incluem notas dos alunos, características demográficas, sociais e relacionadas com a escola) usado para **classificação (Regressão)**
Clustering on TurkiyeStudentEvaluation	[DS9]	Dados de 8 características com diferente variáveis	Prevemos o abandono universitário utilizando florestas aleatórias baseadas em **árvores de inferência condicional** e modelamos a decisão de abandono como uma **classificação binária (licenciado ou desistente)**
A Dataset of Dropout Rates and Other School-Level Variables in Louisiana Public High Schools	[DS10]	Dados foi recolhido numa duração de cinco anos letivos (2014-2015 a 2018-2019) com as características a nível escolar, taxas de abandono escolar, demográficos socioeconómicos, variáveis financeiras, dimensão das turmas	Este enorme conjunto de dados de variáveis escolares foi originalmente compilado com a intenção de identificar os factores que se correlacionam com o abandono escolar nas escolas secundárias públicas do Louisiana foi usado com o modelo de classificação

Capítulo 3

3. Fundamento de aprendizagem automática

Aprendizagem automática envolve o desenvolvimento de modelos ou algoritmos que aprendem com dados históricos para fazer previsões ou tomar ações. Estes modelos são treinados com dados rotulados ou não rotulados, e o seu desempenho melhora à medida que são expostos a mais dados e *feedback* [24]. A compreensão dos conceitos, técnicas e tarefas essenciais do processo é um dos pilares da aprendizagem automática. O agrupamento, as árvores de decisão, SVM e a floresta aleatória estão entre essas abordagens neste estudo.

Os fundamentos de aprendizagem automática neste capítulo. Ele aborda tanto métodos não supervisionados (como *Kprototypes* e suas métricas de validade) quanto técnicas supervisionadas (como SVM, Árvores de decisão e Florestas aleatórias). Apresentamos como essas técnicas funcionam com as suas respetivas métricas como precisão, acurácia, *recall*, pontuação F1 e matriz de confusão. Além disso, exploramos o método SMOTE para destacar as diferenças entre dados balanceados e não balanceados.

3.1 Método não supervisionado

Aprendizagem automática não supervisionada é uma técnica de aprendizagem automática em que os utilizadores não precisam supervisionar o modelo. Ao invés disso, permite que o modelo trabalhe por conta própria para descobrir padrões e informações que não foram detetados anteriormente, pois lida com dados não rotulados. Os dados rotulados poderiam ser descritos como um conjunto de dados que possui uma identificação, uma etiqueta para as observações [25].

3.1.1 Técnica de *Clustering*

Clustering é uma técnica estatística usada para classificar elementos em grupos, de forma que elementos num um mesmo cluster sejam muito parecidos, e os elementos em diferentes clusters sejam distintos entre si. O *clustering*, ou agrupamento, e uma técnica base, pois nenhuma suposição e feita a respeito do número de grupos ou da estrutura dos grupos.

O agrupamento e feito com base em semelhanças ou distâncias (dissimilaridade). As entradas necessárias são medidas de semelhança ou dados a partir dos quais as semelhanças podem ser calculadas [25].

a) Clustering Kprototypes

- Conceitos

Entre os referidos que são *K-means* e *Kmodes* apesar de ser agrupamento eficaz [26]. Por outro lado, os seus métodos são inadequados quando se aplica dados com variáveis categóricas, no caso do *Kmeans*, este problema surge quando a função de custo em *K-Means* é calculada usando a distância euclidiana, que só é adequada para dados numéricos. Por outro, o algoritmo *Kmodes* não é aplicável a tipos de dados mistos, embora seja limitado a dados categóricos. O algoritmo *Kprototypes* combina variáveis numéricas e categóricas para oferecer uma abordagem inovadora para agrupar dados mistos. Huang estabeleceu a base para lidar com variáveis categóricas no contexto de agrupamento, adaptando o método *K-Means* tradicional no seu estudo.

No seu estudo do ano subsequente [27], o algoritmo combina os princípios *Kmodes* e *K-Means* e oferece uma solução eficaz para o problema de agrupar dados que inclui uma variedade de tipos de variáveis. Esses dois artigos estabelecem uma base importante para a pesquisa de *clustering* e apresentam uma abordagem útil para a análise de dados mistos.

Nestes artigos, apresentam o algoritmo *Kprototypes* que se baseia no paradigma *k-means*, mas remove a limitação numérica de dados, preservando a sua eficiência. No algoritmo, os objetos são agrupados contra protótipos k. Um método é desenvolvido para atualizar dinamicamente os protótipos k, a fim de maximizar a semelhança intra-cluster de objetos. Quando aplicado a dados numéricos, o algoritmo é idêntico às médias k. Para auxiliar a interpretação de clusters, usando algoritmos de indução de árvore de decisão para criar regras para clusters. Essas regras, com outras estatísticas sobre clusters, podem ajudar os mineradores de dados a entender e identificar clusters interessantes.

- Processo de algoritmo *Kprototypes*

O algoritmo *Kprototype* foi inicialmente proposto por Zhexue Huang no seu artigo intitulado *"Clustering large data sets with mixed numeric and categorical values"* [26]. Em 1998, ele foi modificado no outro artigo [27] e na sua dissertação de doutorado. Para

melhorar a eficiência e a versatilidade do algoritmo, vários ajustes e melhorias foram feitos desde então. Aqui está um resumo do progresso e possíveis mudanças que podem ocorrer ao longo dos anos:

- Desenvolvimento Inicial (1997): Huang sugeriu o algoritmo *Kprototype* como uma extensão do *K-Means* para lidar com dados com variáveis categóricas e numéricas. Ao combinar medidas de distância Euclidiana para características numéricas e medidas de dissimilaridade para características categóricas, ele introduziu o conceito de distância híbrida.
- Implementações e aplicações desde o ano 2000: O algoritmo *Kprototype* começou a ser implementado em várias bibliotecas de *machine learning* e *software* de análise de dados nos anos seguintes à proposta inicial. Isso permitiu que ele fosse usado numa ampla gama de situações.
- Aumento da eficiência (ano de 2010): O foco em melhorar a eficiência do algoritmo aumentou nos últimos dez anos, especialmente para conjuntos de dados de grande escala. Para aproveitar o poder computacional do *hardware* moderno, incluiu estratégias de inicialização de centroides, paralelização de algoritmos e otimizações de cálculo de distâncias.
- Extensões e mudanças nos últimos anos: várias extensões e variações foram propostas para lidar com problemas específicos em vários tipos de conjuntos de dados, além do *Kprototype* básico. Inclui adaptações para lidar com dados textuais, temporais e de alta dimensionalidade.

Em resumo, o algoritmo *Kprototype* foi desenvolvido, otimizado e adaptado ao longo dos anos para atender às demandas de agrupamento em conjuntos de dados cada vez mais heterogéneos e de grande escala.

- Função de custo

Kprototypes inclui valores numéricos e categóricos. Aborda o desafio de definir medidas de similaridade/distância para valores categóricos, que não possuem uma ordem natural. As medidas de similaridade existentes para valores categóricos não reconhecem as diferenças entre os tipos nominais e binários e a importância variável dos valores num atributo binário. O reconhecimento destas diferenças pode levar a melhores resultados de agrupamento. Uma abordagem é determinar os centroides iniciais de forma adaptativa com base na densidade e distância, o que ajuda a determinar o número de clusters e melhora a precisão e estabilidade do agrupamento [28]. As fórmulas de

funcionamento para *clustering Kprototypes* baseadas nos artigos [26], [27] mencionados anteriormente estão abaixo. Função de Custo de Algoritmo Kmeans

$$P(W,Q)\cdot = \sum_{i=1}^{k}\sum_{i=1}^{n} w_{i,j}d(X_i,Q_i)$$

$$P(W,Q)\cdot = \sum_{i=1}^{k} w_{i,j} = 1, \quad 1 \le i \le n$$

$$w_{i,j} \in \{0,1\}, \qquad 1 \le i \le n, 1 \le i \le k$$

Onde W é uma matriz de partição, $n \times k$, $Q = \{Q1, Q2, \dots, Qx\}$ é um conjunto de objetos no mesmo domínio de objeto e d (,) é a distância euclidiana quadrada entre dois objetos. O problema P pode generalizado para permitir (w_{ij}) onde w ∈ [0,1], a≥ 1 [29], [30]

Função de Custo de Algoritmo Kmodes

Em princípio a formulação do problema P na fórmula do *kmeans* também é válida para objetos categóricos e de tipo misto. A causa pela qual o algoritmo *k-means* não pode agrupar objetos categóricos é sua medida de dissimilaridade e o método utilizado para resolver o problema P2.

Devido à limitação da medida de dissimilaridade pelo algoritmo K-means tradicional, ela não pode ser usada para agrupar conjuntos de dados categóricos. O algoritmo de agrupamento de modos K é baseado no paradigma *K-means*, mas remove a limitação de dados numéricos, preservando a sua eficiência. O algoritmo *Kmodes* [27] estende o paradigma *K-means* para agrupar dados categóricos, removendo a barreira imposta pelos *K-means* essas barreiras podem ser removidas fazendo as seguintes modificações no algoritmo *k-means*:

- Usando uma medida de dissimilaridade de correspondência simples para objetos categóricos,
- Substituição de meios de clusters por modos, e
- Usar um método baseado em frequência para encontrar os modos para resolver o problema P2. Esta seção discute essas modificações.

Quando é usado como medida de dissimilaridade para objetos categóricos, a função de custo torna-se

$$P(W,Q)\cdot = \sum_{i=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m} w_{i,j}\delta(X_{ij},q_{ij})$$

Onde $w_{ij} \in W$ e $Q_i = [q_{i2}, q_{i2}, \ldots., q_{im},] \in Q$.

Para minimizar a função de custo, o algoritmo básico k-means pode ser modificado usando a medida de dissimilaridade de correspondência simples para resolver P1, usando modos para clusters em vez de médias e selecionando modos conforme o Teorema 1 para resolver P2. No algoritmo básico precisamos calcular o custo total P em relação a todo o conjunto de dados cada vez que um novo Q ou W é obtido.

Função de Custo de algoritmo *Kprototypes*

A dissimilaridade entre dois objetos de tipo misto X e Y atributos descritos por $A_1^r, A_2^r, \ldots., A_p^r, A_{p+1}^c, \ldots., A_m^c$ pode ser medido por como foi proposto pelo Huang [27]

$$d_2(X,Y)\cdot = \sum_{j=1}^{p}(x_j - y_j)^2 + \sum_{j=p+1}^{m}\delta(x_j, y_j)$$

onde o primeiro termo é a medida quadrada da distância euclidiana nos atributos numéricos e o segundo termo é a medida de dissimilaridade de correspondência simples nos atributos categóricos. O peso ÿ é usado para evitar favorecer qualquer tipo de atributo.

$$P(W,Q)\cdot = \sum_{i=1}^{k}(\sum_{i=1}^{n} w_{i,j}\sum_{j=1}^{p}(x_{i,j} - q_{i,j})^2 + y\sum_{i=1}^{n} w_{i,j}\sum_{j=p+1}^{m}\delta(x_{i,j}, q_{i,j}))$$

$$P_1^r = \sum_{i=1}^{n} w\sum_{j=1}^{p}(x_{i,j} - q_{i,j})^2$$

e

$$P_1^r = y\sum_{i=1}^{n} w_{i,j}\sum_{j=p+1}^{m}\delta(x_{i,j}, q_{i,j})$$

Reescrevendo $P(W,Q)\cdot$ Como

$$P(W,Q) = \sum_{i=1}^{k}(P_1^r + P_1^c)$$

3.1.2 Métricas de validade

O desempenho do algoritmo de agrupamento *Kprototype* está intrinsecamente ligado à qualidade dos clusters formados, o que torna a tarefa de determinar o número ideal de clusters um desafio considerável. No entanto, existem várias maneiras de descobrir essa quantidade ideal. Mas agora nos concentraremos no método mais adequado para encontrar o valor de K. [24]. O nosso trabalho baseou-se em três métodos são dados abaixo:

a) Método do Cotovelo (Elbow Method)

O método do cotovelo é uma das maneiras mais populares de encontrar o número ideal de clusters. Esse método usa o conceito de valor WCSS que significa **Within Cluster Sum of Squares**, que define as variações totais num cluster [24]. A fórmula para calcular o valor do WCSS (por exemplo, para 3 clusters) é dada abaixo:

$$\text{WCSS} = \sum P\ i\ na\ distancia\ do\ cluster(PiC1)^2 + \sum P\ i\ na\ distancia\ do\ cluster(PiC2)^2 + \sum P\ i\ na\ distancia\ do\ cluster(PiCn)^2$$

Na fórmula acima do WCSS, $\sum_{\text{Pi na distância Cluster1}}(P_i\ C_1)^2$: É a soma do quadrado das distâncias entre cada ponto de dados e o seu centroide num cluster1 e o mesmo para os outros dois termos.

Para medir a distância entre os pontos de dados e o centroide, podemos usar qualquer método, como a distância euclidiana ou a distância de Manhattan.
Para encontrar o valor ideal dos clusters, o método do cotovelo segue as etapas abaixo:

- Ele executa o agrupamento Kprototype num determinado conjunto de dados para diferentes valores K (intervalos de 2 a 40).
- Para cada valor de K, calcula o valor WCSS.
- Traça uma curva entre os valores WCSS calculados e o número de clusters K.
- O ponto afiado de curvatura ou um ponto do gráfico parece um braço, então esse ponto é considerado como o melhor valor de K.

b) Método Silhouette_Score

A análise de silhuetas pode ser utilizada para estudar a distância de separação entre os clusters resultantes. O gráfico de silhuetas apresenta uma medida da proximidade entre

cada ponto de um agrupamento e os pontos dos agrupamentos vizinhos, fornecendo assim uma forma de avaliar visualmente parâmetros como o número de agrupamentos [24]. Esta medida tem um intervalo de [-1, 1].

Distancia Média Intra-cluster (a_i)

$$a_i = \frac{1}{|C_i| - 1} \sum j \in C_{i,}\ i \neq j\ d(i,j)$$

Distancia Média Inter-cluster (b_i)

$$b_i = min_{k \neq i} \frac{1}{|C_k|} \sum j \in C_{k,}\ d(i,j)$$

Silhouette Score (S_i)

$$S_i = \frac{b_i - a_i}{\max(b_i, a_i)}$$

Onde d(i,j) é a distância entre os pontos i e j, Ci é o cluster ao qual i pertence, e |c| é o número de pontos no cluster C. Os coeficientes de silhueta (como são designados estes valores) próximos de +1 indicam que a amostra está muito afastada dos agregados vizinhos. Um valor de 0 indica que a amostra está na fronteira de decisão entre dois agregados vizinhos ou muito perto dela e valores negativos indicam que essas amostras podem ter sido atribuídas ao agregado errado.[2]

c) Método Davies_bouldin

Tem como objetivo avaliar quão "bem separados" foram os clusters. O resultado índice é similaridade padrão entre os grupos, cujo resultado pode variar de 0 até o infinito positivo [24]. Como queremos que os nossos grupos sejam o mais bem particionados quanto possível, quanto mais próximo de 0 for o nosso resultado, melhor.
Abaixo, temos a fórmula da métrica.

$$DB = \frac{1}{\mathrm{k}} \sum \max((R_{ij})$$

$$R_{ij} = \frac{S_i + S_j}{D_{ij}}$$

Onde:

- DB → Índice de Davies_bouldin.
- k → O número de clusters criados pelo algoritmo.
- S → A distância média entre cada instância do grupo e o seu centroide.

[2] https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

- D → Distância entre os centroides dos grupos sendo analisados.
- R → Índice de similaridade entre clusters.

Em suma, nós calculamos o quão parecidos os grupos são um dos outros, somamos, e depois dividimos pelo número total de grupos, resultado na “similaridade padrão”.

3.2 Método supervisionado

De acordo com site [24], aprendizagem supervisionada, os dados rotulados de amostra são fornecidos ao sistema de aprendizagem automática para treino e, em seguida, o sistema prevê a saída com base nos dados de treino. O sistema usa dados rotulados para criar um modelo que entende os conjuntos de dados e aprende sobre cada um. Depois que o treino e o processamento são feitos, testado o modelo com dados de amostra para ver se ele pode prever com precisão a saída. O mapeamento dos dados de entrada para os dados de saída é o objetivo da aprendizagem supervisionada.

3.2.1 *Support Vector Machine*

Atualmente, SVM é o algoritmo de aprendizagem automática mais utilizado. Classifica os dados através da construção de um hiperplano (HP) no espaço de características de elevada dimensão [31], com o hiperplano a dividir um conjunto de dados em duas classes, como mostra a Figura 1.2.

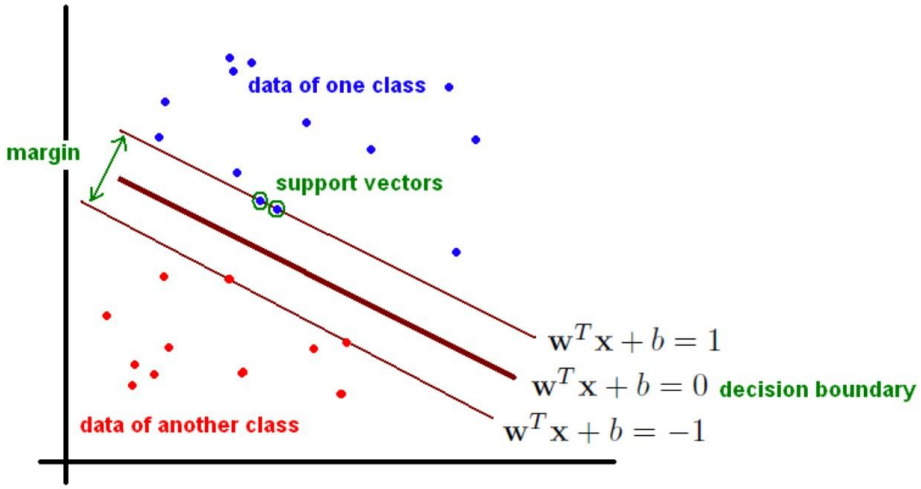


Figura 1: Modelo SVM. [32]

3.2.2 Árvore de Decisão

Em geral, uma árvore de decisão é uma técnica indutiva de aprendizagem automática para a extração de dados. São aplicados principalmente dois tipos de árvores de decisão para resolver problemas de extração de dados: Árvore de classificação e Árvore de regressão. Resumidamente, uma árvore de decisão é uma classificação expressa como uma divisão repetitiva do espaço de exemplo. No início, a árvore de decisão cria o nó raiz que constitui o nó folha. Cada nó tem o nó de saída e o nó de entrada, exceto o nó raiz, e divide-se em duas ou mais sub-árvores que dependem do nó de entrada [33]. A Figura 1.1 mostra uma forma geral de uma árvore de decisão.

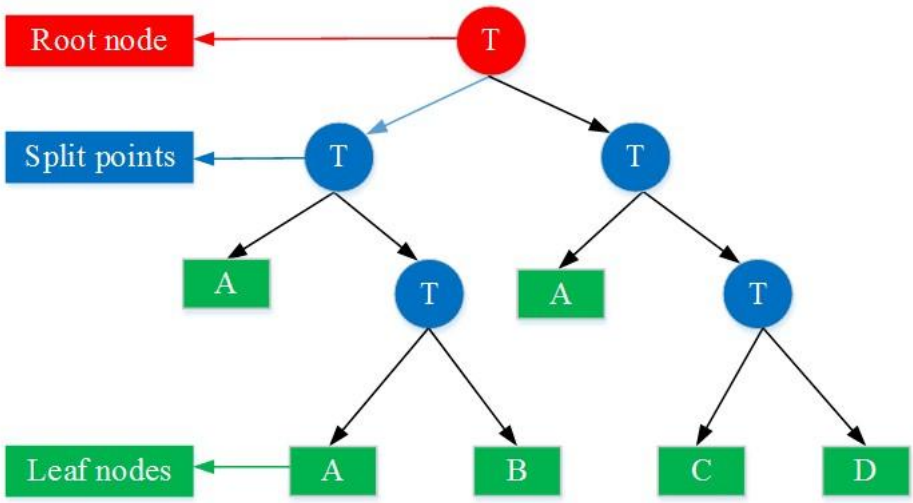


Figura 2: Modelo de Árvore de decisão [33]

3.2.3 Floresta Aleatória

A floresta aleatória é um método de aprendizagem automática flexível e de fácil utilização que produz excelentes resultados sem necessidade de afinação de hiper-parâmetros. É um dos algoritmos mais utilizados, devido à sua simplicidade e diversidade [34].

Cada árvore de classificação utiliza amostras do conjunto de dados inicial como entrada. Em cada nó, as características são escolhidas ao acaso e utilizadas para construir a árvore. Cada árvore da floresta não deve ser podada até que o exercício esteja concluído, quando a previsão é alcançada de forma decisiva. Nesta abordagem, a floresta aleatória permite que qualquer classificador fracamente correlacionado se torne um classificador poderoso. [25]

A floresta aleatória é uma abordagem baseada em árvores de decisão para prever resultados e analisar comportamentos [35]. Inclui inúmeros árvores de decisão que representam uma instância única da classificação dos dados introduzidos na floresta aleatória. O método da floresta aleatória analisa cada instância separadamente, selecionando a que tem a maioria dos votos como a previsão selecionada. A Figura 2.4 mostra uma forma geral de um algoritmo de floresta aleatória.

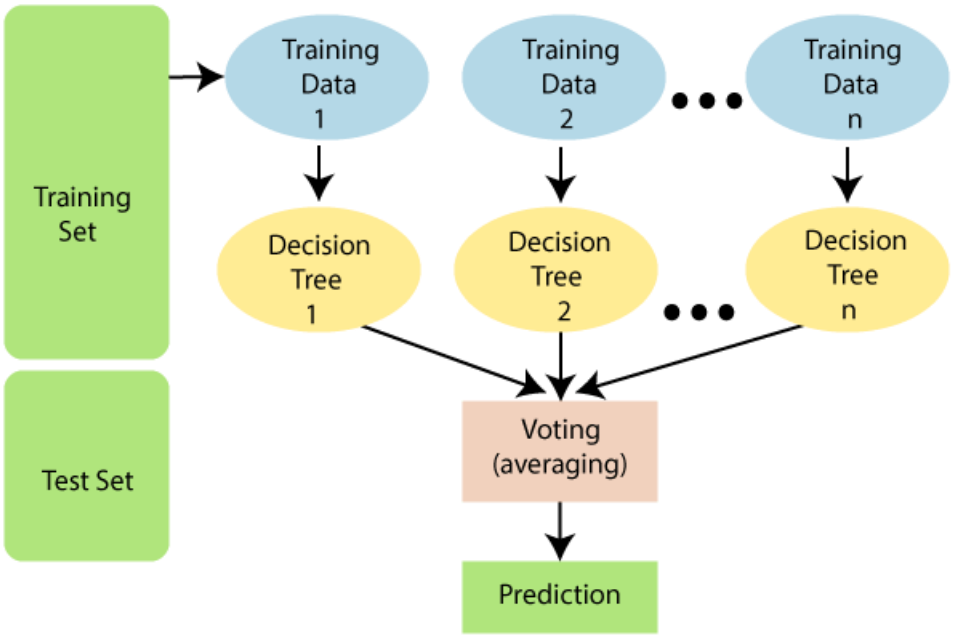


Figura 3: Floresta Aleatória

Fonte: https://www.javatpoint.com/machine-learning-random-forest-algorithm

3.2.4 Métricas de desempenho

As métricas de desempenho que descrevem a avaliação dos algoritmos de aprendizagem automática. No capítulo seguinte, é revelado a metodologia e configuração experimental que define a implementação do algoritmo durante as experiências. Para este trabalho, o desempenho do sistema foi medido através da *Accuracy*, Precisão, *Recall*, *F1-Measure* e Matriz de confusão

a) A matriz de confusão

É uma matriz utilizada para determinar o desempenho dos modelos de classificação para um determinado conjunto de dados de teste. Só pode ser determinada se os valores reais dos dados de teste forem conhecidos. A matriz em si pode ser facilmente compreendida, mas as terminologias relacionadas podem ser confusas. Uma vez que mostra os erros no

desempenho do modelo sob a forma de uma matriz, é também conhecida como matriz de erros [24]. Algumas características da matriz de confusão são apresentadas a seguida:

- Para as 2 classes de previsão dos classificadores, a matriz é uma tabela 2*2, para 3 classes, é uma tabela 3*3, e assim por diante.
- A matriz está dividida em duas dimensões, sendo os valores previstos e os valores efetivos, com o número total de previsões.
- Os valores previstos são os valores previstos pelo modelo e os valores efetivos são os valores reais para as observações dadas.[3]

O aspeto é o da tabela seguinte:

Tabela 6: Confusion Matrix for Classification

	Atual		
Predict		Positive	Negative
	Positive	TP: True Positive	FN: False Negative
	Negative	FP: False Positive	TN: True Negative

b) Classification Accuracy

É um dos parâmetros importantes para determinar a exatidão dos problemas de classificação. Define a frequência com que o modelo prevê o resultado correto. Pode ser calculada como a relação entre o número de previsões corretas feitas pelo classificador e o número total de previsões feitas pelos classificadores [24]. A fórmula é apresentada a seguir:

$$\text{Accuracy} = \frac{TP + TN}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

c) Precision

É definida como o número de resultados corretos fornecidos pelo modelo ou, de todas as classes positivas que foram corretamente previstas pelo modelo, quantas delas eram realmente verdadeiras. É calculada utilizando a fórmula seguinte:

$$\text{Precision} = \frac{TP}{\text{TP} + \text{FP}}$$

[3] https://www.javatpoint.com/confusion-matrix-in-machine-learning

d) F1-Measure

Se dois modelos tiverem uma precisão baixa e uma recuperação alta ou vice-versa, é difícil comparar esses modelos. Assim, para este efeito, podemos utilizar a pontuação F. Esta pontuação ajuda-nos a avaliar a recuperação e a precisão ao mesmo tempo. A pontuação F é máxima se igualar à precisão [24]. É calculada utilizando a fórmula seguinte:

$$\text{F1} - \text{Measure} = \frac{2 * (\text{Recall} * \text{Precision})}{\text{Recall} + \text{Precision}}$$

e) Recall

É definido como o número de classes positivas que o nosso modelo previu corretamente. A recuperação deve ser tão elevada quanto possível.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

3.3 Padronização

- Abordagem

As características do conjunto de dados dado flutuam significativamente dentro dos seus intervalos ou são registadas em várias unidades de medida, temos a necessidade de padronizar. Os dados são dimensionados para uma variância de 1 depois que a média é reduzida para 0 via padronização. Mas ao determinar a média empírica dos dados e o desvio padrão, os *outliers* presentes nos dados têm um impacto significativo que reduz o espectro de valores característicos [24].

Muitos algoritmos de aprendizagem automática podem encontrar problemas devido a essas variações nos recursos iniciais. Para algoritmos que calculam distância, por exemplo, se qualquer um dos recursos do conjunto de dados tiver valores com intervalos grandes ou completamente diferentes, esse recurso específico do conjunto de dados controlará o cálculo da distância.

A função *StandardScaler* do *sklearn* baseia-se na teoria de que as variáveis do conjunto de dados cujos valores estão em diferentes intervalos não têm uma contribuição igual para os parâmetros de ajuste e função de treino do modelo e podem até levar a viés nas previsões feitas com esse modelo [24].

Portanto, antes de incluir os recursos no modelo de aprendizagem automática, devemos padronizar os dados (μ = 0, σ = 1). A padronização na engenharia de recursos é comumente empregue para resolver esse problema potencial.

- Justificativa

  - Preservação da Escala Original: A padronização preserva a escala original dos dados; isso pode ser crucial para a interpretação dos resultados educacionais. Isso torna os resultados fáceis de entender e conectar às métricas e variáveis originais da universidade.
  - Robustez em algoritmos baseados em distância: algoritmos que usam medidas de distância, como o *Kprototype* e o SVM, podem ser sensíveis à escala dos dados. A padronização, que dimensiona os dados pela variância e centraliza os dados em torno de zero, pode melhorar a desempenho desses algoritmos, garantindo uma representação mais equilibrada e comparável entre as variáveis.
  - Maior Interpretabilidade: Os resultados do modelo estão diretamente relacionados às características e métricas coletadas na universidade porque a escala original dos dados foi mantida. Isso facilita a comunicação com partes interessadas, como educadores e gestores escolares, e a interpretação dos resultados.

3.4 Balanceamento de dados

Em Ciência de Dados e Aprendizagem automática, recorremos frequentemente a um termo chamado Distribuição de Dados desbalanceados, que acontece quando as perceções em uma das classes são muito maiores ou menores do que em classes diferentes. Os cálculos de aprendizagem automática aumentam geralmente a exatidão diminuindo o erro, para que eles não pensem na transmissão da classe. Esse problema é predominante em modelos, por exemplo, Detecção de Fraude, Detecção de Anomalias, Reconhecimento Facial e assim por diante [24].

Os procedimentos padrões de aprendizagem automática, por exemplo, Árvore de Decisão e Regressão Logística, tendem para a classe da maior parte, e muitas vezes ignoram a classe minoritária. Eles tendem a antecipar a classe maior, assim, tendo um erro significativo de classificação da classe minoritária no exame com a maioria da classe.

Quanto a técnicas de tratamento de dados desbalanceados, existem principalmente 2 cálculos predominantemente que são amplamente utilizados para lidar com o transporte de classe desequilibrado que são SMOTE e Algoritmo de Near Miss[4]. No nosso caso de estudo trabalhamos com a técnica SMOTE.

a. SMOTE – Oversampling

A estratégia de superamostragem de minorias fabricadas SMOTE é uma das técnicas de superamostragem mais utilizadas para resolver a questão da irregularidade. Ele planeia ajustar a transmissão de classe expandindo arbitrariamente os modelos de classe minoritária, duplicando-os. Destroyed incorpora novos exemplos minoritários entre casos minoritários existentes. Produz os registos de preparação virtual por adição direta para a classe minoritária. Esses registos de preparação projetados são produzidos escolhendo arbitrariamente pelo menos um dos vizinhos mais próximos para cada modelo na classe minoritária. Após o sistema de super amostragem, as informações são refeitas e alguns modelos de ordem podem ser aplicados para as informações manipuladas [36].

Procedimento de trabalho do algoritmo SMOTE

- **Estágio 1:** Classe minoritária Configuração, conjunto A, para cada um, os k-vizinhos mais próximos de x são obtidos trabalhando a distância euclidiana entre x e cada exemplo no conjunto A.
- **Estágio 2:** A taxa de teste N é definida pela extensão desequilibrada. Para cada um, N modelos (x1, x2, ... xn) são escolhidos arbitrariamente dos seus vizinhos k-mais próximos, e eles constroem o conjunto.
- **Estágio 3:** Para cada modelo (k= 1, 2, 3 ..... N), a equação que a acompanha é utilizada para produzir outro modelo: rand (0, 1) aborda o número irregular em algum lugar na faixa de 0 e 1.

3.5 Considerações e a justificação da escolha de modelo

Esta pesquisa utilizou uma abordagem metodológica de aprendizagem automática para tipificar o risco de abandono escolar e as taxas de reprovações. O algoritmo *Kprototype*

[4] https://www.javatpoint.com/handling-imbalanced-data-in-python-with-smote-algorithm-and-near-miss-algorithm

foi inicialmente usado para agrupar os dados. Essa abordagem combina *K-Means* com uma abordagem específica para variáveis categóricas, permitindo uma representação mais precisa e completa das características dos alunos. Após a fase de agrupamento, vários algoritmos de classificação, como floresta aleatória, árvore de decisão e SVM, foram escolhidos. A necessidade de capturar uma variedade de aspetos e nuances dos dados, além de garantir a generalização e a robustez do modelo, foi o que levou a esta decisão.

As árvores de decisão foram escolhidas por sua interpretação e capacidade de capturar relações não lineares entre variáveis, enquanto o SVM foi escolhido por sua eficácia na identificação de fronteiras de decisão complexas em conjuntos de dados de alta dimensionalidade. Além disso, as florestas aleatórias foram incluídas como um método de grupo para aumentar a precisão da classificação e reduzir o *overfitting*.

A combinação desses algoritmos fornece uma abordagem abrangente para a tipificação da taxa de reprovação e do risco de abandono escolar. Isso permite a implementação de intervenções e políticas educacionais direcionadas para identificar os alunos mais vulneráveis. Este método pretende melhorar a compreensão dos fatores que contribuem para o abandono escolar e a reprovação, bem como fornecer informações úteis para a criação de planos de prevenção e apoio aos alunos em risco.

Capítulo 4

4. Metodologia

Neste capítulo, apresentamos a metodologia adotada para abordar a tipificação do risco de abandono escolar e da taxa de reprovações num contexto académico. A abordagem de arquitetura, a caixa de ferramentas de IA para tipificar o risco de abandono de aluno, esquema de dados de aluno e as considerações finais são todos componentes da abordagem integrada.

Desenvolver um modelo para o sistema educacional angolano, principalmente para a UNIKIVI na província do Uíge, para ilustrar um caso prático, relevante por três razões:

- Em primeiro lugar, a preocupação nas instituições e sociedade na redução do abandono escolar;
- Em segundo lugar, a qualidade e autenticidade dos dados permite análises sofisticadas para a abordagem de aprendizagem automática; e,
- Finalmente, Angola é um país de rendimento médio, pelo que esta despesa poderia ser útil para outras instituições de ensino superior da província e do país em geral.

Esta metodologia produz passos necessários para desenvolver um modelo robusto para estimar o risco individual de cada aluno abandonar a escola, gerando aplicações para apoiar a tomada de decisão de políticas públicas. Como avanço de pesquisas anteriores, esta proposta centra-se na criação e análise de perfis individuais de alunos com modelo de *clustering*, incorporando o perfil criado na classificação para prever o abandono com diferentes modelos.

Os métodos de pesquisa foram descritivos, exploratórios e quantitativos. envolveram a coleta, análise e interpretação de dados visando identificar as correlações entre as variáveis e os fatores causais da evasão escolar e a elevada taxa de reprovações. Foi utilizada pesquisa bibliográfica a partir de materiais já desenvolvidos e publicados em livros e artigos científicos como base teórica [37].

No trabalho de Gil, afirma que a pesquisa exploratória é justificada porque permite novas descobertas e maior flexibilidade para estudar e analisar exemplos de problemas. O autor

afirma que, relativamente aos objetivos específicos da investigação, a pesquisa descritiva é adequada porque o objetivo principal da pesquisa é determinar as características do grupo de estudo e determinar como as variáveis estão ligadas entre si [38].

Na mesma linha de pensamento de Gil, a pesquisa experimental inclui a escolha de um objeto de estudo, a seleção de varáveis que podem influenciá-lo e a determinação de métodos de controle e observação. As conclusões da investigação atual se enquadram nessas características. O uso da abordagem quantitativa é razoável porque permite a análise de dados estatísticos relacionados às relações sociais [39].

4.1 Arquitetura do projeto

Neste ponto apresentam-se as principais análises de diferentes métodos e ferramentas e métricas utilizadas no projeto. As etapas do estudo estão ilustradas na Figura 4.

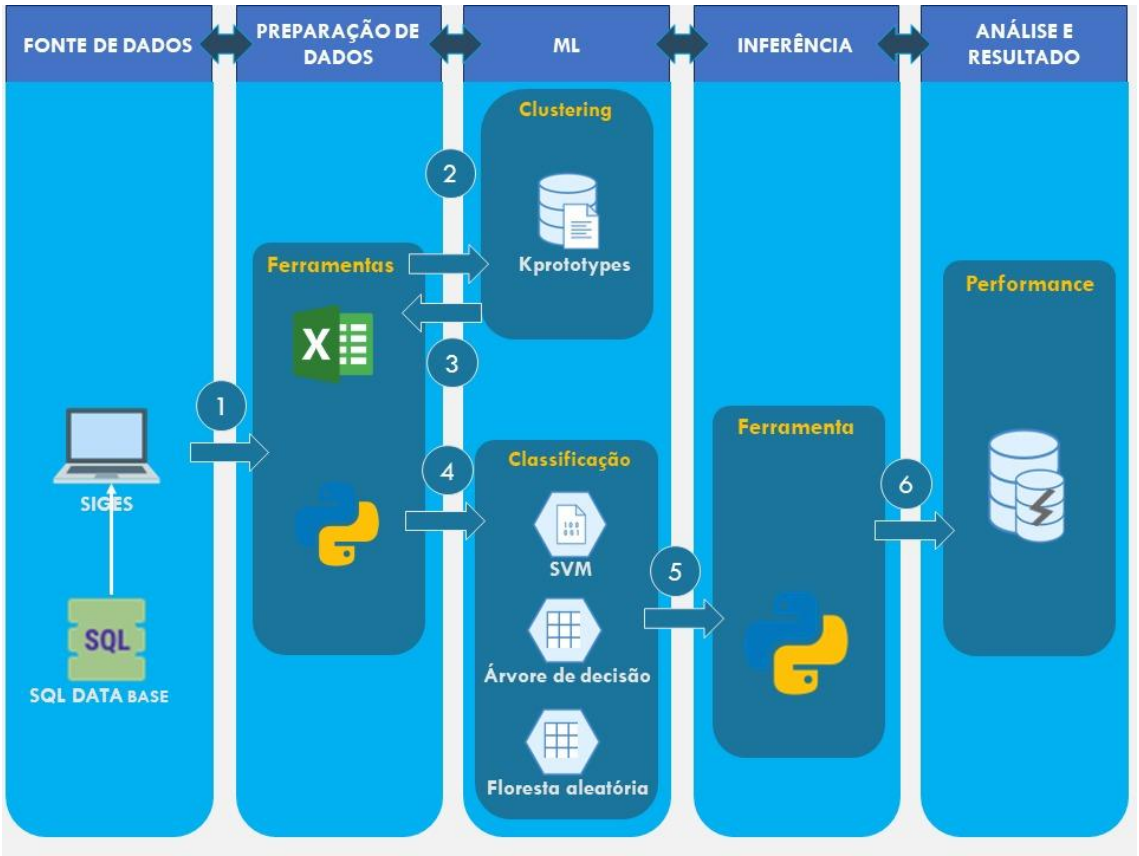


Figura 4: Arquitetura do projecto

4.2 Perfis de alunos utilizando aglomeração

4.2.1 Fonte de dados

Para avaliação das diferentes técnicas de ML usados nesta dissertação, foi considerado um conjunto de dados relativo aos estudantes que frequentaram os cursos de Engenharia Informática, Contabilidade e Gestão, Hidráulica e Saneamento de Água, Agronomia e Enfermagem Geral na UNIKIVI, nos anos académicos de 2016 ao 2022/2023. O conjunto de dados é constituído por 20056 registos (alunos com vários registos histórico) com 20 atributos que foram extraídos a partir do aplicativo académico utilizado na UNIKIVI designado SIGES.

4.2.2 Preparação de dados

Após a colheita de dados, fez-se a preparação dos mesmos no intuito de combinar, estruturar e organizar os para serem usado:

- Para criar modelos de aprendizagem automática;
- Para fins de *Business Intelligence* na Universidade;
- Em aplicações de análise e visualização de dados.

Utilizando ferramentas estatísticas e algoritmos de aprendizagem automática, para descobrir a tendência dos alunos que arriscam abandonar a escola.

4.2.3 Treinamento de aprendizagem automática

Os algoritmos de aprendizagem automática sugeridos neste estudo, que nos permitiram tipificar o abandono escolar dos alunos, foram treinados usando a linguagem *Python* e a biblioteca *SciKit-Learn*.

- *Python*: É uma linguagem de programação de alto nível amplamente utilizada para várias aplicações, incluindo *machine Learning* [40].
- *Scikit-Learn*: É um módulo Python que integra algoritmos de aprendizagem automática de última geração para problemas supervisionados e não supervisionados, com foco na facilidade de uso e desempenho [40].

4.2.4 Inferência de aprendizagem automática

Os modelos de aprendizagem automática não supervisionada e supervisionada recebem pontos de dados inéditos durante o processo de inferência. Para começar, usamos o modelo *Kprototypes* para agrupar os alunos de acordo com os seus perfis ou características semelhantes. Depois disso, a variável agrupada, ou cluster, é adicionada ao novo *dataset* para ser alimentada no treino para os modelos de aprendizagem automática supervisionada que calculam a probabilidade dos alunos que abandonam a escola. O registo de modelos facilita o acompanhamento de modelos treinados na universidade.

4.2.2 Estrutura de dados dos alunos

As informações relevantes para o modelo de abandono escolar dos alunos incluem elementos que impactam o comportamento dos alunos. A Tabela 3-1 lista os elementos de dados que afetam os padrões de atrito e retenção dos alunos.

Tabela 3-1: Estrutura de Dados dos alunos da UNIKIVI

Característica	***Tipo***	***Descrição***
CodAluno	Número	Código do aluno
Nome completo	Categórico	Nome completo do aluno
Id. Institucional	Categórica	Código da Instituição atual
Sexo	Categórica	Sexo do aluno (Masculino e Feminino)
Idade	Número	Idade do aluno (19 a 75 anos)
Ano Lectivo Ingresso	Número	Ano de Ingresso na Universidade
Turma acesso	Categórica	Turma (TL101, TL201 e TL202)
Ano(classe)	Número	Nível do aluno (1º ano ao 4º ano), exceto curso de Enfermagem e Hidráulica que estende ao 5º ano
Morada	Categórica	Endereço do aluno
Instituição de Proveniência	Categórica	Código da Instituição da Proveniência
Desc. habilitação anterior	Categórica	Descrição da habilitação anterior do aluno comprida entre 11º, 12º ou 13º Classe
Nome instituição curso	Categórica	Nome da instituição do curso anterior no secundário
Período	Categórica	Período que aluno frequentou (Manhã e Tarde)
Esta_civil	Categórica	Estado civil do aluno (Solteiro, Casado e Divorciado)
Naturalidade	Categórica	Naturalidade do aluno
Habilitações literária pai	Categórica	Habilitação literária do pai (Não estudou, Técnico básico, Técnico Médio, Licenciado, Mestrado, Doutor e Técnico Profissional)

Habilitações literária mãe	Categórica	Habilitação literária da mãe (Não estudou, Técnico básico, Técnico Médio, Licenciado, Mestrado, Doutor e Técnico Profissional)
Semestre	Categórico	Histórico de semestre (1º e 2 semestre)
Curso	Categórica	Curso do aluno (Informática, Enfermagem, Contabilidade e Gestão, Agronomia e Hidráulica e saneamento de água)
Ano_Acad	Número	Ano académico (dados recolhidos de 2016 a 2022)

De acordo com [41], o pré-processamento foi realizado para "melhorar a qualidade dos dados por meio da eliminação ou minimização dos problemas", incluindo ruídos, valores incorretos, inconsistentes ou ausentes. Um perfil de aluno é criado e classificado conforme o risco de abandono escolar usando técnica de aprendizagem automática.
Por meio do pré-processamento de dados, os elementos considerados irrelevantes foram removidos manualmente e o que colocou em risco a privacidade dos dados dos alunos. Os autores afirmam claramente que, quando um atributo não contribui para a estimativa do valor do atributo alvo, é considerado irrelevante. Neste caso, dos 20 atributos do conjunto de dados inicial, 8 atributos foram removidos, e ficaram os 12 outros atributos que são ano de conclusão, naturalidade, estado civil, código do aluno, sexo, idade, ano de ingresso, ano (classe), habilitação do pai, habilitação da mãe, instituição de origem e curso que frequenta.

Após análise exploratória foram adicionados os atributos seguintes a situação de cada aluno, incluindo o número de inscrições do aluno, o ano da primeira e última matrícula do aluno, o número de vezes que o aluno foi reprovado em cada classe e a situação de abandono, ou o número de anos abandonados. Isso permitiu-nos analisar por classe e por curso para descobrir quais cursos são mais frequentes na instituição de origem, estabelecendo perfis e outras características. Ficamos com 4354 registos (alunos) e 21 registos, a tabela mostra os atributos restantes.

Tabela 7: Atributos Selecionados

Característica	***Descrição***
CodAluno	Código do aluno
Sexo	Sexo do aluno
Idade	Idade do aluno
Ano_ingres	Ano de ingresso do aluno (Primeira matrícula do aluno)
Ulti_matri	Última matrícula ou último registo na faculdade do aluno (Última matrícula do aluno)
Num_repro	Número de vez que aluno já reprovou durante o seu ciclo
Dura_Curso	Duração do curso que é de 4 anos, exceto curso de Enfermagem e Hidráulica e saneamento de água que são 5 anos de duração
N_inscri	Número de vez foi inscrito, visto que por regra um aluno deve se inscrever pelo menos uma vez por ano, mas no caso de dever disciplinas noutra classe, ele deve se inscrever mais neste caso duas vez num ano assim por diante

N_abando	Número de vez que aluno abandonou a faculdade
Clas_atual	Classe que está frequentar ou frequentou atualmente
Semestre	Semestre que pode ser 1 ou 2 semestre
Período	Período que aluno frequentou
Esta_civil	Estado civil do aluno
Naturalidade	Naturalidade do aluno
Habil_pai	Habilitação do pai
Habil_mãe	Habilitação da mãe
Insti_Proven	Instituição de Proveniência do aluno
Corresp_inst_curso	Código da correspondência do curso de proveniências (valor 1 representa a correspondência do curso frequentado na instituição anterior e 0 caso contrário)
Curso	Curso do aluno (Informática, Enfermagem, Contabilidade e Gestão, Agronomia e Hidráulica e saneamento de água)
Ano_Acad	Ano académico (dados recolhidos de 2016 a 2022)
Status	Situação do aluno que pode ser "Cursando", "Concluído" ou "Abandonado", neste caso olhou-se últimos registos (Ano académico 2022) dos dados recolhidos.

Os algoritmos de agrupamento *Kprototypes* e modelos de classificação como DT, SVM e RF foram usados como métodos de aprendizagem automática para resolver o problema de abandono e reprovação na universidade. As técnicas de pré-processamento utilizadas incluíram o tratamento de dados desbalanceados, ruídos, incompletos, redundâncias e conversão de dados categóricos em números.

4.3 Os perfis de alunos utilizando aglomeração

a) Contextualização

Para formar grupos de perfis de alunos em risco de abandono escolar com aprendizagem automática, é importante levar em consideração uma variedade de elementos. O estudo enfatiza a importância de examinar informações como Sexo, Correspondência da formação anterior, Idade, Ano de ingresso, Duração do curso, Estado civil, Período, Naturalidade, Habilitação do pai, Habilitação da mãe, Ano académico, Instituição de Proveniência, última matrícula, Número de reprovações por aluno, Curso, Número de inscrições por aluno, Número de vez que aluno abandonou, Classe atual e Status ou Situação do aluno. O objetivo desta fase é fornecer um sistema de alerta precoce para um grupo de aluno em risco de evasão e aumentar a nossa compreensão dos diferentes perfis de alunos e o seu comportamento dinâmico ao longo do tempo.

b) Método para encontrar o valor ideal de K

Para determinar o número ideal de clusters (k), empregamos a técnica de repetição 20 vezes de k=2 a 40. Calculamos a média de cada k para evitar resultados aleatórios. Utilizamos métodos de validação, como método *Elbow*, *Silhueta* e *Davies_bouldin*, para selecionar o k ideal. Posteriormente, aplicamos o algoritmo *Kprototypes* para segmentar os alunos em perfis distintos.

A abordagem *Elbow*, *Silhouette Score* e *Davies_bouldin Method* com múltiplas iterações médias é uma abordagem sólida e criteriosa para encontrar o valor ideal de k. Além disso, os resultados são comparados. Uma estratégia é usar a técnica de iterar várias vezes dentro dos clusters e calcular a média para encontrar o valor ideal de k. Isso aumenta a confiabilidade dos resultados porque as iterações podem variar e a média ajuda a reduzir as flutuações. Isso leva em consideração a variação e a estabilidade entre vários valores k ao longo das iterações. Nas imagens abaixo, o valor k mais adequado é 5, pois mostra a pontuação de silhueta mais alta apesar de termos mais também k=4 ideal, mas optamos por 5 por razões de ser igual resultado com outro método *Silhouette*. A qualidade do agrupamento pode ser avaliada por meio de uma pontuação de *silhueta*, onde um valor mais alto indica que os agrupamentos estão bem separados e distintos.

A Pontuação de Silhueta para *Clustering Kprototypes* é representada por uma linha azul na imagem. Os detalhes são os seguintes: Ele é identificado como "Número de Clusters (k)", que varia de 5 a 40 no Eixo X. O eixo Y é representado por médias de "pontuação de silhueta" de cada cluster com 20 iteração e tem uma variação de aproximadamente 0,08 a 0,11. O método do Elbow, ou "cotovelo" da curva, mostra onde a curva começa a ser tornar mais plana. O gráfico mostra a média da soma dos quadrados dentro do cluster (WCSS) relativamente ao número de clusters (k). A queda inicial é rápida até atingir o ponto k=5, após o qual a taxa de declínio diminui significativamente, indicando um cotovelo. Portanto, K=5 seria uma escolha apropriada para esse criar perfis de aluno da UNIKIVI.

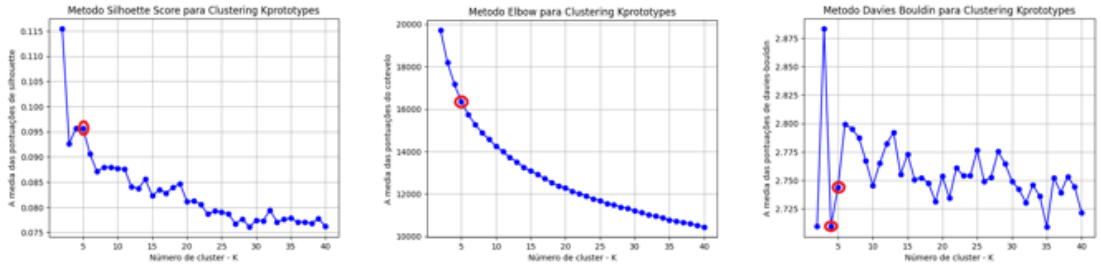


Figura 5: Método de Validação

c) Visualizar os perfis criados após o treinamento do modelo *kprototype*

Após a seleção do valor k ideal usando os métodos acima mencionados, aplicamos o algoritmo *Kprototypes* para segmentar os alunos em perfis distintos. Os cinco perfis criados com as suas diferentes características constam em anexo. A Figura 6 mostra a repartição de aluno distribuído em diferente *cluster*, optamos de duas formas para ter mais resultados para a nossa hipótese, criamos *cluster* com variável *Status* e outro sem.

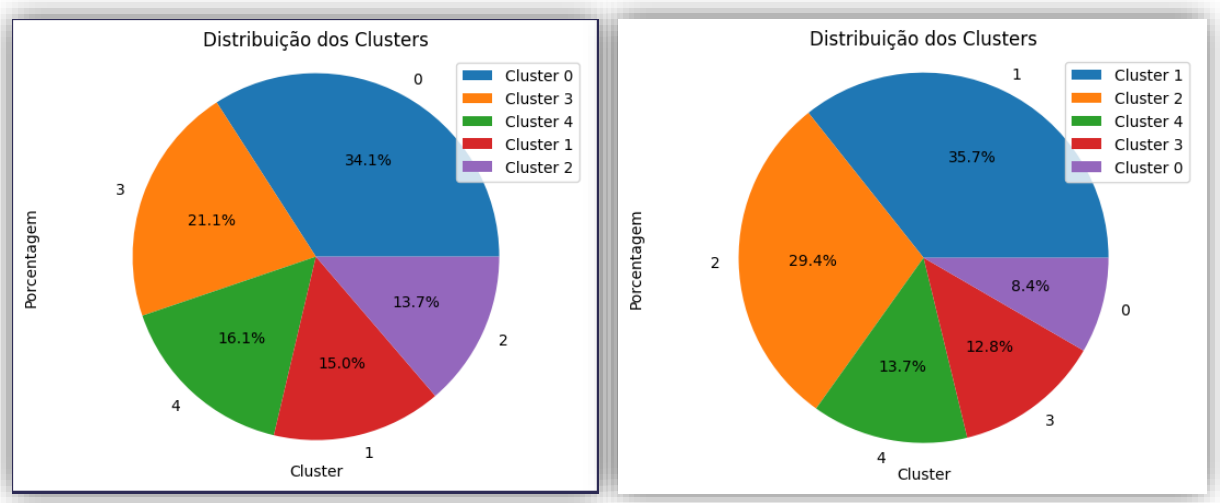


Figura 6: Distribuição de aluno em cluster com variável 'Status' e sem 'Status'

Após segmentar os alunos com o modelo *Kprototypes*, obtemos os seguintes resultados:

a) Cluster com Status

- Cluster 0: 1485 alunos;
- Cluster 1: 652 alunos;
- Cluster 2: 597 alunos;
- Cluster 3: 919 alunos;
- Cluster 4: 701 alunos.

Essa organização permite uma análise mais detalhada dos grupos de alunos com características semelhantes. Cada cluster representa um conjunto específico de alunos com base nas suas características. É uma abordagem útil para personalizar estratégias de ensino e oferecer suporte adequado a cada grupo.

- Em termos de género, os grupos são predominantemente masculinos. O cluster 0 tem 78,3%, o grupo 1 tem 68,2%, o grupo 2 tem uma distribuição ligeiramente equilibrada entre 51,7% para homens e 48,2% para mulheres, o grupo 3 tem uma proporção ligeiramente maior de 58% e o grupo 4 é o mais forte em género com 85%.
- Os grupos foram distribuídos conforme a faixa etária: o grupo 0 tem a maioria de jovens de 19 a 29 anos, com 55,6%, seguido de 30 a 39 anos, com 37,7%. O grupo 1 tem a maioria de 30 a 39 anos, com 61,2%. O grupo 3 é dominado por jovens de 19 a 29 anos, com 76%, e o grupo 4 tem a maioria de 30 a 39 anos, com 63,6%.
- Os alunos do grupo 0 têm uma proporção baixa, relativamente de 38%, os alunos do grupo 1 têm 40%, os alunos do grupo 2 têm 45%, os alunos do grupo 3 têm 37% e os alunos do grupo 4 têm 39%. Esses cursos são comparáveis aos cursos de faculdade.
- Nos anos ou classes, o grupo 0 é dominado pelo terceiro ano com 69%, o grupo 1 e o grupo 2 são fortes no terceiro ano com 78,7%, o grupo 3 é forte no terceiro ano com 50% e o grupo 4 é forte no segundo e quarto ano com 36,6%.
- No curso, os grupos 0 e 1 foram compostos principalmente por alunos de contabilidade, com 42,7% e 27,5% concentrados em hidráulica; o grupo 2 foi composto principalmente por alunos de enfermagem, com 97,7%; o grupo 3 teve uma combinação de cursos relacionados à saúde e hidráulica com 61,6% e 38%; e o grupo 4 foi dividido entre Agronomia e Informática com 29,7% e 70%.
- Relativamente ao estado civil, os solteiros tiveram o maior desempenho com 99,4%, 96,5%, 100%, 99,8% e 99,3%, respetivamente.
- Formação da mãe: os grupos 0,1,2 foram dominados pela mãe com formação básica, com 64,7%, 76,4% e 79,1%, respetivamente. O grupo 3, por outro lado, estava equilibrado entre formação básica e não estudou, com 40,2% e 35,8%, respetivamente.
- Formação do pai: O grupo 0 tem formação técnica média de 46,8 e formação básica de 34,8%; o grupo 1,2 tem formação técnica média de 80% e 68%; o grupo 3 tem um pouco de equilíbrio entre formação técnica média e formação não realizada, com 27,9 por cento e 31,6 por cento, respetivamente; e o grupo 4 tem

formação distribuída entre 44,8 por cento formação básica e 44,2 por cento formação.

- No que diz respeito ao período, o grupo 0 tem uma distribuição equilibrada entre o período manhã com 53,3% e o período tarde com 47,7%; o grupo 1,2 tem o período predominante com 66,9%, 62,7% e 71%, respetivamente; e o grupo 4 tem uma distribuição equilibrada entre o período manhã com 49,5% e tarde com 50,5%.
- Os grupos que deixaram a faculdade pelo menos uma vez receberam 31,7%, 20,9%, 19,5%, 16,5% e 11,2%, respetivamente.
- Conforme a situação dos alunos, o grupo 0 apresentou a maior taxa de abandono com 87,5% e 49,1% concluídos; o grupo 1 concluiu com 33,9% e uma taxa de abandono de menos de 1%; o grupo 2 concluiu com 26,9% e uma taxa de abandono de 11,3%; o grupo 3 concluiu com menos de 2,1% e uma taxa de abandono de 50%; e o grupo 4 concluiu com 36,6% e uma taxa de abandono de 13%.

A imagem abaixo descreve de forma clara aglomeração criada.

Grupo	Perfil 1	Perfil 2	Perfil 3	Perfil 4	Perfil 5
Genero	Esse perfil tem uma maioria significativa de homens com 78,32%.	Esse perfil tem uma maioria masculina, mas com uma proporção mais equilibrada com 68,25%	Esse perfil apresenta uma distribuição quase igualitária entre homens e mulheres com 51,76% (M) e 48,24% (F).	Esse perfil tem uma proporção de homens é ligeiramente maior com 58%.	Esse perfil tem uma maioria esmagadora de homens com 85,31%.
Faixeitaria	Predominantemente jovens, com a maioria na faixa dos 19 aos 39 anos. Proporcionado por 19-29 tem 55,6% e 30-39 com 37,7%	Equilíbrio entre diferentes faixas etárias, com destaque para a faixa dos 30 aos 39 anos com 30-39: 61,2%. e 40-49 tem 20,4%.	Grupo diversificado, com representantes de várias faixas etárias com o um dominio de Idade 30-39: 43,7%.	Dominado por pessoas mais jovens, com a maioria na faixa dos 19 aos 29 anos com 76,1%.	Uma distribuição equilibrada, com destaque para a faixa dos 30 aos 39 anos que tem 63,6%.
Correspondencia com curso anterior	Uma proporção de correspondência relativamente baixa com curso secundário com 38,0%.	Apresenta uma correspondência média com curso secundário com 40%.	Apresenta uma correspondência mais alta, indicando uma forte relação com curso secundário tem 45%.	Apresenta uma correspondência um pouco abaixo da média em relação com curso secundário e tem 37%.	Apresenta uma correspondência está próxima da média em relação ao curso secundário e tem 39%.
Classe	Esse perfil é predominantemente pela 3ºano com 69,56%.	Esse perfl é mais proeminente é a 4º e 2ºano.	Esse perfil é fortemente caracterizado pela 5ºano com 78,73%.	Nesse perfil, a 3ºano também tem uma presença significativa com 49,95%.	A 4ºano é proeminente junto 2ºano nesse perfil com 36,66%.
Curso	Esse perfil tem uma combinação de cursos variada, com destaque para Hidráulica e Saneamento das Águas 42,7% e Contabilidade com 27,5%	Esse cluster é composto exclusivamente por estudantes de Contabilidade e Gestão	A maioria dos estudantes pertence ao curso de Enfermagem com 97,7%	Esse perfil tem uma combinação de cursos relacionados à saúde (61,6%) e meio ambiente(38,3%)	Esse perfil tem uma divisão entre Agronomia (29,7%) e Informática (70,3%)
Estado Civil	Predominante por solteiro com 99,4%	Combinação de casadas e solteiras, com a maioria sendo solteira (96,5%)	Maioria é solteira	Predominância de solteiros com 99,8%	Predominância de solteiros com 99,3%
Formação da mãe	Esse perfil é composto principalmente por mães com formação técnico básico com 64,7% e tecnico médio com 18,3%.	Esse perfil é predominantemente por mães com formação técnico básico com 76,4%.	Esse perfil é predominantemente por mães com formação técnico básico com 79,1%	Esse perfil é predominantemente por mães com formação técnico básico com 40,2% e 35,8% Não estudaram	Esse perfil é predominantemente por mães com formação técnico básico com 84,9%
Formação do pai	Habilitação predominado pela formação de Técnico Médio com 46,8% e Básico com 34,8%	Habilitação predominando pela formação de Técnico Médio com 80,4%	A maioria dos pais tem formação em técnico médio com 68,2%	Diversidade de formação de pais com Não estudaram com 31,6%, Técnico Médio (27,9%) e ensino básico (26,8%	Uma divisão entre pais com formação em ensino básico (44,8%) e técnico médio (44,2%)
Periodo	A distribuição equilibrada entre os períodos da manhã (52,3%) e da tarde (47,7%)	Periodo da tarde é predominante com 66,9%	Periodo da tarde é predominante com 62,7%	Periodo de Manhã é predominante com 71%	A distribuição é quase igual entre os dois períodos entre Manhã(49,5%) e Tarde(50,5%)
Situação (Aprovado ou Reprovado)	40,7% de aluno nunca reprovaram enquanto 12,7% já reprovaram pelo menos uma vez e o restante com mais de duas vezes.	Com uma proporção muito reduzida de aluno que nunca reprovaram 15,4%	Predominante com aluno que nunca reprovram com 68,5%	Predominando pelos alunos que nunca reprovaram com 88,6%	Com menos de 21,8% de aluno que nunca reprovaram
Ingresso	Ingresso em 2017	Ingresso em 2012	Ingresso em 2015	Ingresso em 2020	Ingresso em 2019
Última matricula	Última matricula em 2021	Última matricula em 2018	Última matricula em 2019	Última matricula em 2021	Última matricula em 2018
Abanondo	Pelo menos uma vez já abandonou a escola com 31,7%	Pelo menos uma vez já abandonou a escola com 20.9%	Pelo menos uma vez já abandonou a escola com 19.5%	Pelo menos uma vez já abandonou a escola com 16.5%	Pelo menos uma vez já abandonou a escola com 11.2%
Abanondo	Predominante pelo abandonaram com 87.5%	Abandonaram com 0.9%	Abandonaram com 11.3%	Abandorama com 0.13%	Abandorama com 11,3%
Concluido	Concluído com menos de 1%	Concluido com 33.9%	Concluiram com 26.9%	Com menos de 2.1% de conclusão	Concluiram com 36.9%
Cursando	Com 49.1% Cursando	Cursando com menos 1%	Cursando com 1%	Cursando com 50.6%	Cursando com menos de 0,3%

Figura 7: Aglomeração com a variável" Status"

b) Cluster sem a variável Status

- Cluster 0: 364 alunos;
- Cluster 1: 1554 alunos;
- Cluster 2: 1282 alunos;
- Cluster 3: 559 alunos;
- Cluster 4: 364 alunos.

Essa organização permite uma análise mais detalhada dos grupos de alunos com características semelhantes. Cada cluster representa um conjunto específico de alunos com base em suas características. É uma abordagem útil para personalizar estratégias de ensino e oferecer suporte adequado a cada grupo.

- Gênero: Os grupos 0,1,2 e 3 são maioritariamente masculinos, com 86%, 76%, 80% e 62%, respetivamente. O grupo 4 é ligeiramente equilibrado, com 53% para homens e 47% para mulheres.
- Na faixa etária, os grupos são distribuídos da seguinte maneira: o grupo 0 é predominantemente composto por jovens de 19 a 29 anos com 82%, o grupo 1 é predominantemente composto por pessoas de 30 a 39 anos com 53%, o grupo 2 é mais diversificado com um pequeno aumento de 46% entre 30 a 39 anos e 19 a 29 anos com 43%, o grupo 3 é predominantemente composto por jovens de 19 a 29 de idade com 64% e grupo 4 possui uma distribuição com destaque 37% entre 30 a 39 e o grupo que tem taxa maior de 5% com idade com mais de 60 anos.
- Os cursos do ensino médio, que são comparáveis aos cursos da faculdade, tiveram uma proporção relativamente baixa do grupo 0 com 22%, o grupo 1 com 36%, o grupo 2 com 42%, o grupo 3 com 47% e o grupo 4 com 45%.
- Nos anos ou classes, o grupo 0 é dominado pelo 2º e 3º ano com 59%, o grupo 1 tem 3º ano com 40%, o grupo 2 é fortemente dominado pelo 2º e 3º ano com 53%, o grupo 3 tem uma forte presença de 3º ano com 32,7% e o grupo 4 é dominado pelo 3º e 5º ano com 65,21%.
- No curso, o grupo 0 era exclusivamente de hidráulica; o grupo 1 era composto por alunos de contabilidade com 38%, informática com 33% e agronomia com 19%; o grupo 2 era composto por alunos de contabilidade com 24%, informática com 48% e agronomia com 27%; e o grupo 3 e 4 era exclusivamente de enfermagem.
- Estado Civil: Os grupos de solteiro têm a maioria de 100%, 98%, 99%, 99% e 98%, respetivamente.
- Formação da mãe: O grupo 0 é predominantemente composto por mães com formação técnica básica com 48% e que não estudaram com 34%. O grupo 1,2 é

predominantemente composto por mães com formação técnica básica com 77% e 67%, o grupo 3 é predominantemente composto por mães que não estudaram com 45% e com formação técnica secundária e básica com 31% e 19%, e o grupo 4 é predominante pela formação básica.

- Formação de pais: O grupo habilitado predominou a formação de técnico básico com 35%, não estudaram com 28% e médio com 27%. O grupo 1 predominou a formação de técnico médio com 98%, o grupo 2 predominou a formação de técnico básico com 71%, o grupo 4 predominou a formação de pais com não estudaram com 37%, técnico básico com 22% e técnico médio com 19%. 77% dos pais são técnicos superiores.
- Período: O grupo 0 predominou o período com 78%, o grupo 1 equilibrou o período com 57% de tarde e 43% de manhã, o grupo 2 equilibrou um pouco com 61% de tarde e 39% de manhã, o grupo 3 predominou o período de tarde com 64% e o grupo 4 predominou o período de manhã com 68%.
- Conforme o número de reprovações, o grupo 1 tem 82,6% de alunos que nunca reprovaram, o grupo 2 tem uma proporção muito menor de alunos que nunca reprovaram, com 34% e 29% que já reprovaram mais de duas vezes. No grupo 3, os alunos que nunca reprovaram predominam, com 86,7% e 7%, respetivamente, o grupo 4 tem a maioria dos alunos que nunca reprovaram, com 70% e 20% que já.
- A taxa de abandono da faculdade por pelo menos uma vez foi de 5%, 40%, 31%, 5% e 16%, respetivamente.
- Situação dos alunos, com o grupo 0 predominante com 91% de aprovação, concluindo com 9% e 1% de abandono, o grupo 1 teve um pouco de destacamento, concluindo com 55%, alunos abandonando com 20% e cursando com 25%. O grupo 2 teve uma preponderância de alunos, concluindo com 39%, cursando com 36% e abandonando com 25%. O grupo 4 teve preponderância de alunos, concluindo com 66%, cursando com 21% e abandonando com 13%.

A imagem abaixo descreve de forma clara aglomeração criada.

Perfil	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Genero	A maioria significativa de homens com 86%	A maioria masculina, mas com uma proporção mais equilibrada com 76%	A maioria masculina, mas com uma proporção mais equilibrada com 80%	A proporção de homens é ligeiramente maior com 62%.	A distribuição quase igualitária entre homens e mulheres com 53% (M) e 47% (F).
Faixa etária	Predominança de jovens, com a maioria na faixa dos 19 aos 39 anos. Proporcionado por 19-29 tem 82% e 30-39 com 14%	Equilíbrio entre diferentes faixas etárias, com destaque para a faixa dos 19 aos 49 anos com 19-29: 30%. e 30-39 tem 53% e 40-49 tem 12%.	Repartido maioritariamente pela Idade 19-29: 43% e 30-39: 46%.	Dominado por pessoas mais jovens, com a maioria na faixa dos 19 aos 29 anos com 64%.	Uma distribuição mais ou menos equilibrada, com destaque para a faixa dos 30 aos 39 anos que tem 37% e que possui porção maior de 5% de Idade de 60-69 e 40-49 com 18% em relação aos outros grupos.
Curso feito no secundário	Uma proporção de correspondência relativamente baixa com curso secundário com 22%.	Apresenta uma correspondência média com curso secundário com 36%.	Apresenta uma correspondência mais alta, indicando uma forte relação com curso secundário tem 42%.	Apresenta uma correspondência um pouco abaixo da média em relação com curso secundário e tem 47%.	Apresenta uma correspondência está próxima da média em relação ao curso secundário e tem 45%.
Ano	Predominante pela 2 e 3ºano com 59%.	Mais proeminente é a 3º ano com 40%.	Fortemente caracterizado pela 2º e 3º ano com 52%.	O 3ºano também tem uma presença significativa com 32,74%.	O 4º e 5ºano é proeminente nesse perfil com 65,21%.
Curso Universitário	Composto exclusivamente por estudantes de Hidráulica e Saneamento das Águas	Uma combinação de cursos variada, com destaque para Contabilidade com 38%, Informática com 33% e Agronomia com 19%	Uma combinação de cursos variada, com destaque para Informática com 48%, Contabilidade com 24% e Agronomia com 27%	Composto exclusivamente por estudantes de Enfermagem	Composto exclusivamente por estudantes de Enfermagem
Estao civil	Predominante por solteiro com 100%	Combinação de casadas e solteiras, com a maioria sendo solteira (98%)	Combinação de casadas e solteiras, com a maioria sendo solteira (99%)	Predominância de solteiros com 99%	Predominância de solteiros com 98%
Formação da mãe	Composto principalmente por mães com formação técnico básico com 48% e que não estudaram com 34%.	Predominantemente por mães com formação técnico básico com 77%.	Predominantemente por mães com formação técnico básico com 67%	Predominantemente por mães que não estudaram com 45% e com formação técnico secundário e básico com 31% e 19%	Predominantemente por mães com formação técnico básico com 94%
Formação do pai	Habilitação predominado pela formação de Técnico Básico com 35%, Não estudaram com 28% e Médio com 27% e	Habilitação predominando pela formação de Técnico Médio com 98%	A maioria dos pais tem formação em técnico básico com 71%	Diversidade de formação de pais com Não estudaram com 37%, Técnico básico com 22% e Técnico Médio com 19%	A maioria dos pais tem formação em Técnico Secundário com 77%
Periodo	Periodo da manhã é predominante com 78%	Periodo equilibrado com 57% de Tarde e 43% para Manhã	Periodo um pouco equilibrado com 61% de Tarde e 39% para Manhã	Periodo de Tarde é predominante com 64%	Periodo de Manhã é predominante com 68%
Reprovação	82,6% de aluno nunca reprovaram enquanto 13% já reprovaram pelo menos uma vez e o restante com mais de duas vezes.	Com uma proporção muito reduzida de aluno que nunca reprovaram 22% enquanto 26% já reprovaram mais de duas vezes	Predominado ligeiramente com aluno que nunca reprovram com 34% e 29% com mais de duas vezes	Predominando pelos alunos que nunca reprovaram com 86,7% e 7% com uma vez	Predominando pelos alunos que nunca reprovaram com 70% e 20% com uma vez
Ingresso	Ingresso em 2020	Ingresso em 2014	Ingresso em 2016	Ingresso em 2019	Ingresso em 2016
Matricula	Última matricula em 2021	Última matricula em 2019	Última matricula em 2020	Última matricula em 2019	Última matricula em 2020
Abandonado pelo menos uma vez	Pelo menos uma vez já abandonou a escola com 5%	Pelo menos uma vez já abandonou a escola com 40%	Pelo menos uma vez já abandonou a escola com 31%	Pelo menos uma vez já abandonou a escola com 5%	Pelo menos uma vez já abandonou a escola com 16%
Status	Predominante pelos alunos Cursando com 91%, Concluido com 9% e 1% para Abandono	Destacado ligeiramente pelos alunos concluido com 55%, alunos Abandonaram com 20% e Cursando com 25%	Proporcionado com Concluido com 39%, Cursando com 36% e que Abandonaram com 25%	Predominante pelos alunos Cursando com 76% , Concluido com 23% e Abandonaram com 0%	Predominante pelos alunos Concluido com 66% , Cursando com 21% e Abandonaram com 13%

Figura 8; Aglomeração sem a variável" Status"

4.4 As classificações de alunos com modelo de aprendizagem automática

a) Contextualização

A pesquisa alcançou uma precisão de classificação extremamente alta. Usando a validação cruzada, antes de usar outras técnicas anteriores, os resultados não foram muito convincentes; um método apresentou resultados de 99% para SVM e outro de 100%. No entanto, ao usar a validação cruzada, descobrimos que os resultados para SVM é de 93% e Árvore de decisão e floresta aleatória tem 95%.

Modelos de previsão podem ajudar a minimizar perdas na instituição. Para classificar os alunos em risco de abandono escolar com aprendizagem automática, levou-se em consideração uma variedade de elementos. Como já foi mencionado no modelo anterior sobre a escolha de características além das características trabalhado na *clustering* foi adicionado a coluna de "*Cluster*" onde foi criado diferentes grupo de alunos e usar o *target* "*Status*" que contém valor como Cursando, concluído e abandono.

O *dataset* usado contém de forma distribuído concluído tem 43,6% (1899 alunos), cursando tem 39,8% (1732 alunos) e abandono 16,6% (723 alunos), a Figura 10 mostra em detalhes a distribuição de estado de aluno em percentagem. Como estamos a prever a evasão e o sucesso académico de aluno, essa imagem mostra em detalhes os dados que pretendemos prever, onde temos Status de Abandono representado por 0, Status de Concluído por 1 e por fim Status Cursando por 2.

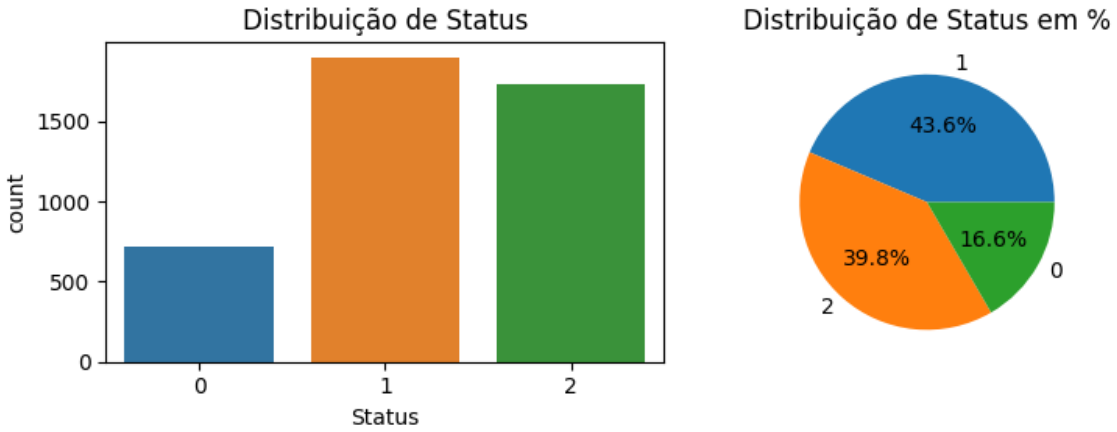


Figura 9: Distribuição de estado

A União Europeia definiu como objetivo para 2030 uma taxa de abandono escolar precoce inferior a 9%. Portanto, uma percentagem razoável para tolerar no abandono escolar seria aquela que se mantém abaixo desse limite. [5]

b) Método de classificação

Uma estratégia é usar três tipos de técnica de classificação para poder encontrar a melhor. Isso aumenta a confiabilidade dos resultados. Ao comparar os resultados desses métodos, chegamos a uma melhor conclusão. Usamos os modelos de aprendizagem automática de SVM, *Random Forest Classifier* e *Decision Tree Classifier*. Para melhor treinar o modelo passaremos por Transformar os dados para a escala padronizada, a razão da escolha de padronização relativamente à normalização visto que SVM é mais sensível na técnica de normalização. *Dataset* possui a classe 1 (43,8%) tem a maior proporção relativamente à classe 0 (16,6%) que é menor e a classe 2 tem uma percentagem aproximadamente da classe 1 que é 39,8%, para melhor obter resultado passamos por experimentar de duas forma primeira forma treinamos com a quantidade de forma inicial e depois passamos por balancear as amostras com o método de SMOTE onde permitiu-nos aumentar as amostras da classe 0 aproximadamente 50% da classe 1 num conjunto de dados sintético.

c) Instrumento de avaliação

Utilizamos métricas de desempenho, como precisão, *recall* e *F1-score*, para avaliar o desempenho dos modelos SVM, árvore de decisão e floresta aleatória na previsão do risco de abandono e taxa de reprovações. Os resultados mostraram uma alta precisão, indicando a eficácia dos modelos desenvolvidos. A ideia principal é encontrar o modelo de previsão mais adequado e comparar as várias métricas de desempenho de previsão de cada classificador aplicado. Começamos por utilizar as amostras não balanceado e depois balanceado para poder comparar os resultados.

1) Previsão sem balanceamento de amostra

Atualmente, vamos examinar os resultados da previsão sem balanceamento de amostra, observando como os modelos funcionam sem ajuste nas proporções das classes. Veremos

[5] https://www.ensino.eu/ensino-magazine/escola/2024/abandono-escolar-em-portugal-aumenta-pela-primeira-vez-desde-2017/#

como esse método pode impactar a precisão, o recall e outras métricas de avaliação. Vamos examinar esse caso e compreender as suas implicações.

a) Resultado das classificações

Os resultados mostram que os três modelos (SVM, Árvore de Decisão e Floresta Aleatória) funcionam muito bem. A acurácia do SVM foi de 97%, a precisão de 96% e a taxa de recall de 100%. A Floresta Aleatória e a Árvore de Decisão alcançaram 100% em todas as métricas. Esse valor não convenceu

Nota: No caso de cluster sem a variável sem Status o resultado quase não mudou em relação experimentação anterior, exceto no caso de Acurácia que baixa para 96% e *Recall* que baixa 98% em 100% relativamente ao anterior. Mas esses resultados não foram convincentes.

b) Matriz de confusão

- Verdadeiros Negativos (VN): Esses são os casos em que o modelo previu a classe negativa (0) corretamente e os dados reais também eram negativos. O SVM tem 138 verdadeiros negativos, enquanto a Árvore de Decisão e a Floresta Aleatória têm 152.
- Falsos Positivos (FP): Esses são os casos em que o modelo previu incorretamente uma classe positiva (1), mas os dados reais eram negativos (0). Em contraste com a Árvore de Decisão e a Floresta Aleatória, não há falsos positivos na SVM (FP = 0).
- Falsos Negativos (FN): Esses são os casos em que o modelo previu a classe negativa (0) incorretamente, mas os dados reais eram positivos (1). Nesse caso, não temos falsos negativos (FN = 0) em nenhum dos modelos.
- Verdadeiros Positivos (VP): Esses são os casos em que o modelo previu corretamente a classe positiva (1) e os dados reais também eram positivos. Todos os modelos apresentam 373 resultados positivos reais.

2) Previsão com balanceamento de amostra (SMOTE)

Examinando os resultados da previsão com balanceamento de amostra, observando como os modelos funcionam agora com ajuste nas proporções das classes. Vimos como esse método pode impactar a precisão, o recall e outras métricas de avaliação

relativamente à previsão sem balanceamento de amostra. Vamos examinar esse caso e compreender suas implicações.

a) Resultado das classificações

Os resultados mostram que os três modelos (SVM, Árvore de Decisão e Floresta Aleatória) funcionam muito bem. A acurácia do SVM foi de 98%, a precisão de 97% e a taxa de recall de 100%. A Floresta Aleatória e a Árvore de Decisão alcançaram 100% em todas as métricas.

b) Matriz de confusão

- Verdadeiros Negativos (VN): Esses são os casos em que o modelo previu a classe negativa (1) corretamente e os dados reais também eram negativos. O SVM tem 189 verdadeiros negativos, enquanto a Árvore de Decisão e a Floresta Aleatória têm 198.
- Falsos Positivos (FP): Esses são os casos em que o modelo previu incorretamente uma classe positiva (9), mas os dados reais eram negativos (0). Em contraste com a Árvore de Decisão e a Floresta Aleatória, não há falsos positivos na SVM (FP = 0).
- Falsos Negativos (FN): Esses são os casos em que o modelo previu a classe negativa (0) incorretamente, mas os dados reais eram positivos (0). Nesse caso, não temos falsos negativos (FN = 0) em nenhum dos modelos.
- Verdadeiros Positivos (VP): Esses são os casos em que o modelo previu corretamente a classe positiva (0) e os dados reais também eram positivos. Todos os modelos apresentam 372 resultados positivos reais.

4.4 Resultado com validação cruzada

Após a análise, ficamos duvidosos relativamente a certos resultados apresentados pelos modelos. Para tentar validar novamente a nossa hipótese, tentamos usar um método alternativo de validação cruzada. Para treino e teste, os dados são divididos aleatoriamente. Isso significa que os dados usados para teste podem agora ser usados para treino e vice-versa. A razão pela qual essa técnica de validação cruzada é chamada de *K-Fold* é porque K representa o número de subdivisões (iguais) que criamos. Por exemplo, no nosso caso, K = 5; *Fold* significa que cada uno dos blocos de cada K É interessante comparar os desempenhos de vários modelos e determinar cual é o mais eficaz.

O resultado foi alterado um pouco mais. No caso de dados não balanceados, o SVM teve uma taxa de 93%, uma árvore de decisão de 95% e uma Floresta Aleatória de 96%, enquanto no caso de dados balanceados, o SVM manteve uma taxa de 93% e árvore de decisão de 95%, quanto a Floresta Aleatória foi de 95%. Esses resultados nos permitem ter uma ideia do nosso resultado. Esses valores mostram que os modelos são excecionalmente precisos na classificação de dados e podem identificar casos positivos.

4.5 Considerações

O abandono escolar dos alunos é um dos desafios mais comuns enfrentados pelas instituições de ensino em todo o mundo. É prevalente em todos os níveis dos sistemas de ensino público e privado, levando a várias consequências [42].

Agrupamos os dados com a variável status, que continha os valores de curso, conclusão e abandono, para atingir o nosso objetivo e confirmar a hipótese. Após a obtenção desses resultados, decidimos fazer um teste diferente excluindo ou isolando a variável *status*. Para experimentar os resultados de diferentes maneiras, após essa experiência, conseguimos obter novos resultados, os quais apresentam as diferenças.

Um problema de agrupamento e classificação binária é a previsão do abandono escolar do aluno. Estes modelos são construídos com dados centrados no aluno, que incluem a sua demografia e histórico académico. A aprendizagem adaptativa é essencial para o sucesso dos alunos, então esta solução funciona na educação. As instituições de ensino podem apoiar os professores após considerar os avanços de cada aluno. Eles também podem adaptar os seus métodos para garantir que todos os alunos tenham a melhor experiência de aprendizagem possível.

A precisão e a robustez dos modelos Random Forest e Decision Tree são notáveis, com o Random Forest geralmente apresentando o melhor desempenho. A precisão de classificação para todos os modelos parece melhorar com o balanceamento dos dados, com erros de classificação diminuindo principalmente para as classes menos representadas.

As eficácias desses resultados podem prover de diferentes maneiras mais uma dela pode ser que as amostras são poucas, outra hipótese é que pode ser adição da variável oriundo do agrupamento influenciou os resultados nessa previsão.

4.5.1 Resultado com algoritmo de aglomeração

Esta tabela 8, apresenta os resultados de vários métodos de validação para previsão de aglomeração usando o algoritmo Kprototypes.

Tabela 8: Resumo dos resultados com modelo não supervisionado

Categoria	Método/Técnica	Valor
Previsão de agrupamento com Kprototypes	Método do Cotovelo	K = 5
	Método do Score Silhouette	K = 5
	Método de Davies_bouldin	K = 4, K = 5
	Média	5
	Frequência (Ideal)	5

4.5.2 Resultado com algoritmo de classificação

a) Dados Não Balanceados

As previsões de um modelo para dados não balanceados para três classes (Classe 0, Classe 1 e Classe 2) são apresentadas nas tabelas 9, 10 e 11. O número de instâncias previstas para cada classe está escrito em cada célula.

- SVM

Cross-validation score: 0.939827124325323

Tabela 9: Matriz de Confusão de SVM com dados não balanceado

	Predito Classe 0	Predito Classe 1	Predito Classe 2
Classe 0	209	28	1
Classe 1	0	556	2
Classe 2	0	2	509

- Decision Tree

Cross-validation score: 0.950849466196867

Tabela 10: Matriz de Confusa de DT com dados não balanceado

	Predito Classe 0	Predito Classe 1	Predito Classe 2
Classe 0	237	0	1
Classe 1	0	556	2
Classe 2	0	3	508

- Random Forest

Cross-validation score: 0.9607231745780382

Tabela 11: Matriz de Confusão de RF com dados não balanceado

	Predito Classe 0	Predito Classe 1	Predito Classe 2
Classe 0	236	1	1
Classe 1	0	558	0
Classe 2	0	1	510

b) Dados Balanceados

As previsões de um modelo para dados balanceados para três classes (Classe 0, Classe 1 e Classe 2) são apresentadas nas tabelas 12, 13 e 14. O número de instâncias previstas para cada classe está escrito em cada célula.

- SVM

Cross-validation score: 0.939827124325323

Tabela 12: Matriz de Confusão de SVM com Balanceado

	Predito Classe 0	Predito Classe 1	Predito Classe 2
Classe 0	605	2	1
Classe 1	12	551	0
Classe 2	1	2	536

- Decision Tree

Cross-validation score: 0.9519975718225846

Tabela 13: Matriz de Confusão de DT com dados Balanceado

	Predito Classe 0	Predito Classe 1	Predito Classe 2
Classe 0	607	0	1
Classe 1	0	563	0

Classe 2	0	1	538

- Random Forest

Cross-validation score: 0.9568196154505986

Tabela 14: Matriz de Confusão de RF Balanceado

	Predito Classe 0	Predito Classe 1	Predito Classe 2
Classe 0	607	0	1
Classe 1	0	563	0
Classe 2	0	0	539

c) Análise das matrizes de confusão para dados balanceados e não balanceados

Na tabela 15 e 16, apresentam os resultados de três modelos de classificação: SVM, DT e RF, os nomes das classes são 0, 1 e 2. O desempenho desses modelos foram analisados em duas situações diferentes: dados balanceados e não balanceados.

Tabela 15: Analise da experiência com matriz de confusão

Modelo / Classe	**Classe 0**	**Classe 1**	**Classe 2**
Dados Não Balanceados			
SVM	209 (0)	556 (30)	509 (3)
Decision Tree	237 (0)	556 (3)	508 (3)
Random Forest	236 (0)	558 (2)	510 (1)
Dados Balanceados			
SVM	605 (13)	551 (4)	536 (1)
Decision Tree	607 (0)	563 (1)	538 (1)
Random Forest	607 (0)	563 (0)	539 (1)

Observação: valor entre parêntesis são valor mal classificado pelos modelos

Tabela 16: Resultado da experiência com matriz de confusão

Modelo	Dados não balanceados	Dados Balanceados
SVM	Um bom desempenho na classe 0, dificuldade nas classes 1.	Acurácia geral boa, mas com dificuldade na classe 1.
DT	Um bom desempenho em todas as classes	Alto desempenho nas classes.
RF	Bom desempenho geral.	Melhor desempenho geral.

Com base nos resultados, o modelo de floresta aleatória teve a melhor matriz de confusão e a melhor pontuação de validação cruzada. Como resultado, o modelo de floresta aleatória é o mais adequado para classificar os dados.

Capítulo 5

5. Discussão

Com base nos resultados estatísticos obtidos, os modelos de aprendizagem automática, incluindo *Kprototypes* que permitiu-nos agrupar os alunos em relação as suas semelhanças e para prever o abandono foi utilizado algoritmo de classificação SVM, Árvore de Decisão e Floresta Aleatória, demonstraram alta precisão e eficácia na classificação de alunos em risco de abandono escolar. Ao utilizar amostras balanceadas, os modelos apresentaram acurácia de 93% (SVM) e 95% (Árvore de Decisão e Floresta Aleatória), juntamente com altas taxas de precisão e recall. Além disso, a análise das características dos alunos revelou insights sobre a distribuição de pais com diferentes níveis de educação, períodos de estudo, histórico de reprovações e taxas de abandono. A preparação dos dados envolveu a coleta, combinação e organização de registos acadêmicos, permitindo a aplicação de técnicas estatísticas e algoritmos de aprendizagem automática para identificar tendências relacionadas ao abandono escolar.

A utilização de métricas como precisão, recall e F1-score revelou a eficácia dos modelos SVM, Árvore de Decisão e Floresta Aleatória. Além disso, a aplicação do algoritmo SMOTE para lidar com dados desbalanceados permitiu melhorar a capacidade dos modelos em prever o abandono escolar. A padronização dos dados também foi crucial para garantir a Interpretabilidade dos resultados e a robustez dos algoritmos baseados em distância, como o Kprototype e o SVM. Os insights obtidos proporcionaram uma compreensão mais profunda dos fatores que contribuem para o abandono escolar e a reprovação, possibilitando a implementação de intervenções e políticas educacionais direcionadas para identificar e apoiar os alunos mais vulneráveis.

A acurácia da predição variou entre 93 e 100%, indicando que, independentemente do modelo utilizado, as características escolhidas neste estudo mostraram-se bem-sucedidas na predição do sucesso ou abandono dos alunos, apesar da limitação do conjunto de dados e do número de características.

Por outro lado, todos os algoritmos alcançaram valores de recordação superiores a 0,93 e simultaneamente uma precisão muito boa. No entanto, os algoritmos de classificação podem apresentar resultado talvez diferente visto esses resultados podem vir por causa de poucos dados, mas podemos ter a garantia depois de usar varias técnicas para verificar

esses resultados além do uso de k com media iterando 20 vezes dentro de k de 2 a 40, apos isto, começamos por agrupar os alunos usando a variável 'Status' alcançar uma precisão muito alta prevendo apenas a classe majoritária depois optamos por utilizar outra forma isolar a variável 'Status' para verificar os resultados de agrupamos com essa duas formas de agrupamento fizemos a previsão com os modelos acima mencionados. Com essas duas técnicas não mudou tanto em relação a classificação usando essas duas experimentações, na primeira com a variável depois prever com a classificação vimos que acurácia foi de 96% para SVM e 100% para outros dois algoritmos enquanto na segunda experimentação subiu de 96% para 98%, mas outro algoritmo não mudou mantendo 100%.

O estudo examinou como três modelos de aprendizado de máquina (SVM, Decision Tree e Random Forest) funcionaram com conjuntos de dados balanceados e não balanceados utilizando validação cruzada. A SVM apresentou um score de validação cruzada de 0,939 e uma matriz de confusão de alta precisão, mas com alguns falsos positivos de Classe 0 com dados não balanceados. Com um score de 0.950 e poucos erros de classificação, a Decision Tree apresentou um desempenho ligeiramente melhor. Ao mesmo tempo, o Random Forest foi o mais eficaz com uma pontuação de 0,960 e excelente precisão e retenção.

Todos os modelos demonstraram melhorias para os dados balanceados. A SVM manteve os pontos, mas melhorou nas classes 1 e 2. Com um score de 0.951, o Decision Tree alcançou uma acurácia quase perfeita. O Random Forest se destacou em precisão e consistência com um score de 0,956. Enquanto as SVMs são boas, mas são mais propensas a cometer erros em classes menos representadas, os modelos Random Forest e Decision Tree são mais robustos e precisos, especialmente com dados balanceados.

No entanto, o estudo de caso apresentado e os seus resultados têm algumas limitações. Como já foi mencionado, o tamanho limitado do conjunto de dados é o primeiro. Em contraste com muitos outros domínios de aplicação de modelos de aprendizagem automática, a quantidade de dados no domínio educacional não pode ser facilmente aumentada através da combinação de diferentes recursos.

A última fraqueza do estudo de caso é uma forma de intervenção adequada, que não é discutida em detalhe no caso de dados fora da instituição que são as dimensões sociais, culturais e económicas regionais, num contexto global. Ao reconhecer a interconexão entre o sucesso individual dos alunos e o desenvolvimento holístico das comunidades.

Este estudo já mostrou que o insucesso do aluno pode ser previsto com base nas categorias de atividades selecionadas. Por outro lado, a razão para este estado pouco lisonjeiro permanece em aberto e requer mais investigação. Seria interessante examinar se outras categorias de atividades deveriam ter um impacto semelhante no envolvimento dos alunos e para que categorias de atividades podem ser trocadas na fase de intervenção.

Capítulo 5

6. Conclusão

A dissertação atual se concentra na aplicação de métodos de aprendizagem automática para prever e prevenir o abandono escolar. Ao longo de sete anos letivos, o estudo utilizou algoritmos de aprendizagem automática, incluindo Kprototypes para aglomeração, Random Forest, Support Vector Machines (SVM) e Decision Tree para classificação. Os dados foram coletados de vários cursos. A pesquisa examinou e previu o risco de abandono escolar entre os alunos usando métodos de aprendizagem automática. As principais etapas metodológicas incluíram a coleta e organização de dados acadêmicos. O tratamento de dados ausentes, a normalização e a transformação de variáveis categóricas foram partes da preparação dos dados.

Kprototypes: são utilizados para agrupamento, identificando grupos de alunos com padrões semelhantes por meio da combinação de informações numéricas e categóricas. Decision Tree, Random Forest e SVM: usados para predição e classificação do risco de abandono escolar. Métricas de Desempenho: Para avaliar a eficácia dos modelos preditivos, foram utilizadas métricas como precisão, recall e pontuações F1. Tratamento de Dados Desbalanceados: A melhoria da capacidade preditiva dos modelos e o tratamento do Desbalanceamento dos dados dependeram da implementação do algoritmo SMOTE, uma técnica sintética de coleta excessiva de minorias.

Por fim, a análise dos modelos de aprendizado automático utilizados em conjuntos de dados balanceados e não balanceados fornece informações significativas sobre suas vantagens e desvantagens. O modelo Random Forest foi o mais confiável e mostrou precisão e consistência superiores, especialmente quando os dados foram balanceados. Além disso, com dados balanceados, a Decision Tree demonstrou um desempenho excecional. É claro que o balanceamento dos dados é necessário para resultados mais confiáveis; no entanto, a SVM apresentou uma maior probabilidade de erros em classes menos representadas, apesar de ser uma SVM eficaz.

Os resultados confirmaram a capacidade dos modelos escolhidos de prever com precisão o sucesso ou a desistência dos alunos. Mas foram identificadas algumas restrições: Tamanho do Conjunto de Dados: um conjunto de dados limitado pode ter afetado os resultados, indicando que mais dados devem ser coletados para validar e generalizar os achados. Exploração de Métricas de Desempenho: Os resultados devem ser complementados com métricas de desempenho mais amplas e intervenções sociais.

Existem muitos benefícios significativos para a educação que o estudo oferece. Um deles é a previsão do risco de abandono escolar por meio da utilização de métodos de aprendizagem automática para prever esse risco. Análise dos Fatores Contributivos: explora fatores socioeconômicos, individuais e familiares para uma análise abrangente dos fenômenos de abandono escolar e taxas de reprovação. Intervenções Educacionais: Possíveis consequências práticas e métodos preventivos para ajudar os alunos em risco e permitir a implementação de políticas educacionais direcionadas.

No geral, esta dissertação enfatiza o uso de aprendizagem automática e análise de dados para melhorar o desempenho dos alunos, reduzir as taxas de abandono escolar e melhorar os resultados acadêmicos. A investigação enfatiza a necessidade de continuar explorando e ampliando a coleção de dados e as métricas de desempenho para aumentar a eficiência dos modelos preditivos e a implementação de intervenções educacionais eficazes.

Referências Bibliográficas

[1] A. Capitão, "Excesso de reprovações no 'Kimpa Vita' leva estudantes à desistência," *Jornal de Angola*, Apr. 2021.

[2] Ango Emprego, "Reprovações leva à desistência de muitos Estudantes (Universidade Kimpa Vita)," *Ango Emprego*, May 2021.

[3] M. Francisco, "Estudantes da Universidade Kimpa Vita, no Uíge, protestam contra decreto presidencial," *Voa português* , Apr. 2021.

[4] C. M. C. Matias, C. : Doutor, N. De, and A. Alves, "ABANDONO ESCOLAR NO 3º CICLO DO ENSINO SUPERIOR: ESTUDO DE CASO."

[5] ciência.ao, "A Qualidade do Ensino Superior em Angola (Do estado real ao estado desejado)," *ciência.ao*, Jan. 2018.

[6] K. Kalegele, "School dropout profiling and prediction approach using machine learning," *International Journal of Information Technology, Communications and Convergence*, vol. 3, no. 4, p. 245, 2020, doi: 10.1504/IJITCC.2020.10034641.

[7] C. F. Rodríguez-Hernández, E. Cascallar, and E. Kyndt, "Socio-economic status and academic performance in higher education: A systematic review," *Educ Res Rev*, vol. 29, p. 100305, Feb. 2020, doi: 10.1016/j.edurev.2019.100305.

[8] Y. D. Li, G. H. Ding, and C. Y. Zhang, "Effects of learner-centred education on academic achievement: a meta-analysis," *Educ Stud*, 2021, doi: 10.1080/03055698.2021.1940874.

[9] N. Iam-On and T. Boongoen, "Generating descriptive model for student dropout: a review of clustering approach," *Human-centric Computing and Information Sciences*, vol. 7, no. 1. Springer Berlin Heidelberg, Dec. 01, 2017. doi: 10.1186/s13673-016-0083-0.

[10] S. Sood and M. Saini, "Hybridization of cluster-based LDA and ANN for student performance prediction and comments evaluation," *Educ Inf Technol (Dordr)*, vol. 26, no. 3, pp. 2863–2878, May 2021, doi: 10.1007/s10639-020-10381-3.

[11] N. Mduma, K. Kalegele, and D. Machuve, "Machine learning approach for reducing students dropout rates," *International Journal of Advanced Computer Research*, vol. 9, no. 42, pp. 156–169, May 2019, doi: 10.19101/ijacr.2018.839045.

[12] S. Kim, E. Choi, Y.-K. Jun, and S. Lee, "Student Dropout Prediction for University with High Precision and Recall," *Applied Sciences*, vol. 13, no. 10, p. 6275, May 2023, doi: 10.3390/app13106275.

[13] J. Liang, X. Zhao, D. Li, F. Cao, and C. Dang, "Determining the number of clusters using information entropy for mixed data," *Pattern Recognit*, vol. 45, no. 6, pp. 2251–2265, Jun. 2012, doi: 10.1016/j.patcog.2011.12.017.

[14] G. Szepannek, "clustMixType: User-Friendly Clustering of Mixed-Type Data in R," *R J*, vol. 10, no. 2, pp. 200–208, 2018, doi: 10.32614/RJ-2018-048.

[15] K. Kalaivani and A. P. V. Raghavendra, "An integrated clustering approach for high dimensional categorical data," in *2013 International Conference on Green High Performance Computing (ICGHPC)*, IEEE, Mar. 2013, pp. 1–4. doi: 10.1109/ICGHPC.2013.6533920.

[16] J. Kabathova and M. Drlik, "Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques," *Applied Sciences*, vol. 11, no. 7, p. 3130, Apr. 2021, doi: 10.3390/app11073130.

[17] L. C. Sorensen, "'Big Data' in Educational Administration: An Application for Predicting School Dropout Risk," *Educational Administration Quarterly*, vol. 55, no. 3, pp. 404–446, Aug. 2019, doi: 10.1177/0013161X18799439.

[18] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting Student Dropout and Academic Success," *Data (Basel)*, vol. 7, no. 11, p. 146, Oct. 2022, doi: 10.3390/data7110146.

[19] A. Behr, M. Giese, H. D. Teguim K, and K. Theune, "Early Prediction of University Dropouts – A Random Forest Approach," *Jahrb Natl Okon Stat*, vol. 240, no. 6, pp. 743–789, Oct. 2020, doi: 10.1515/jbnst-2019-0006.

[20] A. Behr, M. Giese, H. D. Teguim K, and K. Theune, "Early Prediction of University Dropouts – A Random Forest Approach," *Jahrb Natl Okon Stat*, vol. 240, no. 6, pp. 743–789, Oct. 2020, doi: 10.1515/jbnst-2019-0006.

[21] O. K. Oyedotun, S. N. Tackie, E. O. Olaniyi, and A. Khashman, "Data Mining of Students' Performance: Turkish Students as a Case Study," *International Journal of Intelligent Systems and Applications*, vol. 7, no. 9, pp. 20–27, Sep. 2015, doi: 10.5815/ijisa.2015.09.03.

[22] A. B. F. Mansur and N. Yusof, "The Latent of Student Learning Analytic with K-mean Clustering for Student Behaviour Classification," *Journal of Information Systems Engineering and Business Intelligence*, vol. 4, no. 2, p. 156, Oct. 2018, doi: 10.20473/jisebi.4.2.156-161.

[23] A. Majeed and S. O. Hwang, "Quantifying the Vulnerability of Attributes for Effective Privacy Preservation Using Machine Learning," *IEEE Access*, vol. 11, pp. 4400–4411, 2023, doi: 10.1109/ACCESS.2023.3235016.

[24] Develophard.com, “O que é o Machine Learning?” Accessed: Mar. 22, 2024. [Online]. Available: https://www.develophard.com/pt/glossary/machine-learning

[25] M. Vinícius Alves de Araújo and D. Karina Yuriko Yaginuma, “Universidade Federal Fluminense.”

[26] Z. Huang, “Clustering large data sets with mixed numeric and categorical values*,” 1997.

[27] Z. Huang, “Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values,” 1998.

[28] R. Brnawy and N. Shiri, “K-mixed prototypes,” in *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, New York, NY, USA: ACM, Apr. 2019, pp. 542–545. doi: 10.1145/3297280.3297549.

[29] M. R. Anderberg, *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*, vol. 19. Academic press, 2014.

[30] M. Anderberg, *Cluster Analysis for Applications*, AcademicPressl. Elsevier, 1973. doi: 10.1016/C2013-0-06161-0.

[31] S. A. Obaidullah, S. Rato, and L. G. Teresa, “RMID: a novel and efficient image descriptor for mammogram mass classification,” 2018, Accessed: Mar. 14, 2024. [Online]. Available: http://hdl.handle.net/10174/25062

[32] B. Akcesme, “Southeast Europe Journal of Soft Computing Support Vector Machines for Predicting Protein Structural Classes Images Derived From Amino Acid Sequences.” [Online]. Available: http://scjournal.ius.edu.ba

[33] L. Yuan, H. Chen, and J. Gong, “Classifications Based Decision Tree and Random Forests for Fanjing Mountains’ Tea,” in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Aug. 2018. doi: 10.1088/1757-899X/394/5/052002.

[34] S. Islam and S. H. Amin, “Prediction of probable backorder scenarios in the supply chain using Distributed Random Forest and Gradient Boosting Machine learning techniques,” *J Big Data*, vol. 7, no. 1, Dec. 2020, doi: 10.1186/s40537-020-00345-2.

[35] A. Belmokre, M. K. Mihoubi, and D. Santillán, “Analysis of Dam Behavior by Statistical Models: Application of the Random Forest Approach,” *KSCE Journal of Civil Engineering*, vol. 23, no. 11, pp. 4800–4811, Nov. 2019, doi: 10.1007/s12205-019-0339-0.

[36] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” Jun. 2011, doi: 10.1613/jair.953.

[37] C. V. Giordano and R. Antonio De Souza, “Prognóstico da evasão escolar em instituição de educação profissional e tecnológica por meio da inteligência artificial,” 2023.

[38] A. Carlos. Gil, *Como elaborar projetos de pesquisa*. Atlas, 1991.

[39] L. Cristina *et al.*, “Aplicação de técnicas de aprendizado de máquina em um contexto académico com foco na identificação dos alunos evadidos e não evadidos,” 2020.

[40] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” Jan. 2012, [Online]. Available: http://arxiv.org/abs/1201.0490

[41] K. Faceli, A. C. Lorena, J. Gama, and A. C. P. de L. F. de Carvalho, *Inteligência artificial: uma abordagem de aprendizado de máquina*. LTC, 2011.

[42] learn.microsoft.com, “Predict student attrition,” *Microsoft*. 2024. [Online]. Available: https://learn.microsoft.com/pt-pt/azure/architecture/example-scenario/ai/student-attrition-prediction

Referências para conjuntos de dados

DS1- Realinho,Valentim, Vieira Martins,Mónica, Machado,Jorge, and Baptista,Luís. (2021). Predict students' dropout and academic success. UCI Machine Learning Repository. https://doi.org/10.24432/C5MC89.

DS2- Alvarado-Uribe, J.; Mejía-Almada, P.; Masetto Herrera, A.L.; Molontay, R.; Hilliger, I.; Hegde, V.; Montemayor Gallegos, J.E.; Ramírez Díaz, R.A.; Ceballos, H.G. Student Dataset from Tecnologico de Monterrey in Mexico to Predict Dropout in Higher Education. Data 2022, 7, 119. https://doi.org/10.3390/data7090119 doi: 10.3390/data7090119

DS3- Sakshi Sood & Munish Saini (2021). Dados de teste gerados pelo Mockaroo [Conjunto de dados]. Mockaroo. Recuperado em 2020, de https://mockaroo.com/

DS4- PRIYADHARSHINI, C. (2017). MOOC Dataset. Conjunto de dados MOOC. Licenciado sob a licença CC0: Domínio Público. Disponível em: Conjunto de dados MOOC | Kaggle. Acesso em: 2021.

DS5- Yilmaz,Nevriye and Şekeroğlu,Boran. (2023). Higher Education Students Performance Evaluation. UCI Machine Learning Repository. https://doi.org/10.24432/C51G82.

DS6- Niti Witthayawiroj, July 10, 2023, "Rajamangala University of Technology Thanyaburi Dropout Dataset (RDD)", IEEE Dataport, doi: https://dx.doi.org/10.21227/4sgg-2274.

DS7- Realinho,Valentim, Vieira Martins,Mónica, Machado,Jorge, and Baptista,Luís. (2021). Predict students' dropout and academic success. UCI Machine Learning Repository. https://doi.org/10.24432/C5MC89.

DS8- Cortez,Paulo. (2014). Student Performance. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T.

DS9- Behr, Andreas; Giese, Marco; Teguim Kamdjou, Herve Donald; Theune, Katja (2019): Early prediction of university dropouts - a random forest approach. Version: 1. Journal of Economics and Statistics. Dataset. http://dx.doi.org/10.15456/jbnst.2019333.185049

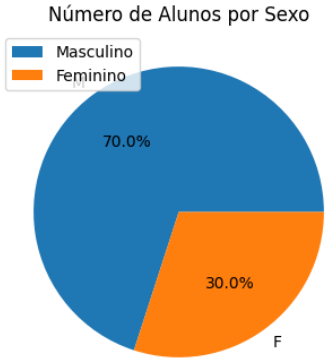
DS10- Michael Stein & Michael Leitner & Jill C. Trepanier & Kory Konsoer, 2022. "A Dataset of Dropout Rates and Other School-Level Variables in Louisiana Public High Schools," Data, MDPI, vol. 7(4), pages 1-10, April.

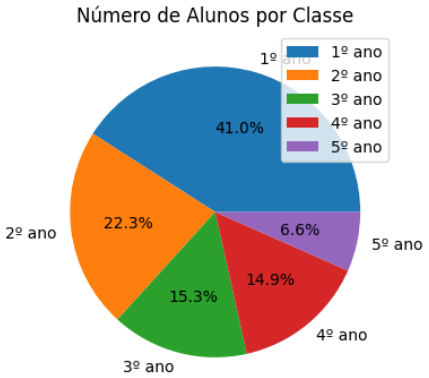
Análise Descritiva

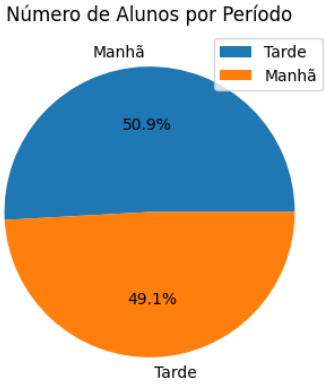
A demografia dos alunos de uma instituição é analisada por quatro gráficos de pizza, como mostrado em anexo 1.1. Observa-se que setenta por cento dos alunos são homens e trinta por cento são mulheres. A maioria das pessoas está no 1o ano (41%), seguida pelo 2o ano (22,3%) e 3o ano (15,3%). As taxas mais baixas são no 4o ano (14,9%) e 5o ano (6,6%). Com 50,9% estudando de manhã e 49,1% estudando à tarde, a distribuição é quase equilibrada. Os alunos são principalmente solteiros, 98,8% são solteiros, 0,6% casados e 0,6% divorciados. O gráfico de barras horizontais mostra a quantidade de alunos por curso em uma instituição; Enfermagem tem o maior número de alunos (1149), seguido por Informática (1127) e Contabilidade e Gestão (1061). O departamento de Agronomia tem 650 alunos, enquanto o departamento de Hidráulica e Saneamento das Águas tem o menor número de alunos, com 367. De acordo com esses dados, as pessoas preferem se inscrever em cursos de enfermagem e informática do que em qualquer outro curso.
O anexo 1.2 mostra um gráfico de linha que mostra o número de alunos por ano académico de 2016 a 2022. A variação é notável ao longo dos anos. Em 2016, havia cerca de 500 alunos; então caiu para cerca de 400 em 2017 e atingiu seu ponto mais baixo em 2018, com cerca de 250 alunos. O número de alunos subiu para cerca de 1500 em 2019, um ponto alto. Após esse ponto alto, o número de alunos voltou a cair; em 2020, chegou a cerca de 500, mas caiu ainda mais para cerca de 250 em 2021. A quantidade de alunos aumentou novamente para aproximadamente 750 em 2022.

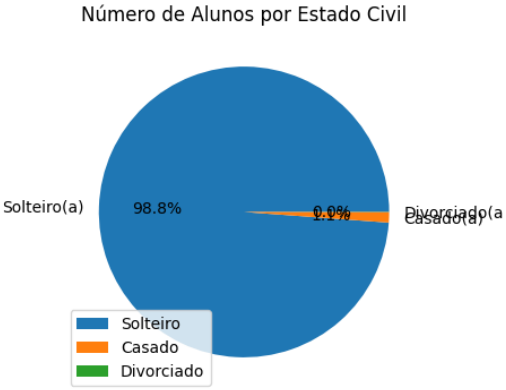
O curso de Contabilidade e Gestão apresenta uma taxa de rejeição superior face a outros cursos, seguido da Informática. Por outro lado, a Informática apresenta a maior taxa de abandono de estudantes, seguida da Agronomia (o curso domina em termo de proporção de abandono em relação aos outros cursos), Contabilidade e Gestão, e Enfermagem. O grande número de alunos em Enfermagem e Informática deve-se às maiores oportunidades de emprego no estado, uma vez que o Ministério da Saúde contrata mais funcionários do que o Ministério da Educação. O menor número de alunos na Agronomia é resultado da perda anual de estudantes na Enfermagem, tornando o campo menos atrativo e menos valorizado no país.

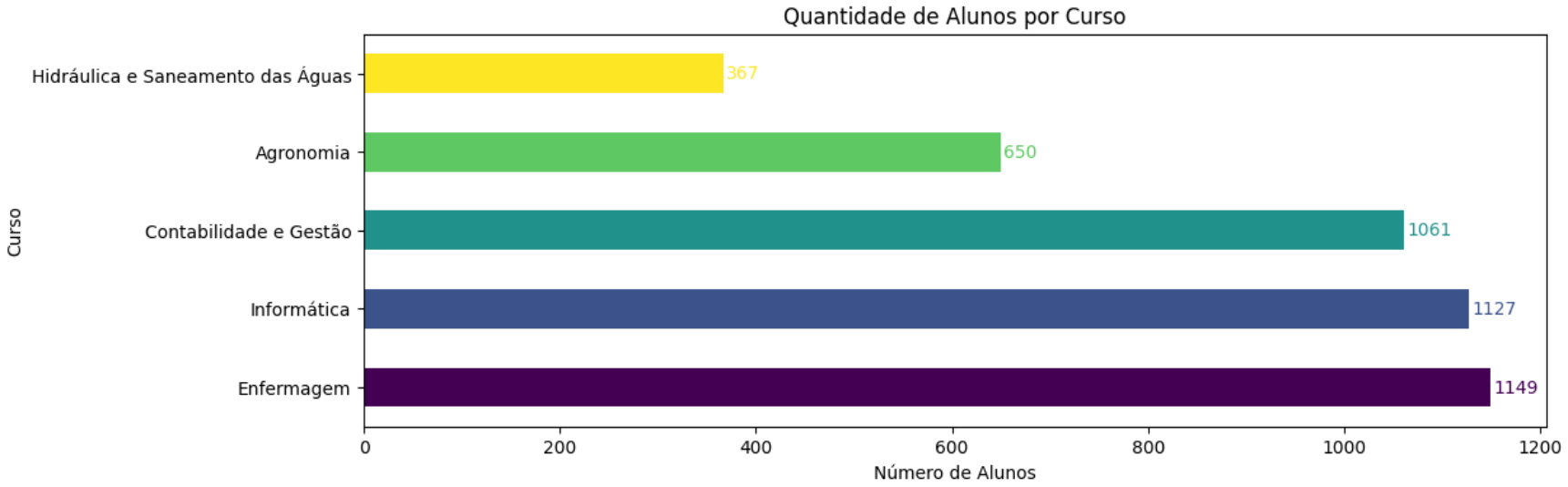
ANEXO 1.1 – A demografia dos alunos de uma instituição



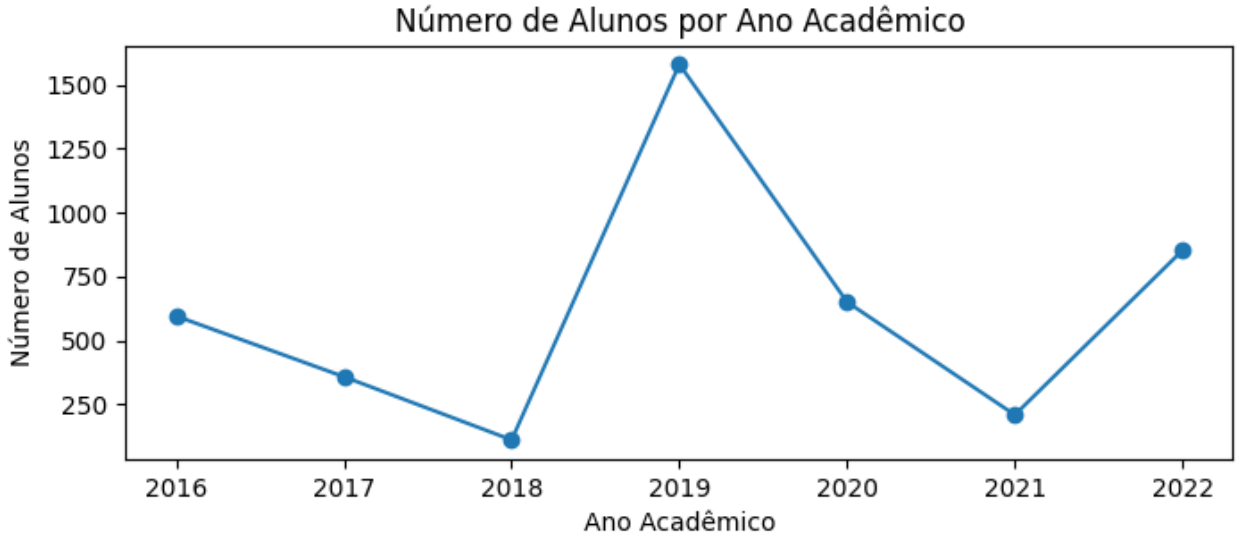


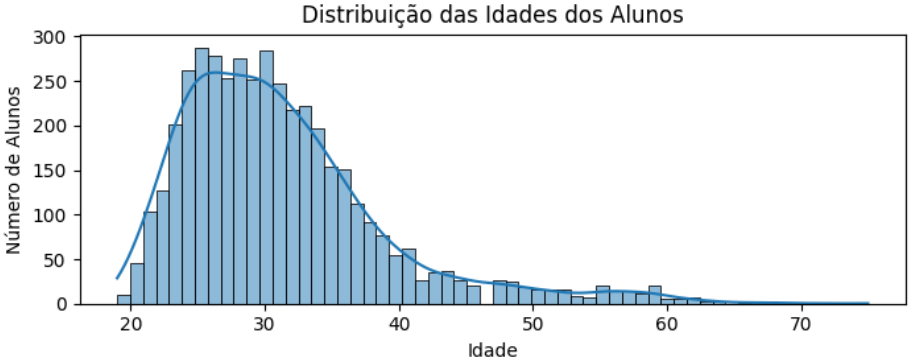


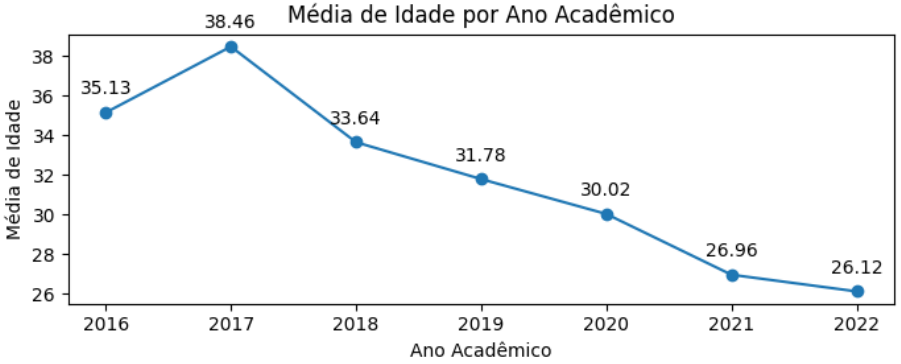




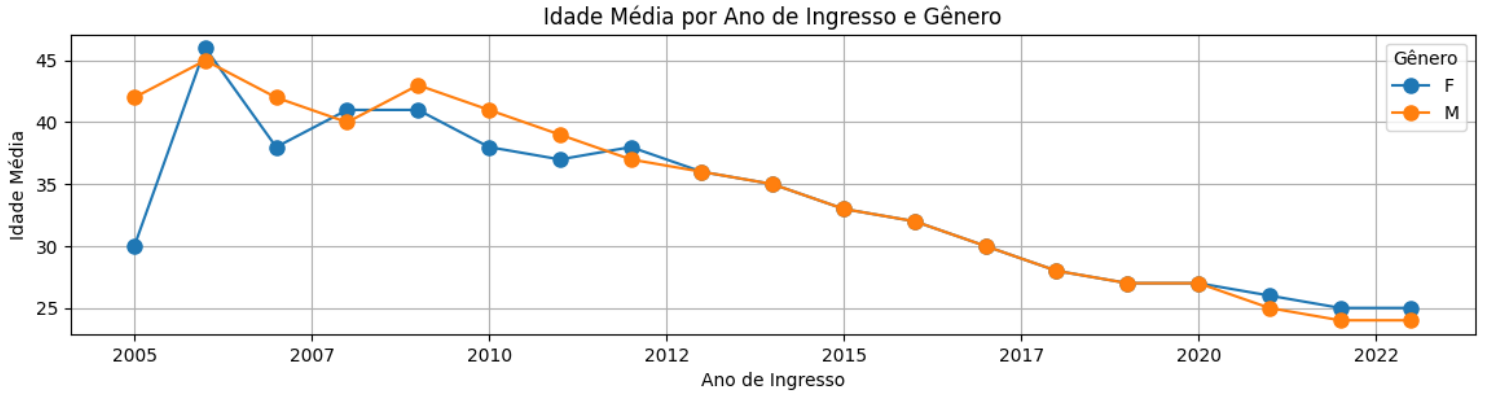
ANEXO 1.2 – Número e media de aluno por Ano académico

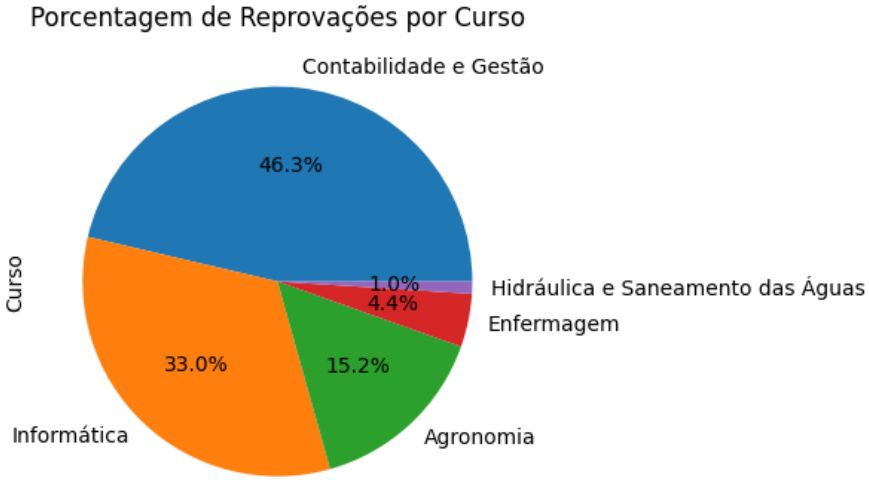


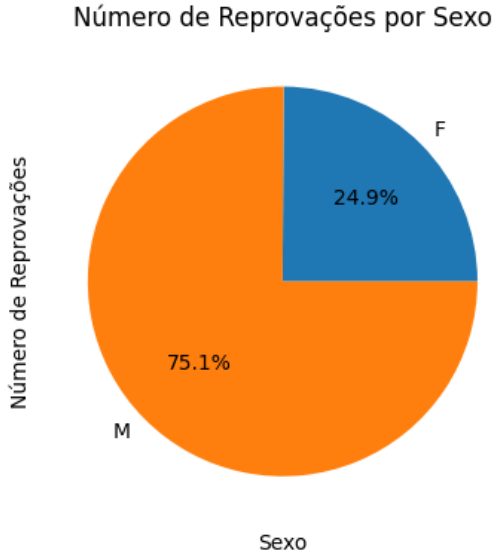




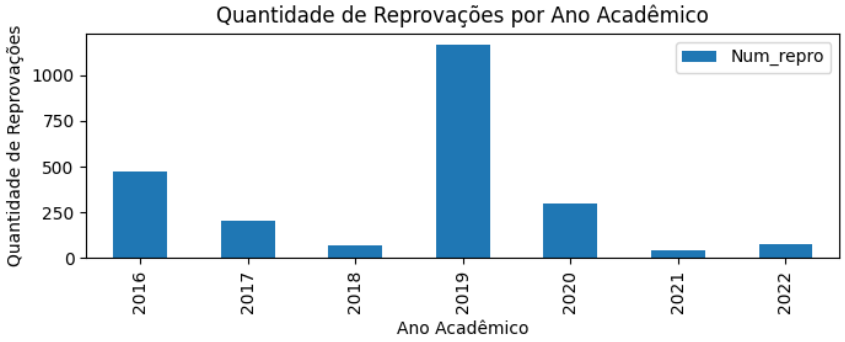
ANEXO 1.3 – Idade media por ano de Ingresso e Taxa de reprovações por sexo

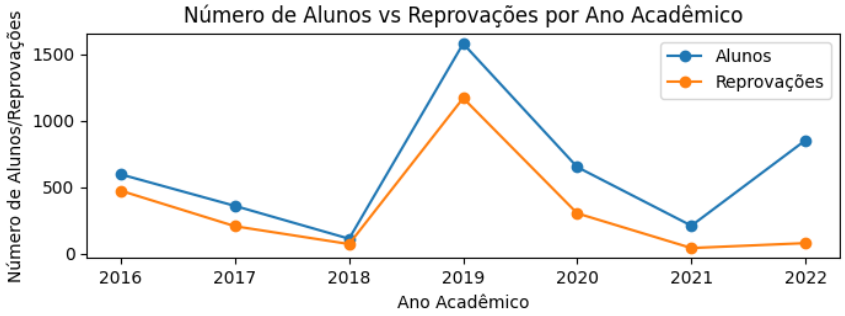


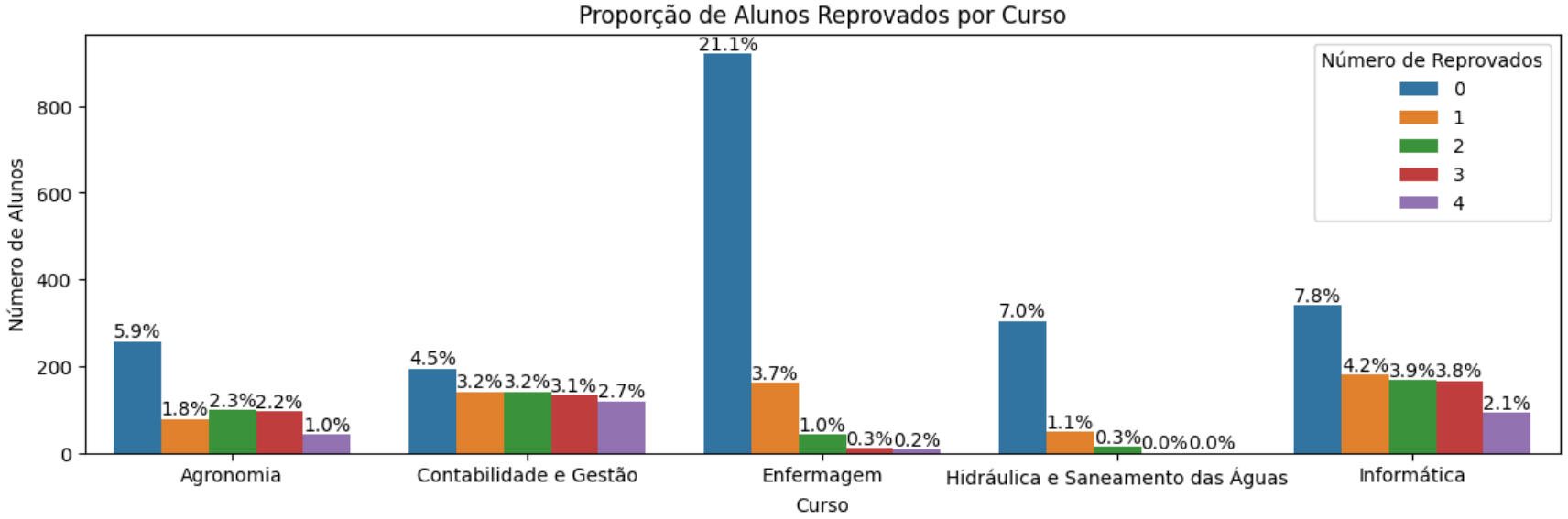




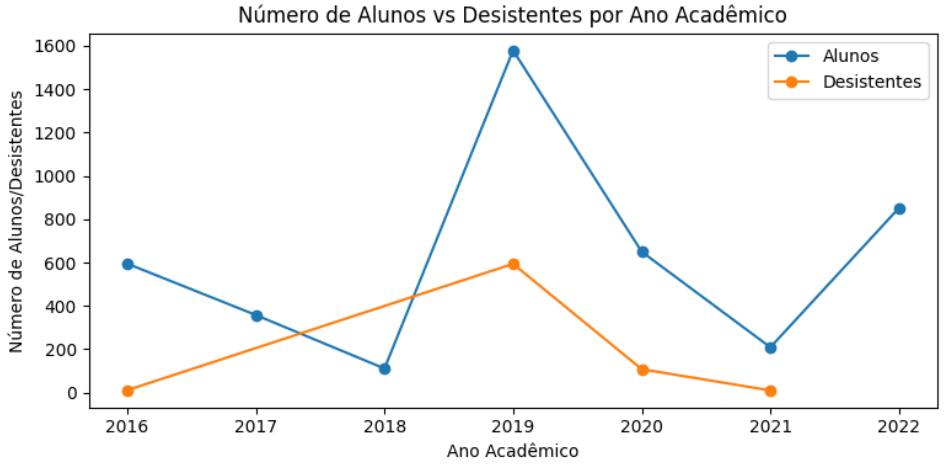
ANEXO 1.4 – Quantidade de reprovados

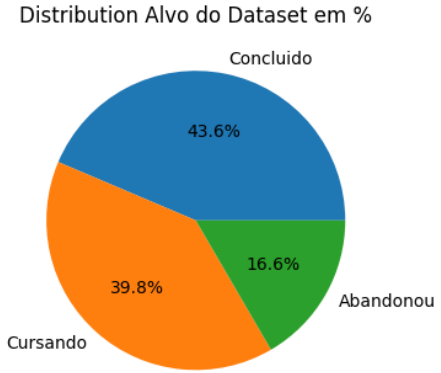


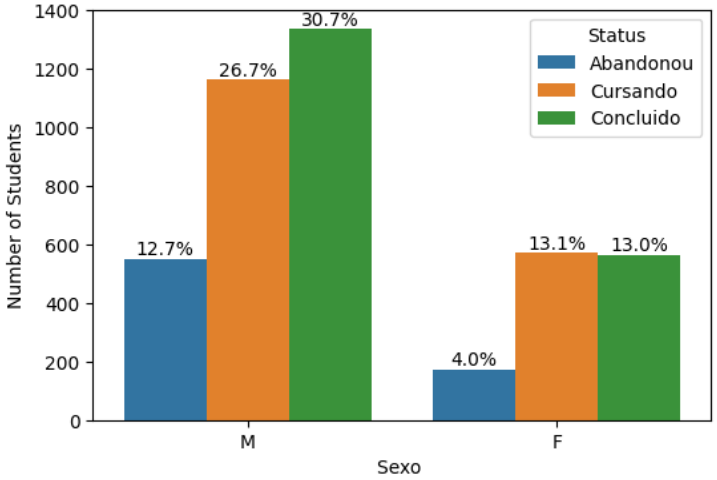


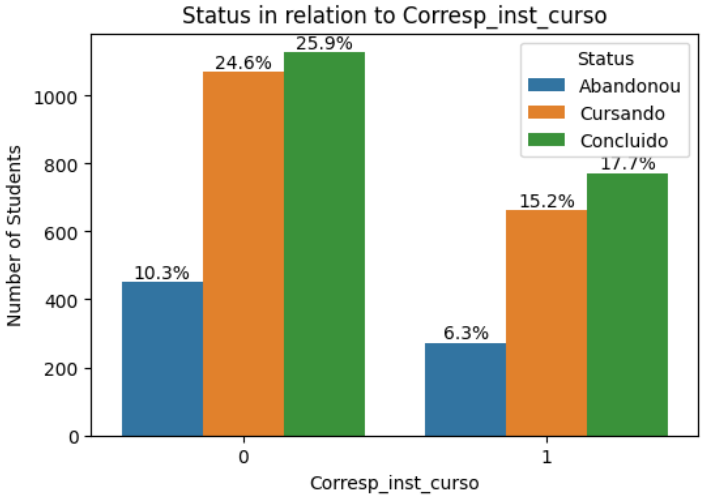


ANEXO 1.5 – Quantidade por situação









ANEXO 1.6 – Status por curso