CENTRO UNIVERSITÁRIO DE ANÁPOLIS – UniEVANGÉLICA
BACHARELADO EM ENGENHARIA DE SOFTWARE

DENIS RODRIGUES DE FARIA
JOSÉ VICTOR ROCHA SILVESTRE
PEDRO MOREIRA DE OLIVEIRA NETO
VICTOR ELIAS PALASIOS SILVA

A UTILIZAÇÃO DE APRENDIZADO DE MÁQUINA SUPERVISIONADO PARA PREDIÇÃO DE EVASÃO NO ENSINO SUPERIOR

ANÁPOLIS
2022

CENTRO UNIVERSITÁRIO DE ANÁPOLIS – UniEVANGÉLICA
BACHARELADO EM ENGENHARIA DE SOFTWARE

DENIS RODRIGUES DE FARIA
JOSÉ VICTOR ROCHA SILVESTRE
PEDRO MOREIRA DE OLIVEIRA NETO
VICTOR ELIAS PALASIOS SILVA

A UTILIZAÇÃO DE APRENDIZADO DE MÁQUINA SUPERVISIONADO PARA PREDIÇÃO DE EVASÃO NO ENSINO SUPERIOR

Pesquisa apresentada como requisito para a conclusão da disciplina de Trabalho de Curso 2 do curso de Bacharelado em Engenharia de Software do Centro Universitário de Anápolis – UniEVANGÉLICA.

Orientadora: Prof. Mestre Walquíria Fernandes Marins

Anápolis
2022

RESUMO

O impreterível avanço da ciência e tecnologia e sua inserção nas mais diversas áreas de conhecimento, culminaram na crescente relação de dependência entre o homem e os sistemas. Observa-se, ainda, uma busca incansável por inovações e praticidade para as tarefas das mais diversas áreas da vida, desde a profissional até a pessoal. Inovações que trazem consigo a possibilidade de trabalho remoto, estudo remoto e diversas outras atividades mediadas por tecnologia. Nesse cenário, a educação, mais especificamente, a educação de nível superior enfrenta um problema que assola boa parte das instituições: a evasão dos alunos, seja migrando para outras instituições, a volatilidade do avanço científico e outros motivos diversos que são objeto de outros estudos, como a renda e escolaridade familiar. A fim de evitar ou amenizar esta ocorrência, este trabalho apresenta uma ferramenta baseada em Mineração de Dados (*Data Mining*) e Aprendizado de Máquina (*Machine Learning*) para monitorar evidências de evasão através de relatórios. Para iniciação do projeto foi realizada uma pesquisa bibliográfica nas áreas de educação, evasão e aprendizado de máquina para compreender melhor o escopo do problema e desafios do estado da arte. Os trabalhos relacionados permitiram identificar o tipo de algoritmo e processo que seriam aplicados neste trabalho, o F1-score. Como principal metodologia de desenvolvimento foi utilizado o KDD (Knowledge Discovery in Databases), que possibilitou alcançar resultados satisfatórios a fim de demonstrar com acurácia a possibilidade da evasão nas instituições, bem como seus principais indicadores. Por fim, o planejamento técnico enfocou nos requisitos, construindo um mapa de personas e um *Product Backlog Building* (PBB), com o intuito de delimitar melhor os processos e estruturar o desenvolvimento. Ademais, a Governança de TI evidencia a importância da ferramenta para a tomada de decisões assertivas das áreas de negócio.

Palavras-Chave: Educação; Evasão; Ensino Superior; Gestão de indicadores; Software; Aprendizado de Máquina.

ABSTRACT

The unavoidable advancement of science and technology and its insertion in the most diverse areas of knowledge, culminated in the growing relationship of dependence between man and systems. There is also a tireless search for innovations and practicality for tasks in the most diverse areas of life, from professional to personal. Innovations that bring with them the possibility of remote work, remote study and several other activities mediated by technology. In this scenario, education, more specifically, higher education, faces a problem that plagues most institutions: student dropout, whether migrating to other institutions, the volatility of scientific progress and other diverse reasons that are the subject of other studies. , such as income and family education. In order to avoid or mitigate this occurrence, this work presents a tool based on Data Mining and Machine Learning to monitor evidence of evasion through reports. To initiate the project, a bibliographical research was carried out in the areas of education, evasion and machine learning to better understand the scope of the problem and challenges of the state of the art. Related works allowed identifying the type of algorithm and process that would be applied in this work, the F1-score. KDD (Knowledge Discovery in Databases) was used as the main development methodology, which made it possible to achieve satisfactory results in order to accurately demonstrate the possibility of evasion in institutions, as well as its main indicators. Finally, the technical planning focused on the requirements, building a map of personas and a Product Backlog Building (PBB), in order to better define the processes and structure the development. Furthermore, IT Governance highlights the importance of the tool for making assertive decisions in the business areas.

Keywords: Education; Evasion; University Education; Indicator management; Software; Machine Learning.

SUMÁRIO

1. INTRODUÇÃO

A modernidade e a evolução humana são marcadas pela crescente da relação entre o ser humano e a tecnologia. Em um contexto histórico-evolutivo, o homem moderno, *homo sapiens*, passa por uma convergência entre o processo de conhecimento evolucional pautado por uma complexidade cultural de dependência, com o avanço tecnológico e científico (NASCIMENTO, 2019).

Com esse avanço no ramo da ciência e tecnologia, a relação de dependência do homem o tornou fadado a buscar constantemente uma nova tecnologia, o que aumentou a competição de mercado em diversas áreas. Um desses ramos é o desenvolvimento de software, à qual se induz a ofertar produtos constantes desta evolução (ALBERTIN, 2017).

Concomitantemente, uma das áreas que sofreram impactos gerados por essa constante evolução é a educação, pois ela precisa se adaptar constantemente em termos de tecnologia. Em um dado momento, o ensino tanto fundamental-infantil, quanto o de nível técnico passaram por avanços induzidos por este crescimento tecnológico. Por conta disso, esta área, neste caso, a educação superior, sofre com um problema que assola, ainda, boa parte das instituições, sendo ela, a evasão dos alunos durante o período corrente, por motivos institucionais ou pessoais dos alunos. (SACCARO, 2019).

A partir do exposto é possível identificar que esta mazela traz como consequência fatores que vão além do controle institucional. Dentre eles, podem ser citados a economia externa, as assiduidades, mas também, sem deixar de levar em consideração, o momento em que o processo foi ocorrido (SACCARO, 2019). Como por exemplo, no momento da pandemia do Covid 19 que afetou o ano de 2020, a saída de alunos foi algo incontrolável perante as universidades superiores, as quais sofreram com esse processo, já que foi uma decisão e um período delicado para ambas as partes (NUNES, 2021).

Neste contexto, uma problemática que se potencializa é a gestão dos dados de evasão e potencial evasão. Uma vez que, os meios de conhecimento e gestão de indicadores não preveem uma mudança brusca de comportamento da curva de evasão do ensino. Para que isso possa ser feito, basta que sejam desenvolvidas e aplicadas ferramentas de monitoramento com base em dados e com possibilidade de predição. Desse modo, cabe a indagação: como a análise e mineração de dados, atrelada aos processos de aprendizado de máquina, podem apontar uma possibilidade de evasão?

Por este motivo, é válida a necessidade da geração de um relatório de controle desses dados. Visto que, boa parte dos controles existentes para a evasão demoram para serem gerados, as pesquisas de campos para monitoramento de resultados geram conflitos de tempo, demandando

uma espera não operante para um problema que pode ser resolvido antecipadamente por tecnologia. Dessa forma, é possível entender o mercado e a existência dos mecanismos que geram essas predições e optar pela melhor escolha para a instituição, a qual conseguirá predizer o melhor momento para incitar os discentes e entender o que está acontecendo no limite sala de aula-professor-aluno.

Como dito, *softwares* que entendem e fazem o processamento destes dados existem, como Regressão logística (RL); Máquina de vetores de suporte (MVS); Árvores de decisão (AD) e Floresta aleatória (FA) (MARTINS et al., 2021; GÉRON, 2019). Entretanto, a variação de acurácia entre eles pode não registrar um relatório de evasão tão confiável. O que abre a perspectiva para a criação de uma ferramenta que analise estes dados minerados e gere relatórios que possam determinar com acurácia o entendimento da evasão discente dos alunos.

Neste sentido, visando o controle dos processos e o conhecimento gerado pelos problemas, pôde-se observar os seguintes objetivos específicos: i) avaliar a base de dados; ii) selecionar qual o melhor tipo de banco para ser trabalhado no processo; iii) realizar a mineração dos dados; iv) aplicar a inteligência dos dados; v) escolher os algoritmos para verificação de acurácia vi) comparar assertividade com uma base real de dados; vii) definir um modelo de relatório.

Em um primeiro momento, como base foram usados os dados de um estudo de caso do Instituto Politécnico de Portalegre, Portalegre, Portugal (MARTINS et al., 2021). Para que assim, se lançasse um modelo inicial do processo, com geração de relatórios, que gerencie de modo visível, e com um linguagem computacional de alto nível, os indicadores que influenciam em tal evasão.

Para o processo de criação do *software* que irá implementar os algoritmos de classificação, teve-se como base o estudo bibliográfico sobre alguns pontos específicos para que o processo de entendimento chegasse em um patamar embasado e de cunho verdadeiro. Como principais pontos, foram trabalhados desde o processo do ensino superior e seus períodos de evasão, até o entendimento e caminhar de uma análise de dados em uma base de dados específica, para que, desse modo, a resposta fosse desejável como fundamentação teórica para o assunto.

Sendo assim, na Seção 2 estão sendo mostradas as pesquisas feitas, declarando os melhores e mais importantes temas deduzidos. A Seção 3 mostra a metodologia utilizada em cada fase do trabalho, desde as bibliografias. Em seguida, a Seção 4 demonstra o desenvolvimento para a escolha e produção do algoritmo como fundamentação para aplicação do aprendizado de máquina sobre a base de dados, além de mostrar como a governança de TI é importante no projeto tecnológico da Instituição. Por fim, a Seção 5 aborda os resultados retirados de todo o complexo proposto, a fim de demonstrar as considerações a respeito de todo o desenvolvimento.

2. FUNDAMENTAÇÃO TEÓRICA

O processo de conhecimento de processos é de cunho necessário em um projeto. Para o embasamento do conhecimento e dos processos de entendimento do projeto foram feitas pesquisas em áreas correlatas a metodologia de atribuição de conhecimento durante a construção. Primeiramente, trata-se de um projeto voltado à educação superior e os seus processos de evasão. Nesse sentido, o ponto a ser correlato em uma abordagem primária são os seus meios de funcionamento, até chegar em um processo mais técnico, onde estuda-se o processo de mineração e como funcionam seus algoritmos, e, para uma finalidade de conhecimento, os trabalhos relacionados mostra o conhecimento adquirido durante toda a desenvoltura do *software*.

2.1 Educação superior

O Brasil desde seu período de descoberta foi uma colônia de exploração, e até então seria usado apenas para fins benéficos à coroa portuguesa, entretanto, durante o Congresso de Viena fez com que o território se tornasse um reino e instaurando assim não só casas mas também processos de ensinos (SOUZA, 2018). No meio deste percurso, as instituições de ensino superior foram criadas para o compartilhamento de conhecimento militar para defesa no Brasil , passando por inúmeras mudanças no decorrer do processo (COELHO, 2009).

Durante a evolução do conhecimento adquirido e os processos da educação superior, o Brasil passou por problemas e por necessidades de inclusão nestas instituições, há algum tempo. De acordo com Rosana Heringer, 2018, o Brasil está em uma posição atrás de outros países da América Latina quando comparado ao processo de entradas no ensino superior. Um motivo citado que possibilita essa posição, e que explica um entendimento dos processos, é a oportunidade, a qual se embasa no entendimento de alguns alunos poderem estar nas escolas públicas e outros em privadas, mas ao olhar o âmbito acadêmico e ao processo de entrada no ensino superior, as universidades públicas se enchem com alunos de ensino privado, tirando as chances dos menos favorecidos.

Este problema é um dos motivos pelo qual foi criado o Programa de Assistência do governo, o PNAES (Programa Nacional de Assistência Estudantil), o qual ainda tenta sanar um outro problema que assola as universidades e faculdades do país, a evasão escolar. Neste contexto, é de necessidade de conhecimento que a evasão não ocorre pelo simples motivo de saída, mas pelo entendimento dos processos que levam o aluno a evadir, tornando a evasão mais do que apenas o processo de perda de dinheiro do governo ao investir em um aluno evadido (SACCARO, 2019).

2.2 Evasão escolar

A evasão em meio educativo, é um assunto demasiado delicado para as instituições educacionais, em meio a tantos fatores que podem causar esta evasão, torna-se complicado entender os motivos da evasão do aluno. Tais podem variar de assuntos pessoais ou a própria estrutura da universidade, fazendo com que, assim, prejuízos tanto ao aluno quanto a instituição afetem boa parte do Ensino Superior, majoritariamente as instituições particulares (DAVID ; CHAYM, 2019).

Estudos identificaram as principais causas de evasão e seus níveis, e também, alguns aspectos que impactam na satisfação dos alunos, com o propósito de estabelecer estratégias e ações gerenciais, que possam contribuir para manter os alunos estudando. Segundo a Comissão Especial de Estudos sobre a Evasão nas Universidades Públicas Brasileiras, a evasão possui níveis onde se permeiam aspectos inerentes que diferenciam os tipos de desligamento. Também foram levantados alguns fatores e causas da evasão, que podem variar de fatores internos e externos tanto dos estudantes e instituições, por causas como econômicas, sociais ou profissionais.

Sendo assim, torna-se necessário verificar as principais causas de evasão de alunos de uma Instituição de Ensino Superior, do ponto de vista dos fatores internos à instituição, com o propósito de estabelecer estratégias e ações gerenciais, que possam contribuir para mantê-los com o estudo. Visto que, em uma população mundial, onde o crescimento tecnológico reside, o conhecimento intelectual passa a ter maior valor agregado em termos de capacitação.

2.3 Governança

A Governança de TI, vem sendo fator decisivo tanto para âmbito empresarial quanto para a gestão de projetos menores, isto se dá ao fato de que ela busca identificar e desenvolver diversos elementos dentro de um projeto ou organização, visando aprimorar fatores de extrema importância como o controle de riscos, relação de gastos e definição da estrutura organizacional, isto tudo tendo como base a gestão da informação (BARBOSA, 2011).

Em termos de Governança de TI, serão impostos os valores de um guia da cultura de governo de TI, o COBIT para parametrização dos processos e suas reais necessidades.

Figura 1 - Princípios do COBIT 5

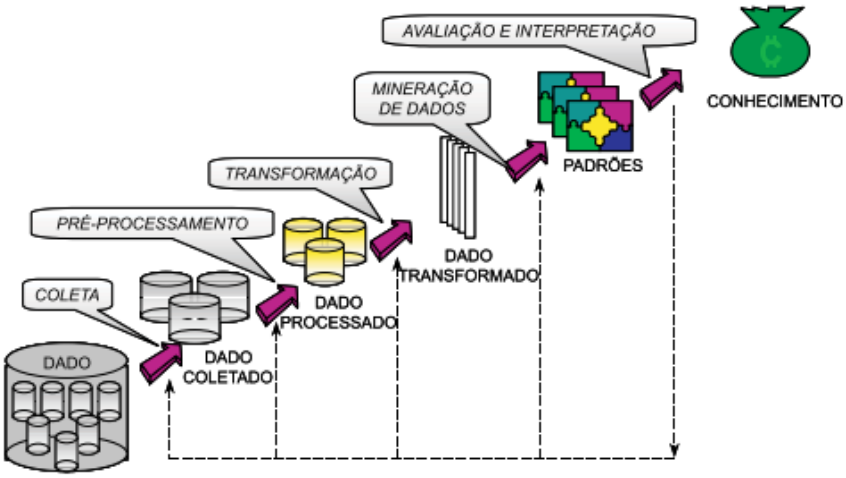
Fonte: ISACA, 2012.

Como pode ser visto na Figura 1, o COBIT 5, última versão do guia, conta com alguns pontos cruciais e que foram desenvolvidos durante todo o trabalho, sendo eles: i) atender as necessidades das partes interessadas; ii) cobrir a empresa de ponta a ponta; iii) aplicar um *framework* único e integrado; iv) permitir uma abordagem holística; v) distinguir a governança de gestão.

2.4 Descoberta de conhecimento em bancos de dados - KDD

Uma variedade de dados são coletados e armazenados em ritmos alarmantes a cada dia. Precisa-se que a cada vez mais surjam ferramentas e processos que possam ajudar a extrair informações realmente úteis (conhecimento) desse amontoado que cresce rapidamente. Esses processos e ferramentas são o tema da área emergente de descoberta de conhecimento em bancos de dados (KDD). Na Figura 2, temos uma representação gráfica de cada processo realizado no KDD.

Figura 2 - Representação das principais fases do processo de KDD



Fonte: BATISTA, GUSTAVO (2003, p. 34).

Pode-se definir em um nível mais abstrato que o KDD está preocupado com o desenvolvimento de métodos e técnicas para dar sentido aos dados (FAYYAD, 1996). As fases do processo de KDD são (BATISTA, 2003):

I. Identificar e entender o problema: Precisa-se entender as necessidades do usuário que possam ser resolvidas através da mineração de dados, pois, frequentemente os usuários descrevem suas demandas de maneira informal e sem conhecimento técnico então cabe a um analista de dados revisar as necessidades e aplicá-la a um dos métodos de mineração de dados;

II. Identificar os dados relevantes: Após a identificação do problema e entendimento da mesma é necessário identificar quais os atributos relevantes para criação do modelo preditivo. Nesta etapa é relevante verificar se os dados que pretende-se utilizar está disponível no banco de dados da a ser utilizado ou em um banco de dados externo ao proposto;

III. Coleta de dados: Coletar os atributos identificados como relevantes e que serão utilizados para a análise do banco de dados. Um dos principais problemas desta etapa é descobrir onde e como estão armazenados os dados na base de dados;

IV. Pré-processar os dados: Nesta fase deve ser aprimorado a qualidade dos dados coletados no passo anterior. Esta fase é essencial porque frequentemente os dados apresentam uma má-qualidade desencadeando diversos tipos de problema como valores desconhecidos, atributos com valores incorretos, atributos que não irão ter

valor agregado à predição e outros casos que podem ser levantados na análise dos dados;

V. Transformar os dados: O principal objetivo desta fase é transformar como os dados coletados são representados, buscando superar o maior número possível de limitações que existem nos algoritmos utilizados para extração de padrões. A decisão de qual transformação utilizar depende de qual algoritmo iremos utilizar na fase de mineração de dados;

VI. Minerar os dados: A fase de mineração dos dados gira em torno de decidir qual ou quais algoritmos serão utilizados nos dados com a finalidade de criar um modelo preditivo. Nesta fase, podemos utilizar diversos tipos de algoritmos que são fruto de interdisciplinaridade como, Aprendizado de Máquina, Estatística e Redes Neurais, mas precisa-se atentar na escolha de qual algoritmo terá o melhor desempenho para o problema proposto;

VII. Avaliar e interpretar os dados: Nesta etapa irá se avaliar e interpretar os resultados obtidos através do modelo preditivo, analisando se o classificador atingiu as expectativas, métricas de taxas de erros, tempo de processamento e quão complexo o modelo é. Por fim, o usuário do sistema deve julgar sobre a aplicabilidade dos resultados obtidos.

2.5 Aprendizado de máquina

Primeiramente, deve ser definido o que é Aprendizado de Máquina, uma definição simplificada para Géron (2019, p. 4) é que o "Aprendizado de Máquina é a ciência (e a arte) da programação de computadores para que eles possam aprender com os dados" e apresenta em seguida uma definição mais completa do pioneiro no ramo do aprendizado de máquina Géron (2019, p. 4 apud AL Samuel, 1959) citando “Aprendizado de Máquina é o campo de estudo que dá aos computadores a habilidade de aprender sem ser explicitamente programado”.

A frase ‘habilidade de aprender sem ser explicitamente programado.’ da última citação é explicada por (AGHABOZORGI; SANTARCANGELO, 2022) fazendo uma reflexão a respeito da abordagem tradicional de 'inteligência', pois ao invés de se escrever um conjunto de regras (como um encadeamento de condições se/senão) para dar inteligência ao sistema ao processar certo conjunto de dados, pode-se construir modelos que irão considerar as características do conjunto de dados e aprender com o padrão desses dados, afinal, ele realiza a identificação desses padrões sem ser programado explicitamente para isso.

2.5.1 Principais tipos de aprendizado de máquina

Existem inúmeros sistemas de aprendizado de máquina, que podem ser classificados de acordo com o tipo de aprendizado utilizado (GÉRON, 2019). Abaixo pode ser observado a lista com os principais tipos de aprendizado e sua maneira de atuação:

- **Aprendizado Supervisionado**: O sistema aprende através de um conjunto de treino devidamente preenchido com dados rotulados e depois tenta identificar um conjunto de teste sem os devidos rótulos com uma a melhor acurácia possível (LEARNED-MILLER, 2014).
- **Aprendizado Não Supervisionado**: Diferente do aprendizado supervisionado, aqui não se diz para o sistema qual a conexão entre os dados e deve-se deixar que ele encontre a conexão entre eles sem 'ajuda' (GÉRON, 2019).
- **Aprendizado por Reforço**: No aprendizado por reforço não possuímos supervisores reais e é também baseado na resposta de retorno fornecido pelo ambiente. No aprendizado por reforço, essa "resposta" geralmente é chamada de recompensa (às vezes, uma resposta negativa é definida como uma punição) e é útil para entender se uma determinada ação realizada em um estado é positiva ou não. A sequência de ações mais úteis é uma política em que o agente tem que aprender, para poder tomar sempre a melhor decisão em termos de obter a maior recompensa imediatamente e cumulativamente. Em outras palavras, uma ação também pode ser imperfeita, mas em termos de uma política global ela deve oferecer a maior recompensa total. Esse conceito se baseia na ideia de que um agente racional sempre persegue os objetivos que podem aumentar sua riqueza (BONACCORSO, 2017).

2.5.2 Tipos de dados

De acordo com Campos (2002), quando se fala de análise de dados deve-se entender como esses dados são representados e como pode-se extrair conhecimento de um conjunto de dados. Os dados basicamente podem ser representados por dois tipos de dados, os numéricos (discretos e contínuos) e os categóricos (ordinais e nominais).

Os dados numéricos apresentam um caráter mais quantitativo podendo ser representado por meio de valores discretos finitos e contáveis onde se utiliza apenas número inteiros, como por exemplo, a quantidade de maçãs em uma fruteira ou a quantidade de alunos em uma sala de aula. Já os valores numéricos contínuos são valores (inteiros e fracionários) que podem assumir qualquer valor em um determinado intervalo, como por exemplo, uma medida de temperatura em graus Celsius ou um valor de item.

Os dados categóricos apresentam um maior valor quantitativo sendo representado por variáveis nominais que não apresentam uma ordem direta em si, como por exemplo, os códigos das unidades federativas (UF) ou a classificação de gêneros. Já os ordinais são dados que podem ser ordenados e classificados entre si, como por exemplo, o sistema de notas escolares de classificação(*grading system*) onde os alunos são avaliados por letras do alfabeto indo da letra A (90% ~ 100%) até a letra F (< 60%).

2.6 Classificação

Como descrito no Seção 2.4.1, existem diversos tipos de aprendizado de máquina, cada um tendo sua utilidade para o tipo de problema proposto. Neste capítulo será explicado acerca dos algoritmos de classificação que são amplamente utilizados no aprendizado de máquina supervisionado.

O autor Santos (2013, p. 25), define os algoritmos de classificação como:

> Classificação consiste em atribuir uma classe a um documento a partir dos seus dados de entrada. Essa classificação é feita através de algoritmos de aprendizagem. Esses algoritmos conseguem "criar conhecimento" sobre o domínio tratado através de dados de entrada. Quando esses dados de entrada já possuem suas classes atribuídas dizemos que essa é uma aprendizagem supervisionada.

A classificação, então, pretende classificar os dados em rótulos por meio dos dados de entrada e realizar uma avaliação do modelo utilizando os dados do conjunto de testes aplicando métricas específicas para tal.

Quando se fala em classificação existem diversos tipos de algoritmos que podemos utilizar para criação do modelo de aprendizado de máquina. Após a pesquisa bibliográfica foi levantado que os algoritmos a seguir que são amplamente utilizados para criar modelos de classificação: Regressão logística; Máquina de vetores de suporte; Árvores de decisão e Floresta aleatória (MARTINS et al., 2021).

2.6.1 Regressão logística

Vale ressaltar que a regressão logística é semelhante a regressão linear, porém, há uma grande diferença em como esses algoritmos podem ser utilizados. Por exemplo, o algoritmo de regressão linear possui uma ampla utilização em para previsão de um valor (resposta) dado um conjunto de dados (variáveis de entrada) (SOUZA, 2020). Em contraponto a regressão logística é mais utilizada para realizar classificações de classes pois ao invés de calcular uma soma ponderada das variáveis de entrada gera um logística das mesmas como apresentado na Equação 1.

Equação 1 - Modelo de regressão logística probabilidade estimada

$$\rho = h_{\theta}(x) = \sigma(\theta^{T}. x)$$

Fonte: Autores, baseado em Géron (2019, p. 140).

A regressão logística segundo Géron (2019), é usada então para estimar a probabilidade de uma determinada classe ocorrer com base nos valores das variáveis de entrada (por exemplo, qual é a probabilidade do aluno que está se graduando concluir o curso?). A regressão logística é então apropriada para estudos em que se tem um conjunto de variáveis explicativas que relacionam-se à variável resposta.

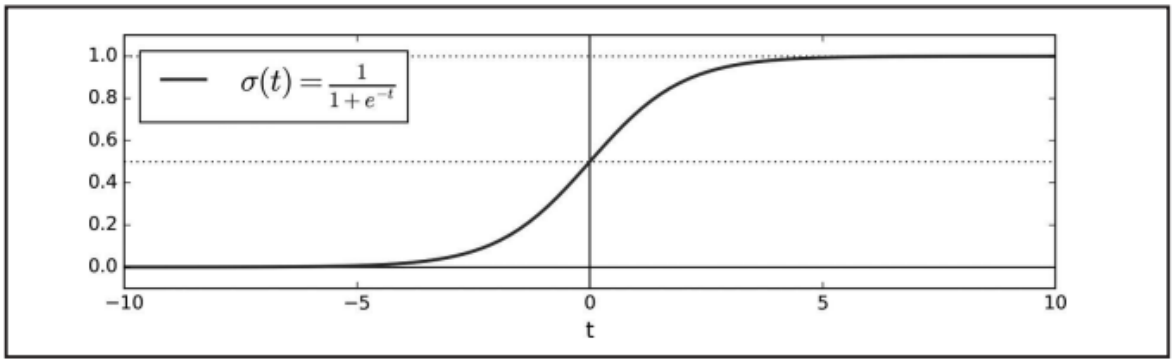
O funcionamento da regressão logística se dá através de uma função matemática sigmóide (que possui formato da letra S) que é representada pela letra σ (sigma) que retornará um valor reposta entre 0 (negativo) e 1 (positivo). A definição dessa função matemática está representada na Equação 2 e em sua representação gráfica na Figura 3.

Equação 2 - Função Logística

$$\sigma(t) = \frac{1}{1 + exp(-t)}$$

Fonte: Autores, com base em GÉRON (2019, p. 140).

Figura 3 - Gráfico da Função Logística



Fonte: (GÉRON, 2019, p. 140).

Uma vez estimada a probabilidade com $\rho = h_{\theta}(x)$ onde a instância x pertence a uma classe positiva é possível fazer sua previsão y. A previsão y receberá o resultado da condicional:

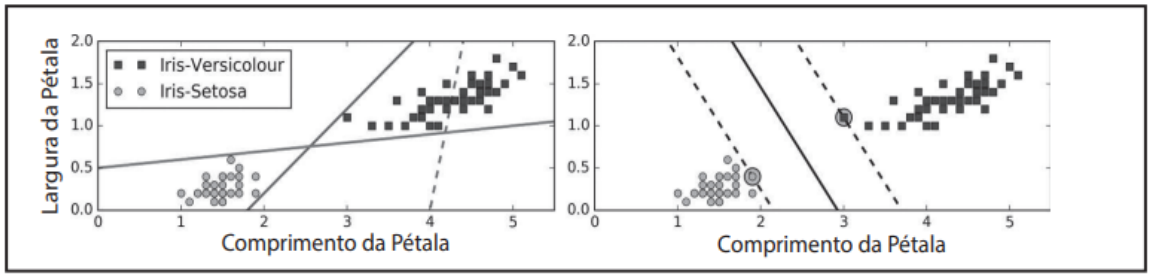
$\sigma(t) < 0{,}5$ quando $t < 0$, e $\sigma(t) \geq 0{,}5$ quando $t \geq 0$. Então um modelo de regressão logística prevê 1 se $\theta^T . x$ for positivo, e 0 se for negativo (GÉRON, 2019, p. 140).

2.6.2 Máquina de vetores de suporte

Uma máquina de vetores de suporte (MVS) é uma técnica de aprendizado de máquina que tenta encontrar os limites de um conjunto de dados. Esse tipo de aprendizado de máquina utiliza-se de um modelo matemático para encontrar esses limites e os pontos onde os dados estarão melhor alinhados. A MVS cria então uma linha que separa o conjunto de dados em dois grupos em um plano bidimensional, aprendendo a dividir os dados por meio dos padrões encontrados (SANTOS, 2003).

Na Figura 4, pode-se ver um exemplo da utilização do MVS para a classificação de duas plantas do gênero Iris. O modelo classifica as plantas em duas classes possíveis: a Iris-Versicolor e a Iris-Setosa, utilizando como parâmetro a largura e comprimento da pétala.

Figura 4 - Classificação das plantas do gênero Iris com MVS



Fonte: (GÉRON, 2019, p. 150).

De acordo com Géron (2019), o MVS possui grande diversidade, sendo capaz de realizar classificações lineares e não lineares, de regressão e detecção de isolamentos. Apesar de ser bastante diverso, neste trabalho será utilizado e analisada a sua aplicação para a classificação do tema proposto. Algumas das principais características que o tornam o seu uso atrativo e tão diversos são as seguintes (LORENA, 2003):

- **Capacidade de generalização:** Classificadores gerados por MVSs geralmente conseguem uma boa generalização. A generalização de um classificador é medida por sua eficiência em classificar dados que não pertencem ao conjunto utilizado em seu treinamento. Assim, quando o MVS gera o preditor, o sobreajuste é evitado, ou seja, o preditor não se torna especializado no conjunto de treinamento e seu desempenho se mantém quando confrontado com novos padrões.

- **Robustez em grandes dimensões:** As MVSs são robustas para objetos de grandes dimensões, como imagens. Muitas vezes, os classificadores produzidos por outros métodos inteligentes nesses tipos de dados são passíveis de sobreajuste.
- **Convexidade da função objetivo:** A aplicação do MVS implica na otimização de uma função quadrática que possui apenas um mínimo global. Esta é uma vantagem em relação às redes neurais artificiais, por exemplo, onde existem mínimos locais na função objetivo a serem minimizados.

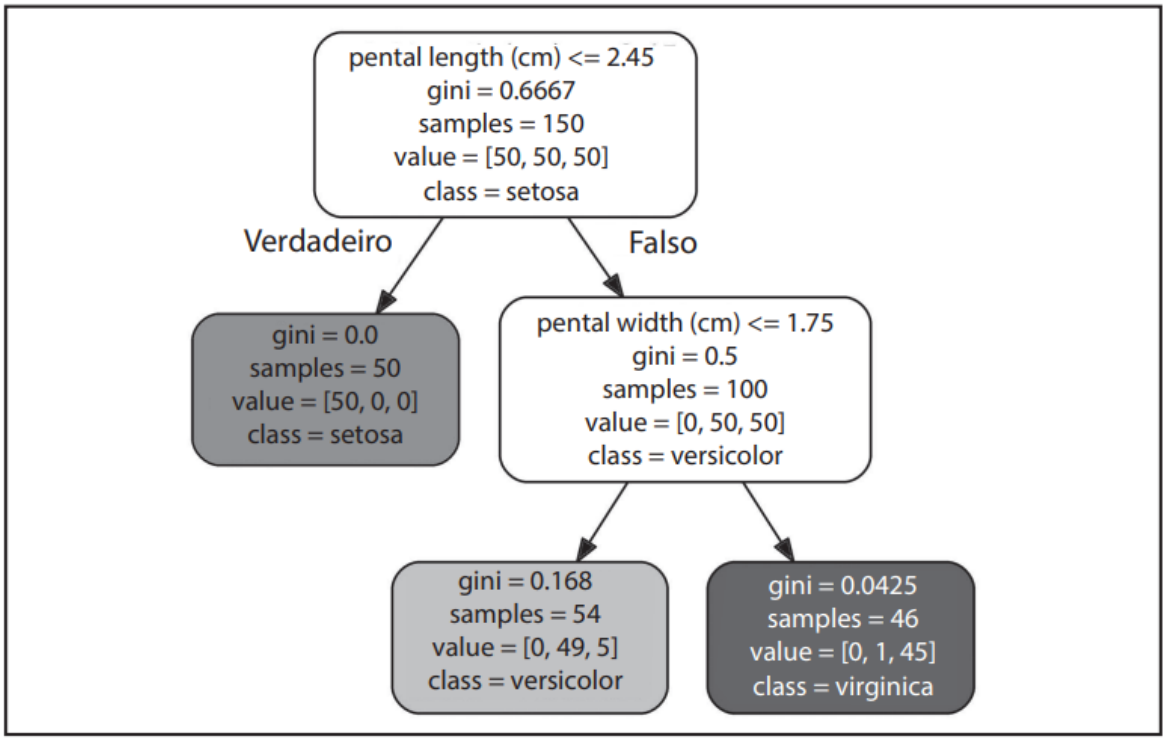
2.6.3 Árvores de decisão

As árvores de decisão (AD) são compostas por algoritmos bastante versáteis tendo diversas aplicações como o MVS. A vantagem de usar esse método é que ele permite que os usuários tomem decisões com base em seus fatores mais importantes. As AD apresentam primeiro as informações importantes, seguidas pelas informações menos significativas. Eles mostram efetivamente quais atributos são mais importantes para um trabalho.

As AD utilizam uma estratégia chamada “dividir e conquistar”. Isso significa que elas pegam um problema grande e complexo e o dividem em subproblemas menores e mais fáceis. Então, eles passam pelo mesmo processo com cada subproblema (LEMOS et al., 2005, p. 229 apud GAMA, 2000).

A maioria das pessoas conseguem entender as AD, porque o método agrupa os dados de uma maneira que faz sentido. As AD geram um gráfico direcionado onde cada nó representa um resultado possível e a direção de cada filial de saída indica uma ação específica a ser tomada. Cada nó corresponde a uma ação específica a ser tomada para alcançar seu possível resultado (LEMOS et al., 2005). Utilizando o mesmo exemplo da Iris da Seção 2.5.2, o algoritmo de classificação de AD geraria a seguinte árvore da Figura 5.

Figura 5 - Árvores de decisão classificando as plantas do gênero Iris



Fonte: (GÉRON, 2019, p. 172).

2.7 Avaliação do classificador

Para avaliar os modelos preditivos de classificação é preciso trabalhar com medidas que possam avaliar a eficácia da classificação realizada, basicamente é necessário verificar qual é a taxa de erro e acerto do classificador. As medidas a serem explanadas neste capítulo são amplamente utilizadas nos trabalhos com tema semelhante a este.

Antes de prosseguir-se para o estudo das medidas de desempenho, deve-se entender que um objeto que será analisado pode ser classificado em 4 nomenclaturas diferentes que irão formar a Matriz de Confusão apresentada na Tabela 1. São eles (HRIPCSAK, 2005):

- **Verdadeiro-Positivo** (VP): Quantidade de objetos que foram classificados como positivos e realmente pertencem à classe de positivos.
- **Falso-positivo** (FP): Quantidade de objetos que foram classificados como positivos e pertencem à classe de negativos.
- **Falso-negativo** (FN): Quantidade de objetos que foram classificados como negativos e pertencem à classe dos positivos.
- **Verdadeiro-negativo** (VN): Quantidade de objetos que foram classificados como negativos e pertencem à classe de negativos.

Tabela 1 - Matriz de Confusão

		Avaliação do Classificador	
		Positivo	Negativo
Avaliação Real	Positivo	VP	VN
	Negativo	FP	FN

Fonte: Autores, baseando-se nos trabalhos de HRIPCSAK, 2005 e SANTOS, 2003.

Com essas nomenclaturas expostas pode-se verificar que o VP e VN são objetos que foram classificados corretamente, enquanto o FP e FN são o oposto, ou seja, objetos que foram classificados de maneira incorreta.

2.7.1 Acurácia

A acurácia irá medir o desempenho das classificações realizadas pelo modelo preditivo utilizando a equação matemática representada na Equação 3:

Equação 3 - Acurácia

$$Acurácia = \frac{VP + VN}{VP + FP + FN + VN}$$

Fonte: SANTOS (2003, p. 31) e HRIPCSAK (2005, p. 296-298).

A princípio essa medida pode parecer totalmente ideal para todos os tipos de modelo, porém, ela apresenta algumas falhas. Imagine por exemplo que o objetivo de um modelo seja avaliar uma base de dados com comentários de um aplicativo de vídeos com 100 comentários. Do total de 100 comentários 95 são comentários negativos e 5 são comentários positivos e que a classe que se tem interesse nessa avaliação seja a de comentários positivos.

Supondo que o modelo classifique todos os 100 comentários como sendo negativos, esse modelo ainda assim apresentaria uma acurácia de 95% sendo que não avaliou nenhum comentário positivo corretamente. Isso ocorreu por que a medida de acurácia não levou em consideração de maneira isolada os 5 comentários que deveriam ter sido classificados como positivos, verificando apenas o contexto geral da avaliação. Levando em consideração esse ponto as próximas medidas a serem apresentadas neste capítulo poderá se observar a diferença da acurácia.

2.7.2 Precisão

A precisão examina a quantidade de documentos que foram corretamente classificados como positivos em comparação com todos os outros documentos que foram classificados como

positivos. Utiliza-se uma divisão entre o número de documentos classificados corretamente (VP) pelo número de documentos classificados como positivos (VP + FP), representada pela Equação 4:

Equação 4 - Precisão

$$Precisão = \frac{VP}{VP + FP}$$

Fonte: SANTOS (2003, p. 32) e HRIPCSAK (2005, p. 296-298).

2.7.3 *Recall*

O *recall* é uma medida contraposta a precisão, ela relaciona a quantidade de documentos que foram corretamente classificados com positivos em comparação com todos os documentos que são realmente da classe positiva. Utiliza-se uma divisão entre o número de documentos classificados corretamente (VP) pelo número de documentos classificados que realmente são positivos (VP + FN), representada pela Equação 5:

Equação 5 - *Recall*

$$Recall = \frac{VP}{VP + FN}$$

Fonte: SANTOS (2003, p. 32) e HRIPCSAK (2005, p. 296-298).

Utilizando-se do exemplo dos comentários do Seção 2.6.1 e as medidas das Seções 2.6.2 e 2.6.3, teria-se então uma precisão de 0% e um *recall* de 0%, pelo fato do modelo preditivo não ter feito nenhuma classificação positiva correta e ainda errou a classificação dos 5 que eram da classe positiva.

2.7.4 *F-Measure*

Embora as medidas de precisão e *recall* sejam muito úteis para a avaliação do desempenho do classificador, às vezes é necessária uma única medida para que dois ou mais classificadores possam ser comparados diretamente. Este é o conceito da métrica F_{β}, que é uma medida baseada na média harmônica de precisão e *recall*. Pode-se observar então a Equação 6:

Equação 6 - *F-Score*

$$F\beta = \frac{(1+\beta^2) * Precisão * Recall}{\beta^2 * Precisão + Recall}$$

Fonte: SANTOS (2003, p. 33) e HRIPCSAK (2005, p. 296-298).

Na fórmula apresentada pode-se observar que dependendo do valor da constante β, a medida F_β dará maior importância para a medida de Precisão caso ($0 < \beta < 1$) ou dará maior importância para a medida de *Recall* caso ($\beta > 1$). Frente a isso há um caso de utilização da medida F, sendo ela a medida F_1 medida, com a constante $\beta = 1$, assim a medida F dará a mesma importância para Precisão e *Recall*, resultando na Equação 7:

Equação 7 - *F1-Score*

$$F1 = \frac{2 * Precisão * Recall}{Precisão + Recall}$$

Fonte: SANTOS (2003, p. 33) e HRIPCSAK (2005, p. 296-298).

2.8 Trabalhos relacionados

Com o crescimento da criação e da utilização de ambientes virtuais por parte das instituições nos últimos anos, foi possibilitado a extração de uma gama de dados que influência em diferentes estudos, entre eles o que mais se destaca é a mineração de dados educacionais (MDE). Tal busca otimiza diferentes elementos acadêmicos com base na análise nos dados coletados (MASCHIO; et al; 2018).

Com a utilização da mineração de dados educacionais (MDE), é possível identificar que existem estatísticas e predições como utilização de estudos, tendo como base comparações feitas visando dados de alunos concludentes e evadidos. Assim, analisando elementos como idade, endereço, tempo relacionado a conclusão do ensino médio até o ingresso no ensino superior e principalmente o coeficiente de rendimento, foram geradas estatísticas referentes a possíveis evasões com 90,7% de eficiência na instituição ao qual foi aplicada (LANES and ALCÂNTARA, 2018).

Nos ambientes virtuais de aprendizagem (AVAs), são gerados, também, diversos tipos de dados que podem servir de estudo para a predição de evasão, sendo um deles, as relações entre interações e seus elementos, como a média e mediana da quantidade de *logins* realizados diariamente e semanalmente (QUEIROGA; CECHINEL; ARAÚJO, 2017).

Por fim, é feita a preparação dos dados visando a comparação de alunos que finalizaram o curso e os evadidos através de rótulos onde é feito o treinamento. Em conjunto ao aprendizado de

máquina e ao algoritmo de mineração de dados, entendeu-se que os dados que mais se destacaram na pesquisa foram as interações e os em formato de texto oriundos de fóruns e bate-papo online, isto é, pela quantidade e pela influência no resultado final. Desse modo, com a utilização de diversos classificadores se obteve uma acurácia acima de 0,86 o que é uma alta taxa de acerto (RAMOS; et al; 2018).

Utilizando da ferramenta Azure machine learning studio que fornece estrutura para iteração de elementos de aprendizado de máquina, assim fornecendo assistência ao mesmo através de uma oficina visual com diversas funcionalidades. Além do Azure MLS se utilizou do *Visual Studio Data Tools* para desenvolver o sistema de SQL, já que a ferramenta fornece suporte para esse tipo de sistema, se aproveitando principalmente do módulo de SQL Server Integration Services – SSIS. Tendo isso em vista foi feita a preparação dos dados com o auxílio do modelo de processo CRISP-DM para estruturar o projeto em si (SOUZA, 2020).

Durante o processo foram realizados testes com diferentes algoritmos, entretanto o que mais se destacou foi as árvores de decisão de duas classes, com uma matriz de confusão para realizar a classificação de cada registro e obter métricas de desempenho, foi possível alcançar valores com taxa de acurácia de 0.964 e AUC de 0,994, assim configurando um ótimo resultado (SOUZA, 2020).

Já tendo como base um banco de dados onde os principais dados são interligados com a pandemia e aplicando a estrutura CRISP-DM para auxiliar nas etapas do processo, foi testando diferentes classificadores como MLP e árvores de decisão, contudo obteve-se maior sucesso o *XGBoost* que se mostrou bastante eficiente principalmente no *F1-Score*, onde é uma média composta pelas métricas de precisão e *recall*. Assim para a decisão final decidiu-se utilizar o *XGBoost* junto a hiperparâmetros implementados manualmente, com isso alterando apenas alguns parâmetros para alcançar maior eficiência, e assim gerando valores acima de 0.90 no *F1-Score* (PRIMÃO, 2022).

Demonstrando uma alternativa diferente ao qual se utiliza de uma base de dados principalmente advindos do AVA, a pesquisa realizada em uma universidade privada não especificada em Bangladesh obteve maior eficiência com a aplicação do classificador de votação ponderada onde chegou a alcançar 0.93 de acurácia e 0.90 de AUC, assim se sobressaindo em relação a concorrentes como as árvores de decisão, regressão logística e MLP no qual possuem vertentes bastantes difundidas em pesquisas dessa natureza (ZULFIKER; et al; 2020).

Tendo em vista essas pesquisas e suas aplicações temos como principais limitadores identificados a falta de opinião dos próprios aluno em relação ao que poderia levá-los a evadir, isso através de questionários ou outras formas, além da ausência de outras variáveis que poderia

aumentar a precisão dos sistemas como aspectos comportamentais e engajamento familiar. Mas o principal limitador observado demonstrou-se ser a falta de acesso ao banco de dados de outras instituições, para que assim possam produzir relatórios com base na comparação de seus resultados, sendo assim busca-se desenvolver o projeto encontrando formas de driblar essa limitação aos quais algumas espelham-se na pesquisa proposta, além de extrair o maior potencial possível do algoritmo utilizado.

Em comparação ao trabalho desenvolvido é de cunho explicativo que para a construção e o apoio declarado nos expostos aqui descritos, é de interesse a construção da metodologia e dos materiais em uso. Desse modo, o sucesso da produção se consagra em cunhos tanto bibliográficos quanto cronogramados e pensados.

3. PROCESSO METODOLÓGICO

O processo metodológico será realizado utilizando do conceito de pesquisa chamada: pesquisa exploratória. Esse tipo de pesquisa pode ser considerada como um primeiro passo para a realização de uma pesquisa mais longa e aprofundada. A pesquisa exploratória foi escolhida, pois, deseja-se examinar um conjunto de fenômenos e anomalias acerca do problema proposto para que se descubra ideias a fim de deixar uma base para trabalhos futuros. Apesar de se tratar de um trabalho de cunho exploratório, isso não limita a utilização de outros tipos de pesquisa, formando assim uma combinação de diferentes técnicas (WAZLAWICK, 2009).

Foi utilizado o processo de KDD para auxiliar no processo metodológico e da extração de conhecimento, a partir disso, houve o início da primeira e da segunda fase do processo, onde busca-se identificar e entender o problema e quais são os dados relevantes para esse problema. Para concluir esses passos foi realizada uma pesquisa bibliográfica exploratória e quantitativa para aprofundar os conhecimentos na área escolhida, buscando trabalhos renomados no assunto para adquirir o máximo de informações importantes sobre o tema proposto.

Para realização das pesquisas bibliográficas utilizou-se websites de repositórios de artigos científicos e trabalhos acadêmicos que possuem fontes confiáveis e bem fundamentadas, tais como: *Scientific Electronic Library Online* (SciELO); Biblioteca Digital de Teses e Dissertações da USP; Google Acadêmico. Para auxiliar no processo de pesquisa foram utilizadas diversas palavras-chaves a fim de se filtrar trabalhos que sejam relevantes para o referencial bibliográfico. Na Tabela 2 pode-se visualizar as palavras-chave que foram utilizadas.

Tabela 2 - Tabela das palavras-chave utilizadas para a pesquisa bibliográfica.

Machine Learning + Dropout	*Machine Learning + Types*	*Machine Learning*	*Data Mining Techniques*	*Pre-processing + Machine Learning*	*Algorithm + Classification*	*Supervised Machine Learning*
Evasão	Evasão + Brasil	Evasão + Educação	Evasão + Ensino Superior	*Software* + Ensino Superior	*Unsupervised Machine Learning*	*Machine learning + Decision Three*
Aprendizado de máquina + Máquina de vetores de suporte	Aprendizado de máquina + Árvores de Decisão					

Fonte: Autores.

Após entender-se o problema e levantar quais dados serão relevantes para o trabalho deu-se continuidade a fase de coleta dos dados. Primeiramente, foi realizada uma nova pesquisa bibliográfica para encontrar uma base de dados que se encaixasse e possuísse os dados necessários para o tema proposto. Nessa pesquisa, foi encontrada a base de dados disponibilizada por um estudo de caso do Instituto Politécnico De Portalegre, Portalegre, Portugal, a qual tornou-se a principal fonte das informações de desenvolvimento (MARTINS et al., 2021).

Após a coleta da base de dados iniciou-se as fases de pré-processamento e transformação de dados. Na primeira etapa, o conjunto de dados foi analisado várias vezes e processado para melhorar sua qualidade e eliminar ao máximo a possibilidade de impactar negativamente nosso modelo e na etapa de transformação irá se trabalhar como os dados são representados para que possam ser utilizados em sua melhor forma.

Com os dados devidamente processados e transformados nas etapas anteriores deu-se continuidade a fase de mineração de dados, em que se faz necessária a construção de um *software* que irá implementar os algoritmos de aprendizado de máquina classificatórios. Os algoritmos que foram escolhidos para serem implementados a fim de se construir o modelo preditivo são: Regressão logística, Máquina de vetores de suporte e Árvores de decisão.

Com os modelos preditivos construídos, o processo caminhou para a última fase do processo de KDD: a avaliação e interpretação dos dados obtidos através desses modelos

classificadores. Nesta avaliação foram aplicadas diversas métricas de avaliação (Acurácia, Precisão, *Recall* e *F1-Score*) a fim de se calcular a precisão, efetividade e aplicabilidade dos modelos em um cenário real e comparar os resultados desse trabalho com os resultados obtidos em pesquisas com um tema relacionado, verificando assim se a metodologia aqui pesquisada e utilizada é válida para resolver o problema proposto.

4. DESENVOLVIMENTO

Neste capítulo será mostrado o desenvolvimento das construções teóricas e dos experimentos realizados apresentando dados; testes; e gráficos para validação do trabalho e a contribuição do mesmo para o conhecimento científico. Este capítulo não se foca em mostrar como é um *software* ou um passo a passo de como construir uma ferramenta específica para o problema proposto.

Primeiramente será abordado os pontos da Engenharia de *Software* em sua total conjuntura, passando pelos ramos de requisitos, e governança de TI, a fim de demonstrar os processos utilizados na Engenharia de Requisitos e Governo de TI. Após a iniciação dos procedimentos técnicos, serão demonstrados, principalmente, o modo como ocorreu o desenvolvimento do programa, trazendo incitações e os principais focos nos artefatos produzidos.

4.1 Engenharia de Software

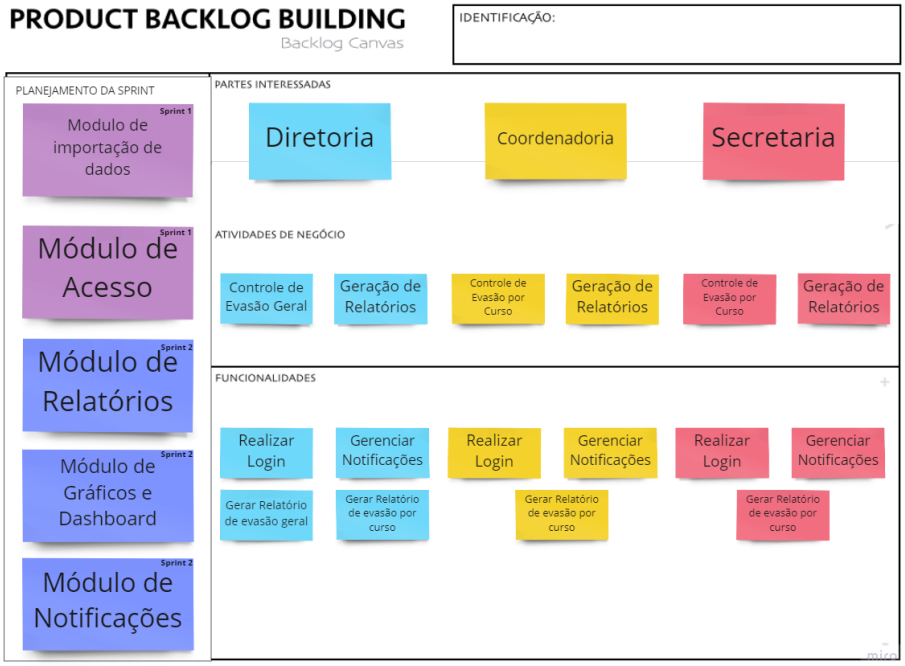
Para maior embasamento prático da engenharia de *software*, no início do desenvolvimento do projeto, foram descobertos os artefatos que possibilitaram a gestão do *backlog* e a definição de funcionalidades do projeto, as quais serão explicadas mais à frente. Por último, a Governança foi pensada e colocada como centro decisório entre Instituição e TI, a qual será o cerne diferencial para organização dos pontos específicos entre os resultados obtidos pelos relatórios para com a motivação das reitorias das universidade em entender que o ponto central de decisão é institucional.

4.1.1 Engenharia de Requisitos

Primeiramente, como citado, na criação e análise de requisitos foi predisposto que as funcionalidades teriam maior relevância nos processos durante o projeto, sendo assim, foram instanciados as histórias de usuário, com o intuito de credibilizar as personas e possibilitar a metrificação de uma maneira mais forte. Sendo assim, foram distribuídos em um *backlog* os requisitos mais importantes e as formas pelas quais seriam tratados os processos.

Como pode ser visto na Figura 6, foi utilizado o PBB (*Product Backlog Building*) para criação e metrificação de personas, atividades, funcionalidades e planejamento das duas *sprints* de produção.

Figura 6 - *Product Backlog Building*

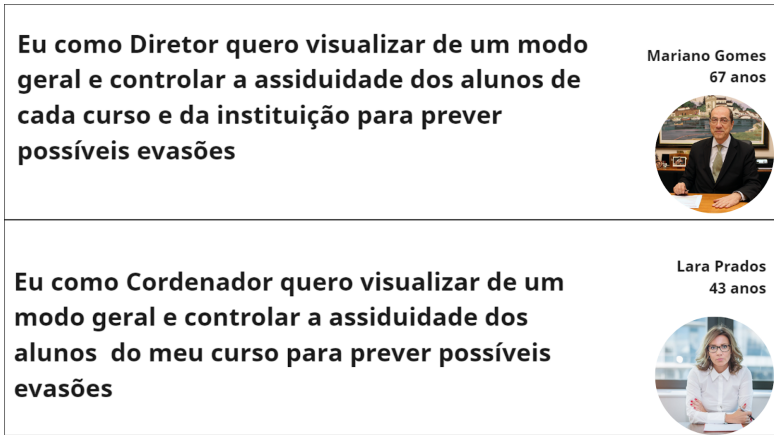


Fonte: Adaptado de Product Backlog Building.

Em que, é possível identificar, à esquerda, como serão desenvolvidos os módulos e suas prioridades em *sprints*. Além de visualizar os interessados, os quais são também as personas do projeto, que tem a ferramenta para manuseio e, por fim, a diferença de visualização de cada instância.

Após a definição do *Backlog*, foi feita a produção das personas do projeto a fim de trazer maior fidelidade aos processos. Dessa forma, é possível identificar na Figura 7 as personas e suas atribuições dentro do projeto.

Figura 7 - Personas



Fonte: Elaborado pelos autores.

Analisando a imagem, os 2 interessados são definidos como diretor e coordenador, respectivamente, onde, cada um deles tem seus interesses e suas previsões para com o produto, a fim de chegar na diminuição dos teores de evasão, o que se evidencia como a solução proposta pelo *software*.

Sendo assim, após a análise e priorização de requisitos, fase importante para o projeto, onde foram parametrizados, definidos e pensados os processos, foi feito o planejamento da Governança, mostrando a importância dela para a criação do projeto como um empreendimento futuro.

4.1.3 Governança de TI

Para fins de desenvolvimento do projeto futuramente, foram utilizados três princípios, que se destacaram na definição de estrutura organizacional da equipe e dos processos em si, que são: i) atender as necessidades das partes interessadas; ii) aplicar um *framework* único e integrado; iii) permitir a abordagem holística. Estes pontos nortearam os momentos de adentrar com as definições de governança e organização em sua totalidade dentro do ambiente de TI.

De modo crucial, a governança desenvolveu o projeto de forma a se pensar em quais seriam as principais ações a serem tomadas após o desenvolvimento geral do sistema, visto que, a sua venda e comercialização, como informado, será em um empreendimento futuro. Pontos estes que precisam incitar os gastos, ganhos e importância do projeto para as instituições de modo geral.

Sendo assim, o enfoque na adaptação e utilização do guia principal, COBIT, foi a fim de reduzir os problemas que seriam acarretados, caso não tivesse ocorrido um planejamento e uma ideia de finanças bem definidas, pontos que são bem colocados em todo o processo, e, por estar em finalização de desenvolvimento e iniciação no mercado consumidor, o produto atende a

preocupação com a organização da governança, para que assim, se implemente os ramos mais técnicos de controle gerais como ITIL, CMMI, entre outros guias e instruções durante sua comercialização.

Por fim, é importante que seja colocado que o estudo principal junto à Governança foi de torná-la o ponto focal de finalização do projeto, visto que, por ter um teor decisório, os resultados obtidos a partir das gerações dos relatórios fazem com que se tenha informações cruciais em mãos, das quais podem ser tomadas por todos os ramos da instituição. Desse modo, como fator de gerenciamento dos dados, a decisão final deve ser institucional, fazendo com que o papel da Governança não esteja apenas na definição dos riscos, manutenções ou finanças, mas também, em como proceder em conversas junto às reitorias para trazer os resultados e mostrar que o trabalho a partir deles são das áreas da universidade que solicitou a pesquisa.

4.2 Ambiente

A ferramenta a ser utilizada para montagem do ambiente de desenvolvimento do software será o Google Colab. O Google Colab é uma ferramenta baseada no software de código aberto chamada Jupyter Notebook e está disponível por meio de SaaS para os usuários do Google. A ferramenta foi escolhida por conta de suas diversas vantagens para realização de pesquisas científicas colaborativas, entre elas estão:

- Possui um ambiente robusto para desenvolvimento totalmente gratuito;
- Possui GPU;
- Possui suporte ao Python 2.7 a 3.9 (Utilizaremos a versão 3.8);
- Suporte a todas bibliotecas nativas do Python;
- Bibliotecas de aprendizados de máquina são pré-instaladas, como: Scikit-learn, Matplotlib, PyTorch, TensorFlow, Keras, OpenCV;
- Ambiente colaborativo com conexão ao Google Drive.
- Facilitação no processo de compartilhamento do documento (*notebook*) que será utilizado para realização do trabalho.

4.3 Coleta dos dados

A base de dados coletada para esse trabalho foi encontrada após uma longa pesquisa bibliográfica e foi fornecida através de um arquivo com extensão .csv onde contém dados institucionais relacionados aos alunos do ensino superior do Instituto Politécnico De Portalegre, Portalegre, Portugal. A base de dados contém tuplas referente aos alunos que foram matriculados durante os períodos acadêmicos de 2008/2009 a 2018/2019 e de diferentes cursos de graduação

tal como: agronomia, design, áreas de educação, enfermagem, jornalismo, administração, serviço social e áreas de tecnologias (MARTINS et al., 2021).

A primeira etapa é transformar essa base de dados utilizando a biblioteca pandas onde o arquivo .csv é transformado em um *DataFrame* (objeto da biblioteca) que permitirá a manipulação dos dados como: editar, adicionar, excluir e visualizar. Essa manipulação ocorre através de métodos disponibilizados pela biblioteca, assim permitindo que a fase de análise de dados seja realizada de maneira mais dinâmica e eficiente.

Na Tabela 3 que foi obtida através do método *describre* do *DataFrame* que extrai informações das colunas presentes no mesmo. A tabela possui três colunas: a primeira apresenta a posição do campo no *DataFrame*; a segunda coluna apresenta o nome da coluna; e a terceira apresenta o tipo do dado daquela coluna sendo, float64, int64 ou object. Outros dados que podem ser visualizado é que o *DataFrame* contém 4.424 tuplas e 37 colunas, sendo 36 colunas independentes numéricas (7 do tipo float64 e 29 do tipo int64) e 1 coluna dependente categórica (com o tipo *object*) com o nome de Classificação.

Tabela 3 - Informações das colunas presentes no DataFrame transformadas em uma tabela.

Posição	Coluna	Tipo de Dado
0	Estado civil	int64
1	Modo de aplicação	int64
2	Ordem de aplicação	int64
3	Curso	int64
4	Atendimento diurno/noturno	int64
5	Qualificação anterior	int64
6	Qualificação anterior (Média Nota)	float64
7	Nacionalidade	int64
8	Qualificação Materna	int64
9	Qualificação Paterna	int64
10	Profissão Materna	int64
11	Profissão Paterna	int64
12	Nota de Admissão	float64
13	Refugiado	int64
14	Necessidades educacionais especiais	int64
15	Devedor	int64
16	Mensalidades em dia	int64
17	Gênero	int64
18	Bolsista	int64
19	Idade de inscrição	int64
20	Estrangeiro	int64
21	Créditos curriculares 1º sem. (Creditado)	int64
22	Créditos curriculares 1º sem. (Matriculado)	int64
23	Créditos curriculares 1º sem. (Avaliações)	int64

Posição	Coluna	Tipo de Dado
24	Créditos curriculares 1º sem. (Aprovado)	int64
25	Créditos curriculares 1º sem. (Nota)	float64
26	Créditos curriculares 1º sem. (Sem avaliações)	int64
27	Créditos curriculares 2º sem. (Creditado)	int64
28	Créditos curriculares 2º sem. (Matriculado)	int64
29	Créditos curriculares 2º sem. (Avaliações)	int64
30	Créditos curriculares 2º sem. (Aprovado)	int64
31	Créditos curriculares 2º sem. (Nota)	float64
32	Créditos curriculares 2º sem. (Sem avaliações)	int64
33	Taxa de desemprego	float64
34	Taxa de inflação	float64
35	PIB (Produto Interno Bruto)	float64
36	Classificação	object
Total: 37 colunas Total de linhas: 4424 - (0 to 4423)		

Fonte: Elaborado pelos autores.

4.4 Pré-processamento de dados

Com os dados devidamente coletados é necessário iniciar-se a fase de pré-processamento. A etapa de pré-processamento é uma das fases que mais demanda tempo no processo de KDD, além de ser uma das fases mais importantes a ser realizada no processo. Nesta fase serão realizadas diversas análises no conjunto de dados e processamento no conjunto de dados a fim de melhorar a qualidade do mesmo e remover ao máximo a possibilidade de impactos negativos no nosso modelo preditivo.

4.4.1 Nulidade dos dados

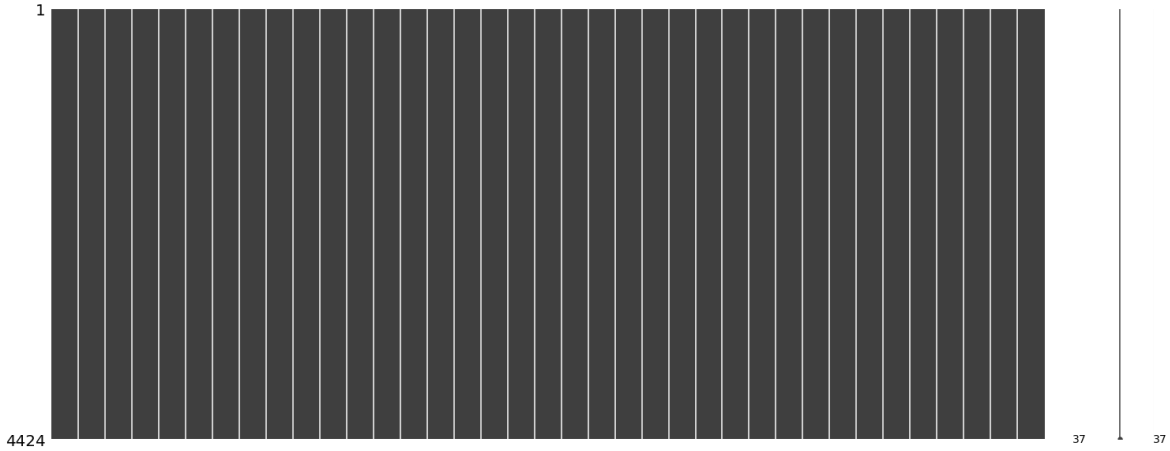
A primeira verificação que realizou-se na etapa de pré-processamento de dados foi a de nulidade no conjunto de dados, a fim de verificar a consistência do conjunto de dados em questão de campos não preenchidos. Para esta etapa foi utilizada a biblioteca Missingno disponível para o Python que criará um gráfico de matriz de densidade dos dados permitindo assim uma fácil visualização da nulidade dos dados.

O gráfico da Figura 8 traz informações relevantes sobre o conjunto de dados. No gráfico de barras pode-se observar que as barras estão completamente preenchidas com a cor preta e sem ruídos que seriam preenchidos pela cor branca, isso indica que todas as linhas de 1 a 4424 estão preenchidas. Pode-se verificar também que o gráfico de *sparkline* a direta não possui nenhum

ponto de variação, estando completamente reto, indicando assim que o conjunto de dados está com todos os campos preenchidos e não possuem campos nulos ou vazios.

.

Figura 8 - Gráfico de densidade dos dados gerados pela biblioteca Missingno



Fonte: Execução de algoritmo dos próprios autores.

Como todas as linhas e todas as colunas estão preenchidas e não possuem dados nulos ou vazios, não se faz necessário nenhum processamento nos dados ao ônus de nulidade, mas vale ressaltar que há diversas estratégias para processar esse tipo de ruído em dados. Para campos numéricos uma das estratégias mais utilizadas é imputar nos atributos nulos a média ou mediana dos dados corretamente preenchidos na coluna. No caso de campos texto a estratégia mais simples calcula qual o valor mais frequente na coluna e aplica ao atributo que está nulo ou vazio.

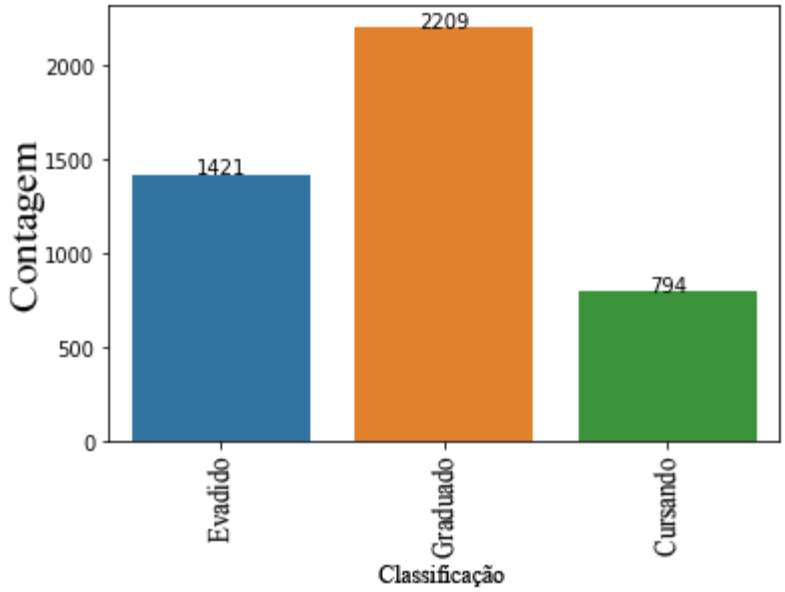
4.4.2 Separação de dados

Ao realizar uma análise na coluna Classificação, pode-ser ver que os valores presente nas linhas desta coluna apresentados na Figura 9, onde foi gerado um relatório de contagem dos valores utilizando os *frameworks* Pandas e Seaborn, apresenta a seguinte informação sobre a coluna: 1.421 alunos estão classificados com o valor de Evadidos; 794 alunos estão classificados com o valor de Cursando; 2.209 alunos estão classificados com o valor de Graduados; totalizando assim as 4.424 linhas.

Precisa-se então realizar um processamento nesta coluna que será separar todas as linhas que possuem o valor de Cursando para um subconjunto separado do principal. Isso deve ser feito para que o classificador tenha somente duas classes possíveis, sendo elas: Graduado e Evadido. Posteriormente será utilizado o subconjunto com as linhas classificadas como Cursando para criar

um novo *DataFrame* e utilizar o mesmo para verificar se os alunos que estão cursando irão concluir o curso ou evadir através dos modelos classificadores gerados.

Figura 9 - Gráfico com a contagem dos valores presente na coluna Classificação



Fonte: Execução de algoritmo dos próprios autores utilizando a biblioteca Pandas para contagem dos dados e a biblioteca Seaborn para geração do gráfico.

Para realizar esse processamento deve-se utilizar a biblioteca Numpy que irá auxiliar no processo de percorrer o *DataFrame* e aplicar uma condicional em todas as linhas da coluna Classificação (valor da linha é igual 'Cursando'?) a fim de criar o novo *DataFrame* secundário que será um subconjunto com as linhas onde a condicional for satisfeita e também remover as mesmas linhas do conjunto principal que deixará o *DataFrame* principal com o total de 3.630 linhas e o *DataFrame* secundário ficará com 794 linhas.

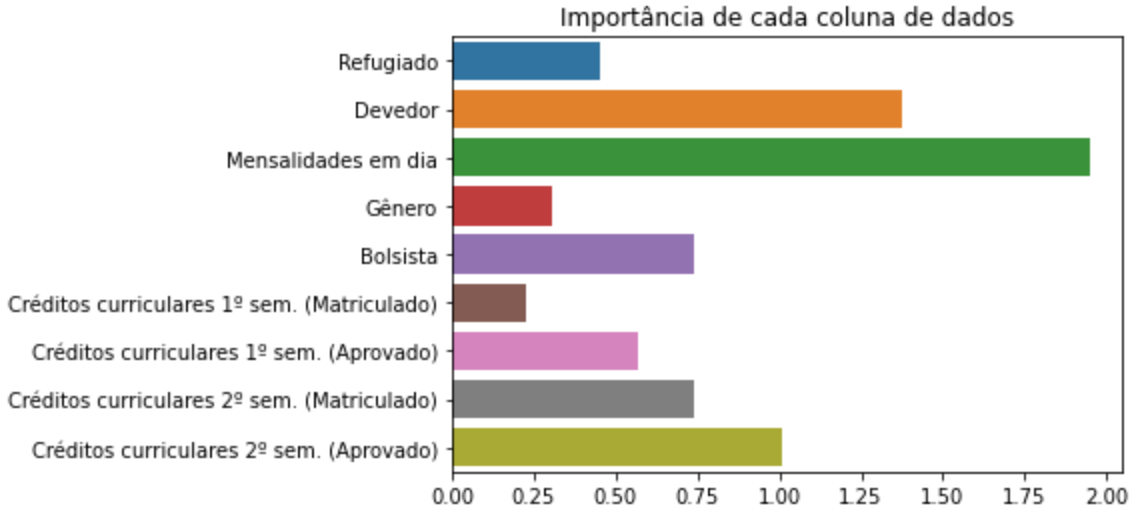
4.4.3 Identificar atributos relevantes

Nesta etapa tem-se a intenção de avaliar e selecionar os atributos (colunas) que realmente contém dados representativos para o conjunto de dados. Utilizou-se então do objeto *SelectFromModel* da biblioteca do *scikit-learn* que basicamente calcula valores de "pesos" para cada uma das colunas e depois remove todas as colunas que não atingirem um certo valor limite pré-estabelecido também conhecido como *threshold*, restando assim somente as colunas com "pesos" relevantes.

Com o processo realizado e as colunas removidas, serão gerados relatórios a partir das colunas remanescentes e seus respectivos pesos. Vale ressaltar que cada algoritmo de classificação

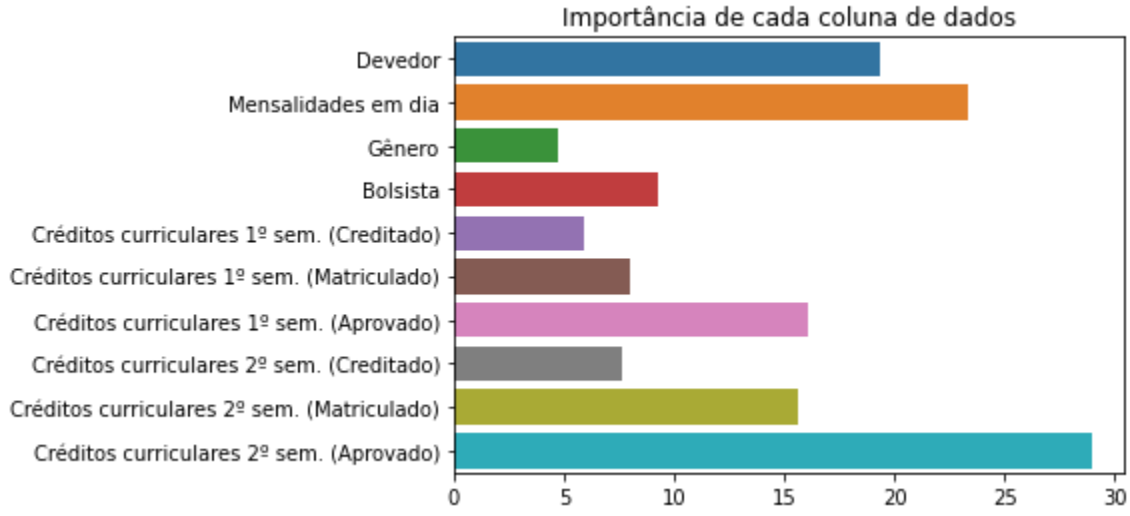
pode dar importância a colunas de maneira diferente de acordo com suas características, então foi gerado ao total 3 gráficos onde cada um deles irá representar as colunas selecionadas para os algoritmos descritos na Seção 2.6, são eles: Figura 10, Figura 11 e a Figura 12.

Figura 10 - Gráfico com as colunas mais relevantes para o algoritmo de regressão logística



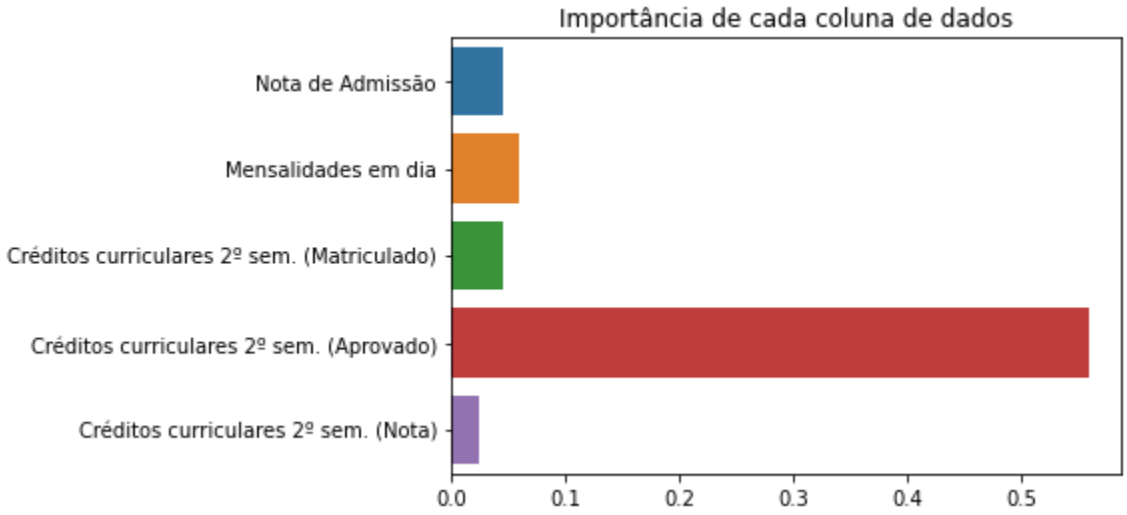
Fonte: Execução de algoritmo dos próprios autores utilizando a biblioteca Seaborn para geração do gráfico.

Figura 11 - Gráfico com as colunas mais relevantes para o algoritmo de máquina de vetores de suporte



Fonte: Execução de algoritmo dos próprios autores utilizando a biblioteca Seaborn para geração do gráfico.

Figura 12 - Gráfico com as colunas mais relevantes para o algoritmo de árvores de decisão



Fonte: Execução de algoritmo dos próprios autores utilizando a biblioteca Seaborn para geração do gráfico.

Analisando-se os gráficos: Figura 10, Figura 11 e Figura 12, podemos verificar que do total de 36 colunas independentes do conjunto de dados obtivemos uma redução significativa de colunas que foram consideradas como não relevantes. Importante se observar que em todos os algoritmos analisados as colunas: Mensalidade em dia e os Créditos curriculares do 2º semestre, se apresentaram como relevantes e que nos algoritmos de regressão logística e máquina de vetores de suporte obteve-se uma grande gama de colunas em comum, como: Devedor, Gênero, Bolsista e alguns dados dos Créditos curriculares do 1º semestre.

Com esses dados em mãos será realizada então a remoção das colunas que se mostraram não relevantes e usar somente as colunas selecionadas durante o processo de treinamento do modelo de classificação na etapa de mineração de dados, lembrando-se que a remoção será feita de maneira seletiva a partir dos dados obtidos para o algoritmo em específico que está sendo treinado.

4.5 Transformação dos dados

Após se realizar o pré-processamento dos dados iniciará a etapa de transformação dos dados. O objetivo desta fase é transformar a representação dos dados para um formato em que possa-se ser utilizado, buscando assim superar o maior número possível de limitações que existem nos algoritmos utilizados para extração de padrões.

4.5.1 Tratamento de dados categóricos

O tratamento de dados categóricos se faz necessário pois como pode ser visto anteriormente nas Seções 2.4 e 2.5 a criação de modelos de aprendizado de máquina trabalham (em seu âmbito básico) com a matemática. Então dados categóricos formados por texto devem ser convertidos em dados numéricos, sem que se perca a relevância e lógica que esse dado proporciona dentro da modelagem dos dados.

Com isso em mente pode ser visto na Seção 4.5, onde se realizou a coleta dos dados, que a única coluna que possui seus dados como categóricos (sendo eles textos) é a coluna Classificação. Os dados desta coluna são de caráter nominal e não possuem uma ordenação ou "peso" entre si, tendo em seus valores possíveis: Evadido, Cursando e Graduado. Uma das melhores técnicas para conversão desse tipo de dados é a *Label Encoder* fornecida pela biblioteca Pandas.

A estratégia do *Label Encoder* consiste em transformar uma quantidade X de dados categóricos presentes em uma coluna, em valores de 0 a X-1 que irão representar cada categoria, onde seus possíveis valores irão ser nesse caso: 0 (Evadido) ou 1 (Graduado). Como pode-se ver na Figura 13, os dados da coluna Classificação foram transformados em 2 novos valores que indicam se aquela linha possui ou não determinada categoria, mantendo assim lógica dos dados para o modelo e permitindo que eles possam ser utilizados através de cálculos matemáticos.

Figura 13 - Resultados com os novos valores da coluna Classificação

	Classificação
0	0
1	1
2	0
3	1
4	1
...	...
3625	1
3626	0
3627	0
3628	1
3629	1

3630 rows × 1 columns

Fonte: Elaborado pelos autores.

4.6 Mineração de dados

Para iniciar-se a extração de conhecimento com a mineração de dados se faz necessário realizar uma separação do banco de dados em dois conjuntos, X e Y, onde o conjunto X conterá as 36 colunas que contém as características da linha e Y conterá a coluna dependente que contém o rótulo da linha.

Com esses dois conjuntos criados deve-se realizar então um parcelamento desses dados a fim de criar os conjuntos de treino e teste totalizando quatro conjuntos: X de Treino e Y de Treino; X de Teste e Y de Teste. Esse parcelamento será realizado na seguinte proporção: o conjunto de testes contendo a parcela de 75% dos dados e a outra parcela com a proporção de 25% dos dados serão utilizados para teste e avaliação do conjunto.

Esse parcelamento se faz necessário pois o conjunto de testes será totalmente isolado do treinamento, ou seja, o modelo preditor receberá dados totalmente desconhecidos para classificar no momento da avaliação, deixando assim o processo mais preciso o possível, pois se o modelo fosse predizer dados que já tivessem sido utilizados anteriormente para treinamento ele sempre apresentaria uma ótima avaliação por já conhecer aqueles dados.

Com todos os dados já devidamente separados e parcelados devemos então começar a mineração de dados em si, utilizando os algoritmos descritos na Seção 3. Os conjuntos de treinamento serão aplicados aos objetos da biblioteca Pandas que contém a implementação dos algoritmos que serão utilizados, sendo os objetos: *LogisticRegression* (Regressão logística), *DecisionTreeClassifier* (Árvores de decisão) e *SVC* (Máquina de vetores de suporte).

5. RESULTADOS E AVALIAÇÃO

Nesta seção será realizada a avaliação dos modelos preditivos construídos na etapa de desenvolvimento. Irá se aplicar as métricas de, Acurácia, Precisão, *Recall* e *F1-Score* para realizar a comparação dos resultados a fim de verificar qual dos modelos obteve a melhor performance ao problema proposto. Para auxiliar no processo de aplicação das métricas de avaliação será utilizada a biblioteca Pandas por possuir métodos que auxiliam a gerar relatórios dos modelos preditivos criados na Seção 4.8.

5.1 Avaliação dos modelos

Utilizando-se dos relatórios gerados pela biblioteca Pandas apresentados na Figura 14 (onde pode-se ver a avaliação dos três modelos preditivos utilizados), pode-se comparar então as avaliações para obter qual modelo performou melhor e como as avaliações obtidas se relacionam

com outros trabalhos que utilizam técnicas semelhantes a fim de constatar se há uma relevância e aplicabilidade dos mesmos.

Figura 14 - Gráfico com a avaliação de acurácia, precisão, recall e *F1-Score* dos modelos preditivos

Árvore de decisão:

	Precisão	Recall	F1-Score
0	0.79	0.81	0.80
1	0.88	0.86	0.87
Acurácia			0.84
Média Macro	0.84	0.84	0.84

Regressão logística:

	Precisão	Recall	F1-Score
0	0.91	0.85	0.88
1	0.91	0.95	0.93
Acurácia			0.91
Média Macro	0.91	0.90	0.90

Máquina de vetores de suporte:

	Precisão	Recall	F1-Score
0	0.92	0.80	0.85
1	0.88	0.95	0.92
Acurácia			0.89
Média Macro	0.90	0.88	0.89

Fonte: Elaborado pelos autores utilizando o pacote *metrics* da biblioteca *scikit-learn*.

Observando os relatórios da Figura 14, é possível analisar os resultados das métricas de acurácia, precisão, recall e *F1-Score* aplicadas nos modelos preditivos que estão separados pelo nome do modelo e pelas possíveis classificações: 0 (Evadido) e 1 (Graduado). Ambos os modelos apresentaram uma performance considerável nas métricas aplicadas, em especial no *F1-Score*, a qual será a métrica principal utilizada para avaliação.

Em todos os modelos foi obtido uma porcentagem média maior que 75% no *F1-Score*, que, segundo Souza (2020), é um nível satisfatório para algoritmos de classificação. Em comparação com o relatório de um estudo de caso na mesma área de Martins (2021), presente na Figura 15, mostra-se que os mesmos modelos apresentaram um *F1-Score* entre 60% e 62% (sem a aplicação de nenhuma técnica para performar esses modelos).

Figura 15 - Gráfico com a avaliação dos modelos preditivos do estudo de caso

	Regressão Logística	Máquina de vetores de suporte	Árvores de Decisão	Floresta Aleatória
F1-Score Falhou	0.63	0.53	0.63	0.66
F1-Score Sucesso Relativo	0.41	0.31	0.39	0.37
F1-Score Sucesso	0.69	0.71	0.75	0.82
Média F1-Score	0.58	0.52	0.59	**0.62**
Acurácia	0.61	0.60	0.65	**0.72**

Fonte: Adaptado e traduzido de (MARTINS et al., 2021).

Pode-se observar que os modelos criados seguindo todas as etapas do KDD (Seção 2.3) onde os dados foram devidamente coletados, processados, transformados e avaliados a fim de se extrair conhecimento do conjunto de dados, conseguiram obter uma performance satisfatória em comparação com um modelo que foi criado sem a utilização de um processo específico para extração de conhecimento.

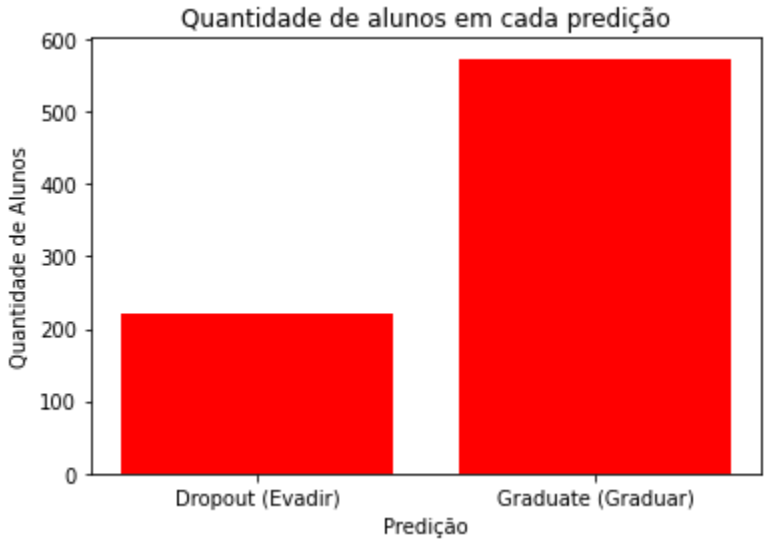
Para se realizar uma segunda avaliação foi utilizado o *DataFrame* secundário que contém o conjunto de dados de todos os alunos que possuem o status de Cursando separados na Seção 4.6.2. As linhas desse DataFrame foram preditas (em Evadido ou Graduado) pelo modelo preditivo de regressão logística, que foi o modelo classificador que obteve a melhor avaliação no *F1-Score*. A predição desse conjunto será comparada com a percentagem real de evasão dos estudos de David (2019), para que se possa obter uma comparação do modelo com um cenário real de evasão.

No que diz respeito de evasão, segundo David (2019), verifica-se que os alunos do curso de Enfermagem representaram a maioria dos casos de evasão com 20,68%, seguidos dos alunos de Administração com 19,28%, Medicina Veterinária com 19,08%, Educação Física com 15,46%, Serviço Social com 12,65%, Engenharia de Produção com 6,63%, e dos alunos de Engenharia Agrícola e Ambiental com 6,23%. Realizando uma média aritmética das porcentagem de evasão teremos uma média de evasão de 14,28%, com isso em mente podemos agora aplicar o conjunto de dados as predições.

Como pode-se observar na Figura 16, o modelo preditivo de regressão logística avaliou que 221 alunos irão evadir e que 573 continuarão nos cursos até sua graduação, totalizando assim as 794 linhas. A percentagem de evasão nesse cenário foi de 27,83%, colocando em contraponto com

a percentagem do estudo de caso real que é em média de 14,28% tem-se uma diferença de 13,55 pontos percentuais. Em uma análise inicial pode parecer que a diferença é gritante, mas deve ser levado em consideração dois fatores: a realidade perante os dados e a percentagem de assertividade do teste.

Figura 16 - Gráfico com os totais de predição do modelo preditivo de regressão logística



Fonte: Elaborado pelos autores.

O fator realidade, pode ser entendido de modo que o conjunto de dados que foi utilizado para treinamento do modelo, não condiz necessariamente com a realidade de todas as universidades e cursos existentes, e, por tratarem de dados que podem variar de instituição para instituição, a metodologia abordada neste trabalho deve ser aplicada a seu contexto, como já exposto anteriormente. O fator percentual, tem sua verdade pautada de modo que, mesmo o *F1-Score* sendo o modelo de regressão linear com maior taxa de acurácia, 90%, ele ainda indica que há uma taxa de erro de 10%, podendo assim, causar essa variação na porcentagem das predições.

6. CONCLUSÕES

Com a pesquisa exploratória realizada e com a aplicação prática dessas pesquisas através da construção de um *software* (descrito na Seção 4) foi possível avaliar e analisar os dados (Seção 5). O objetivo do trabalho dá-se por concluído, pois, mostra-se que é possível a utilização de aprendizado de máquina supervisionado para predizer risco de evasão dos alunos do ensino superior com uma boa precisão através da aplicação da metodologia de extração de conhecimento (KDD) em conjunto com os modelos classificatórios (Seção 2.6).

É válido ressaltar que o modelo de classificação de regressão logística foi aquele que se destacou dos demais, segundo Jaggi (2021) isso se deve ao fato desse algoritmo ser menos inclinado a se "acostumar" somente com os dados utilizados no treino e por trabalhar bem com conjunto de dados com poucas features. Também é extremamente recomendado que sejam utilizados os demais modelos aqui apresentados, pois seus resultados costumam ser mutáveis de um conjunto de dados para outro, assim se demonstrou necessário a comparação de seus resultados visando encontrar aquele que mais se adequa a instituição aplicada.

Por hora, houve a identificação de que alunos com risco de evasão podem contribuir para tomadas de decisões institucionais, levando em consideração os fatores que mais influenciam nessas desistências, sendo que entre os atributos aqueles que mais se destacaram na verificação dos classificadores foram, Mensalidade em dia, Devedor e Créditos curriculares 2º semestre (como se pode observar na Seção 4.4.3). Pensando de modo Governamental, estes dados fornecem um panorama para a análise de custos e controle de mudanças, assim, subsidiando um suporte para a instituição ao qual foi aplicado, de forma a demonstrar uma preocupação humana institucional para com os discentes e corpo docente.

Dessa forma, é importante salientar que cada indicação e problema é uma realidade daquela instituição em específico, cada uma tem suas características, e como exposto, precisa de um estudo próprio. Sendo assim, cada banco de dados terá um padrão de respostas motivacional para a evasão diferente, tanto em porcentagem quanto em motivação, fazendo com que a decisão que deva ser tomada torne-se pessoal do corpo reitor, necessitando da Governança de TI como alicerce do cerne do problema, dando suporte nas motivações e incitando as características retiradas do algoritmo, mas não como o campo decisor da mudança.

7. TRABALHOS FUTUROS

Pretende-se utilizar todo o compilado bibliográfico levantado através das pesquisas exploratórias em conjunto com as ideias que foram interpretadas, organizadas e desenvolvidas nesta pesquisa, para a criação de um *software* que irá analisar os dados de alunos de um ambiente de aprendizado de uma universidade para que possa-se validar a eficácia dos modelos classificatórios em um contexto real no mercado empreendedor, assim como descrito durante a Seção de Engenharia de Software (Seção 4.1).

Cabe-se dizer que esta pesquisa além de evolução e construção própria dos integrantes, serve como um grande ponto de apoio para estudos que desejam iniciar uma pesquisa futura mais abrangente de algum tema aqui abordado, mas também, que desejam construir um *software* para um ambiente corporativo com o intuito de descobrir novas perspectivas em sua empresa, trazendo

pontos aprofundados em: Evasão Escolar; Extração de Conhecimento de Dados; Aprendizado de Máquina; Tipos de Aprendizado de Máquina; Algoritmos de Classificação; Avaliação de Modelos Classificatórios e etc.

8. REFERÊNCIAS BIBLIOGRÁFICAS

AGHABOZORGI, Saeed; SANTARCANGELO, Joseph. Machine Learning with Python. [S.l.], 2022. Disponível em: https://www.coursera.org/learn/machine-learning-with-python. Acesso em: 29 abr. 2022.

Aguiar, Fábio ; Caroli, Paulo. Product Backlog Building: Concepção de um Product Backlog evetivo. 1ª edição. São Paulo, Editora Caroli, 2020.

ALBERTIN, Alberto. "Gestão e Cultura para o Jovem Administrador". vol. 16, n. 2. 2017.

BARBOSA, Andressa Munhoz; LIMA, Valter. Governança em TI: COBIT; ITIL. Revista científica Eletrônica de Administração, 2011.

BATISTA, Gustavo Enrique de Almeida Prado et al. Pré-processamento de dados em aprendizado de máquina supervisionado. 2003. Tese de Doutorado. Universidade de São Paulo.

BONACCORSO, Giuseppe. Machine learning algorithms. Packt Publishing Ltd, 2017.

CAMPOS, Geraldo Maia. Estatística prática para docentes e pós-graduandos. 2002.

COELHO, Sintia; VASCONCELOS, Maria. A CRIAÇÃO DAS INSTITUIÇÕES DE ENSINO SUPERIOR NO BRASIL: O DESAFIO TARDIO NA AMÉRICA LATINA. Florianópolis, 2009.

DAVID, Lamartine ; CHAYM, Carlos. "Evasão Universitária: Um Modelo para Diagnóstico e Gerenciamento de Instituições de Ensino Superior". Vol. 9, n. 1. 2019.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. AI magazine, v. 17, n. 3, p. 37-37, 1996.

JAGGI, Mukul et al. Text mining of stocktwits data for predicting stock prices. Applied System Innovation, v. 4, n. 1, p. 13, 2021.

GÉRON, Aurélien. Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow. Alta Books, 2019.

GIAMPANI, Alberto; TOLEDO, Dario; CUETO, Isaac. "Database Schema", 2020. Disponível em: <https://docs.moodle.org/dev/Database_Schema>. Acesso em: 05 de Junho de 2022

HERINGER, Rosana. "Democratização da educação superior no Brasil: das metas de inclusão ao sucesso acadêmico". Revista Brasileira de Orientação Profissional Vol. 19, No. 1, p.7-17. Rio de Janeiro. 2018.

HRIPCSAK, George; ROTHSCHILD, Adam S. Agreement, the f-measure, and reliability in information retrieval. Journal of the American medical informatics association, v. 12, n. 3, p. 296-298, 2005.

ISACA. COBIT 5: Um modelo corporativo para governança e gestão de TI da organização. [s. l.], p. 98, 2012.

LANES, Mariele; ALCÂNTARA, Cleber. Predição de Alunos com Risco de Evasão: estudo de caso usando mineração de dados. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE). 2018. p. 1921.

LEARNED-MILLER, Erik G. Introduction to supervised learning. I: Department of Computer Science, University of Massachusetts, p. 3, 2014.

LEMOS, Eliane Prezepiorski; STEINER, Maria Teresinha Arns; NIEVOLA, Julio César. Análise de crédito bancário por meio de redes neurais e árvores de decisão: uma aplicação simples de data mining. Revista de Administração-RAUSP, v. 40, n. 3, p. 225-234, 2005.

LORENA, Ana Carolina; DE CARVALHO, André CPLF. Introdução às máquinas de vetores suporte. Sao Carlos-SP, 2003.

MARTINS, Mónica V. et al. Early Prediction of student's Performance in Higher Education: A Case Study. In: World Conference on Information Systems and Technologies. Springer, Cham, 2021. p. 166-175.

MASCHIO, Pedro et al. Um panorama acerca da mineração de dados educacionais no Brasil. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE). 2018. p. 1936.

NASCIMENTO, Marcus. "O USO DE FERRAMENTAS DE INTELIGÊNCIA ARTIFICIAL EM BASE DE DADOS DE SAÚDE MILITARES" Rio de Janeiro. 2019.

NUNES, Renata Cristina. Um olhar sobre a evasão de estudantes universitários durante os estudos remotos provocados pela pandemia do COVID-19. Research, Society and Development, v. 10, n. 3, p. e1410313022-e1410313022, 2021.

PRIMÃO, Aline Pacheco et al. Uso de algoritmos de machine learning para prever a evasão escolar no ensino superior: um estudo no Instituto Federal de Santa Catarina. 2022.

QUEIROGA, Emanuel; CECHINEL, Cristian; ARAÚJO, Ricardo. Predição de estudantes com risco de evasão em cursos técnicos a distância. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE). 2017. p. 1547.

RAMOS, Jorge Luis Cavalcanti et al. Um estudo comparativo de classificadores na previsão da evasão de alunos em EAD. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE). 2018. p. 1463.

SACCARO, Alice; FRANÇA, Marco Túlio Aniceto; JACINTO, Paulo de Andrade. Fatores Associados à Evasão no Ensino Superior Brasileiro: um estudo de análise de sobrevivência para os cursos das áreas de Ciência, Matemática e Computação e de Engenharia, Produção e Construção em instituições públicas e privadas. Estudos Econômicos (São Paulo), v. 49, p. 337-373, 2019.

SANTOS, Fernando Leandro dos. Mineração de opinião em textos opinativos utilizando algoritmos de classificação. 2013.

SOUZA, Alex Marques de. Machine learning e a evasão escolar: análise preditiva no suporte à tomada de decisão. 2020. Tese de Doutorado. Mestrado em Sistemas de Informação e Gestão do Conhecimento.

SOUZA, Glauco. A FORMAÇÃO DO ESTADO BRASILEIRO – O QUE FEZ OU FAZ O BRASIL SER O BRASIL? UM OLHAR HISTÓRICO SOBRE CULTURA E COMPORTAMENTO. Cad. de Pesq. Interdisc. em Psicologia: Fund. teóricos, históricos e epistemológicos do pensamento psicológico. Registro, vol. 2, p. 44-53, ag. 2018.

WAZLAWICK, Raul Sidnei. Metodologia de pesquisa para ciência da computação. Elsevier, 2009.

ZULFIKER, Md Sabab et al. Predicting students' performance of the private universities of Bangladesh using machine learning approaches. International Journal of Advanced Computer Science and Applications, v. 11, n. 3, 2020.