**Evasão de alunos concluintes UFJF**

**Caio Vinicius Monteiro Martins**

**Filipe Carvalho Lacerda**

**Igor Procópio do Carmo**

**Orientador: Ricardo Silva Campos**

**Resumo**

A evasão de alunos concluintes demonstra ser um grave problema para instituições de ensino superior, por fatores como, o orçamento recebido pelas universidades que dependem da quantidade de alunos matriculados e de concluintes, assim como prejuízos na perda de programas e planos do governo.

O objetivo do projeto é descrever os principais fatores que causam a evasão de alunos concluintes da Universidade Federal de Juiz de Fora, utilizando algoritmos de IA. A investigação do artigo busca chegar em modelos de previsão para entender o comportamento de futuras evasões de alunos.

Utilizando um *dataset* contendo 53.098 registros como treinamento para o algoritmo, iniciou-se a utilização de algoritmos de aprendizagem de máquina para realizar análises que buscam padrões para fazer uma previsão de evasões.

**Palavras-Chave:** Evasão Escolar, Aprendizagem de Máquina, Mineração de Dados

**Abstract**

The dropout of graduating students proves to be a serious problem for higher education institutions for factors such as the budget received by universities that depends on the number of enrolled students and graduates, as well as damnification in the loss of government programs and plans.

The objective of the project is to define the main factors that cause the dropout of graduating students at the Federal University of Juiz de Fora, using AI algorithms. The investigation of the article seeks to reach predictive models for the behavior of future dropouts.

Using a *dataset* containing 53.098 entries as training for the algorithm, the employment of machine learning algorithms initiates the search for patterns to predict school dropout.

**Keywords:** School Dropout, Machine Learning, Data Mining.

**1-Introdução**

O trabalho tem como objetivo identificar os perfis de alunos universitários concluintes que evadem de seus cursos, assim como implementar modelos de previsão para futuras evasões por alunos concluintes. Com o uso de algoritmos de aprendizagem supervisionada, é possível analisar os critérios e gerar estimativas para maior compreensão de cada cenário.

Através da análise dos dados de todos os alunos concluintes da UFJF (Universidade Federal de Juiz de Fora) entre os anos de 2003-2020, aplicando algoritmos de IA e utilizando métodos de identificação de padrões em conjuntos de dados, é possível obter informações relevantes para auxiliar a tomada de alguma decisão (FEYYAD, SHAPIRO, SMYTH, & UTHURUSAMY, 1996).

A Universidade Federal de Juiz de Fora (UFJF) é uma universidade pública. Possui um campus sede localizado em Juiz de Fora (MG) e um campus avançado em Governador Valadares (MG). A UFJF oferta cerca de 93 cursos de graduação, 45 cursos de mestrado e 24 de doutorado, em todas as áreas do conhecimento. Ao todo, possui mais de 20 mil alunos na modalidade presencial e cerca de três mil na modalidade à distância.

Para administrar suas informações institucionais, a UFJF tem implementado desde 2003 um sistema próprio, chamado de Sistema Integrado de Gestão Acadêmica (SIGA). Ele foi desenvolvido completamente em software livre, reduzindo custos e promovendo o alinhamento estratégico com o Governo Federal. O SIGA é composto atualmente por 43 módulos, distribuídos em quatro grandes áreas ou sistemas: Acadêmico, Administrativo, Gestão de pessoas, e Sistemas de apoio[[1]](#footnote-0).

A evasão de universitários pode causar prejuízos significativos para a economia, para a gestão institucional e também para diversos aspectos sociais. A verba recebida pelas universidades depende da quantidade de alunos matriculados e concluintes. As instituições de ensino públicas quando não atingem o objetivo institucional de formar os alunos, sofrem prejuízos como a perda de programas e planos fornecidos pela IES (Instituições de Ensino Superior) e além disso os professores são prejudicados por não conseguirem exercer seu papel (PRESTES, 2014).

De acordo com De Oliveira Junior (2017), “Nesse contexto, surge a importância de identificar antecipadamente quais estudantes não terão êxito na conclusão do curso, pois essa informação pode auxiliar na tomada de decisões para que essa previsão possa ser modificada.”

**1.1 - Objetivos**

**1.1.1- Objetivo Geral**

O intuito da realização do projeto é descrever os principais fatores que causam a evasão de alunos concluintes da Universidade Federal de Juiz de Fora nos últimos anos, dessa forma através da utilização de aprendizagem de máquina buscar atingir o desenvolvimento de modelos de previsão capazes de entregar perfis para futuras evasões.

**1.1.2- Objetivos Específicos**

Este projeto tem como foco dois objetivos específicos, sendo eles à análise exploratória dos dados das variáveis do *dataset* com maior relevância para a utilização das mesmas nos modelos de previsão e também a correlação entre essas variáveis, assim como utilizar algoritmos de aprendizagem de máquina para assim então determinar as previsões de evasão e com esses resultados das previsões entregar hipóteses concisas que possam fornecer informações benéficas à instituição com o fim de buscar pelos padrões mais recorrentes nos aspectos sociais para mapear os grupos com maior risco de evasão.

**1.2- Justificativa**

A evasão de alunos (principalmente os concluintes) é um grande problema para as escolas, universidades federais, estaduais e privadas, de acordo com Santana (1996), “a evasão é um dos maiores e mais preocupantes desafios do Sistema Educacional, pois é fator de desequilíbrio, desarmonia e desajustes dos objetivos educacionais pretendidos”.

Segundo Fialho (2014) a fórmula que possibilita o cálculo do prejuízo anual causado pela evasão estudantil é dada da seguinte forma:

Tabela 1: Explicação da fórmula de perda anual.

| Perda Anual | |
| --- | --- |
| Termos | Significado |
|  | Perda anual. |
|  | Nº total de alunos que abandonaram os cursos dividido pelo total de anos analisados. |
|  | Média do tempo de permanência dos alunos nos cursos. |
|  | Gasto direto por aluno. |

Fonte: Fialho, 2014.

Uma das alternativas viáveis para entender melhor a causa deste problema é através da análise de dados utilizando algoritmos de classificação supervisionada, fornecendo os dados necessários para o treinamento do algoritmo. Através da utilização destes algoritmos, cujos dados utilizados contém amostras de um cenário desejado e desta forma a IA busca por padrões entre os dados.

**2- Referencial Teórico**

De acordo com Brito Junior (2019), o impacto gerado pela evasão nos cursos de graduação é significativo, visto que esse fenômeno atrapalha o desenvolvimento do mercado de trabalho criando um déficit de mão de obra qualificada para os setores com maior evasão de cursos superiores, bem como uma diminuição da verba arrecadada tanto pelas universidades públicas quanto pelas universidades privadas a cada aluno evadido. “A evasão destes alunos gera custos públicos e privados para o país.” (PEREIRA, 2003).

Segundo Vasquez (2003), é possível separar a evasão em três categorias devido a causalidades diferentes: 1) evasão precoce: indivíduo que foi aprovado no curso, porém não realizou a matrícula, 2) evasão inicial: aluno que evade do curso nos primeiros quatro períodos e 3) evasão tardia: aluno que evade a partir do quinto período.

A partir dos estudos de Costa e Campos (2000) conclui-se que, existe uma grande necessidade de entender a forma como a evasão do ensino superior é tratada nos dias atuais, principalmente devido a atual escassez de profissionais qualificados com ensino superior no mercado de trabalho e a necessidade do país de qualificar sua população.

Os índices de evasão em cursos de graduação no país estão em cerca de 18,4%[[2]](#footnote-1) apontados pelo SEMESP em uma de suas pesquisas mais recentes no ano de 2019, dentro da rede privada a taxa de evasão neste mesmo ano o curso de sistemas de informação apresenta o maior índice liderando a tabela com 37,6%, enquanto com a menor porcentagem medicina demonstra-se com 6,8% de evasão dentro da rede privada.

Estudos evidenciados por Pereira (2003) apontam que, existem dois tipos de fatores capazes de influenciarem a decisão do aluno para a evasão, os caracterizando como: 1) fatores internos da instituição, que consistem em infra-estrutura deficitária, acervo desatualizado, métodos de avaliação docente e deficiência didático pedagógica dos professores e 2) fatores externos, que consistem em dificuldades financeiras, escolha equivocada do curso, falta de base para acompanhar o curso escolhido, fato de ter sido admitido em um curso que não foi sua primeira opção e também assim como uma série de outros fatores que podem se englobar nos fatores externos sejam eles os variados motivos que o estudante possa estar enfrentando.

No Brasil, já foram realizados diversos estudos relacionados para buscar entender os motivos mais frequentes que levam os alunos a evadirem de seus cursos durante sua graduação. Buscando informações que possam levar a uma estimativa para contribuir com a resolução deste problema.

Brito Junior et al. (2019) desenvolveram um estudo com o uso da mineração de dados para identificar os perfis dos graduandos que evadem do curso de sistemas de informação, utilizando os dados da UFRN. Santos et al.(2020) utilizando a Aprendizagem de Máquina, verifica se os classificadores gerados por algoritmos são capazes de distinguir entre alunos que concluem seus cursos e alunos propensos a evadir. Assim como De Jesus et al.(2021) que através do aprendizado supervisionado e do uso de redes neurais artificiais, criou um modelo preditivo para prever os alunos com risco de evadir do curso de Licenciatura em Computação da Universidade do Estado do Amazonas.

Neste artigo, utilizaremos dados específicos da Universidade Federal de Juiz de Fora (UFJF) para desenvolver um estudo semelhantes aos citados acima, porém além de mensurar sobre os alunos que evadem de seus cursos, iremos também aplicar o modelo aos alunos concluintes, para chegar a estimativas mais precisas e relevantes para a análise do problema.

**3- Metodologia**

O projeto tem uma natureza descritiva aplicada devido a isso utilizará uma pesquisa de abordagem mista buscando seus objetivos descritivos através de procedimentos bibliográficos, de um banco de dados obtido através do portal Fala.BR**[[3]](#footnote-2)** e de metodologias de aprendizagem de máquina.

Os Algoritmos de IA utilizados para o desenvolvimento do projeto foram: A árvore de decisão é um dos tipos de algoritmos utilizados no projeto, é um algoritmo de decisão por aprendizagem de máquina de tipo supervisionado que se baseia na divisão de dados em grupos de maneira homogênea, os dados então se tornam classificados, dessa forma esse algoritmo proporciona um gráfico em formato de árvore capaz de entregar um modelo que contempla decisões e suas possíveis consequências, seu objetivo é encontrar o atributo que gera a melhor divisão dos dados possível com o maior grau de precisão.

O KNN (Vizinhos mais Próximos de K) é um algoritmo de aprendizagem supervisionada de classificação não paramétrica usada para a classificação dos dados, na qual a entrada de dados é o K sendo o exemplo de treinamento mais próximo ao conjunto de dados, desencadeando na saída a apresentação dos valores mais relacionados com o K inicial, assim sua forma de aprendizagem se baseia em quão similar é um vetor em relação aos outros em busca de padrões entre os mais próximos, é um algoritmo que necessita de uma análise mais apurada posteriormente aos resultados apresentados, visto que não houveram dados de treinamento inicial para gerar o modelo.

Outro importante algoritmo utilizado no projeto é a regressão logística que é definida segundo Gonzalez (2018) como um algoritmo de estatística que busca prever valores a partir de sua correlação logística, na regressão logística recebemos valores que são ou que possam ser binários (0 ou 1, true ou false, sucesso ou falha, etc). O algoritmo de regressão logística permite que se estime a probabilidade de um evento ocorrer aleatoriamente comparando com a probabilidade do mesmo não ocorrer, assim como consegue prever o impacto de uma variável sobre uma outra. Os resultados deste algoritmo são apresentados no final por forma de probabilidade, o que o torna de extrema importância devido ao fato de assim ser necessário poucas suposições para realizar as predições, além de que os resultados possuem um alto grau de confiabilidade.

O algoritmo de *random forest* (Floresta Aleatória) parte de uma origem semelhante ao do modelo de árvore de decisão a questão aqui posta é que neste modelo a aleatoriedade irá demandar o comportamento no qual o algoritmo irá seguir com base na seleção do subconjunto de variáveis, e a cada nova árvore criada a próxima é escolhida aleatoriamente novamente com base em suas variáveis para decidir o próximo nó, dessa forma o algoritmo monta diversas árvores de decisões menores, neste modelo de aprendizagem de máquina a predição por classificação passa então ser demonstrada com seus resultados através de cada árvore que foi criada.

As variáveis a seguir são as utilizadas nos algoritmos já mencionados, são com elas as quais os algoritmos nos oferecem resultados a serem interpretados após a utilização dos mesmos de acordo com a Tabela 2:

Tabela 2: Explicação de variáveis utilizadas nas análises dos *datasets*.

| Conceito de Variaveis Importantes | |
| --- | --- |
| Variaveis | Representações |
| cota\_numerica | Sistema de cotas convertidas em valores ordinais por ordem de vulnerabilidade sendo 9 o mais vulnerável e 1 o menos vulnerável, mais detalhes adiante. |
| diff\_periodos | Referente ao percentual de períodos cursados em relação ao prazo ideal |
| etnia\_AMARELA | Pessoas que se identificam amarelas |
| etnia\_BRANCA | Pessoas que se identificam brancas |
| etnia\_INDIGENA | Pessoas que se identificam indígenas |
| etnia\_OUTRA | Pessoas que se identificam como outra etnia não listada |
| etnia\_PARDA | Pessoas que se identificam pardas |
| etnia\_PRETA | Pessoas que se identificam pretas |
| GAP | Diferença entre os anos levados entre o ensino médio à entrada do ensino superior de um aluno |
| grandearea\_Ciências Biológicas | Todos os cursos que se encaixam dentro da área de ciências biológicas |
| grandearea\_Ciências da Saúde | Todos os cursos que se encaixam dentro da área de ciências da saúde |
| grandearea\_Ciências Exatas e da Terra | Todos os cursos que se encaixam dentro da área de ciências exatas e da terra |
| grandearea\_Ciências Humanas | Todos os cursos que se encaixam dentro da área de ciências humanas |
| grandearea\_Ciências Sociais Aplicadas | Todos os cursos que se encaixam dentro da área de ciências sociais aplicadas |
| grandearea\_Engenharias | Todos os cursos que se encaixam dentro da área de engenharia |
| grandearea\_Lingüística, Letras e Artes | Todos os cursos que se encaixam dentro da área de linguística, letras e artes |
| grupoira | Índice de rendimento acadêmico. |
| idadeSaida | Referente a idade de saída de determinado aluno do curso |
| mesma\_cidade | Coluna que se refere se o aluno faz o curso em sua cidade natal |
| montccest\_bool | Determina se o curso contém em sua grade monografia, tcc ou estágio obrigatório. |
| sexo\_F | Pessoas que se identificam pelo sexo feminino |
| sexo\_M | Pessoas que se identificam pelo sexo masculino |
| Situacao\_num | Determina se o aluno é evadido ou concluinte. |
| turno\_INTEGRAL | Cursos que acontecem em tempo integral |
| turno\_MATUTINO | Cursos que acontecem em período matinal |
| turno\_NOTURNO | Cursos que acontecem em período noturno |

Dito isto, o projeto se inicia utilizando a linguagem de programação em Python na IDE Jupyter Notebook, começa-se importando a biblioteca Pandas e logo após importa-se o dataframe recolhido do site FalaBr² para dentro do projeto.

Muitas das vezes dataframes contém erros, sejam esses valores ausentes ou até mesmo valores errados, devido a isso é necessário uma limpeza desse dataframe, retirar não somente valores ausentes das colunas como também valores que possam estar errados como colunas numéricas contendo dados com caracteres, esta limpeza chamamos de tratamento dos dados e isso deve ser feito logo após a inserção do dataframe no projeto.

Assim que ocorreu a importação do dataframe foi necessário tratar os dados das colunas que continham “NULL” como registro, nas colunas “Classificação” e “Etnia” haviam registros desse porte, desta forma usando a biblioteca pandas foi possível no código trocar os valores “NULL” de “Classificação” e “Etnia” para “0” e “Outra” respectivamente.

Outra parte importante do tratamento é lidar com os registros em branco das colunas, a lógica adotada neste caso é apagar uma linha inteira caso haja um registro em branco, após essa parte do tratamento notou-se a existência de um grupo de cota F no dataframe, este seria um grupo criado com um prazo de somente 2 anos, visto que ainda não houve tempo o suficiente para pessoas do grupo F chegarem próximas da conclusão de seu curso sendo assim não haviam dados concretos sobre evasão de concluintes notou-se a necessidade da retirada do grupo F do dataframe.

Na continuação do tratamento dos dados na coluna “GrandeArea” que contém os registros dos tipos de áreas de estudos como humanas e exatas haviam linhas com registros “graduação presencial” estes dados foram limpos do dataframe, na coluna “GAP” que é referente ao tempo entre o término do aluno no ensino médio e a sua entrada no ensino superior em anos haviam números incoerentes, devido a este fato foi adicionado um limitador para essa coluna, no qual se houvesse um “GAP” com valores muito altos seriam retirados esses registros, na coluna “idadeSaida” que é referente a idade na qual o aluno terminou o curso existiam valores negativos que também foram tratados, por fim na coluna “prazoIdeal” que demonstra quantos anos se espera que o curso seja concluído pelo aluno foram retirados registros errados com valores zerados.

Com todos esses registros tratados é necessário também organizar de maneira mais adequada as colunas do dataframe para uma melhor compreensão na hora de interpretar os resultados, como todos os cursos da UFJF são de tipagem presencial para se iniciar os modelos de previsão a coluna “tipoCurso” não foi adicionada ao código, assim como houve a necessidade de criar uma lista para a divisão de variáveis, entre numéricas e nominais.

Um dos fatores centrais para a compreensão do projeto em relação a metodologia e aos resultados é a noção dos grupos prioritários utilizados como parâmetro dentro da universidade, visto que estes dados ao serem interpretados entregam análises precisas sobre o fato em estudo, é sabido então que entre 2006 e 2012 era aplicado o seguinte sistema :

1) Grupo A - antigo: Quatro séries do Ensino Fundamental e todo o Ensino Médio em escola pública e autodeclara-se negro;

2) Grupo B - antigo: Quatro séries do Ensino Fundamental e todo o Ensino Médio em escola pública;

3) Grupo C - antigo: Não-optante pelo Sistema de Cota.

A partir de 2013 foi adotado o seguinte sistema:

1) Grupo A: Candidatos com renda igual ou inferior a um salário mínimo e meio per capita familiar mensal, que tenham cursado o Ensino Médio integralmente em escola pública e que se declarem pretos, pardos ou indígenas;

2) Grupo B: Candidatos com renda igual ou inferior a um salário mínimo e meio per capita familiar mensal, que tenham cursado o Ensino Médio integralmente em escola pública, independentemente de outra declaração;

3) Grupo C: Vagas de amplo acesso, independentemente de renda ou de escola;

4) Grupo D: Candidatos que tenham cursado o Ensino Médio integralmente em escola pública e que se declarem pretos, pardos ou indígenas, independentemente de renda;

5) Grupo E: Candidatos que tenham cursado o Ensino Médio integralmente em escola pública, independentemente de renda ou autodeclaração.

E em 2017, as cotas para pessoas com deficiência foram adicionadas:

1) Grupo A1: Candidatos com deficiência autodeclarados pretos, pardos ou indígenas, que tenham renda familiar bruta per capita igual ou inferior a 1,5 salário mínimo e que tenham cursado integralmente o ensino médio em escolas públicas;

2) Grupo B1: Candidatos com deficiência que tenham renda familiar bruta per capita igual ou inferior a 1,5 salário mínimo e que tenham cursado integralmente o ensino médio em escolas públicas;

3) Grupo D1: Candidatos com deficiência autodeclarados pretos, pardos ou indígenas que, independentemente da renda, tenham cursado integralmente o ensino médio em escolas públicas;

4) Grupo E1: Candidatos com deficiência que, independentemente da renda, tenham cursado integralmente o ensino médio em escolas públicas.

Com base nesses parâmetros de cotas o projeto busca uma padronização em valores ordinais para esse sistema, dessa forma desenvolveu-se um sistema como uma ordem de vulnerabilidade, na qual o maior número representa um grupo que demonstra maior vulnerabilidade social e o menor grupo a ampla concorrência, dessa forma encontra-se então:

Grupo A1:9, Grupo A:8, Grupo B1:7, Grupo B:6, Grupo D1:5,Grupo D:4, Grupo A - antigo: 4, Grupo E1:3, Grupo E:2, Grupo B - antigo: 2, Grupo C:1, Grupo C - antigo:1

Ao classificar os grupos desta maneira os dados se tornam então tratados para serem manuseados com maior precisão no código.

Após a organização e tratamento de como utilizar numericamente o sistema de grupos houve a necessidade de criar certas colunas no dataframe que seriam de suma importância para os modelos de predição de IA, visto isso adicionou-se uma coluna “mesmaCidade” que se refere se a cidade natal do aluno é a mesma onde é localizada o campus da faculdade onde ele cursou sua graduação, além disso a coluna “montccest\_bool” foi criada na qual seus registros iriam apresentar resultados de verdadeiro ou falso em caso de o curso apresentar monografia, tcc ou estágio obrigatório em seu currículo.

Algumas outras colunas foram adicionadas ao *dataframe* em busca de uma melhor precisão nos resultados, sendo elas “diff\_periodos” que se refere ao percentual de períodos cursados em relação ao prazo ideal em busca de compreender se o número de períodos cursados pelo aluno causa determinado impacto na evasão e “Situacao\_num” um booleano que que demonstra se o aluno em questão está em estado de concluinte ou evadido pelos registros.

Uma coluna essencial para que o algoritmos obtivessem melhores resultados é a coluna “grupoira” que se refere ao índice de rendimento acadêmico, podendo ser descrito em 4 intervalos iguais sejam eles: o Grupo 1: 0 a 25; Grupo 2: 25 a 50; Grupo 3: 50 a 75; Grupo 3: 75 a 100; dessa forma será possível analisar se o IRA do aluno é impactante em sua decisão de evasão.

Por estar lidando com algoritmos é de suma importância transformar os dados nominais em numéricos, desta forma é necessário tratar as variáveis categóricas nominais, transformando-as em quantitativas discretas, após isso temos um *dataframe* lidando exclusivamente com registros numéricos tornando viável a manipulação desses registros pelos algoritmos.

Como o foco do projeto é voltado para os alunos concluintes que estão em situação de evasão foram retirados os alunos ativos do dataframe tratado, pois eles não podem ser usados para o treinamento dos algoritmos.

Dando finalização a esta parte de tratamento e organização do dataframe, o mesmo foi separado em dois *datasets*, um *dataset* contendo os alunos os concluintes (+70% do curso concluído) e o outro *dataset* com todos os alunos, vale salientar também que a amostra utilizada para todos os casos de teste é de um total de 30% de utilização de todos os registros contidos naquele *dataset* utilizado.

**4- Resultados**

**4.1 - Análise Exploratória**

O *dataset* disponibilizado possui as seguintes colunas conforme a Figura 1 e valores, com um total de 46914 registros totais após a limpeza dos dados, incluindo as situações de alunos concluídos, evadidos e ativos.

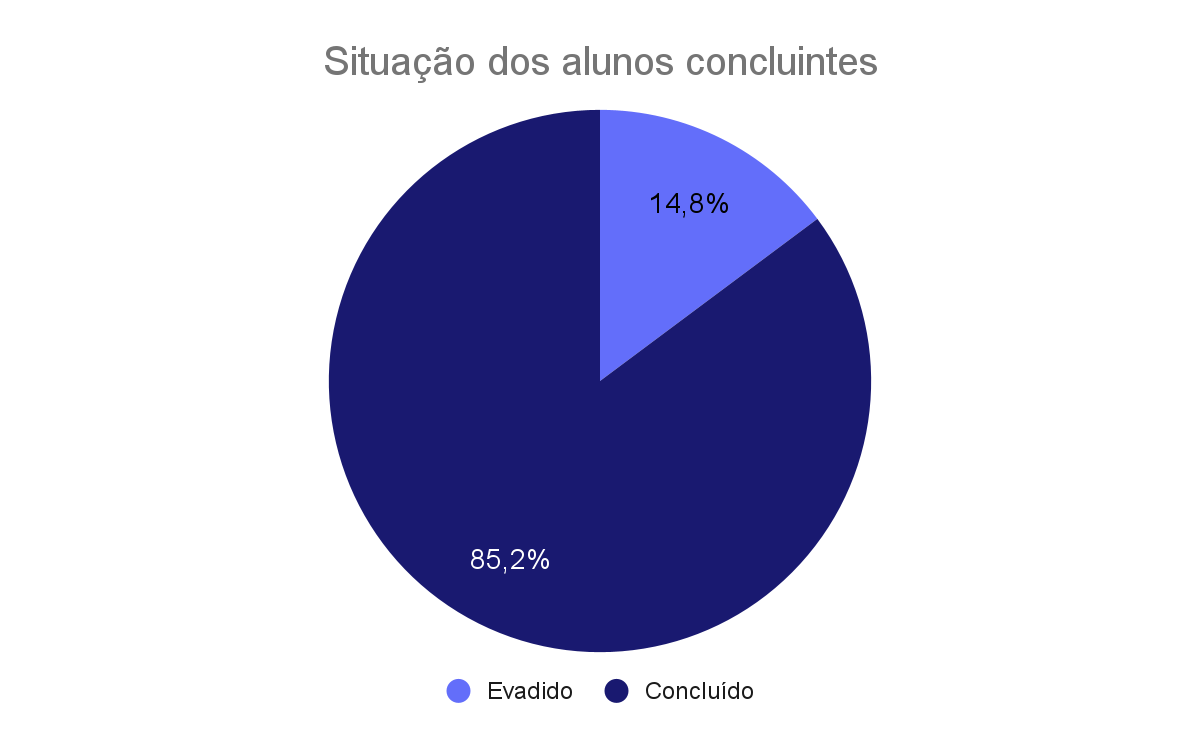
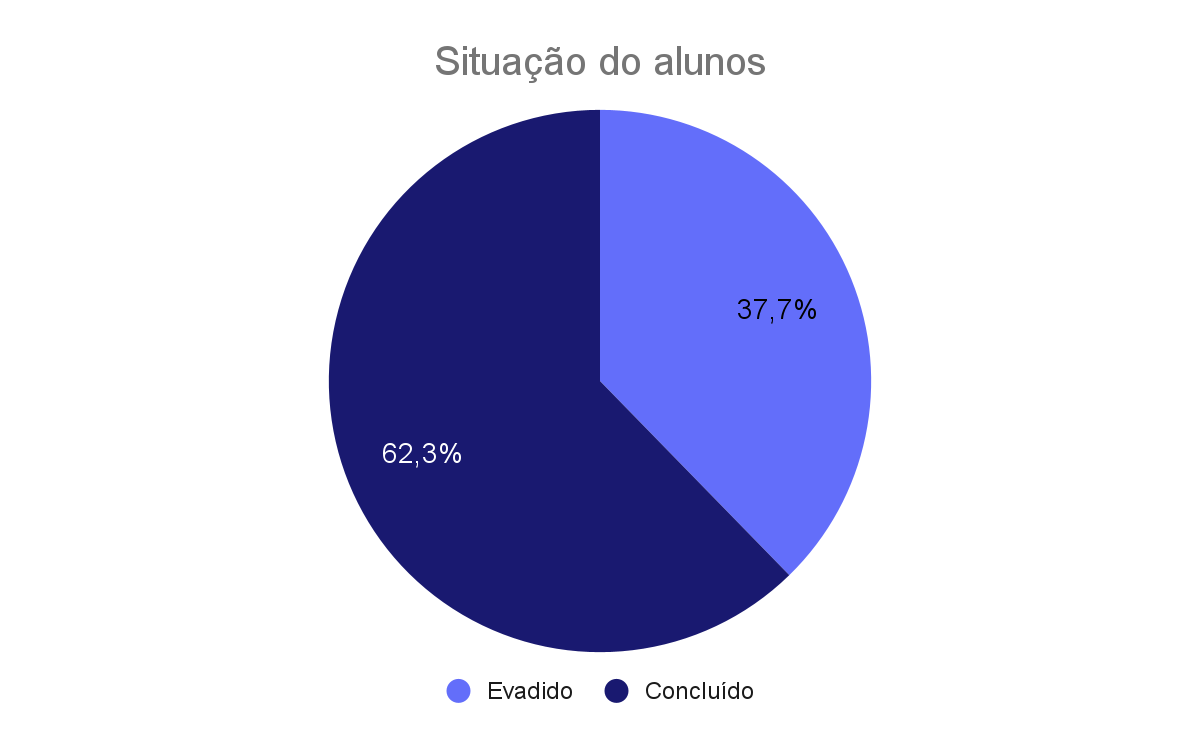
Figura 1: Algumas colunas e linhas contidas no *dataset* utilizado no trabalho.



Para o estudo ser realizado foi necessário filtrar o *dataset* mantendo apenas os concluídos e evadidos totalizando 39600 registros, devido ao fato de que os alunos ativos não seriam úteis para o estudo, vindo da necessidade de encontrar padrões entre os alunos que evadem e que concluem e as discrepâncias entre os mesmos.

Em uma análise inicial ao *dataset*, dividindo os alunos pela situação acadêmica para apenas os evadidos e concluídos, em comparação com o *dataset* com apenas os alunos com mais de 70% de conclusão do curso, temos o seguinte resultado de acordo com a Figura 2:

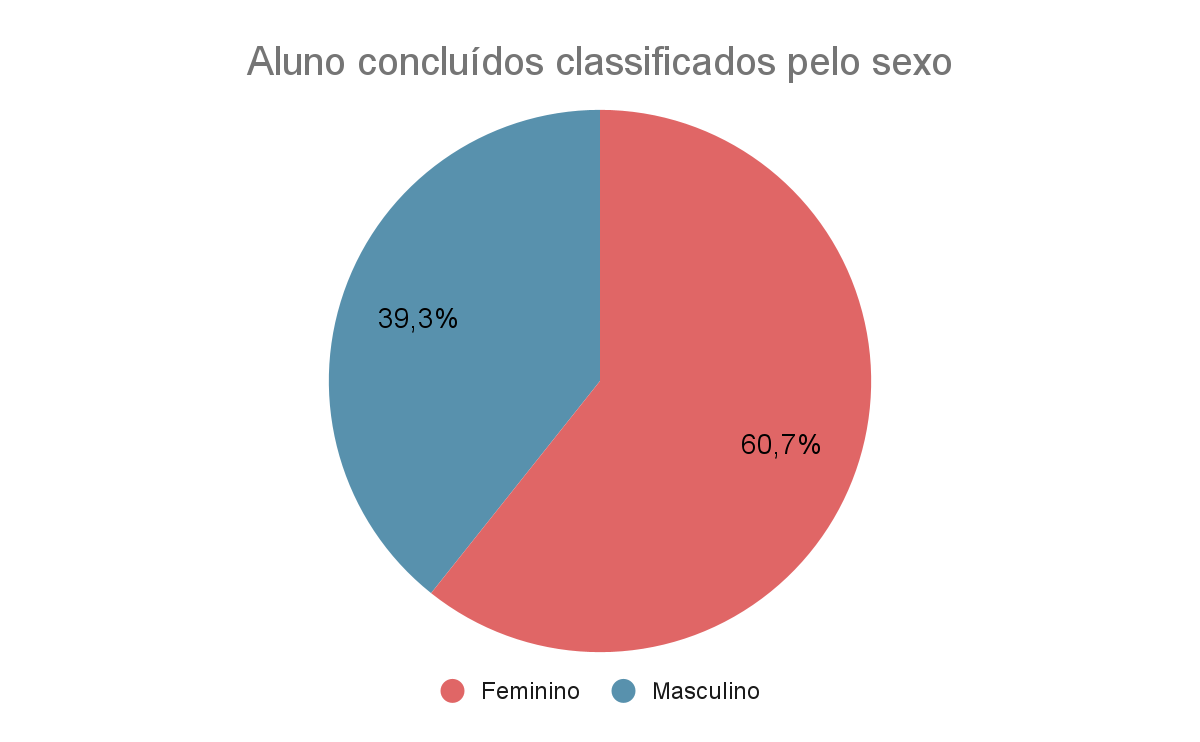
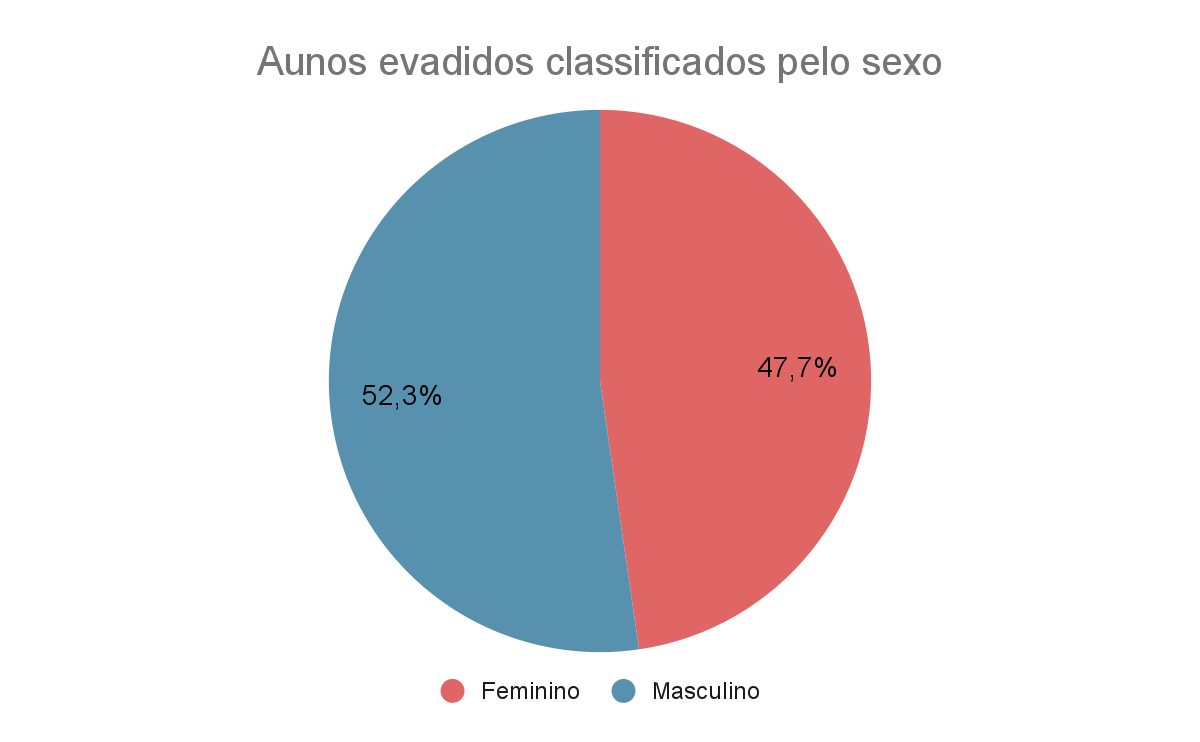
Figura 2: Gráfico da situação dos alunos utilizados, nos datasets geral e de concluídos.



É possível perceber que dentre os alunos concluídos e evadidos (sem os alunos ativos) 37,7% dos alunos abandonaram seus cursos, a maioria desses alunos evadem nos primeiros períodos, porém alguns deles só deixaram de fazer parte da universidade no final do curso (com mais de 70% do curso concluído), ou seja, dos alunos concluintes 14,8% tendem a evadir.

Outra análise interessante a ser levada em conta é a porcentagem de homens e mulheres que evadem e a porcentagem dos que concluem os cursos de acordo com a Figura 3:

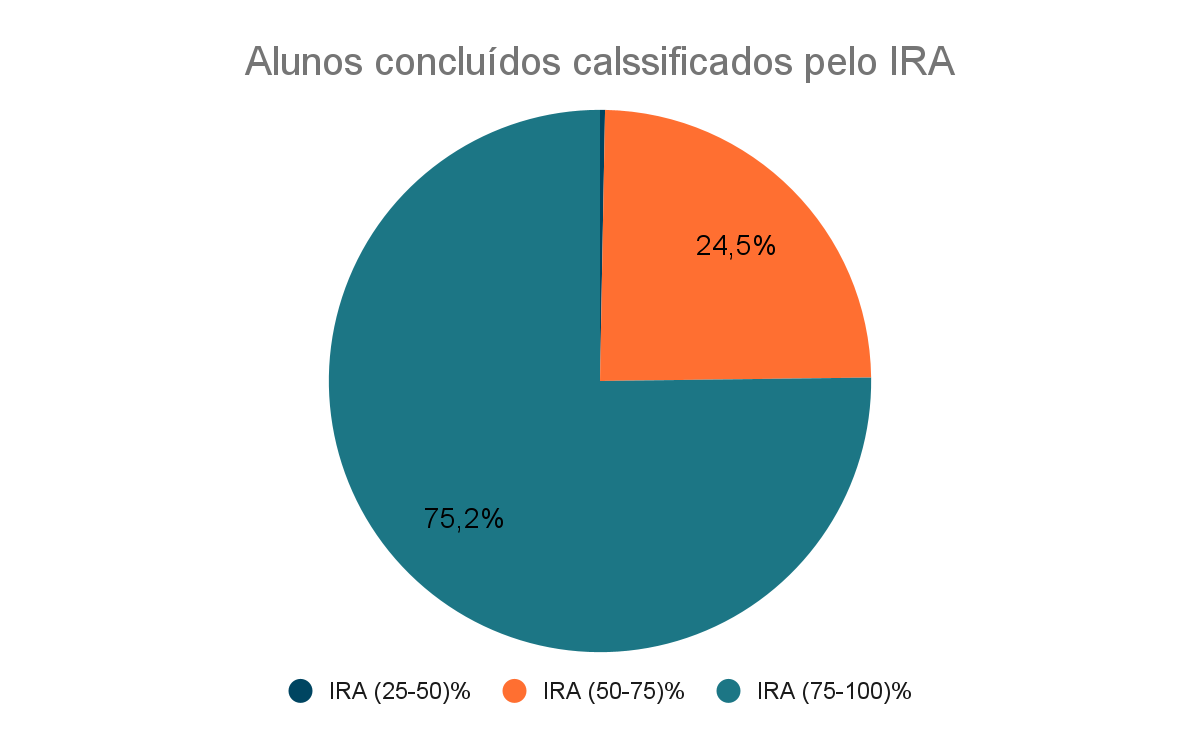
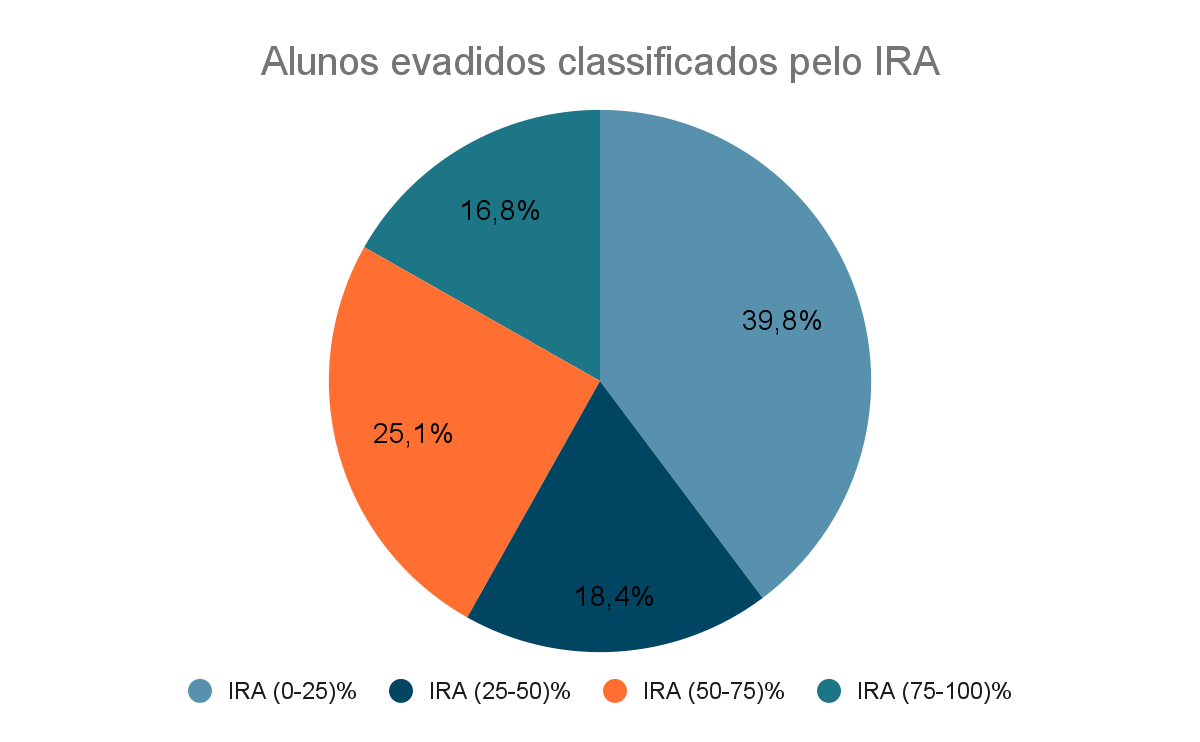
Figura 3: Gráfico dos alunos evadidos e concluídos classificados pelo sexo.

****

Através dessa análise podemos perceber que existe uma quantidade maior de homens evadindo do que mulheres, sendo que 52,3% dos alunos evadidos são homens e 47,7% são mulheres 60,7% dos alunos que concluíram seus cursos 60,7% são mulheres.

Filtrando o *dataset* pelo grupo IRA dos alunos dividindo o IRA em quatro níveis, um a cada 25% de pontuação. No gráfico da esquerda temos o grupo IRA dos alunos evadidos e no da direita dos alunos concluídos, conforme a Figura 4.

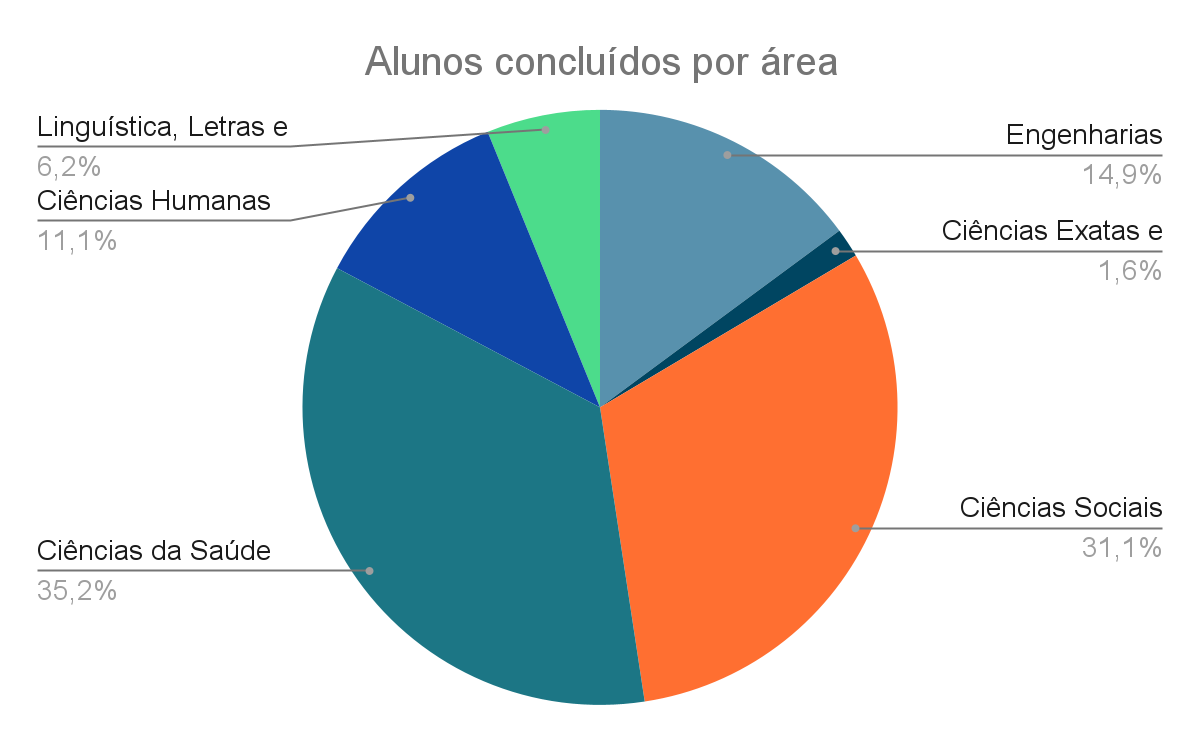
Figura 4: Gráfico dos alunos evadidos e concluídos classificados pelo grupo IRA.

****

Observando o gráfico do IRA médio de todos alunos evadidos é possível perceber que os alunos com IRA 0-25% possuem a maior taxa de evasão (39,8%) e os alunos com o IRA mais alto possuem a menor (16,8%). Analisando o gráfico de concluídos, a grande maioria (75,2%) dos alunos possui o IRA mais alto do que 75% e 24,5% possui o IRA entre 50% e 75%.

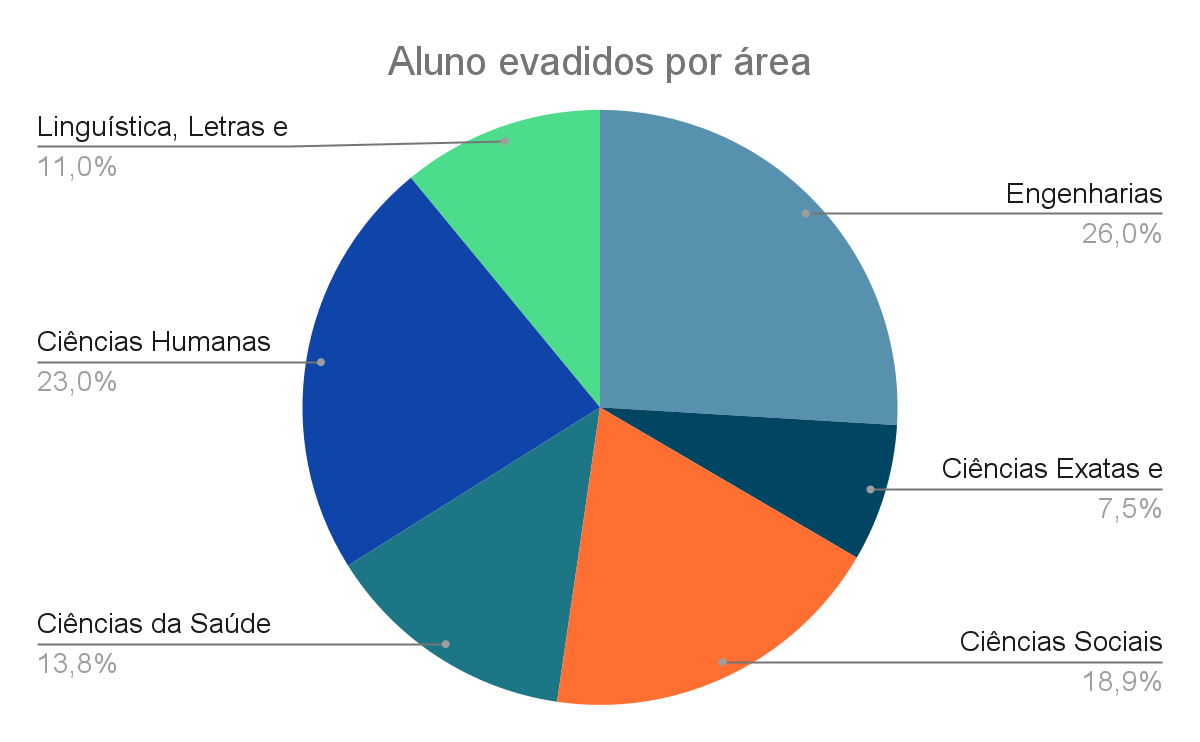
Separando os alunos concluídos por área de estudo de acordo com a Figura 5, utilizando o dataset dos alunos com mais de 70% do curso concluído. Os estudantes dos cursos da área de ciências da saúde possuem a maior porcentagem de concluídos de 35,2%.

Figura 5: Gráfico dos alunos concluídos classificados pela área de estudo.

****

Gerando o gráfico dos alunos evadidos pela área de estudo conforme a Figura 6, é possível perceber que os alunos dos cursos da área de engenharia tem a maior porcentagem de evadidos totalizando 26% comparado com as outras áreas.

Figura 6: Gráfico dos alunos evadidos classificados pela área de estudo.

****

Realizando a leitura do *dataset* filtrado pela situação dos alunos evadidos e concluídos, gerando uma correlação entre as variáveis e o mapa de calor, obtivemos o seguinte resultado, sendo vermelho mais próximo (existe uma correlação maior) e azul o mais distante (existe uma correlação menor) conforme a Tabela 3:

Tabela 3: Coeficiente de correlação entre as variáveis para o dataframe filtrado por concluintes e com o dataframe completo.

|  | Situação - Concluintes | Situação - Geral |
| --- | --- | --- |
| etnia\_INDIGENA | 0,00 | -0,01 |
| etnia\_OUTRA | -0,01 | -0,05 |
| etnia\_AMARELA | 0,01 | 0,00 |
| grandearea\_Ciências Humanas | -0,02 | -0,02 |
| grandearea\_Lingüística, Letras e Artes | -0,02 | 0,03 |
| etnia\_PARDA | -0,03 | -0,07 |
| grandearea\_Ciências Biológicas | -0,03 | 0,05 |
| mesma\_cidade | -0,03 | 0,14 |
| turno\_MATUTINO | 0,05 | 0,02 |
| cota\_numerica | -0,05 | -0,01 |
| etnia\_BRANCA | 0,06 | 0,08 |
| etnia\_PRETA | -0,06 | -0,01 |
| turno\_INTEGRAL | 0,07 | 0,15 |
| idadeSaida | -0,08 | 0,17 |
| montccest\_bool | 0,08 | 0,09 |
| grandearea\_Ciências Sociais Aplicadas | 0,09 | 0,04 |
| GAP | -0,09 | -0,05 |
| turno\_NOTURNO | -0,10 | -0,16 |
| grandearea\_Ciências da Saúde | 0,10 | 0,07 |
| sexo\_F | 0,10 | 0,13 |
| sexo\_M | -0,10 | -0,13 |
| grandearea\_Ciências Exatas e da Terra | -0,13 | -0,15 |
| grandearea\_Engenharias | -0,16 | -0,07 |
| diff\_periodos | -0,09 | 0,55 |
| grupoira | 0,62 | 0,65 |
| Situacao\_num | 1,00 | 1,00 |

Observando a tabela de correlação de todas as variáveis com a variável situação, que pode chegar no máximo 1 e no mínimo -1, de forma que os valores mais altos (quanto mais próximo de 1) apresentam uma correlação mais forte que essas variáveis possuem uma tendência compartilhada, ou seja, se o valor de uma cresce a outra também cresce. Os valores mais próximos de 0 indicam não ter correlação, que as variáveis não possuem uma tendência em comum. Já para os valores mais baixos (mais próximos de -1) também possuem uma correlação forte, porém oposta de forma que quando uma cresce a outra diminui.

É importante ressaltar que essa análise é uma correlação meramente linear, o que não é uma relação direta de causa ou consequência. Analisando os dados obtidos fica evidente que o IRA (Índice de Rendimento Acadêmico) é o critério que possui a maior correlação com a situação do aluno, isso significa que quanto maior o IRA pode existir uma probabilidade maior do aluno concluir o curso, isso é observado para os dois casos, considerando apenas os concluintes e também considerando todos os alunos. O tempo de permanência (diff\_periodos) tem uma correlação muito maior para os alunos que concluíram menos de 70% do curso, indicando que a taxa de evasão é realmente maior no início do curso.

**4.2 - Algoritmos de Aprendizagem de Máquina**

Partindo desses dados já tratados e analisados foram aplicados algoritmos de aprendizagem de máquina supervisionada para a realização das predições em cima desses dados para servirem como treinamento para o algoritmo.

Utilizando o algoritmo Decision Tree (árvore de decisão) obtivemos as seguintes porcentagens de precisão de acordo com a Tabela 4:

Tabela 4: Porcentagens de acerto da previsão do algoritmo de árvore de decisão relacionadas a cada situação, para o dataframe geral e dos concluintes.

| Decision Tree | | |
| --- | --- | --- |
| Situação | Acurácia - Concluintes | Acurácia - Geral |
| Concluído | 96% | 94% |
| Evadido | 66% | 95% |

Os resultados obtidos através do algoritmo de árvore de decisão utilizando o dataframe com todos os aluno obteve uma precisão muito alta com a média ponderada equivalente a aproximadamente 95%, os resultados utilizando o dataframe apenas dos alunos concluintes não foram satisfatórios para prever os alunos evadidos obtendo apenas 66% de precisão, já para os alunos concluídos obteve bons resultados e média ponderada equivalente a aproximadamente 92%.

Para o algoritmo KNN (Vizinhos mais próximos de K) os resultados foram os seguintes de acordo com a Tabela 5:

Tabela 5: Porcentagens de acertos da previsão do algoritmo KNN relacionadas a cada situação, para o dataframe geral e dos concluintes.

| KNN | | |
| --- | --- | --- |
| Situação | Acurácia - Concluintes | Acurácia - Geral |
| Concluído | 92% | 89% |
| Evadido | 90% | 96% |

Através do KNN, utilizando o dataframe de alunos concluintes, foi obtido uma precisão de 92% para os alunos concluídos e 90% para os alunos evadidos, o que torna o desempenho do KNN satisfatório para esse dataframe com média ponderada igual a 91%, já para o dataframe com todos os alunos o algoritmo obteve 89% de precisão para os concluídos e 96% para os evadidos, obtendo uma precisão melhor para os evadidos comparado o dataframe de concluintes.

Utilizando do algoritmo Logistic Regression (Regressão Logística), obtivemos o seguinte resultado de acordo com a Tabela 6:

Tabela 6: Porcentagens de acertos da previsão do algoritmo de regressão logística relacionadas a cada situação, para o dataframe geral e dos concluintes.

| Logistic Regression | | |
| --- | --- | --- |
| Situação | Acurácia - Concluintes | Acurácia - Geral |
| Concluído | 93% | 93% |
| Evadido | 96% | 97% |

A regressão logística para o dataframe de concluintes obteve 93% de precisão para os alunos concluídos e 96% para os alunos evadidos, que é um percentual alto para os alunos evadidos considerando que o número de alunos é muito menor devido ao *dataset* ter sido filtrado pelos alunos com mais de 70% do curso concluído e 30% da amostra para teste, para o dataframe completo obteve uma porcentagem de precisão de 93% para os concluídos e 97% para os evadidos, o que torna a regressão logística um algoritmo satisfatório para ambos casos.

Aplicando o algoritmo Random Forest (Floresta Aleatória) obtivemos o seguinte resultado de acordo com a Tabela 7:

Tabela 7: Porcentagens de acertos da previsão do algoritmo Random Forest relacionadas a cada situação, para o dataframe geral e dos concluintes.

| Random Forest | | |
| --- | --- | --- |
| Situação | Acurácia - Concluintes | Acurácia - Geral |
| Concluído | 90% | 89% |
| Evadido | 98% | 98% |

O Random Forest utilizando do dataframe com os alunos concluintes obteve uma precisão de 90% para os alunos concluídos e 98% para os alunos evadidos, com uma média ponderada de aproximadamente 92%, o algoritmo manteve uma boa média para o dataframe completo com a porcentagem de precisão de 89% para os concluintes e 98% para os evadidos, o que é uma taxa de acerto muito grande para os evadidos e por serem o foco do trabalho o algoritmo se desempenhou muito bem.

**5- Conclusão**

A partir dos resultados obtidos através das previsões dos algoritmos utilizados podemos concluir que o algoritmo Random Forest obteve o melhor desempenho com o *dataset* utilizado, o algoritmo teve precisão de 98% para alunos evadidos para ambos *datasets* usados como base de treinamento e teste, tanto o dataset filtrado pelos alunos com mais de 70% do curso concluído, tanto para o *dataset* completo utilizando todos os alunos (fora os ativos). O algoritmo de árvore de decisão foi o que obteve o pior desempenho para prever os alunos evadidos, conseguindo uma acurácia de apenas 66% o que torna o resultado não confiável.

**Referências Bibliográficas**

GÉRON, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media, 2019.

BRITO Junior, Inamar et al. Uso de mineração de dados educacionais para a classificação e identificação de perfis de evasão de graduandos em Sistemas de Informação. In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação. 2019. p. 159.

COSTA, Anna Edith B; CAMPOS, Helton Rocha. Disponibilidade de informação e

comportamento de evasão. Avaliação Institucional: Belo Horizonte. UFMG,

2000

DE JESUS, Hiago Oliveira; RODRIGUEZ, Luis Cuevas; JUNIOR, Almir de Oliveira Costa. Predição de Evasão Escolar na Licenciatura em Computação. Revista Brasileira de Informática na Educação, v. 29, p. 255-272, 2021.

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

FEYYAD, U. M. Data mining and knowledge discovery: Making sense out of data. IEEE expert, v. 11, n. 5, p. 20-25, 1996.

FIALHO, Marillia Gabriella Duarte. A evasão escolar e a gestão universitária: o caso da Universidade Federal da Paraíba. 2014. 107 f. Dissertação (Mestrado em Gestão de Organizações Aprendentes) - Universidade Federal da Paraí­ba, João Pessoa, 2014.

GONZALEZ, Leandro de Azevedo. Regressão logística e suas aplicações. São Luís, 2018.

PEREIRA, Fernanda Cristina Barbosa et al. Determinantes da evasão de alunos e os custos ocultos para as instituições de ensino superior: uma aplicação na Universidade do Extremo Sul Catarinense. 2003.

PRESTES, Emília Maria da T., Marillia Gabriella D. FIALHO, and Dietmar K. PFEIFFER. "A evasão no ensino superior globalizado e suas repercussões na gestão universitária." João Pessoa, 2014.

SANTANA, A. P.; PEROSSO, J. E. C.; MACEDO, K. L. O.; FARIAS, S. P. D de.(1996)

Evasão escolar em escolas públicas municipais rurais localizadas em Montes

Claros. Trabalho de Conclusão de Curso. Universidade Estadual de Montes Claros.

Montes Claros: 1996.

VASQUEZ, Johanna & Castaño, Elkin & Gallón, Santiago & Gomez, Karoll. (2003). Determinantes de la deserción estudiantil en la Universidad de Antioquia.

SANTOS, Patrícia, e Denise Goya. "Aprendizado de Máquina Aplicado à Análise de Evasão em Cursos de Sistemas de Informação." Anais Estendidos do XVI Simpósio Brasileiro de Sistemas de Informação. SBC, 2020.

1. Disponível em: <<https://www2.ufjf.br/ufjf/>> Acesso em 01 de novembro de 2021 [↑](#footnote-ref-0)
2. Disponível em: <<https://www.semesp.org.br/mapa-do-ensino-superior/edicao-11/dados-brasil/evasao/>> Acesso em 20 de agosto de 2021. [↑](#footnote-ref-1)
3. Disponível em: <<https://falabr.cgu.gov.br>> Acesso em 30 de abril de 2021. [↑](#footnote-ref-2)