REVISTA
CADERNO
VIRTUAL

Previsão de Evasão em Cursos de Graduação Utilizando *Machine Learning:* Uma Análise com Alunos de uma Instituição de Ensino Superior particular

Kelvin Gomes Pimentel Rodrigo dos Santos

Eduardo Conde Pires

Palavras-chave: *machine learning*, predição, evasão, desempenho acadêmico, *Business Intelligence (BI).*

Resumo

O objetivo central deste artigo é explorar a aplicação efetiva de técnicas *de Machine Learning*, aliadas ao poder analítico do *Business Intelligence (BI)*, para prever a evasão em cursos de graduação. Os dados são obtidos através de um formulário desenvolvido pelos alunos e fornecidos pela instituição de ensino IDP. Baseando-se em padrões identificados em pesquisas correlatas, o estudo busca identificar e analisar relações entre os principais atributos acadêmicos e sociais dos alunos. A finalidade é construir um modelo preditivo capaz de antecipar possíveis situações de evasão acadêmica, utilizando tanto as técnicas avançadas de *Machine Learning* quanto as capacidades analíticas do *BI*. Essa abordagem visa proporcionar entendimentos estratégicos para a tomada de decisões e intervenções proativas na gestão acadêmica.

ABSTRACT

The central objective of this article is to explore the effective application of Machine Learning techniques, combined with the analytical power of Business Intelligence (BI), to predict dropout rates in undergraduate courses. The data is obtained through a form developed by students and provided by the educational institution IDP. Building on

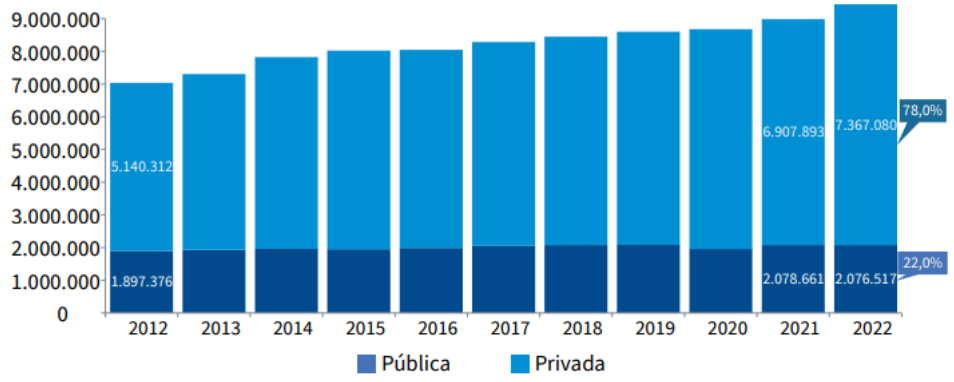
patterns identified in related research, the study seeks to identify and analyze relationships between key academic and social attributes of students. The aim is to construct a predictive model capable of anticipating potential academic dropout situations, utilizing both advanced Machine Learning techniques and the analytical capabilities of BI. This approach aims to provide strategic insights for decision-making and proactive interventions in academic management.

INTRODUÇÃO

Durante as últimas décadas, a formação profissional começou a ser um atributo relevante nas empresas, assim como a qualificação profissional tem aparecido como um requisito aos que buscam se inserir no mercado de trabalho (Martins, B. V., & Oliveira, S. R. D. (2017)). Desse modo, as instituições de ensino superior, tem o papel fundamental social de promover o pensamento crítico, desenvolvimento social e pessoal de seus estudantes, uma vez que, essas características são fundamentais no campo profissional (Monteiro, A., Gonçalves, C., & Santos, P. J. (2019)).

No Brasil, 89% das universidades, são particulares, segundo o Censo do Ensino Superior de 2022 (INEP, 2022), a distribuição das matrículas é de 78% na rede privada, enquanto a rede pública tem 22% das matrículas, a Figura 1 representa com a dispersão.

Figura 1



PERCENTUAL DE MATRÍCULAS EM CURSOS DE GRADUAÇÃO, POR CATEGORIA ADMINISTRATIVA – 2012-2022

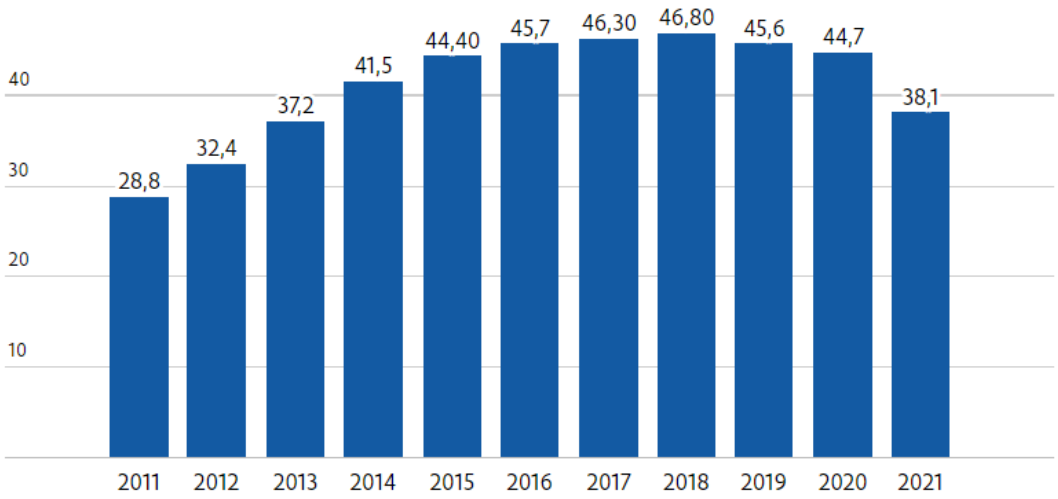
Fonte: Elaborado pela Deed/Inep com base em microdados do Censo da Educação Superior (Brasil. Inep, 2023a).

Projetos Atuais do Governo Federal com intuito de democratizar mais a educação, como PROUNI (Souza, M. R. D. A., & Menezes, M. (2014)) e o FIES (Queiroz, J. C. (2018)) têm financiado alunos em universidades privadas, resultando em uma diminuição na evasão universitária, por questões financeiras (Teixeira, R. D. C. P., Mentges, M. J., & Kampff, A. J. C. (2019)). Em 2021, o Prouni representou apenas 17% das matrículas, Cerca de 2 milhões de matrículas em 2021 eram apoiadas por bolsas ou financiamentos, sendo 221.589 pelo Fies e 478.651 pelo Prouni (dados divulgados em 2021 pelo INEP e PROUNI). O PROUNI, nos últimos anos, proporcionou conquistas ao crescimento da educação do país, especialmente para a população de baixa renda, que possui desafios para entrar no ensino superior (Mendes, G. M., & Mulin, H. D. P. (2017)). A Figura 2 destaca a relação anual entre a porcentagem anual de alunos matriculados com algum tipo de bolsa. A Figura 3 mostra esse aumento de matriculados, durante os últimos anos.

Figura 2

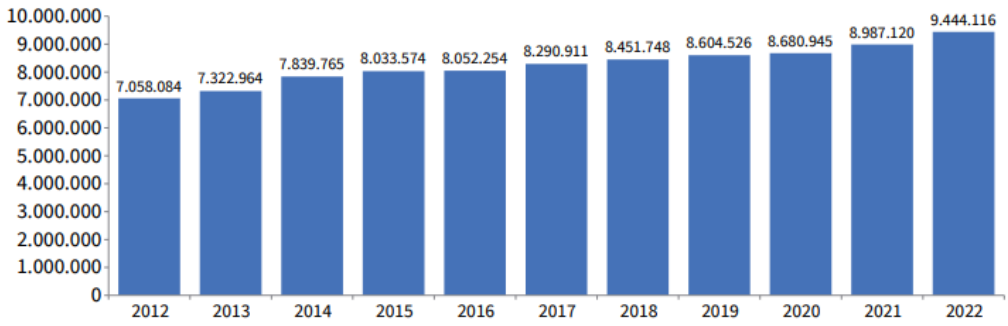
Financiamento do ensino superior é o menor desde 2013

Ano % de matrículas de graduação na rede privada com algum tipo de financiamento/bolsa



Fonte: Censo do Ensino Superior 2021/Inep

Figura 3



NÚMERO DE MATRÍCULAS NA EDUCAÇÃO SUPERIOR (GRADUAÇÃO E SEQUENCIAL) – 2012-2022

Fonte: Elaborado pela Deed/Inep com base em microdados do Censo da Educação Superior (Brasil. Inep, 2023a).

A evasão universitária se refere à desistência dos estudantes de concluírem seus cursos, após terem iniciado seus estudos (Silva Filho, R. L. L., Motejunas, P. R., Hipólito, O., & Lobo, M. B. D. C. M. (2007)). No Brasil, a média de evasão no ensino superior é de 22%. Entretanto, cursos como os da área de Computação, apresentam taxas ainda mais elevadas. Conforme o Sindicato das Entidades Mantenedoras de Estabelecimentos de Ensino Superior no Estado de São Paulo (SEMESP), apenas um terço dos alunos dos cursos de Computação, alcançam a formatura (Hoed, R. M. (2016)). Embora a falta de recursos financeiros possam ser um dos fatores que contribuem para a evasão, outros atributos podem contribuir para desistência do aluno, como o contexto social, cultural, político e econômico em que a instituição está incluída (Baggi, C. A. D. S., & Lopes, D. A. (2011)).

Anualmente, cerca de 30% dos alunos Rque fazem curso superior estão matriculados em cursos que não se encaixam em suas primeiras escolhas vocacionais. Dessa forma, contribui para a compreensão do crescente número de alunos insatisfeitos com o curso, principalmente nas disciplinas de Computação (Almeida, Soares, & Ferreira, 2002).

Portanto, a evasão escolar acarreta consequências importantes para instituições de ensino públicas e privadas. No âmbito econômico, ela gera perdas de receita e orçamento, prejudicando o funcionamento da instituição (PRESTES, E. M. D. T., FIALHO, M., & PFEIFER, D. (2016)).

No âmbito social, ela compromete a função da instituição de formar pessoas para contribuir para a sociedade e o mercado de trabalho. Em caso de instituições privadas, a evasão pode levar à falência, dependendo do prejuízo causado (Fialho, M. G. D. (2014)). Dessa forma, as desistências dos alunos nos cursos de graduação podem prejudicar a economia brasileira pela diminuição da geração de inovações tecnológicas e produtividade, além disso, o mercado atual exige maior diferencial competitivo, devido à crescente globalização e avanços tecnológicos. (Assunção, Y. B., & Goulart, I. B. (2016)).

A tecnologia está cada vez mais no cotidiano das pessoas. Em decorrência da inteligência artificial, sua particularidade essencial para qualquer sistema inteligente é a aprendizagem. Aprender é o método de aprimorar o desempenho ou o comportamento por meio da execução repetitiva, ou experiência (dos Santos, F. A. B. (2021)).

O aprendizado de máquina busca padrões em um conjunto de variáveis com o propósito de prever um resultado. Utiliza algoritmos com o procedimento lógico de inteligência artificial, são três fases: pré-processamento, treinamento e avaliação do modelo. A primeira fase envolve a organização do banco de dados, a segunda fase é responsável por determinar a pergunta de pesquisa e dividir os dados em treinamento e teste, por fim, a avaliação se o modelo é eficiente (Paixão, G. M. D. M., Santos, B. C., Araujo, R. M. D., Ribeiro, M. H., Moraes, J. L. D., & Ribeiro, A. L. (2022)). O aprendizado do teste pode ser dividido em supervisionado e não-supervisionado. No supervisionado os rótulos da classe já são conhecidos. Porém, no aprendizado não-supervisionado, analisa-se exemplos fornecidos e tenta determinar agrupamentos, e fazer uma análise para determinar o que cada agrupamento significa no cenário do problema (Monard, M. C., & Baranauskas, J. A. (2003).).

Técnicas de *Machine Learning* são empregadas no âmbito dos procedimentos de *Knowledge Discovery in Databases (KDD)*. KDD é o processo de extração de dados com intuito da obtenção de informação, transformadas em uma base de dados comum. Depois, são pré-processados e seguem para a etapa de mineração de dados, que desenvolve uma saída no formato de modelo (BELENKE DOS SANTOS, J. C. (2021)).

A previsão da evasão escolar pode ser relacionada a encargo da mineração de dados denominada classificação, que tem como proposito a atribuição de uma categoria a cada elemento examinado, a partir de um grupo de características, intrínseco ao mesmo elemento (de Oliveira Júnior, J. G., Noronha, R. V., & Kaestner, C. A. A. (2017)). Dessa forma, é uma das escolhas mais eficientes para extrair conhecimento a partir abundância de dados, encontrando relações ocultas, padrões e gerando regras para predizer e correlacionar dados, que podem auxiliar as instituições nas tomadas de decisões (Galvão, N. D., & Marin, H. D. F. (2009).).

Objetivo principal dessa pesquisa é desenvolver um modelo de previsão com técnicas de *Machine Learning* e procedimentos KDD, capaz de identificar os principais fatores que levam à evasão escolar e prever quais alunos têm maior tendência a sair dos cursos de graduação de uma instituição de ensino superior privada. Dessa forma, permite à instituição a implementação de ações preventivas para evitar a problemática.

Trabalhos Relacionados

Os conceitos de mineração de dados têm sido empregados por pesquisadores da área de Computação na Educação, com o propósito de detectar padrões na busca de modelos de previsão da evasão de curso.

É o caso, do trabalho de (BELENKE DOS SANTOS, J. C. (2021)), com objetivo de utilizar algoritmos de *Machine Learning* para prever evasão dos estudantes do Instituto Federal de Santa Catarina – IFSC. Logo, foram utilizadas 2 as técnicas, Árvore de Decisão obtendo 87% de precisão na detecção de evasão e Redes Neurais com 78% de precisão. Portando a Árvore de Decisão, obteve melhor resultado.

Com o Propósito de descobrir as variáveis significativas que influenciam evasão e permanência no curso de graduação. Aplicando no banco de dados da a UFRJ (Universidade Federal do Rio de Janeiro), 6 tipos de algoritmos de *Machine Learning*. Dessa maneira, o que demonstrou melhor resultado foi o *Naïve Bayes* com a acurácia por volta de 80% (Manhães, L. M. B., da Cruz, S. M. S., Costa, R. J. M., Zavaleta, J., & Zimbrão, G. (2012, May)).

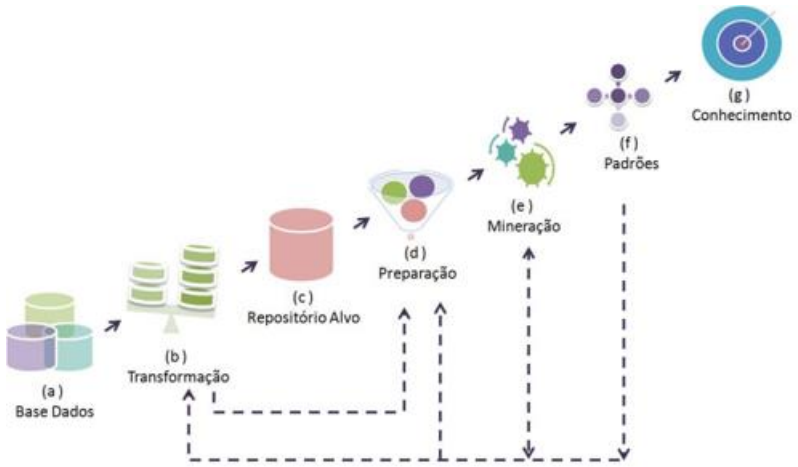
O trabalho de (de Oliveira Júnior, J. G., Noronha, R. V., & Kaestner, C. A. A. (2016)) explora técnicas de mineração de dados educacionais da Universidade Tecnológica Federal do Paraná. O objetivo principal propõe um modelo de previsão do abandono escolar, empregando classificação, criação e seleção de atributos. Dessa forma, os resultados foram verificados que aproximadamente 80% da evasão de curso ocorre no 3° semestre, com a acurácia, com resultados entre 83 e 87%.

MÉTODOS

Esse estudo emprega metodologia KDD. Pois, sua vantagem é permitir integração de maneira mais fácil e rápido absorção de informações e transformá-las em conhecimento (Boente, A. N. P., & Rosa, J. L. D. A. (2007)).

O procedimento KDD (Winck, A. T., Machado, K. S., Ruiz, D. D., & de Souza, O. N. (2010)). é uma série de etapas interativas e iterativas voltada à tomada de decisão. De maneira que, a principal técnica de análise envolvido no processo é a mineração de dados (BELENKE DOS SANTOS, J. C. (2021)). Dessa forma, com os dados estruturados na base de dados alvo, serão minerados, passam pelas seguintes etapas de acordo com a Figura 4.

Figura 4



As etapas (a), (b) e (c) são usadas na construção de uma base de dados alvo, usando as seguintes técnicas:

Limpeza de dados: São retirados os dados que estão discrepantes na base de dados ou fora do padrão, ou se utiliza de alguma técnica para tentar solucionar o problema de dados faltantes.

Integração de dados: Várias fontes de dados podem se tornar integradas, de maneira a enriquecer os dados originais e proporcionar alternativas à obtenção de informações.

Seleção dos dados: Os dados significativos para a alternativa do problema são coletados desta base de dados.

As etapas (d), (e), (f) e (g) executadas na mineração de dados.

Preparação nos dados: Consolidação dos dados.

Mineração de Dados e Padrões: Essa etapa, utiliza técnicas de *Machine Learning* para classificação, como algoritmo de classificação *bayesiano*, de vizinho mais próximo e árvore de decisão.

Conhecimento: Representação e visualização do conhecimento para apresentar conhecimento obtido.

Para prosseguir com a metodologia KDD, devemos primeiramente obter ou criar um conjunto de dados, conforme mostra a Figura 2. Para isso, foram escolhidos diversos atributos relevantes dos alunos, por meio das seguintes análises de artigos sobre a educação e evasão escolar.

Conforme os resultados do trabalho (Stratton, L. S., O'Toole, D. M., & Wetzel, J. N. (2008)), as variáveis mais significativas forma:

Endereço, pela distância, da casa do aluno até o instituto de ensino. Alunos que moram em maiores distâncias, podem evadir mais facilmente;

Renda pessoal, alunos de baixa renda, tendem a evadir do curso;

Idade, Homens mais velhos são mais propensos do que os mais jovens a desistir;

Estado civil, caso o aluno for casado, propende a evadir;

Segundo as descobertas da pesquisa Sampaio, B., Sampaio, Y., de Mello, E. P., & Melo, A. S. (2011)), as variáveis importantes, foram;

Renda dos pais, quanto menor a renda dos responsáveis, maior a probabilidade de o estudante evadir.

Casado, alunos casados inclinam a saída do curso;

Gênero, os resultados mostram que mulheres têm menor possibilidade de evasão;

Idade, quanto mais velho, maior a propensão a evasão;

desempenho acadêmico, quanto pior o desempenho, maiores as chances de o estudante evadir;

Na Estudo (Maria, W., Damiani, J. L., & Pereira, M. (2016, November)) as variáveis estudas foram:

Cidade, onde o aluno mora;

Tipo de vaga;

Raça;

Órfão;

Tipo de bolsa;

Frequência, quantidade de faltas em porcentagem;

Faixa etária;

Lotação da turma;

Situação do aluno, caso ele tenha evadido ou não;

Sexo;

Situação ocupacional, para identificar a situação do aluno se no momento ele trabalha ou está aposentado, autônomo, busca de emprego;

Dessa forma, as variáveis atribuídas para essa pesquisa, serão escolhidas conforme as variáveis mais relevantes de pesquisas correlacionadas apresentadas.

As variáveis escolhidas foram:

Nome;

Registro Acadêmico;

Curso;

Semestre;

Data de aniversário;

Sexo;

Etnia;

Bairro;

Mora com os pais;

Quantidade de filhos;

Estado Civil;

Situação Ocupacional (se tem algum emprego ou estágio)

Renda Própria;

Renda dos responsáveis;

Escolaridade dos Responsáveis;

Frequência;

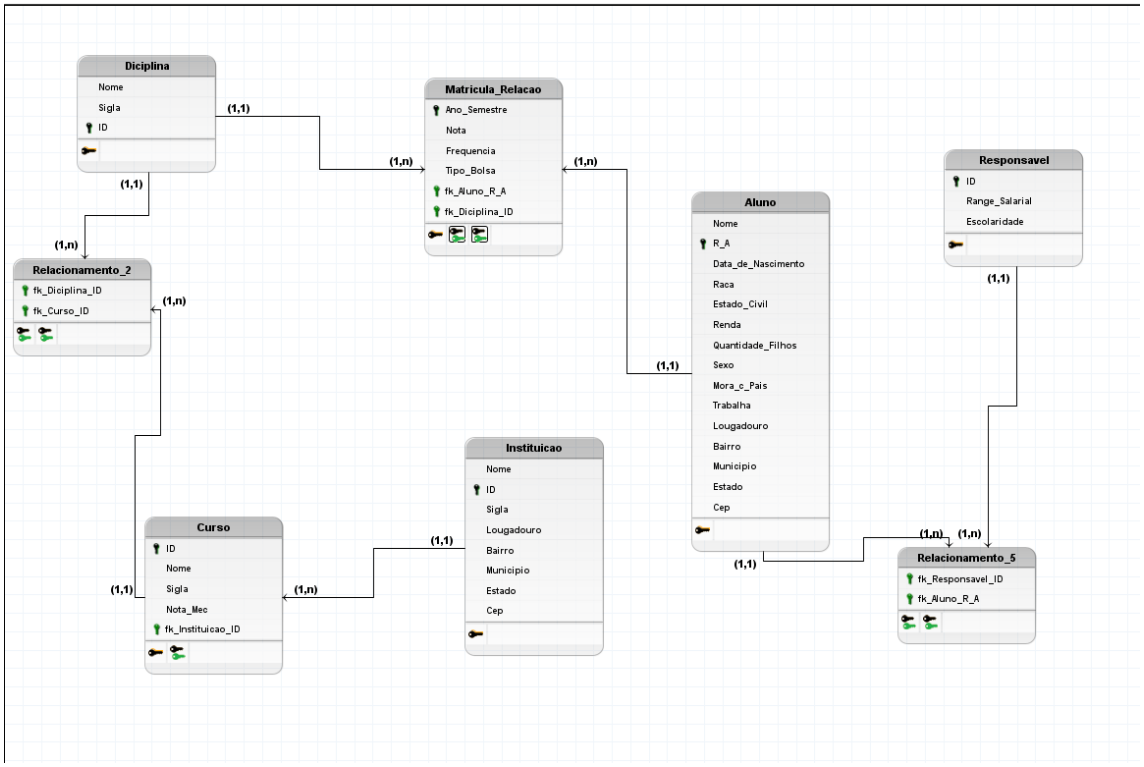
Matérias cursadas;

Notas Bimestrais;

Tipo de bolsa;

A modelagem de dados procura representar e estruturar o conhecimento na criação de uma estrutura de dados. É fundamental permitir a transformação dos dados originais em diversos resultados como formulários, relatórios ou gráficos (Silva, M. B. D. (2011)).
A manipulação dos atributos será necessária, pois precisamos de uma estruturação para o processo de mineração de dados. Dessa forma a modelagem, foi organizada de maneira que as tabelas tenham conexões entre elas, o que favorece a coleta e armazenagem dos dados. A seguir, a Figura 5 abaixo ilustra o modelo.

Figura 5



As informações de cada tabela, serão coletadas na base de dados da instituição. Entretanto, pela A LEI DE PROTEÇÃO DE DADOS, por se tratar de dados pessoais sensíveis, deverá ter consentimento dos alunos a utilização de seus dados que serão tratados para o projeto de pesquisa (Mendes, L. S., & Doneda, D. (2018)).

Porém, informações dos campos ‘Aluno’ e ‘Responsável’ não se encontram na base de dados do Instituto onde será realizada a pesquisa, os atributos que faltam são:

Etnia;

Estado;

Município;

Logradouro;

Bairro;

Mora com os pais;

Quantidade de filhos;

Situação Ocupacional (se tem algum emprego ou estágio)

Estado Civil;

Renda Própria;

Renda dos responsáveis;

Escolaridade dos Responsáveis;

Trabalha;

Dessa forma, para obter a autorização dos alunos, do uso de seus dados na base de dados da faculdade e obter variáveis faltantes, será necessário o desenvolver de um formulário, contendo essas informações necessárias para o progresso do projeto.

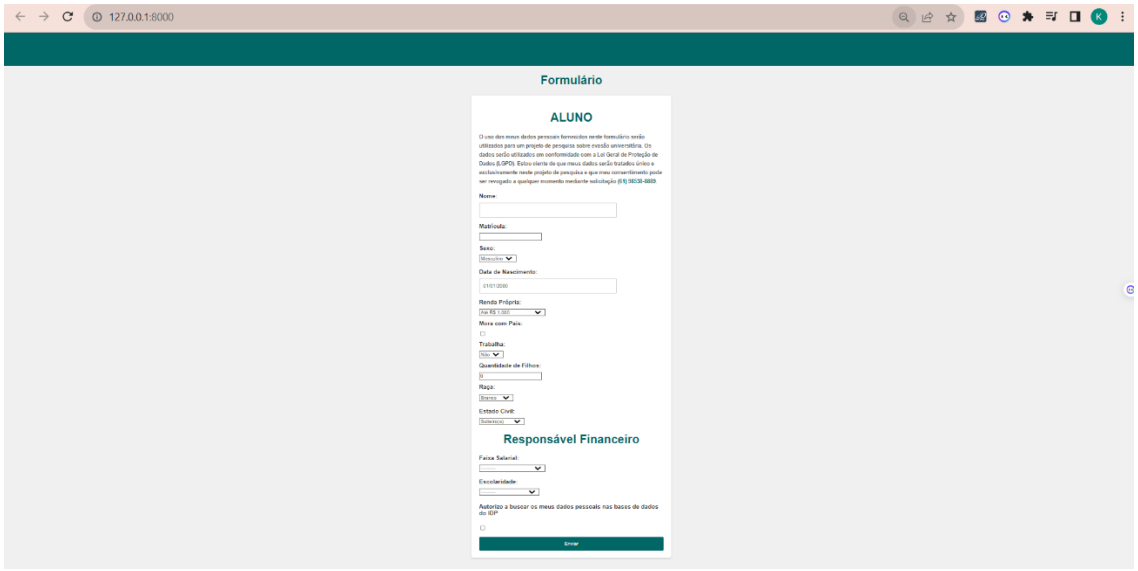
A coleta de dados é um dos passos mais importantes da realização de uma pesquisa, pois obtém as informações necessárias para o desenvolvimento do estudo. Porém, para coletar corretamente as informações, é desafio ao pesquisador escolher as ferramentas de coleta de dados que concedam aos seus objetivos, pois é necessário certificar a confiabilidade da pesquisa (Oliveira, J. C. P. D., Oliveira, A. D., Morais, F. D. A. M., Silva, G. M. D., & Silva, C. N. M. (2016, October)).

Para a coleta de dados, foi necessária uma criação do formulário eficiente, contendo as informações faltantes, de forma que possam ter opções pré-determinadas, evitando esforços nos primeiros processos do KDD, a limpeza de dados, pois todos os dados terão nomes padronizados.

Esse formulário, foi desenvolvido para uma aplicação web, feito no *framework Django* do *Python*. Logo, apresenta apenas as informações que devem ser preenchidas obrigatoriamente pelo usuário, evitando redundâncias e potencializando o processo de inserção de dados. Assim, é possível a automatização da análise instantânea, somente visitando a *URL* do *website* que estará hospedado em um serviço de nuvem e introduzindo dados de entrada esperados para análise.

Foi desenvolvido o formulário, e sua interface principal para os alunos preencher os dados, ficou conforme a Figura 6:

Figura 6

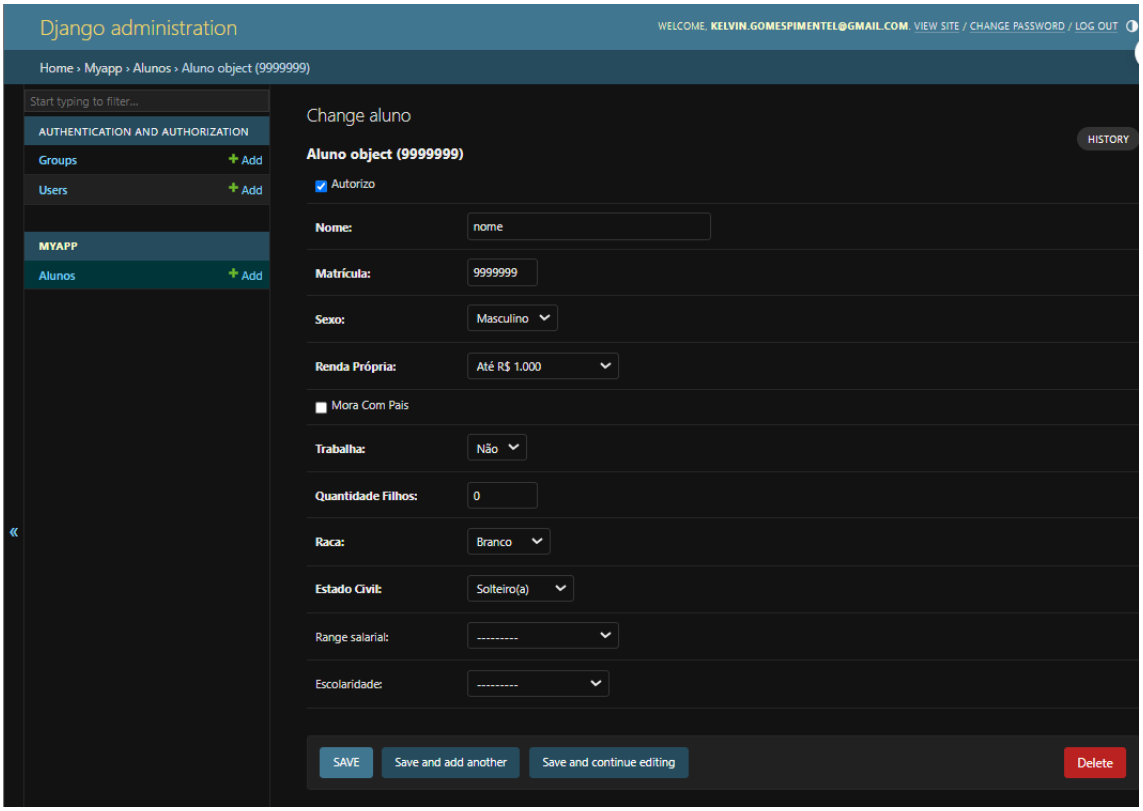


Após preenchimento, o estudante deve clicar no botão 'enviar' que o levará para uma página de agradecimento, conforme a Figura 7, automaticamente o *Django* alimenta o seu banco de dados *SQLITE* com as novas informações preenchidas. Seu banco de dados no final pode ser extraído em *Excel*, o que poderá facilitar ao conectar com dados que a instituição disponibilizará apenas para os alunos que optaram em aceitar a autorização do uso de seus dados para a pesquisa.

Figura 7

O Empregadas em bancos de dados relacionais (BDR), o CRUD, são operações de inserção, de consulta, de atualização e de exclusão de dados, acrônimo de (*Create, Retrieve, Update and delete*) (Arend, F. G. (2011).). A vantagem da ferramenta *Django*, é a possibilidade de utilizar o CRUD. A Figura 8 abaixo, mostra uma área administrativa do *Django* que foi feita o formulário, a qual pode usar as operações do CRUD.

Figura 8



Visando integrar diferentes ambientes, a *Microsoft* lançou um sistema operacional destinado à nuvem, denominado *Windows Azure*. Dessa forma, os datacenters da Microsoft, disponibiliza a capacidade e recursos de programação para as aplicações (CORTIJO, P. D. CURSO DE ESPECIALIZAÇÃO EM ENGENHARIA DE SOFTWARE).

Entretanto, ao buscar criar e implantar um formulário usando *Django* em uma máquina virtual na plataforma *Azure*, ocorreram desafios técnicos que impactaram o sucesso da implementação. Essa experiência destaca a importância de compreender a integração entre *frameworks web* e ambientes em nuvem, e como as complexidades podem surgir mesmo em cenários que inicialmente parecem diretos (CORTIJO, P. D. CURSO DE ESPECIALIZAÇÃO EM ENGENHARIA DE SOFTWARE).

O *Google Forms* oferece uma solução mais simples e prontamente acessível para a criação de formulários online. Sua interface intuitiva e a integração nativa com outros serviços do Google proporcionaram uma alternativa eficaz, permitindo-me concentrar mais na coleta de dados do que na resolução de questões técnicas. Dessa forma, foi a segunda plataforma escolhida para coleta de dados. Com isso, conseguimos coletar dados de 99 estudantes. Porém, seis estudantes não aceitaram que seus dados fossem buscados no banco de dados da instituição.

Um conjunto de dados com desequilíbrio nas variáveis alvo, seja de forma intensa ou escassa, algoritmos de Aprendizado de Máquina podem ter dificuldades em distinguir classes menos e mais frequentes. Isso resulta em rótulos excessivos na classe majoritária, levando a uma acurácia enganosamente alta. Para lidar com esse desafio, o balanceamento da quantidade de seleção das características se destaca como uma abordagem eficaz, identificando características cruciais que contribuem para reduzir os efeitos adversos do desequilíbrio de classes no desempenho da classificação (Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., & Seliya, N. (2019)).

Na fase de Integração de dados do KDD, além do formulário destinado aos alunos ainda matriculados na faculdade, a instituição de ensino disponibilizou os dados dos alunos que consentiram, marcando "sim", para a autorização de acesso ao banco de dados institucional. Os dados fornecidos foram, Data de Nascimento, Semestre, Curso, Disciplinas cursadas, notas das Avaliações 1 e 2 (Av1 e Av2) destinadas ao 1° bimestre

E 2° bimestre, quantidade de faltas, tipo de ingresso, se possuem Bolsa, e a percentagem de Bolsa concedida (% bolsa). No entanto, foi fornecido também as mesmas informações referentes aos alunos que evadiram, mantendo o cuidado de excluir nomes ou matrículas a fim de estar em conformidade com as leis de proteção de dados.

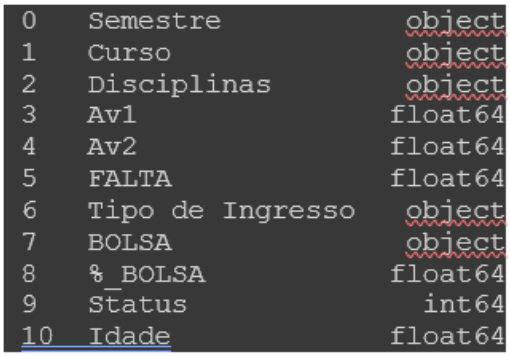
Dessa forma, foram obtidas três planilhas: a primeira, composta pelos dados do formulário; a segunda, derivada dos alunos que autorizaram o acesso aos dados institucionais; e a terceira, referente aos alunos evadidos. Além disso, foi criada uma quarta planilha que contém informações tanto do formulário quanto da base de dados da faculdade. Essa integração foi realizada por meio de junções utilizando as matrículas dos alunos, possibilitando a criação de uma planilha mais enriquecida.

A planilha referente aos alunos evadidos passou por um processo de filtragem. Essa abordagem consistiu na seleção de uma quantidade específica de alunos evadidos, equilibrando-a com a quantidade de alunos presentes no respectivo curso. Tal medida promove um balanceamento adequado entre as duas categorias, facilitando análises mais precisas e minimizando possíveis vieses na modelagem preditiva.

Durante o processo de preparação dos conjuntos de dados, ocorreram duas combinações para obter na mesma planilha dados dos alunos que evadiram e os que estão presentes na faculdade. A primeira ocorreu entre a planilha fornecida pela instituição dos alunos nas quais autorizaram o acesso aos dados e a planilha dos alunos evadidos, garantindo uma uniformidade, pois as colunas entre ambas eram iguais, a Figura 9 mostra a Tabela 1, na qual foi usada. Posteriormente, ocorreu uma segunda junção, envolvendo a planilha resultante da fusão dos dados do formulário com a base da instituição, juntamente com a planilha dos alunos evadido, representada na Figura10, na qual mostra a Tabela 2. Entretanto, devido à ausência de dados fornecidos pelos alunos evadidos no formulário, há lacunas nessa junção, resultando em informações incompletas para esse subconjunto de dados. Com isso, temos dois conjuntos de dados.

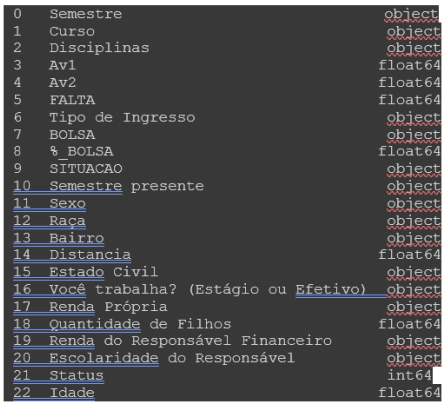
Em seguida foram implementadas medidas nos dois conjuntos de dados, como a subtração da data de nascimento pela data atual foi efetuada, resultando na obtenção da idade dos alunos, assim foi feito a exclusão da coluna de data de nascimento. Com isso, contribuiu para a simplificação da estrutura da planilha, eliminando redundâncias e facilitando a manipulação dos dados.

Figura 9 Figura 10



0	Semestre	object
1	Curso	object
2	Disciplinas	object
3	Av1	float64
4	Av2	float64
5	FALTA	float64
6	Tipo de Ingresso	object
7	BOLSA	object
8	%_BOLSA	float64
9	Status	int64
10	Idade	float64

Tabela 1



0	Semestre	object
1	Curso	object
2	Disciplinas	object
3	Av1	float64
4	Av2	float64
5	FALTA	float64
6	Tipo de Ingresso	object
7	BOLSA	object
8	%_BOLSA	float64
9	SITUACAO	object
10	Semestre presente	object
11	Sexo	object
12	Raça	object
13	Bairro	object
14	Distancia	float64
15	Estado Civil	object
16	Você trabalha? (Estágio ou Efetivo)	object
17	Renda Própria	object
18	Quantidade de Filhos	float64
19	Renda do Responsável Financeiro	object
20	Escolaridade do Responsável	object
21	Status	int64
22	Idade	float64

Tabela 2

A abordagem mais comum para converter características categóricas para um formato adequado para uso como entrada em um modelo de aprendizado de máquina é a codificação *One-hot*. Uma vez que os dados usando a codificação *One-hot* são de natureza numérica, um modelo de aprendizado de máquina pode incorporar facilmente informações de características categóricas aprendendo um parâmetro separado, para cada dimensão. (Seger, C. (2018)).

Para utilizar modelos de ML, é necessário converter variáveis categóricas em variáveis numéricas, uma vez que muitos modelos ML operam apenas com dados numéricos. O *Python* tem a pandas oferece métodos como o *Get_Dummies* para realizar essa conversão (Development of a single retention time prediction model integrating multiple liquid chromatography systems: Application to new psychoactive substance).

A etapa de categorização representou um dos últimos passos no processo de pré-processamento dos dados, focalizando a conversão de variáveis categóricas, tais como Semestre, Curso, Disciplinas, Tipo de Ingresso e Bolsa, para representações numéricas. Durante esse procedimento, cada categoria específica dentro de uma coluna foi mapeada para um valor numérico, sendo atribuído o valor 1 à categoria correspondente, enquanto as demais categorias na mesma coluna receberam o valor 0. Essa técnica de transformação

foi aplicada utilizando o método *Get_Dummies*, permitindo uma representação adequada das variáveis categóricas nos conjuntos de dados.

Ao lidar com a planilha resultante da fusão dos dados provenientes do formulário, da base da instituição e da planilha dos alunos evadidos, que incluiu informações adicionais, como Sexo, Raça, Bairro, Distância, Estado Civil, você trabalha? (Estágio ou Efetivo), Renda Própria, Quantidade de Filhos, Renda do Responsável Financeiro, Escolaridade do Responsável, Status e Idade, uma abordagem semelhante de categorização foi aplicada, porém foi usado apenas o *Get_Dummies*. Cada uma dessas variáveis categóricas foi convertida para representações numéricas, seguindo o mesmo princípio de atribuir o valor 1 à categoria específica e 0 às demais, destacando a principal coluna, a Status, em que 0 são os alunos que estão cursando na instituição e 1 representa alunos evadidos. Isso foi feito para manter a consistência na representação dos dados e garantir a aplicabilidade dos algoritmos de aprendizado de máquina posteriormente.

A Classificação envolve o uso de um conjunto de dados que contém exemplos de entrada e saída, permitindo que o modelo aprenda. O modelo utiliza o conjunto de dados de treinamento para determinar a melhor forma de associar exemplos de dados a rótulos específicos de classe. Portanto, é essencial que o conjunto de dados de treinamento seja representativo do problema, incluindo diversos exemplos de cada classe. No contexto de *Machine Learning*, a classificação refere-se a um problema de modelo preditivo no qual uma classe, rótulo ou etiqueta é prevista ou atribuída a dados de entrada específicos (Garzillo, M. J. W. (2022)).

A matriz de confusão é uma tabela que visualiza acertos e erros em cada classe, mostrando se as classificações foram corretas. Em problemas de classificação binária, essa matriz tem duas linhas e colunas, representando verdadeiro-negativo, falso-positivo, verdadeiro-positivo e falso-negativo. A diagonal principal exibe os verdadeiros acertos, enquanto a diagonal oposta mostra os erros de classificação. Por exemplo, a Figura 11 mostra Tabela 3 apresenta verdadeiros-positivos e verdadeiros-negativos corretos, além de falso-positivos e falso-negativos indicando classificações incorretas feitas pelo modelo (BELENKE DOS SANTOS, J. C. (2021)).

Figura 11

Classificação correta	Classificado como	
	+	-
+	Verdadeiro Positivo (VP)	Falso Negativo(FN)
-	Falso Positivo (FP)	Verdadeiro Negativo(VN)

Tabela 3

O *Overfitting* é um desafio frequente e recorrente na fase de treinamento de algoritmos de *Machine Learning*. Esse problema surge quando um modelo atinge um desempenho notavelmente alto durante o treinamento, mas, ao ser confrontado com dados desconhecidos, sua performance decai significativamente. Isso evidencia a falta de capacidade de generalização do modelo, assemelhando a uma situação em que o modelo simplesmente "memoriza" os dados de treinamento (BELENKE DOS SANTOS, J. C. (2021).).

Árvore de Decisão é uma técnica de mineração de dados conhecida como considerada uma forma simples de representar relações em conjuntos de dados, sendo amplamente utilizada para fins de classificação (Souza, N. D. C., Pitombo, C., Cunha, A. L., Larocca, A. P. C., & de Almeida, G. S. (2017)). O treinamento ocorre por meio de exemplos supervisionados, sendo as árvores consideradas não-paramétricas. No cenário supervisionado, o método tem acesso tanto aos exemplos quanto à resposta final do problema a ser solucionado. Porém, quanto à não-paramétrico, não é necessário possuir conhecimento antecipado sobre a forma da função de mapeamento (Baptista, D. F. F. (2019)).

A nomenclatura das árvores de decisão é adotada de modo semelhante à estrutura de uma árvore. A interpretação dessa árvore ocorre de forma descendente, iniciando com o atributo mais importante no nó principal, o primeiro nó. Os próximos atributos são identificados como subnós ou nós internos, ao passo que as escolhas a serem feitas são expressas nos nós terminais, chamados de folhas. A técnica de poda é uma abordagem

que procura diminuir as dimensões da árvore ao remover subnós menos relevantes, simplificando-a e prevenindo *Overfitting* (Abreu, L. E. D. (2022)).

Para iniciar a mineração de dados, os conjuntos de dados foram divididos entre treinamento (80%) e teste (20%), permitindo a avaliação comparativa do desempenho dos modelos. O *DecisionTreeClassifier* do *Scikit-learn*, na linguagem Python, foi empregado para todas as 2 abordagens.

Na primeira mineração, com o *Dataset* Tabela 1, utiliza uma profundidade máxima de 4 níveis da árvore de decisão para evitar complexidade excessiva e prevenir possíveis problemas de *Overfitting*. O treinamento do modelo resultou em uma acurácia de aproximadamente 77%, com uma precisão de 73% para a classe 0 (alunos cursando) e 90% para alunos evadidos.

No segundo treinamento, referente a Tabela 2, na qual foram aplicadas aos dados resultantes da fusão entre as informações do formulário e a base de dados da instituição, junto com a planilha dos alunos evadidos. No entanto, identificou-se uma situação de *Overfitting* devido à ausência de dados dos alunos evadidos no formulário, causando disparidades nos conjuntos de treinamento e teste.

Ao avaliar o desempenho do modelo, uma acurácia surpreendente de 100% foi observada, indicando suspeitas de *Overfitting* devido à falta de representatividade dos alunos evadidos nos dados do formulário. A matriz de confusão mostrou previsões corretas para todas as instâncias no conjunto de teste, sugerindo uma precisão de 100% para ambas as classes. No entanto, é crucial destacar que o modelo pode ter se ajustado excessivamente a padrões específicos nos dados de treinamento, comprometendo sua generalização para novos casos.

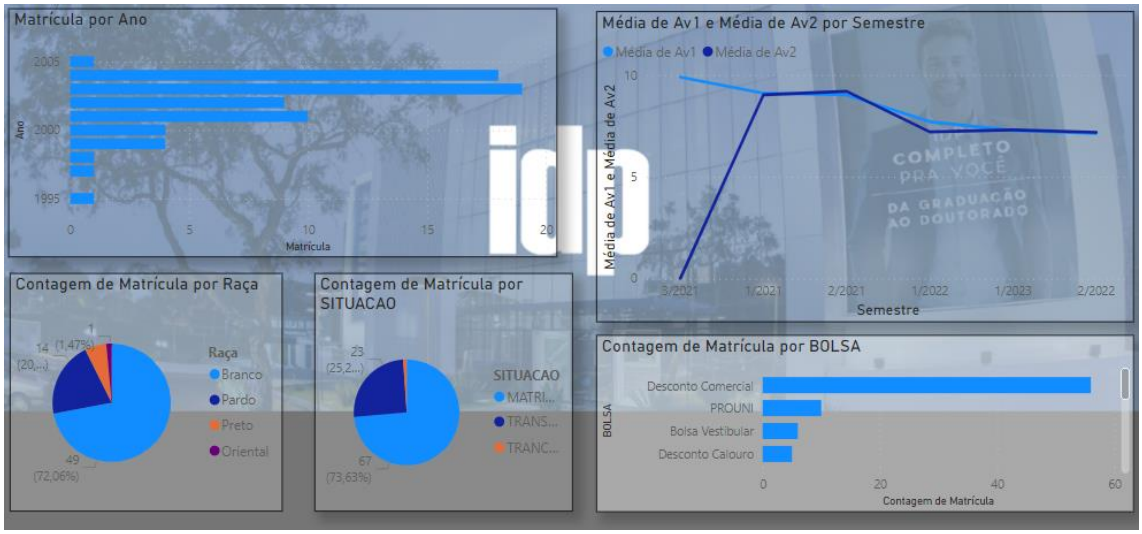
Compreender a Estatística vai além da elaboração de gráficos e cálculos de médias. O objetivo da obtenção de informações numéricas é fundamentar a tomada de decisões. Dessa forma, a estatística se apresenta como um conjunto de técnicas para planejar experimentos, coletar, organizar, resumir, analisar, interpretar dados e extrair conclusões significativas. A análise exploratória de dados oferece uma variedade de métodos para uma investigação detalhada antes de realizar ajustes nos dados (MEDRI, W. (2011)).

Em meio a esse contexto, os sistemas de *Business Intelligence* (BI) se destacam ao utilizar os dados disponíveis nas organizações para fornecer informações relevantes no processo de tomada de decisões. Esses sistemas abrangem a coleta e integração de dados, o armazenamento desses dados em bases de dados, e a gestão do conhecimento por meio de diversas ferramentas analíticas e desempenham um papel crucial ao possibilitar a extração de informações valiosas a partir dos dados coletados, configurando-se como a componente analítica essencial para o processo decisório (Conceição, L. F. M. D. S. (2020).).

A exploração dos dados dos alunos que completaram o formulário, realizada através da plataforma de *Business Intelligence*, *Power BI*, proporcionou insights sobre diversos aspectos do perfil estudantil na instituição. Os resultados oferecem uma visão detalhada das características demográficas, padrões de ingresso, concessões de descontos, desempenho acadêmico e outras métricas relevantes. A Figura 12 apresenta graficamente os dados para análise visual.

Assim, foi desenvolvido um painel de *Business Intelligence* contendo a tabela na qual os alunos preencheram os formulários, permitindo uma análise de alguns atributos para a identificação do perfil tanto do aluno quanto da faculdade.

Figura 12

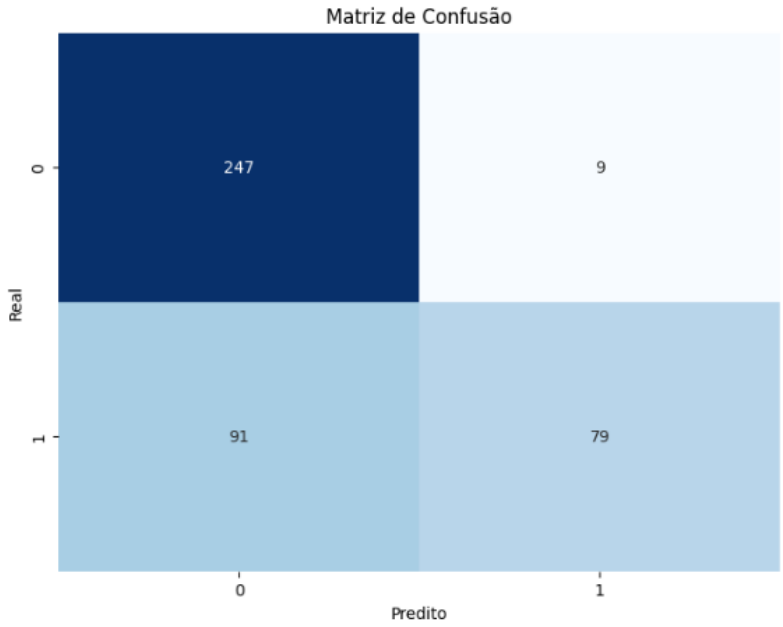


Resultados

O formulário inicialmente preenchido por 99 alunos teve 6 optando por não autorizar o uso de seus dados, resultando em 93 alunos utilizados no projeto. Contudo, 20 alunos matriculados ainda não registrados no sistema não tiveram seus dados utilizados, totalizando 73 alunos no projeto. Os alunos evadidos, sem matrícula e nome por questões de privacidade, foram selecionados (73 alunos) para testar o modelo preditivo, equilibrando os *Datasets* por curso.

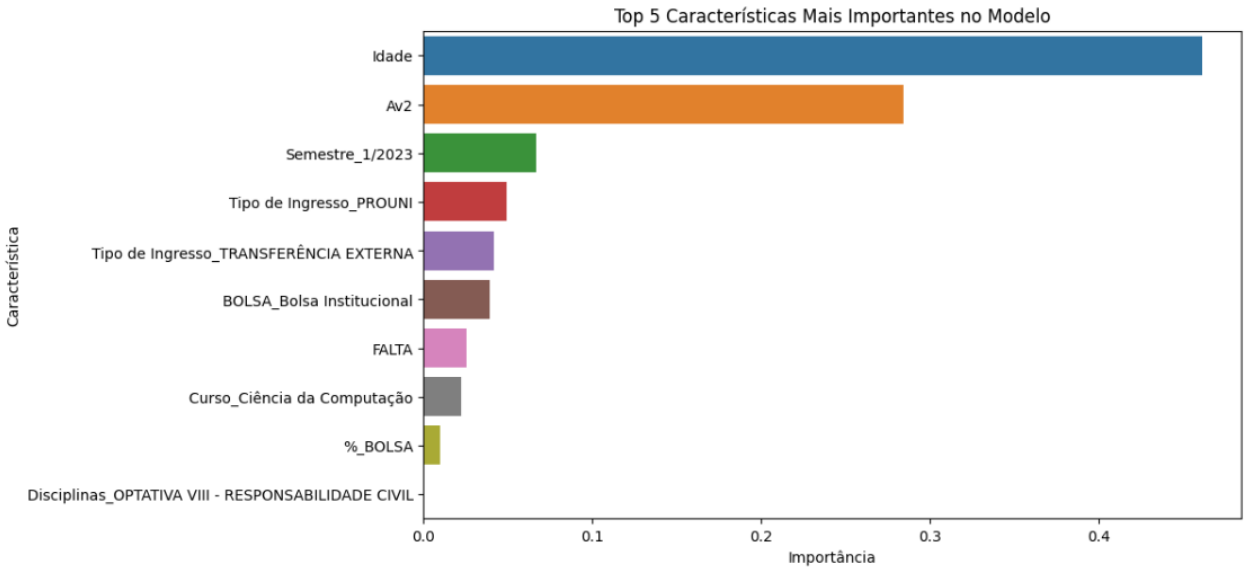
Os resultados do modelo de previsão da Tabela 1 (sem os dados do formulário) apresentaram boa precisão na previsão de evasão (acima de 73%). A acurácia do modelo foi de 0.77. A Matriz de Decisão do modelo revelou 257 verdadeiros positivos, 91 verdadeiros negativos, 79 falsos positivos e 9 falsos negativos. A Figura 13 a seguir apresenta a Matriz de Confusão.

Figura 13



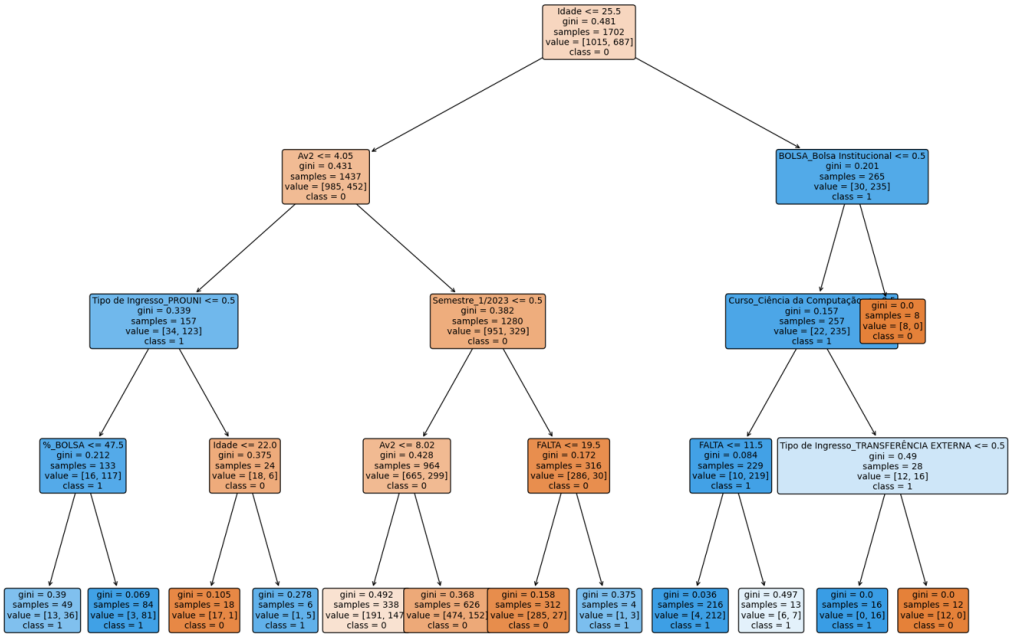
As principais colunas, ordenadas por importância, são: Idade, que representa 46% de relevância; seguida por Av2, com 28% de importância. Na sequência, destacam, o primeiro semestre, o tipo de Ingresso pelo PROUNI, Tipo de Ingresso transferência externa, Bolsa Institucional, Falta, Curso Ciência da Computação e porcentagem da BOLSA. A imagem ilustra essa importância das variáveis mais importantes para o aprendizado do modelo de previsão. A Figura 14 exibe a ordem das colunas mais relevantes.

Figura 14



A análise revela que a idade acima de 25 foi o primeiro nó mais relevante, seguido pela nota Av2. Alunos com notas abaixo de 4,05, bolsa institucional abaixo de 50% e que fazem Ciência da Computação tendem a sair. O segundo nó importante inclui bolsa menor que 47%, idade acima de 22, Av2 abaixo de 8 e falta abaixo de 19, indicando fatores de evasão. A Figura 15 exibe os nós e subnós da Árvore de Descisão.

Figura 15



Na Análise Exploratória de Dados dos estudantes que preencheram o formulário, é possível identificar uma compreensão do perfil e desempenho dos estudantes, fornecendo informações para tomadas de decisão estratégicas na instituição de ensino.

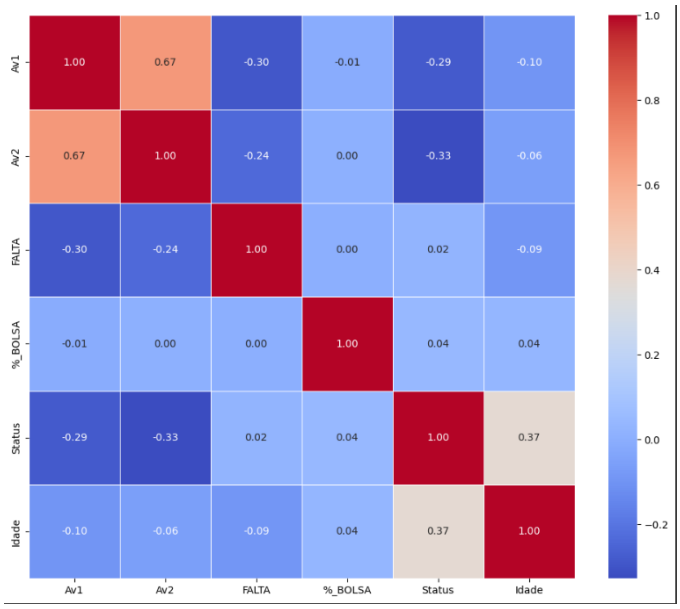
A relação entre duas variáveis é caracterizada pela correlação de dados, que se manifesta por meio de um número capaz de resumir o grau de interdependência entre essas variáveis (Baba, R. K., Vaz, M. S. M. G., & Costa, J. D. (2014)). Entretanto para ser considerada uma robusta correlação, deve ter um grau de correlação igual ou superior a 0,85. (Baba, R. K., Vaz, M. S. M. G., & Costa, J. D. (2014)). De acordo com a Figura 16, expõe a categoria das correlações.

Figura 16

Valor de ρ (+ ou -)	Interpretação
0.00 a 0.19	Correlação muito fraca
0.20 a 0.39	Correlação fraca
0.40 a 0.69	Correlação moderada
0.70 a 0.89	Correlação forte
0.90 a 1.00	Correlação muito forte

Na análise inicial dos estudantes, foi realizada uma correlação entre todas as colunas. No entanto, não foi observada uma correlação forte em nenhuma das relações. A correlação mais significativa foi identificada entre as notas Av1 e Av2. Portanto, a conclusão é que o desempenho positivo no primeiro semestre tende a estar associado a um bom desempenho no segundo semestre. A Figura 17, demonstra esse resultado.

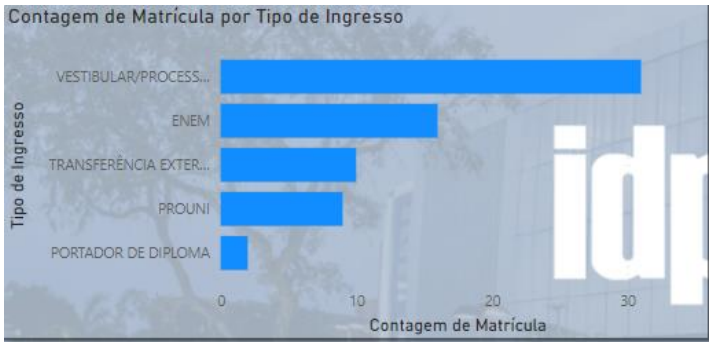
Figura 17



No que diz respeito aos métodos de ingresso, verificou que a maioria dos estudantes, representando 45%, optou por ingressar na instituição por meio do vestibular seletivo. Em seguida, o ingresso via ENEM correspondeu a 23%, transferência a 14%,

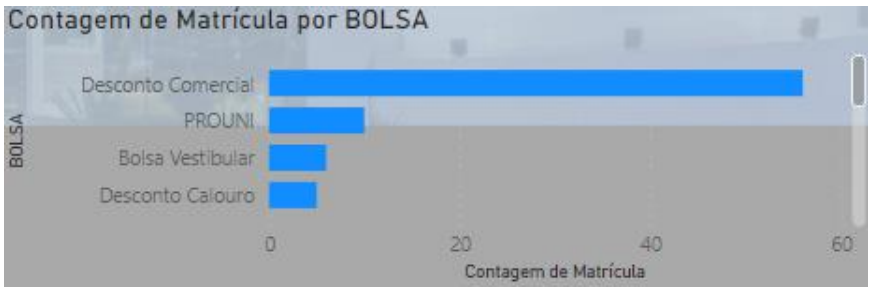
PROUNI a cerca de 13%, e, por último, portador de diploma com 3%. A Figura 18, expõe a análise.

Figura 18



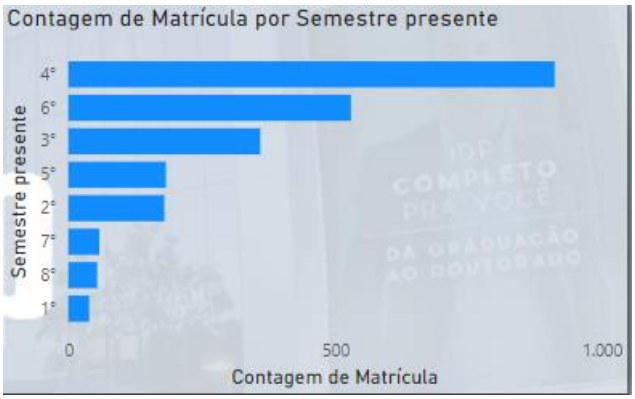
Quanto aos descontos, observou-se que a modalidade comercial foi a mais comum, representando 56% do total, seguida pelo PROUNI, que contribuiu com 11% dos descontos. Essa análise estratégia de captação e retenção de alunos, destacando as preferências dos estudantes em relação aos métodos de ingresso e benefícios financeiros oferecidos. Representada pela Figura 19.

Figura 19



A distribuição dos estudantes pelos semestres indicou que a maioria está no 4° semestre, seguido pelo 6° e 3°, enquanto o primeiro semestre apresentou o menor número de alunos. Entretando não podemos considerar ser uma relevância grande, pois não foram todos os estudantes da instituição que preencheram o formulário. A seguir, a Figura 20 ilustra a distribuição.

Figura 20

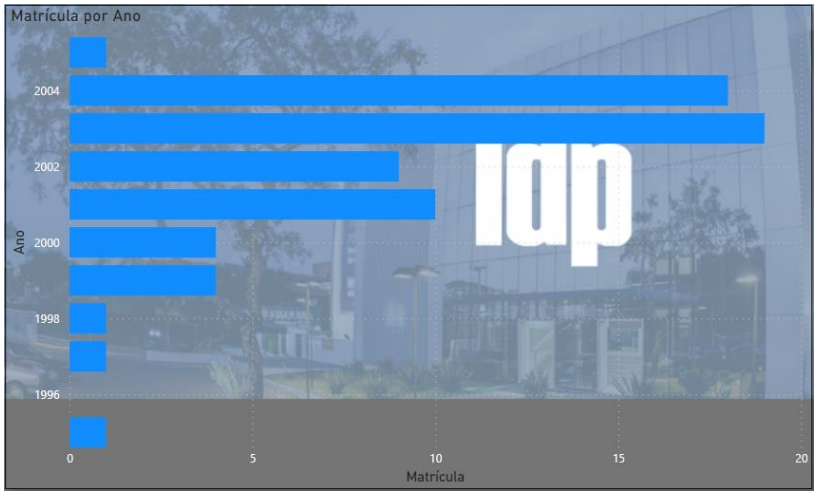


Um aspecto interessante diz respeito à distribuição das notas entre os bimestres. A média da nota da AV1 (primeiro bimestre) foi mais alta do que a AV2 (segundo bimestre) para a maioria dos estudantes. Esse padrão indica o desempenho dos alunos no segundo bimestre é menor, sendo uma informação valiosa para a equipe acadêmica. A Figura 21, indica o gráfico de linhas da média das notas por semestre.

Figura 21

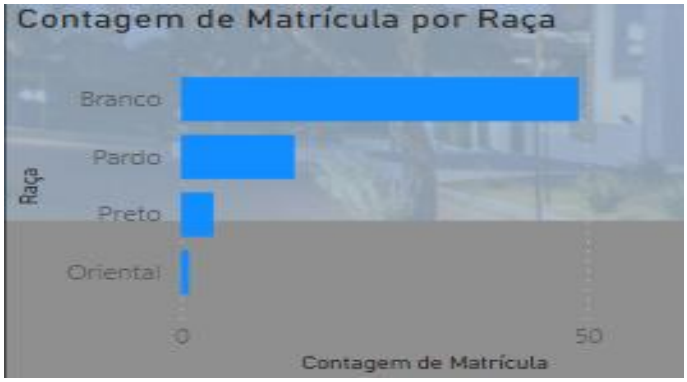
Analisando o ano de nascimento dos estudantes, destacam-se as porcentagens significativas para os anos de 2003 (27%) e 2001 (22%). Os anos de 2002 e 2004 contribuíram com 14% e 17%, respectivamente, enquanto os anos de 2005, 1995 e 1997 apresentaram as menores representatividades, cada um com 1%. Essa distribuição detalhada da faixa etária dos estudantes na instituição, pela Figura 22.

Figura 22



A questão da etnia também foi abordada na análise, indicando que a maioria dos estudantes se autodeclara branca, representando quase 50%. Pardos aparecem em segundo lugar, com aproximadamente 20%, seguidos pelos pretos, que representam cerca de 6%. Essa informação contribui para a compreensão da diversidade étnica. A Figura 23, destaca a questão da distribuição da etnia.

Figura 23



A maior parcela da renda dos responsáveis dos alunos está acima dos 15 mil reais mensais, enquanto a segunda maior faixa compreende aqueles com renda entre 10 mil e 15 mil reais. Em relação à escolaridade dos responsáveis, a maioria, representando 34%, possui ensino superior, seguido por 32% que têm pós-graduação e 12% que completaram apenas o ensino médio. Apenas 3% não concluíram o ensino médio. As Figuras 24 e 25, apresentam essas parcelas.

Figura 24

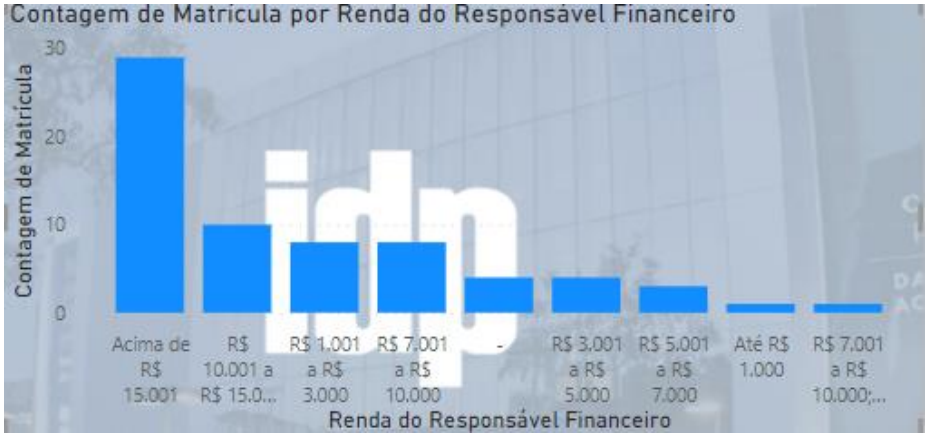
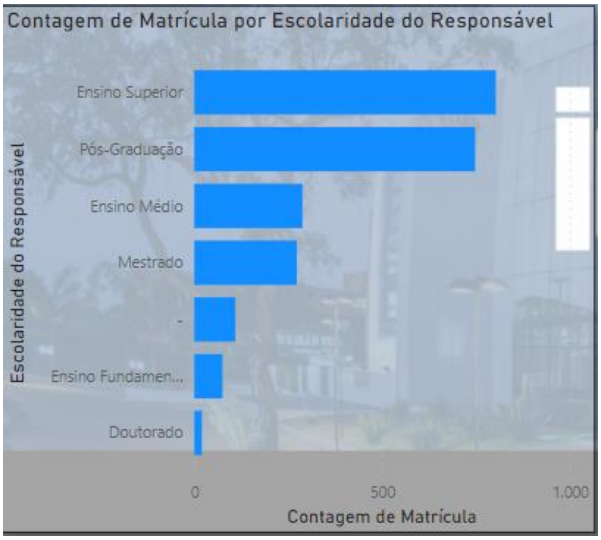
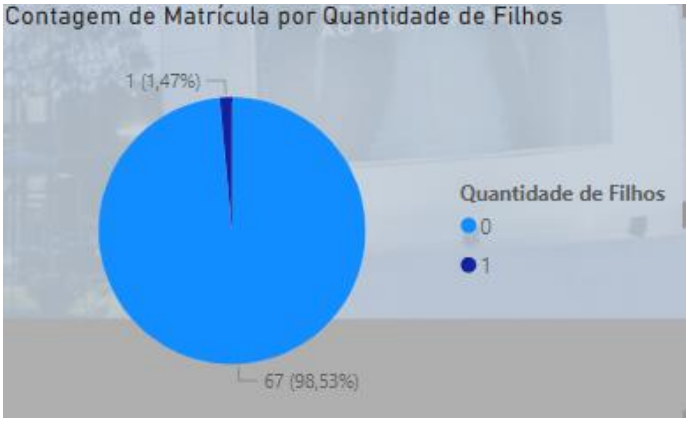


Figura 25



Outros pontos abordados incluíram a quantidade de filhos, onde 99% dos estudantes declararam não ter filhos. A Figura 26, apresenta a quantidade de filhos.

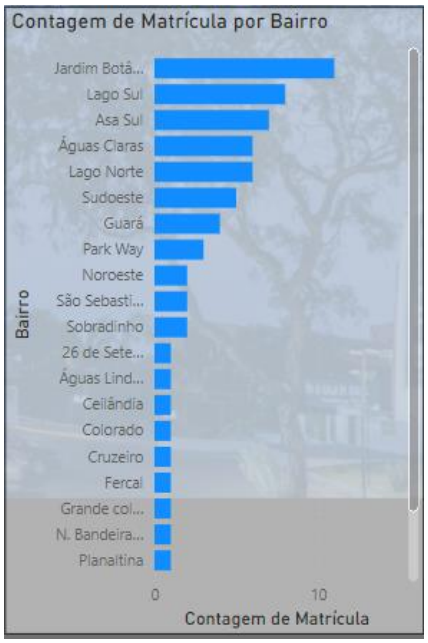
Figura 26



A análise também incluiu informações sobre a localização geográfica dos alunos, com destaque para o bairro Jardim Botânico, onde reside a maior parte dos estudantes. A distância média da faculdade foi calculada para diversos bairros, destacando-se Jardim Botânico com aproximadamente 21 km, Lago Sul com 10 km e Asa Sul com 5 km. Por outro lado, bairros como Santa Maria e Vicente Pires apresentaram distâncias maiores,

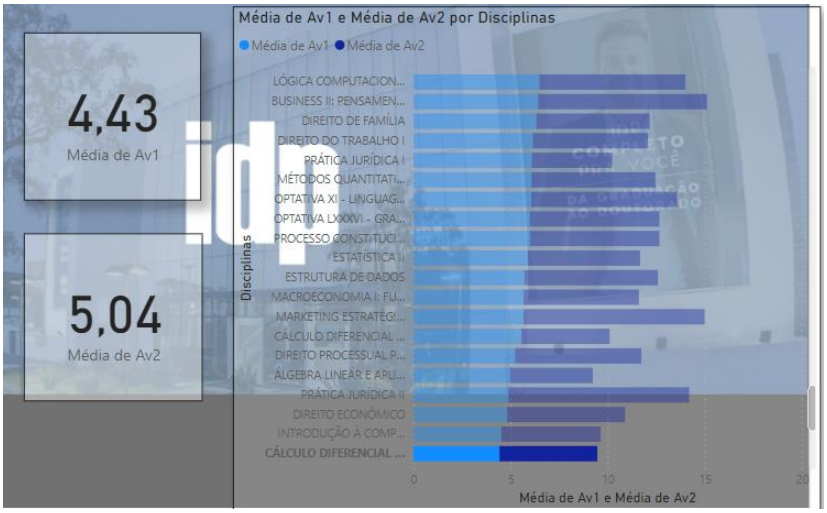
cerca de 32 km e 22 km, respectivamente. Na Figura 27, apresenta a relação de bairro e alunos.

Figura 27



Por fim, foram identificadas matérias com médias de notas mais baixas, tanto na AV1 quanto na AV2, incluindo Cálculo 2, Introdução à Computação e Direito Econômico. Essas informações podem ser úteis para aprimorar estratégias de ensino ou oferecer suporte adicional aos estudantes nessas disciplinas específicas. A Figura 28 representa as médias das notas Av1 e Av2 para cada matéria.

Figura 28



Conclusão:

Na conclusão deste estudo, destacamos resultados expressivos na análise preditiva da Tabela 1, a qual apresentou uma previsão eficaz tanto para alunos evadidos quanto para os que permaneceram. Entretanto, ao ampliar nossa análise para a Tabela 2, nos deparamos com um fenômeno de *Overfitting*, indicando que o modelo não foi eficiente. Esse desafio foi atribuído à ausência de dados nos formulários para os alunos evadidos no banco de dados da instituição

Olhando para o futuro, sugere que a instituição implemente a prática de coletar respostas dos alunos por meio do formulário utilizado na pesquisa, no início de cada semestre. Isso não apenas enriquecerá os dados disponíveis, mas também mitigará o risco de *Overfitting* ao evitar lacunas nas informações dos alunos, caso evadam do curso. Essa abordagem fortalecerá o modelo preditivo, contribuindo para previsões mais precisas e generalizáveis.

Entretanto, a instituição pode adotar medidas preventivas mais específicas, concentrando-se nos atributos com maiores pesos, como idade acima de 25 anos, nota da Av2 abaixo de 4,05 e bolsa institucional abaixo de 50%. Identificar esses fatores críticos permite uma participação mais ativa na vida acadêmica desses alunos, possibilitando maior controle sobre a evasão e, consequentemente, a redução de custos e despesas

financeiras associadas à evasão. Essa abordagem proativa proporcionaria uma gestão mais eficaz dos recursos institucionais.

Além disso, as categorias de renda, bairro, renda dos responsáveis, sexo, trabalho e escolaridade dos pais, presentes no formulário, serão analisadas para compreender como impactam nas decisões dos estudantes em evadir ou permanecer na faculdade. A implementação de um sistema de *Business Intelligence* (BI) permitiria que professores e coordenadores visualizem e analisem de forma eficiente dados relevantes, identificando tendências específicas para cada curso. Essa visão detalhada forneceria à instituição informações valiosas para orientar ações preventivas e estratégias, visando reduzir a evasão e promover o sucesso acadêmico dos alunos.

Referências:

MARTINS, Bibiana Volkmer; OLIVEIRA, Sidinei Rocha de. Qualificação profissional, mercado de trabalho e mobilidade social: cursos superiores de tecnologia. **Sociedade, Contabilidade e Gestão**, v. 12, n. 2, 2017.

MONTEIRO, André; GONÇALVES, Carlos; SANTOS, Paulo Jorge. Jovens: Do ensino superior para o mercado de trabalho. 2019.

SOUZA, Márcio Rodrigo de Araújo; MENEZES, Monique. Programa Universidade para Todos (PROUNI): quem ganha o quê, como e quando?. **Ensaio: Avaliação e Políticas Públicas em Educação**, v. 22, n. 84, p. 609-633, 2014.

QUEIROZ, Jacqueline Clara. Fundo de financiamento estudantil (Fies)-2010 a 2015: mecanismo de financiamento da democratização do acesso e permanência na educação superior. 2018.

https://dadosabertos.mec.gov.br/prouni.pdf

https://www.gov.br/secom/pt-br/assuntos/noticias/2023/10/censo-da-educacao-superior-2022-reforca-preocupacao-com-excesso-de-cursos-a-distancia-e-com-a-formacao-de-professores#:~:text=INTERIORIZA%C3%87%C3%83O%20%E2%80%93%20Segundo%20o%20Censo%2C%20o,eram%20privadas%20e%20312%2C%20p%C3%BAblicas.pdf

TEIXEIRA, Rita de Cássia Petrarca; MENTGES, Manuir José; KAMPFF, Adriana Justin Cerveira. Evasão no ensino superior: um estudo sistemático. **Publicação em final de outubro, 2019, Brasil.**, 2019.

MENDES, Gabriela Mesquita; MULIN, Heloise de Pinho. Avaliação e desempenho do Programa Universidade para Todos (PROUNI). 2017.

SILVA FILHO, Roberto Leal Lobo et al. A evasão no ensino superior brasileiro. **Cadernos de pesquisa**, v. 37, p. 641-659, 2007.

HOED, Raphael Magalhães. Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de Computação. 2016.

BAGGI, Cristiane Aparecida dos Santos; LOPES, Doraci Alves. Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, v. 16, n. 02, p. 355-374, 2011.

BOENTE, Alfredo Nazareno Pereira; ROSA, José Luiz Dos Anjos. Utilização de ferramentas de KDD para Integração de aprendizagem e tecnologia em busca da gestão estratégica do conhecimento na empresa. **Anais do Simpósio de Excelência em Gestão e Tecnologia**, v. 1, p. 123-132, 2007.

PRESTES, Emília Maria da T.; FIALHO, M. G.; PFEIFER, D. K. A Evasão no ensino superior globalizado e suas repercussões na gestão universitária. **Paraíba. Acesso em**, v. 13, 2016.

BELENKE DOS SANTOS, JÚLIO CÉSAR. Usando mineração de dados para predição da evasão escolar. 2021.

DE OLIVEIRA JÚNIOR, José Gonçalves; NORONHA, Robinson Vida; KAESTNER, Celso Antonio Alves. Método de seleção de atributos aplicados na previsão da evasão de cursos de graduação. **Revista de informática aplicada**, v. 13, n. 2, 2017.

FIALHO, Marillia Gabriella Duarte et al. A evasão escolar e a gestão universitária: o caso da Universidade Federal da Paraíba. 2014.

ASSUNÇÃO, Yluska Bambirra; GOULART, Iris Barbosa. Qualificação Profissional ou Competências para o Mercado Futuro?. **Future Studies Research Journal: Trends and Strategies**, v. 8, n. 1, p. 175-207, 2016.

DOS SANTOS, Fábio Alexandre Bértolo. **Algoritmos de Machine Learning para previsão da procura**. 2021. Tese de Doutorado. Universidade do Minho (Portugal).

PAIXÃO, Gabriela Miana de Mattos et al. Machine Learning na Medicina: Revisão e Aplicabilidade. **Arquivos Brasileiros de Cardiologia**, v. 118, p. 95-102, 2022.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. **Sistemas inteligentes-Fundamentos e aplicações**, v. 1, n. 1, p. 32, 2003.

GALVÃO, Noemi Dreyer; MARIN, Heimar de Fátima. Técnica de mineração de dados: uma revisão da literatura. **Acta Paulista de Enfermagem**, v. 22, p. 686-690, 2009.

BELENKE DOS SANTOS, JÚLIO CÉSAR. Usando mineração de dados para predição da evasão escolar. 2021.

MANHÃES, Laci M. Barbosa et al. Identificação dos fatores que influenciam a evasão em cursos de graduação através de sistemas baseados em mineração de dados: Uma abordagem quantitativa. In: **Anais do VIII Simpósio Brasileiro de Sistemas de Informação**. SBC, 2012. p. 284-295.

WINCK, Ana T. et al. Processo de KDD aplicado à bioinformática. **Sociedade Brasileira de Computação**, 2010.

BELENKE DOS SANTOS, JÚLIO CÉSAR. Usando mineração de dados para predição da evasão escolar. 2021.

STRATTON, Leslie S.; O'TOOLE, Dennis M.; WETZEL, James N. A multinomial logit model of college stopout and dropout behavior. **Economics of education review**, v. 27, n. 3, p. 319-331, 2008.

SILVA, Márcio Bezerra da et al. A teoria da classificação facetada na modelagem de dados em banco de dados computacionais. 2011.

SAMPAIO, Breno et al. Desempenho no vestibular, background familiar e evasão: evidências da UFPE. **Economia Aplicada**, v. 15, p. 287-309, 2011.

MARIA, Willian; DAMIANI, João Luccas; PEREIRA, Max. Rede bayesiana para previsao de evasao escolar. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. 2016. p. 920.

MENDES, Laura Schertel; DONEDA, Danilo. Comentário à nova Lei de Proteção de Dados (Lei 13.709/2018), o novo paradigma da proteção de dados no Brasil. **Revista de Direito do Consumidor**, v. 120, p. 555, 2018.

OLIVEIRA, José Clovis Pereira de et al. O questionário, o formulário e a entrevista como instrumentos de coleta de dados: vantagens e desvantagens do seu uso na pesquisa de campo em ciências humanas. In: **III Congresso Nacional de Educação**. 2016. p. 1-13.

AREND, Felipe Gabriel. Geração de operações CRUD a partir de metadados. 2011.
CORTIJO, PRISCILLA DOMINGUES. CURSO DE ESPECIALIZAÇÃO EM ENGENHARIA DE SOFTWARE.

HASANIN, T. et al. Examining characteristics of predictive models with imbalanced big data. J. Big Data 6 (69), 1–21. 2019.

SEGER, Cedric. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing. 2018.

GARZILLO, Monique Joaquim Witt. **Classificação de tumores cerebrais com algoritmos de machine learning**. 2022. Tese de Doutorado. Instituto Politécnico de Lisboa, Escola Superior de Tecnologia da Saúde de Lisboa.

BELENKE DOS SANTOS, JÚLIO CÉSAR. Usando mineração de dados para predição da evasão escolar. 2021.

SOUZA, Natália da Costa et al. Modelo de classificação de processos erosivos lineares ao longo de ferrovias através de algoritmo de árvore de decisão e geotecnologias. **Boletim de Ciências Geodésicas**, v. 23, p. 72-86, 2017.

BAPTISTA, David Felice Falivene. **Equalização de canais baseada em Árvores de Decisão**. 2019. Tese de Doutorado. [sn].

MEDRI, Waldir. Análise exploratória de dados. **Londrina: Universidade Estadual de Londrina**, 2011.

CONCEIÇÃO, Luís Filipe Marques dos Santos. A Importância do Business Intelligence na tomada de decisão. 2020.

ABREU, Leonardo Evangelista de. People Analytics: uso de árvores de decisão na retenção de talentos. 2022.

BABA, Ricardo Kazuo; VAZ, Maria Salete Marcon Gomes; COSTA, Jéssica da. Correção de dados agrometeorológicos utilizando métodos estatísticos. **Revista Brasileira de Meteorologia**, v. 29, p. 515-526, 2014.

BABA, Ricardo Kazuo; VAZ, Maria Salete Marcon Gomes; COSTA, Jéssica da. Correção de dados agrometeorológicos utilizando métodos estatísticos. **Revista Brasileira de Meteorologia**, v. 29, p. 515-526, 2014.