Análise do Censo Educacional Brasileiro

Julio Cesar do Valle Modesto
Matheus dos Santos Pereira
Paulo Jorge Gonçalves Júnior
Orientador: Ricardo Silva Campos

Resumo

Em decorrência do início e avanço da pandemia causada pelo vírus Covid-19, muitas dificuldades têm sido recorrentes em muitos setores e níveis da sociedade, incluindo a educação. Com a adoção do modelo remoto no ensino básico, devido à urgência na mudança no modelo educacional ocasionada pela quarentena, uma grande barreira foi colocada entre os alunos e instituições de ensino, os quais em grande parte e principalmente em áreas mais carentes, sofrem com a falta de infraestrutura, acompanhamento e apoio para o prosseguimento das atividades escolares.

Um levantamento recente realizado pela organização Todos pela Educação, mostra que 244 mil crianças de 6 a 14 anos estavam fora da escola no segundo trimestre de 2021, número este que representa um aumento de 171% em comparação a 2019, quando 90 mil crianças estavam fora da escola.

Desta forma, é possível observar que se faz necessário dar atenção, analisar e estudar os motivos que contribuem para a evasão de alunos do ensino básico com o objetivo de encontrar métodos que auxiliem e evitem este resultado, pois, assim como houve a necessidade de mudança nos modelos de ensino devido às restrições causadas pelo Covid-19, futuramente, existe a possibilidade de situações semelhantes ocorrerem, onde haverá uma mudança brusca e repentina nos modelos de ensino e devemos estar preparados para tais.

O objetivo deste projeto constitui-se no estudo dos fatores que estejam envolvidos no fenômeno de evasão escolar, sendo o seu diferencial, a realização de um estudo mais recente e aprofundado com dados dos anos de 2019, 2020 e 2021, que foram afetados pela pandemia do Covid-19.

Palavras-Chave: Evasão Escolar, Pandemia, Base de Dados, Tecnologia da Informação.

1-Introdução

Graças aos avanços tecnológicos e a integração da sociedade com sistemas automatizados sendo cada vez mais frequentes, nos dias de hoje, possuímos uma vasta base de

dados relacionados a educação, que motiva a realização de um estudo muito mais aprofundado e com maior exatidão, o que nos possibilita identificar os motivos e fatores que estejam causando a evasão escolar. Com o crescimento desta base de dados, é possível utilizar técnicas desenvolvidas no ambiente de Tecnologia da Informação, que permite, de forma automatizada, analisarmos os dados e constituir fontes de informações para o contínuo acompanhamento dos resultados obtidos relacionados às matrículas e evasão de alunos e tendências a serem tratadas.

Para a realização dos estudos, serão utilizadas técnicas de Mineração de Dados (Data Mining) e Aprendizado de Máquina (Machine Learning). Estas técnicas têm como objetivo possibilitar a realização da análise dos dados de forma automatizada, pois, devido ao volume das bases de dados coletadas, não seria possível a realização do estudo de forma manual, sem a integração com Sistemas de Informação. Técnicas de Mineração de Dados são aperfeiçoadas e utilizadas para o processo de explorar grandes quantidades de dados à procura de padrões consistentes, através do uso de algoritmos de aprendizagem ou classificação baseados em redes neurais e estatística. Estes são capazes de explorar um conjunto de dados, extraindo ou ajudando a evidenciar padrões nestes dados e auxiliando na descoberta do conhecimento.

Com a exploração e extração dos padrões na base de dados, utilizaremos técnicas de Aprendizado de Máquina, que são sistemas desenvolvidos que podem modificar seu comportamento automaticamente tendo como base a sua própria existência. Tal modificação comportamental, consiste, basicamente, no estabelecimento de regras lógicas, que visam melhorar o desempenho de uma tarefa, ou, tomar a decisão mais apropriada para o contexto. Essas regras são geradas com base no reconhecimento de padrões dentro dos dados analisados.

Este projeto, está estruturado de forma a separar e descrever os temas separadamente em seções. Na primeira seção, serão identificados elementos do problema de evasão escolar, sendo descritos em maiores detalhes algumas das suas implicações no ambiente educacional. Em seguida, serão discutidas abordagens para solução do problema da evasão escolar utilizando a Mineração dos Dados Educacionais, bem como aspectos importantes a serem ampliados neste contexto. Na próxima seção serão analisados trabalhos relacionados e em seguida estará descrita a abordagem sugerida neste projeto, bem como os sistemas, algoritmos e ferramentas utilizados. Os resultados do estudo de caso com a abordagem sugerida serão descritos na próxima seção. No final do projeto, em sua última seção, será apresentado uma conclusão do estudo e trabalhos futuros.

1.1- Justificativa

A evasão escolar possui consequências que se distribuem para todo o corpo da sociedade, acarretando principalmente no aumento da desigualdade social, comprometimento do desenvolvimento cognitivo, intelectual e cultural. Quanto maior for a evasão escolar, consequentemente mais desqualificado profissionalmente será a população de uma nação.

O indivíduo que abandona os estudos, provavelmente terá dificuldades de se colocar no mercado de trabalho, levando em consideração que sem o certificado de conclusão do ensino médio, o que se é possível obter normalmente são trabalhos com baixa remuneração.

Também é possível observar que devido a ocorrência da recente pandemia causada pelo vírus COVID-19, que se iniciou no Brasil em Fevereiro de 2020, a evasão/abandono escolar dos estudantes se tornou muito mais manifesta entre o período inicial e o momento presente da pandemia com relação aos anos anteriores. A Covid-19 é um agravante à evasão escolar, evidenciando a necessidade de ações de enfrentamento para combatê-la, especialmente nesse contexto de incertezas e diante dos impactos da pandemia sobre a educação. (PEREIRA DE SOUZA, 2020).

1.2 - Objetivos

1.1.1- Objetivo Geral

O projeto têm como objetivo realizar um estudo aprofundado das bases de dados disponibilizadas pelas entidades governamentais que possuem uma série de informações que podem nos auxiliar a identificar previamente, fatores que possam estar corroborando para a evasão escolar, e com os resultados deste estudo, podermos auxiliar nas tomadas de decisões que irão contribuir para que a evasão seja menos recorrente.

1.1.2- Objetivos Específicos

O projeto desenvolvido tem como objetivo identificar os principais fatores que reforçam o crescimento da evasão escolar, dado sua natureza descritiva e expositiva, buscamos extrair as informações através da análise de correlação entre os dados e dos resultados obtidos através das metodologias de aprendizagem de máquina.

2- Referencial Teórico

O trabalho de Colpani (2018) explica que as perdas ocasionadas pela evasão podem acarretar prejuízos às instituições de ensino, sendo, para o setor público, os recursos investidos sem o devido retorno; para o setor privado, importante perda de receita; para ambos os setores, fonte de ociosidade de professores, funcionários, equipamentos e espaço físico.

O problema da evasão escolar está relacionado não somente às instituições estudantis, mas também nos setores públicos e privados, sendo esta, uma grande preocupação para empresários, diretores, pesquisadores, pais e alunos. (LOBO, 2012; BITTENCOURT; MERCADO, 2014).

Conforme demonstrado por Desjardins et al. (1999), o desempenho escolar satisfatória é capaz de ampliar a retenção do conhecimento, logo, o sucesso no âmbito acadêmico seria o melhor preditor de permanência dos alunos. Assim, os autores comentam que os menores índices de reprovação consequentemente irão levar a uma redução da evasão. Veloso (2015) observa os diversos fatores que causaram a evasão escolar como: fatores sociais, educacionais, de localização e econômicos.

Alguns projetos são desenvolvidos na tentativa de mitigar as causas de evasão escolar, como é o exemplo do projeto desenvolvido por dos Santos(2021). O projeto utilizou a base de dados disponibilizada pela Diretoria de Tecnologia da Informação e Comunicação (DTIC) do Instituto Federal Santa Catarina (IFSC) Câmpus Caçador dos estudantes de graduação. Após a limpeza dos dados, substituição dos dados faltantes, foi feita a codificação dos dados, que pode ser usado quando uma coluna não possui uma ordem ou sequência. Para a realização do trabalho, foi utilizado algoritmos de rede neural e árvore de decisão. Na rede neural, foi definido o número de iterações, criado o objeto do classificador, um treinamento do modelo usando os dados selecionados e um modelo de predição com os dados selecionados. Por sua vez, na árvore de decisão, foi criado o objeto, o treinamento de um conjunto de dados definidos na preparação do *dataset*, e após isso realizado a predição dos dados selecionados. Foi criado um aplicativo para prever a possibilidade de evasão do aluno, em que no aplicativo é mostrado de forma gráfica os resultados obtidos anteriormente com uma interface web , em que após a definição de parâmetros preenchidos pelo usuário, o aplicativo mostra uma predição.

A mineração de dados é responsável pela extração de um conhecimento de uma base de dados, com a finalidade de auxiliar na tomada de decisão. O problema da evasão escolar está

relacionado a uma série de fatores sociais, econômicos, desempenho e escolha. Ainda assim, o uso de recursos tecnológicos auxilia no estudo e acompanhamento dos alunos prevenindo a evasão escolar. (RIGO; CAMBRUZZI; BARBOSA;CAZELLA, 2014)

3- Metodologia

Para o desenvolvimento do projeto, o primeiro passo foi selecionar a base de dados dos anos de 2019, 2020 e 2021 disponibilizadas no site Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep)[1]. Após a obtenção dos *dataset*, foi necessário selecionar quais colunas seriam usadas para o desenvolvimento do projeto, e o tratamento dessas colunas, visto que nem todos os dados apresentam-se preenchidos. Com os dados prontos, foi utilizado o algoritmo de árvore de decisão criado com bibliotecas da linguagem de programação Python, a fim de encontrar a principal variável que contribui para a não efetivação de matrículas e também com esses dados, foi feito a correlação entre as variáveis com a finalidade de confrontar e justificar os resultados da árvore. Posteriormente ao processamento dos dados, foi obtido a variável IN_QUADRA_ESPORTES como nó principal da árvore, que é a principal variável que contribui para a diminuição de matrículas e também a correlação de cada variável.

Após a obtenção das bases de dados através do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep)[1], utilizando a linguagem de programação Python, na plataforma disponibilizada pela Google, Google Colab e fazendo uso também das bibliotecas Pandas, Numpy, Matplotlib, Seaborn, plotly e sklearn foi realizado o processo de escolha das colunas (variáveis) que fariam parte da análise bem como suas respectivas normalizações, onde foram tratados dados ausentes utilizando técnicas de Preenchimento de Dados através da realização de médias entre os valores presentes, Imputação de Valores por meio das células que possuíam maior frequência e Exclusão de valores ausentes, quando o preenchimento/imputação de dados ausentes resultaria em um risco de distorção nos resultados, como é o caso da variável QT_MAS_BAS (Número de Matrículas na Educação Básica).

Realizado o processo de normalização, houve a necessidade de criação da coluna “SALDO_MATRICULAS_POSITIVO_MEDIAUF” (Valor Categórico que indica se o número de matrículas em cada escola é superior ou inferior à média de matrículas de sua respectiva Unidade da Federação - UF), visto que para que o Algoritmo de Árvore de Decisão traga os resultados esperados e possa realizar as previsões, é necessário que haja uma variável categórica que indique neste caso, se está ou não ocorrendo a evasão de alunos devido à falta de matrículas, sem a criação desta coluna, não seria possível trabalhar com este parâmetro e definir as previsões

com base nos dados de cada escola.

O Algoritmo utilizado para o desenvolvimento do trabalho foi a Árvore de Decisão, este por sua vez é um algoritmo de decisão por aprendizagem de máquina (*machine learning*) largamente utilizado, com uma estrutura de simples compreensão que costuma apresentar bons resultados em suas previsões. Neste algoritmo, vários pontos de decisão, chamados de nós, são criados e em cada um deles, o resultado irá definir qual caminho será seguido, semelhante a um fluxograma. Neste processo, é realizado pelo algoritmo o cálculo do índice GINI, onde será verificado a distribuição dos dados nas variáveis de acordo com a variação da variável target. A variável preditora com o menor índice Gini será escolhida para o nó principal da árvore, pois um baixo valor do índice indica maior ordem na distribuição dos dados.

Nas figuras a seguir, se encontram as variáveis que foram selecionadas e utilizadas no desenvolvimento do projeto, através delas foram realizados os processos de aprendizado de máquina para obtenção dos resultados. Estas se encontram divididas conforme a instrução dos Dicionários de Dados disponibilizados também pelo INEP em Cadastro Escola, onde são encontradas as informações referente à localização das Escolas, Dados da Escola, onde são disponibilizados os dados referentes a infraestrutura e corpo docente, Dados da Oferta de Matrícula, que contém os processos de realização das matrículas realizadas anualmente e o Número de Matrículas, que neste caso está sendo analisado o Número de Matrículas na Educação Básica.

[1]Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-escolar> Acesso em 30 de agosto de 2022

Figura 1: Relação de variáveis utilizadas no *DataSet* para desenvolvimento do trabalho.

CADASTRO ESCOLA	
NOME VARIAVEL	DESCRIÇÃO
NU_ANO_CENSO	Ano do Censo
NO_REGIAO	Nome da região geografica
CO_REGIAO	Código da região geografica
NO_UF	Nome da Unidade da Federação
SG_UF	Sigla da Unidade da federação
CO_UF	Código da Unidade da Federação
NO_MUNICIPIO	Nome do Município
CO_MUNICIPIO	Código do Município
NO_MICRORREGIAO	Nome da Microrregião
CO_MICRORREGIAO	Código da Microrregião
DADOS DA ESCOLA	
TP_LOCALIZACAO	Localização
TP_SITUACAO_FUNCIONAMENTO	Situação de funcionamento
IN_ENERGIA_INEXISTENTE	Abastecimento de energia elétrica - Não há energia elétrica
IN_ESGOTO_INEXISTENTE	Esgoto sanitário - Não há esgotamento sanitário
IN_BANHEIRO	Dependências físicas existentes e utilizadas na escola - Banheiro
IN_BIBLIOTECA	Dependências físicas existentes e utilizadas na escola - Biblioteca
IN_COZINHA	Dependências físicas existentes e utilizadas na escola - Cozinha
IN_LABORATORIO_CIENCIAS	Dependências físicas existentes e utilizadas na escola - Laboratório de ciências
IN_LABORATORIO_INFORMATICA	Dependências físicas existentes e utilizadas na escola - Laboratório de informática
IN_PARQUE_INFANTIL	Dependências físicas existentes e utilizadas na escola - Parque infantil
IN_QUADRA_ESPORTES	Dependências físicas existentes e utilizadas na escola - Quadra de esportes
IN_REFEITORIO	Dependências físicas existentes e utilizadas na escola - Refeitório
IN_ACESSIBILIDADE_INEXISTENTE	Recursos de acessibilidade para pessoas com deficiência ou mobilidade reduzida.
IN_COMPUTADOR	Equipamentos existentes na escola para uso técnico e administrativo - Computador
IN_DESKTOP_ALUNO	Computadores em uso pelos alunos - Computador de mesa (desktop)
IN_COMP_PORTATIL_ALUNO	Computadores em uso pelos alunos - Computador portátil
IN_TABLET_ALUNO	Computadores em uso pelos alunos - Tablet
IN_INTERNET	Acesso à Internet
IN_ACESSO_INTERNET_COMPUTADOR	Equipamentos que os alunos usam para acessar a internet da escola
IN_PROF_ADMINISTRATIVOS	Profissionais que atuam na escola - Auxiliares de secretaria ou auxiliares administrativos, atendentes
IN_PROF_SERVICOS_GERAIS	Profissionais que atuam na escola - Auxiliar de serviços gerais, porteiro(a), zelador(a), faxineiro(a), horticultor(a), jardineiro(a)
IN_PROF_BIBLIOTECARIO	Profissionais que atuam na escola - Bibliotecário(a), auxiliar de biblioteca ou monitor(a) da sala de leitura
IN_PROF_SAUDE	Profissionais que atuam na escola - Bombeiro(a) brigadista, profissionais de assistência à saúde (urgência e emergência), Enfermeiro(a), Técnico(a) de enfermagem e socorrista
IN_PROF_COORDENADOR	Profissionais que atuam na escola - Coordenador(a) de turno/disciplina
IN_PROF_FONAUDIOLOGO	Profissionais que atuam na escola - Fonoaudiólogo(a)
IN_PROF_PSICOLOGO	Profissionais que atuam na escola - Psicólogo(a) Escolar

Figura 2: Continuação da relação de variáveis utilizadas no *DataSet* para desenvolvimento do trabalho.

IN_PROF_ALIMENTACAO	Profissionais que atuam na escola - Profissionais de preparação e segurança alimentar, cozinheiro(a), merendeiro(a).
IN_PROF_PEDAGOGIA	Profissionais que atuam na escola - Profissionais de apoio e supervisão pedagógica: pedagogo(a), coordenador(a) pedagógico(a), orientador(a) educacional, supervisor(a) escolar e coordenador(a) de área de ensino
IN_PROF_SECRETARIO	Profissionais que atuam na escola - Secretário(a) escolar
IN_PROF_SEGURANCA	Profissionais que atuam na escola - Seguranças, guarda ou segurança patrimonial
IN_PROF_MONITORES	Profissionais que atuam na escola - Técnicos(as), monitores(as), supervisores(as) ou auxiliares de laboratório(s), de apoio a tecnologias educacionais ou em multimeios/multimídias eletrônico/digitais
IN_ALIMENTACAO	Alimentação escolar para os alunos
IN_EXAME_SELECAO	A escola faz exame de seleção para ingresso de seus alunos (Avaliação por prova e/ou análise curricular)
IN_REDES_SOCIAIS	A escola possui site ou blog ou página em redes sociais para comunicação institucional
IN_ORGAO_ASS_PAIS	Órgãos colegiados em funcionamento na escola - Associação de Pais
IN_ORGAO_CONSELHO_ESCOLAR	Órgãos colegiados em funcionamento na escola - Conselho Escolar
DADOS DA OFERTA DE MATRICULA	
TP_AEE	Atendimento Educacional Especializado (AEE)
TP_ATIVIDADE_COMPLEMENTAR	Atividade Complementar
IN_MEDIACAO_PRESENCIAL	Mediação didático-pedagógica oferecida pela escola - Presencial
IN_MEDIACAO_SEMIPRESENCIAL	Mediação didático-pedagógica oferecida pela escola - Semipresencial
IN_MEDIACAO_EAD	Mediação didático-pedagógica oferecida pela escola - Educação a Distância - EAD
IN_DIURNO	Turno - Diurno - Horário de início da turma de escolarização entre 05h e 16h
IN_NOTURNO	Turno - Noturno - Turno de início da turma de escolarização entre 17h e 04h
IN_EAD	Turno Não aplicável para turmas semipresenciais ou educação a distância (EAD)
IN_BAS	Educação Básica (Possui uma ou mais matrículas)
IN_INF	Etapa de Ensino - Educação Infantil (Possui uma ou mais matrículas)
IN_FUND	Etapa de Ensino - Ensino Fundamental (Possui uma ou mais matrículas)
IN_MED	Etapa de Ensino - Ensino Médio (Possui uma ou mais matrículas)
NUMERO DE MATRICULAS	
QT_MAT_BAS	Número de Matrículas na Educação Básica

Dado o estabelecimento e construção das árvores de decisão, também será realizada a análise da correlação entre os atributos utilizados na previsão com o número de matrículas de

cada escola, dado pela variável QT_MAT_BAS (Quantidade de Matrículas do Ensino Básico). Através desta análise, buscamos observar de maneira mais clara e adequada, quais os atributos que contribuem positivamente ou negativamente à efetivação das matrículas. Este processo também tem como objetivo justificar e provar a veracidade das previsões realizadas pelo algoritmo Árvore de Decisão, visto que, poderemos visualizar os principais atributos que constituem os nós principais da árvore. Após a finalização deste processo de análise, também será possível observar os principais fatores associados ao processo de educação durante o cenário de pandemia, como a informatização do processo de ensino e como estes se relacionam para a efetivação das matrículas ou evasão dos alunos.

A correlação quantifica a associação linear entre duas variáveis demonstrando o grau e a direção de relacionamento entre elas. Assim, é dado que duas variáveis X e Y são fortemente relacionadas se um incremento em X causa o mesmo impacto em Y. É possível mensurar essa relação de diversas maneiras. Neste trabalho foi empregada a Correlação de Pearson, (FIGUEIREDO FILHO, 2009), que pode ser usada em diferentes aplicações de mineração de dados (ALQALLAF et al., 2002; REZENDE, 2022). Ela é dada pela equação:

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma\left(x_i - \bar{x}\right)^2 \Sigma\left(y_i - \bar{y}\right)^2}}.$$

O coeficiente traz consigo duas informações: o seu sinal indica a direção do relacionamento e o módulo indica a intensidade da relação. Os coeficientes estão contidos no intervalo entre -1 e 1, sendo que seu módulo igual a 1 indica a correlação perfeita. Por outro lado, o coeficiente igual a 0 indica que não há relação linear entre as variáveis. Assim, pode-se dizer que valores abaixo de 0,5 indicam baixa correlação, e valores mais próximos a 1 indicam relacionamentos fortes.

Existem vários algoritmos de aprendizagem de máquina, em especial para as tarefas de classificação em aprendizagem supervisionada, como o *k-nearest neighbors (KNN),* em que as previsões são classificadas com base na classificação do vizinho mais próximo (Cambronero; Moreno, 2006), Redes Neurais, onde um processador com unidades de processamentos armazena experiências através do ambiente de aprendizagem (Haykin, 2001) e Árvore de decisão, algoritmo usado no trabalho. Existem ainda diversas bibliotecas, como Pandas, Numpy e Matplotlib, que implementam tais algoritmos na linguagem Python (GÉRON, 2019), e aqui serão apresentados os algoritmos utilizados neste trabalho.

O algoritmo Árvore de Decisão e 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.

4 - Resultados Obtidos

O *dataset* disponibilizado possui as seguintes colunas conforme a Figura 3, com um total de aproximadamente 230 mil registros em cada um dos datasets após a realização da limpeza dos dados.

Figura 3: Algumas colunas e linhas contidas no dataset utilizado no trabalho.

	NU_ANO_CENSO	NO_REGIAO	CO_REGIAO	NO_UF	SG_UF	CO_UF	NO_MUNICIPIO	CO_MUNICIPIO	NO_MICRORREGIAO	CO_MICRORREGIAO	...	IN_DIURNO	IN_NOTURNO	IN_EAD	IN_BAS	IN_INF	IN_FUND	IN_MED
0	2019	Norte	1	Rondônia	RO	11	Alta Floresta D'Oeste	1100015	Cacoal	11006	...	1.0	0.0	0.0	1.0	0.0	1.0	0.0
1	2019	Norte	1	Rondônia	RO	11	Alta Floresta D'Oeste	1100015	Cacoal	11006	...	1.0	1.0	0.0	1.0	0.0	0.0	0.0
2	2019	Norte	1	Rondônia	RO	11	Alta Floresta D'Oeste	1100015	Cacoal	11006	...	1.0	0.0	0.0	1.0	0.0	1.0	0.0
3	2019	Norte	1	Rondônia	RO	11	Alta Floresta D'Oeste	1100015	Cacoal	11006	...	1.0	0.0	0.0	1.0	1.0	1.0	0.0
7	2019	Norte	1	Rondônia	RO	11	Alta Floresta D'Oeste	1100015	Cacoal	11006	...	1.0	0.0	0.0	1.0	1.0	1.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
228512	2019	Centro-Oeste	5	Distrito Federal	DF	53	Brasília	5300108	Brasília	53001	...	1.0	0.0	0.0	1.0	1.0	1.0	0.0
228513	2019	Centro-Oeste	5	Distrito Federal	DF	53	Brasília	5300108	Brasília	53001	...	1.0	0.0	0.0	1.0	1.0	1.0	1.0
228514	2019	Centro-Oeste	5	Distrito Federal	DF	53	Brasília	5300108	Brasília	53001	...	1.0	0.0	0.0	1.0	0.0	1.0	1.0
228517	2019	Centro-Oeste	5	Distrito Federal	DF	53	Brasília	5300108	Brasília	53001	...	1.0	0.0	0.0	1.0	1.0	1.0	0.0
228518	2019	Centro-Oeste	5	Distrito Federal	DF	53	Brasília	5300108	Brasília	53001	...	1.0	0.0	0.0	1.0	0.0	1.0	1.0

176726 rows × 61 columns

Para o estudo ser realizado foi necessário filtrar o dataset mantendo apenas os registros que continham a informação referente ao número de matrículas, totalizando aproximadamente 170 mil registros para cada dataset, devido ao fato de que os registros sem a informação referente a quantidade de matrículas efetivadas em respectiva escola não seriam úteis para o estudo, vindo da necessidade de encontrar padrões entre as escolas que possuem maior taxa de evasão e as que possuem mais matrículas efetivadas e as discrepâncias entre as mesmas.

Através do conjunto de dados tratados por suas respectivas técnicas de normalização, realizamos a aplicação do Algoritmo de Árvore de Decisão e foi obtida uma precisão de aproximadamente 80% na assertividade das previsões realizadas a partir de dados simulados de uma escola. Tal porcentagem, é obtida através da comparação das classificações calculadas pela árvore com as classificações originais, neste caso como mencionado anteriormente, o valor que se deseja prever são os da coluna “SALDO_MATRICULAS_POSITIVO_MEDIAUF” (que

indica se o número de matrículas em cada escola é superior ou inferior à média de matrículas de sua respectiva Unidade da Federação - UF). O nó principal é dado pela variável IN_QUADRA_ESPORTES, que indica se a escola possui ou não ambiente direcionado para a Prática de Esportes, visto que este atributo é o que possui maior correlação com a Quantidade de Matrículas no Ensino Básico.

Figura 4: Árvore de Decisão Base de Dados 2019

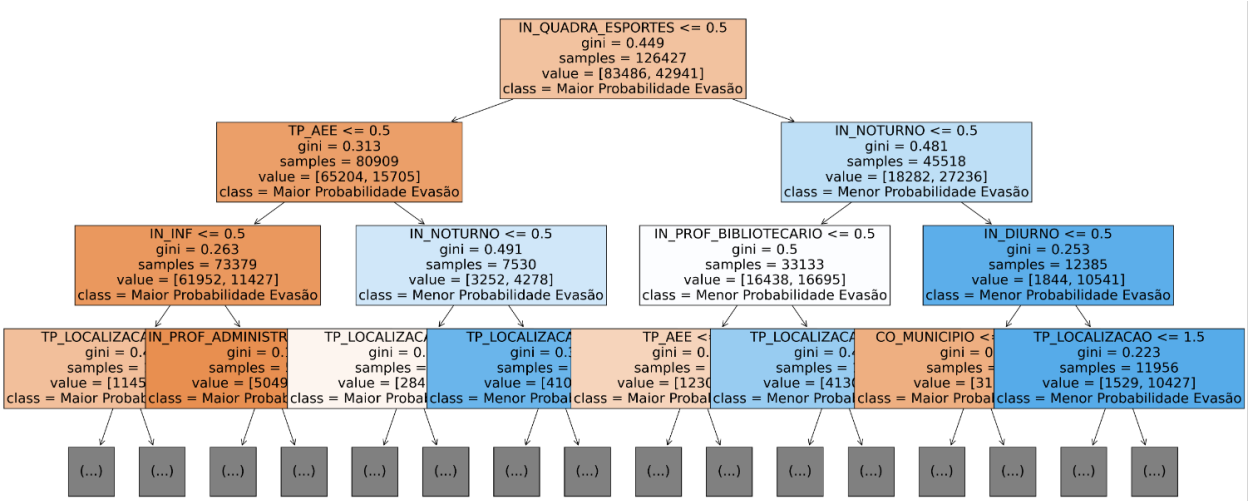


Figura 5: Árvore de Decisão Base de Dados 2020

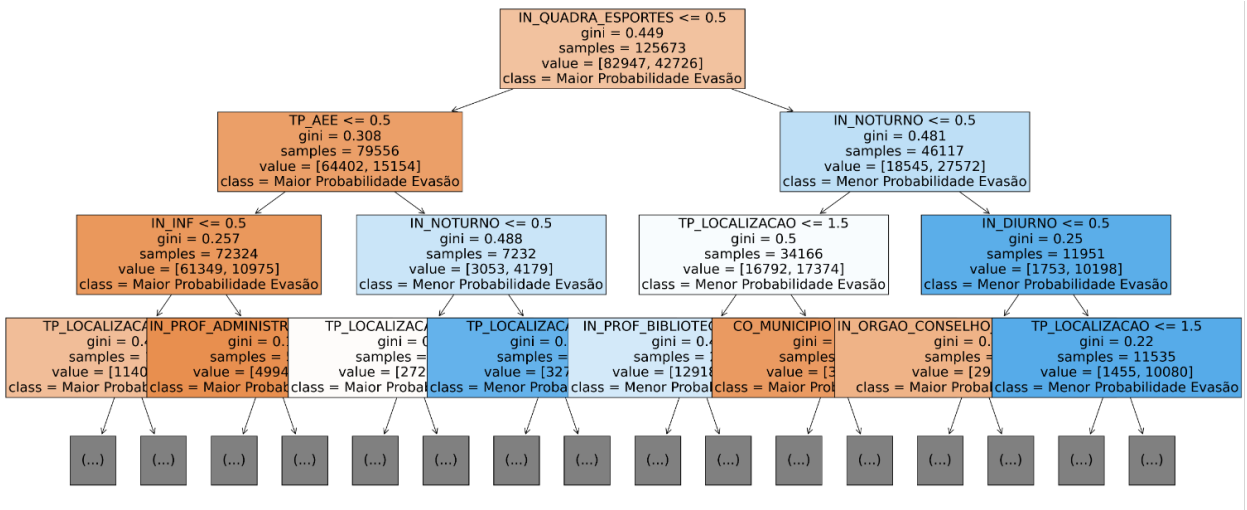
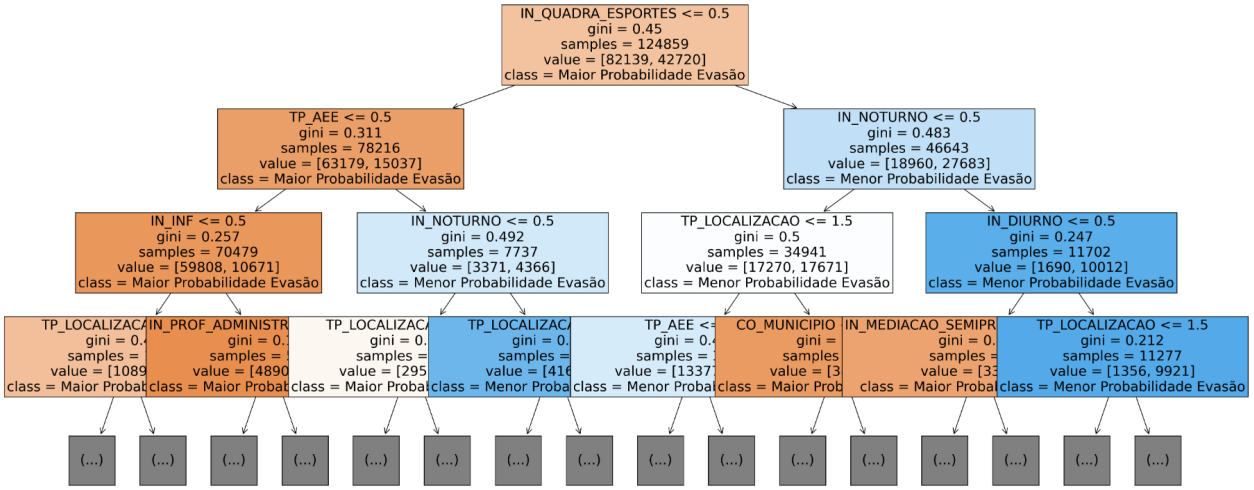


Figura 6: Árvore de Decisão Base de Dados 2021

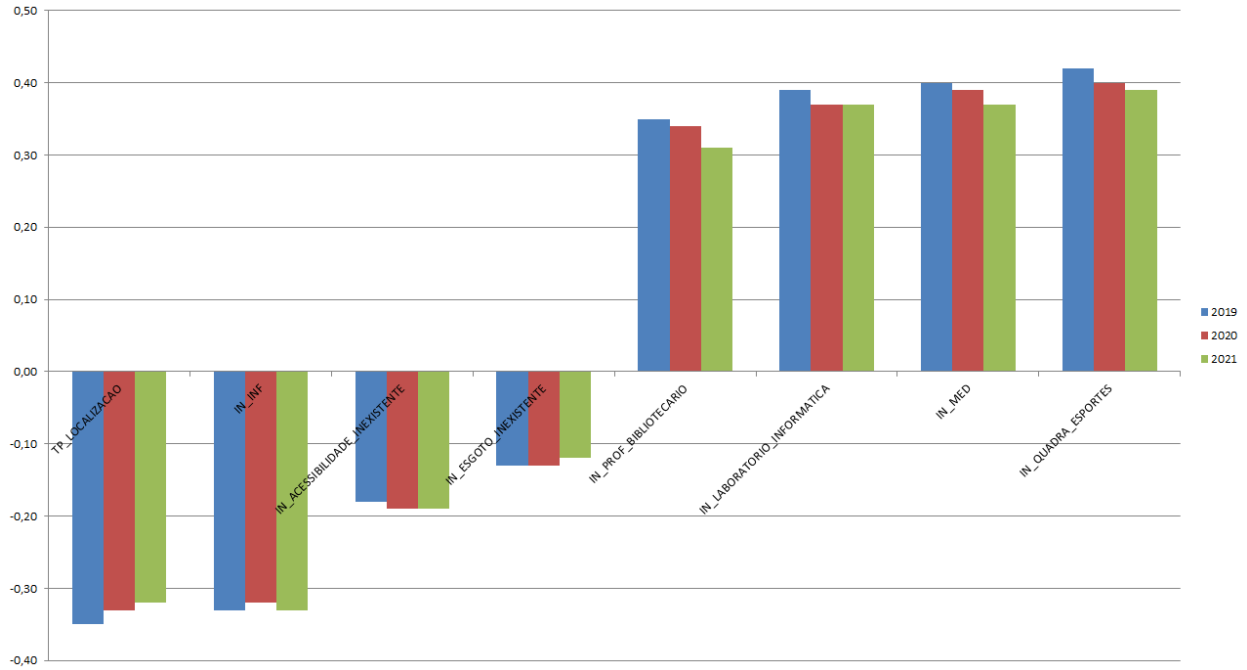


Realizada a análise da correlação entre os atributos utilizados para a construção da Árvore de Decisão e a variável QT_MAT_BAS (Quantidade de Matrículas Ensino Básico), pudemos identificar os principais fatores que estão presentes e contribuem positivamente ou negativamente a efetivação das matrículas no ensino básico, conforme é possível verificar na relação a seguir.

Figura 7: Tabela de Correlação das Principais Variáveis que contribuem positivamente ou negativamente para a efetivação das matrículas.

VARIAVEL	2019	2020	2021
TP_LOCALIZACAO	-0,35	-0,33	-0,32
IN_INF	-0,33	-0,32	-0,33
IN_ACESSIBILIDADE_INEXISTENTE	-0,18	-0,19	-0,19
IN_ESGOTO_INEXISTENTE	-0,13	-0,13	-0,12
IN_PROF_BIBLIOTECARIO	0,35	0,34	0,31
IN_LABORATORIO_INFORMATICA	0,39	0,37	0,37
IN_MED	0,4	0,39	0,37
IN_QUADRA_ESPORTES	0,42	0,4	0,39

Figura 8: Gráfico de Correlação das das Principais Variáveis que contribuem positivamente ou negativamente para a efetivação das matrículas.

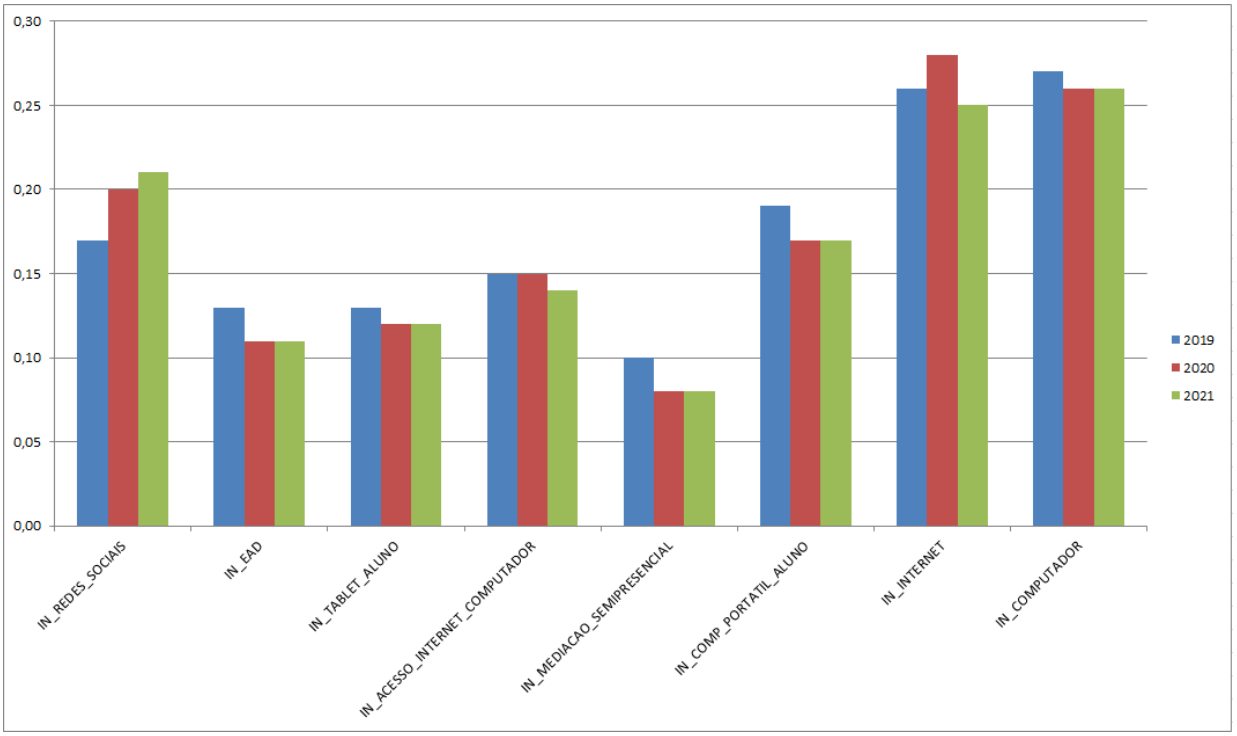


Por meio da análise da correlação da variável QT_MAT_BAS (Quantidade de Matrículas Ensino Básico), também foi possível identificar as variáveis importantes de serem analisadas no contexto de pandemia, como se há ou não posse de recursos tecnológicos como internet, tablets, computadores de uso administrativo ou por aluno, se a escola possui didática pedagógica que contempla o ensino semipresencial. Variáveis essas que, analisando o contexto de pandemia, pesaram mais como uma influência positiva para a não evasão, ou seja, escolas que possuem tais recursos tendem a ter menos alunos evadidos.

Figura 9: Tabela de Correlação das Variáveis utilizada na análise focal do contexto de pandemia

VARIAVEL	2019	2020	2021
IN_REDES_SOCIAIS	0,17	0,20	0,21
IN_EAD	0,13	0,11	0,11
IN_TABLET_ALUNO	0,13	0,12	0,12
IN_ACESSO_INTERNET_COMPUTADOR	0,15	0,15	0,14
IN_MEDIACAO_SEMIPRESENCIAL	0,10	0,08	0,08
IN_COMP_PORTATIL_ALUNO	0,19	0,17	0,17
IN_INTERNET	0,26	0,28	0,25
IN_COMPUTADOR	0,27	0,26	0,26

Figura 10: Gráfico de Correlação das Variáveis utilizada na análise focal do contexto de pandemia



Conclusão

A partir dos resultados obtidos através das previsões do algoritmo de Árvore de Decisão, pudemos concluir que as técnicas de Aprendizado de Máquina tem grande importância e impacto na análise e contribuição para melhorias no processo estudantil, visto que, com base nos dados fornecidos, pudemos ter uma precisão de aproximadamente 80% na previsão das evasões em determinada escola. Tal método ainda possui uma vasta curva de melhorias que podem ser implementadas para o aperfeiçoamento do modelo de previsão e crescimento da taxa de precisão, como por exemplo o fornecimento de mais informações ou atributos para o treinamento do algoritmo, pois a falta de informações de naturezas variadas é o que constitui a taxa de erros do algoritmo, sendo esta, aproximadamente 20%. Quanto mais atributos forem fornecidos ao mesmo, maior será a taxa de assertividade. Também se faz necessário realizar um acompanhamento anual dos dados estudantis para a manutenção do processo, visto que estes, são atualizados anualmente e as previsões sempre devem ser realizadas com base nos anos anteriores ao vigente.

Dado o desenvolvimento da Árvore de Decisão, também foi possível analisarmos as correlações entre os atributos utilizados para a construção e treinamento do algoritmo com a Quantidade de Matrículas no Ensino Básico, desta maneira, pudemos concluir que a Árvore de

Decisão está funcionando conforme o esperado, visto que, os atributos que apresentam maior correlação com a Quantidade de Matrículas são justamente os que tem maior influencia positiva ou negativa na efetivação das matrículas, ou seja, são os nós principais das árvores.

Verificamos que os atributos que mais contribuem para que a efetivação de matrículas seja realizada são a presença de Quadra de Esportes no ambiente escolar, fornecimento da Educação de Nível Médio, presença de Laboratórios de Informática no ambiente escolar e também a presença de um Professor Bibliotecário. Em contrapartida, também verificamos que os atributos que mais contribuem para a não realização das matrículas ou o fenômeno de evasão escolar são a Localização, sendo esta, quando no ambiente Rural apresentando maiores taxas de evasão, Acessibilidade e meios de inclusão social, sendo esta, quando inexistente apresentando maiores taxas de evasão e Rede de Esgoto, sendo esta, quando inexistente apresentando maiores taxas de evasão.

Buscando um foco maior no contexto de pandemia, que afeta principalmente os ambientes de ensino, verificamos através das análises de correlações que para este cenário, alguns atributos contribuem positivamente à efetivação das matrículas no Ensino Básico, evitando assim que o fenômeno de evasão aumente. Sendo estes, a existência de Redes Sociais, sendo estes sites, blogs ou páginas para a comunicação institucional entre docentes, pais e alunos. Modelos de Educação à distância (EAD). Presença de tablets, computadores portáteis e desktop para utilização por parte do aluno no processo de ensino e acesso à internet para utilização no processo de ensino. Através deste processo, pudemos concluir a importância e necessidade da incorporação da informática e de meios de acesso à informação no cenário estudantil, sendo este um motivador essencial para o prosseguimento e conclusão dos estudos.

Em função da indisponibilidade de algumas naturezas de informações para a construção do projeto e do tempo para a conclusão do mesmo, recomenda-se para trabalhos futuros a incorporação de demais naturezas e tipos de dados que são disponibilizados pelas instituições públicas e privadas, bem como o acompanhamento anual e manutenção do projeto. Isto é, conforme novos dados sejam publicados por instituições seguras, os mesmos devem ser submetidos ao algoritmo com o objetivo de obter maior assertividade nas previsões realizadas e possibilitar o estudo dos novos fatores que eventualmente possam surgir e que contribuam para o fenômeno de evasão escolar. Também é possível analisar e discutir outras metodologias e tipos de algoritmos que poderiam contribuir positivamente para a realização das análises.

Referências Bibliográficas

Alqallaf, F. A., Konis, K. P., Martin, R. D., & Zamar, R. H. (2002). Scalable robust covariance and correlation estimates for data mining. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 4-23. [DOI: 10.1145/775047.775050]

COLPANI, R. Educação a Distância: Identificação dos Fatores que contribuíram para a Evasão dos Alunos no Curso de Gestão Empresarial da Faculdade de Tecnologia de Mococa. EAD EM FOCO, [S.l.], v. 8, n. 1, ago. 2018. ISSN 2177-8310. Disponível em: <http://eademfoco.cecierj.edu.br/index.php/Revista/article/view/688>.

DE SOUZA, C. M., Pereira, J. M. ., & Ranke , M. da C. de J. . (2020). Reflexos da Pandemia na evasão/abandono escolar: a democratização do acesso e permanência. Revista Brasileira De Educação Do Campo, 5, e10844.

DOS SANTOS, Júlio César Belenke. Usando Mineração de Dados para Predição da Evasão Escolar. Monografia (Curso de Sistemas de Informação do Campus Caçador) - Instituto Federal de Santa Catarina, Caçador, 2021.

Figueiredo Filho, D. B., & Silva Junior, J. A. (2009). Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r). Revista Política Hoje, 18(1), 115-146.

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

LOBO, M. B. de C. M. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. ABMES Cadernos, n. 25, 2012.

Rezende, C. C. da S., Cantarino, L. A. B., de Souza, P. F., Alves, T. O. M., & Campos, R. S. (2022). O impacto de aspectos socioeconômicos no desempenho de estudantes de Sistemas de Informação no Enade. Revista Brasileira de Informática na Educação, 30, 157-181. [DOI: 10.5753/rbie.2022.2093]

Veloso, L. A. (2015). A Predição da Evasão Escolar dos Cursos Técnicos de Nível Médio: Um Estudo de Caso no Senai. Dissertação (Mestrado), Universidade Católica de Brasília, 2015.

Haykan, S. (2001). Redes Neurais: Princípios e prática

Cambronero, C. G., Moreno, I. G (2006). ALGORITMOS DE APRENDIZAJE: KNN & KMEANS. Inteligencia en Redes de Telecomunicación, Universidad Carlos III de Madrid.