INSTITUTO TECNOLÓGICO DE AERONÁUTICA

Rafael Lima Gonzaga

PERFIS FAMILIARES PRESENTES NOS ENSINOS FUNDAMENTAL E MÉDIO E SUAS INFLUÊNCIAS NA EVASÃO ESCOLAR

Trabalho de Graduação
2021

Curso de Engenharia Civil-Aeronáutica

CDU 37.046.1

Rafael Lima Gonzaga

PERFIS FAMILIARES PRESENTES NOS ENSINOS FUNDAMENTAL E MÉDIO E SUAS INFLUÊNCIAS NA EVASÃO ESCOLAR

Orientadora

Prof. Giovanna Miceli Ronzani Borille (ITA)

ENGENHARIA CIVIL-AERONÁUTICA

São José dos Campos
Instituto Tecnológico de Aeronáutica

2021

Dados Internacionais de Catalogação-na-Publicação (CIP)
Divisão de Informação e Documentação

Gonzaga, Rafael Lima
PERFIS FAMILIARES PRESENTES NOS ENSINOS FUNDAMENTAL E MÉDIO E SUAS INFLUÊNCIAS NA EVASÃO ESCOLAR / Rafael Lima Gonzaga.
São José dos Campos, 2021.
51f.

Trabalho de Graduação – Curso de Engenharia Civil-Aeronáutica– Instituto Tecnológico de Aeronáutica, 2021. Orientadora: Prof. Giovanna Miceli Ronzani Borille.

1. . 2. . 3. . I. Instituto Tecnológico de Aeronáutica. II. Título.

REFERÊNCIA BIBLIOGRÁFICA

GONZAGA, Rafael Lima. **PERFIS FAMILIARES PRESENTES NOS ENSINOS FUNDAMENTAL E MÉDIO E SUAS INFLUÊNCIAS NA EVASÃO ESCOLAR**. 2021. 51f. Trabalho de Conclusão de Curso (Graduação) – Instituto Tecnológico de Aeronáutica, São José dos Campos.

CESSÃO DE DIREITOS

Rafael Lima Gonzaga
Rua Manoel Borba Gato, 900
12.242-270 – São José dos Campos–SP

PERFIS FAMILIARES PRESENTES NOS ENSINOS FUNDAMENTAL E MÉDIO E SUAS INFLUÊNCIAS NA EVASÃO ESCOLAR

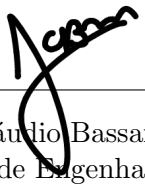
Essa publicação foi aceita como Relatório Final de Trabalho de Graduação

Rafael Lima Gonzaga

Autor

Giovanna Miceli Ronzani Borille (ITA)

Orientadora



Prof. João Cláudio Bassan de Moraes
Coordenador do Curso de Engenharia Civil-Aeronáutica

São José dos Campos, 18 de novembro de 2021.

Dedico este trabalho ao meu avô que de lá de cima me acompanhou durante toda essa jornada.

Agradecimentos

Primeiramente, obrigado meu Deus!

Em segundo lugar, gostaria de agradecer aos meus pais, estes que me suportaram todos os dias sem nunca duvidar das batalhas que viriam pela frente. Por todo o exemplo de ser humano e suporte que me foi dado durante essa caminhada, mainha e painho, obrigado e eu amo vocês!

Gostaria também de agradecer ao meu avô e minha avó. Avô, você me mostrou o quanto eu podia ir longe e que de nada devia temer se seguisse o meu sonho. Vozinha, obrigado por ser minha segunda mãe, sem o seu amor genuíno e suporte ao longo de toda a minha jornada, nada disso seria possível.

Aos meus irmãos Marcelo e Wally Neto, obrigado pela paciência e companheirismo durante essa caminhada. As fases da vida podem nos deixar próximos ou longes, mas nunca separados.

Ao meu Alê, seu tio muitas vezes estará distante aqui na correria. Um dia quando você conseguir ler isso, saiba que nem por um segundo a distância e o tempo diminui o meu amor por você.

Por último, a todos aqueles presentes no meu cotidiano em São José dos Campos, que essa caminhada tenho sido só o começo de uma grande jornada. Que sigamos juntos e para frente, sempre. Nunca um ficará para trás.

"Becoming isn't about arriving somewhere or achieving a certain aim. I see it instead as forward motion, a means of evolving, a way to reach continuously toward a better self."
— Michelle Obama

Resumo

A educação básica tem sido um importante tópico referente ao desenvolvimento de um país na atualidade. No Brasil, o contexto atual tem-se mostrado frágil e preocupante, principalmente devido a como foi enfretada a realidade da pandemia e como estão sendo dados os passos da gestão federal nesse ramo. Dito isso, esse trabalho busca analisar um aspecto crítico na educação básica global que é a evasão escolar. Nesse trabalho de graduação, busca-se encontrar correlações e previsões entre como a família e o aluno atuam na rotina escolar com a possível evasão ou mudança escolar do aluno no final do período letivo. O modelo de estudo aqui projetado utilizou uma escola modelo de São Paulo cuja representatividade não é significativa no âmbito federativo. Propôs-se uma abordagem sistemática e numérica de parâmetros-chave adquiridos da mineração de dados escolares para clusterizar perfis familiares com a presença ou não de ruído e entender qual o impacto dos eventos em uma análise de transferência escolar. Nos modelos aqui implementados abordou-se algoritmos não-supervisionados e supervisionados para as análises de clusterização e de predição, respectivamente.

Palavras-chave: Educação, Ensino Fundamental, Evasão Escolar, Ensino Médio e Algoritmos preditivos

Abstract

Middle high school and high school levels of education have been an important topic when taking into consideration the development of a country. In Brazil, the current educational context is fragile and worrisome, especially when considering how the country responded to the pandemic, and what are the steps taken by the federal administration in the education field. With that being said, this work focuses on the analysis of a critical aspect in the basic levels of education globally, school evasion. In this work, correlations and predictions to the relations between the pair family and student, and how this pair acts on the school routine, taking into consideration a possible evasion scenario at the end of the school year, are done. The model projected here makes use of an existing model of a school in São Paulo, not representative of the country's reality. This work proposes a systematic and numerical approach of key parameters acquired from data mining from schools in order to cluster the family profiles (considering or not noise in the distribution), aiming to understand the impacts of a set of events in the school transfer and dropout rates. In the models discussed and implemented here, non-supervised and supervised algorithms were used in the cluster and prediction phases, respectively.

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

IBGE	Instituto Brasileiro de Geografia e Estatística
Pisa	Programa Internacional de Avaliação de Alunos
OCDE	Organização para a Cooperação e Desenvolvimento Econômico
UNESCO	Organização das Nações Unidas para a Educação, a Ciência e a Cultura
RNA	Rede Neural Artificial
PNUD	Programa das Nações Unidas para o Desenvolvimento
ENEM	Exame Nacional do Ensino Médio
Inep	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
IDEB	Índice de Desenvolvimento da Educação Básica

Sumário

1 Introdução

1.1 Contextualização

A educação no mundo atual é um tema de suma importância. No cenário pós-pandemia, as expectativas e os indicadores de como está o aprendizado e o processo educacional dos jovens no mundo estão nos piores níveis. De acordo com o Índice de Desenvolvimento Humano (IDH), o cenário brasileiro se encontra em uma situação crítica. O Brasil hoje se encontra como o país com a terceira maior taxa de abandono escolar quando comparado com os 100 países de maiores IDH (FILHO; ARAÚJO, 2017).

Diferentes fatores também têm preocupado o cenário educacional brasileiro, não só a evasão escolar, mas também a grande distorção da faixa idade-série que se encontra presente no sistema educacional atual. De acordo com a Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua), somente em torno de 70% dos brasileiros entre 15 a 17 anos se encontram no ensino médio, que é a fase usual para essa faixa etária, percentual ainda distante do proposto como meta para o ano de 2024, que é de 85% (IBGE, 2020). De acordo com o Programa das Nações Unidas para o Desenvolvimento (PNUD), o Brasil foi apontado como o país de menor média de estudo da América Latina, tempo este de 7,2 anos em 2013 que hoje se encontra no patamar de 9,3 anos (INEP, 2019). Ressalta-se que a educação básica brasileira atual é composta por 12 anos.

A evasão e o abandono escolar são um complexo problema que permeia gravemente a educação conforme mencionado. Tal assunto é pauta recorrente no cenário político, este que visa encontrar meios de diminuir essa disparidade existente. O problema educacional brasileiro não é um objeto de estudo que surgiu há poucos anos. Esse entrave no desenvolvimento do Brasil tem sido amplamente estudado em diversas áreas, como neste trabalho explicitado, sendo suportado também pela Constituição Federal de 1988, a qual estabeleceu a obrigatoriedade da elaboração de metas decenais no Art. 2 (BRASIL, 2001) que visam: erradicar o analfabetismo, universalizar o atendimento escolar, melhorar a qualidade do ensino, formar mais empregos; e torna obrigatório a educação no Art. 6.

Dado isso, nota-se que o sistema educacional brasileiro se encontra em uma situação delicada. Mesmo havendo o suporte de instituições privadas e públicas, dos poderes le-

gislativo e judiciário, ainda há um grande abismo de aprendizado presente entre os jovens atuais. Um importante exemplo para esse cenário é o resultado do Programa Internacional de Avaliação de Estudantes (Pisa) de 2018 do Brasil. A evolução da educação brasileira se encontra estagnada desde 2009, não havendo evolução do nível educacional do povo brasileiro em tais provas. No exame realizado pelo Pisa, relatou-se que 68,1% dos estudantes brasileiros com 15 anos de idade não possuem o nível básico de matemática, 55% em ciência e 50% em leitura. Considera-se que ambos os níveis estão aquém do nível mínimo para pleno exercício da cidadania dados os critérios da Organização para a Cooperação e Desenvolvimento Econômico (OCDE) (SCHLEICHER, 2018).

No contexto das provas de acompanhamento de ensino realizadas de forma federal, o Exame Nacional do Ensino Médio (ENEM), vestibular padronizado para o estudante de ensino médio brasileiro, alcançou em 2021 patamares somente atingidos em 2005, elucidando um retrocesso de 16 anos na evolução educacional do Brasil. Ilustra-se a seguir os inscritos do ENEM ao longo dos últimos anos.

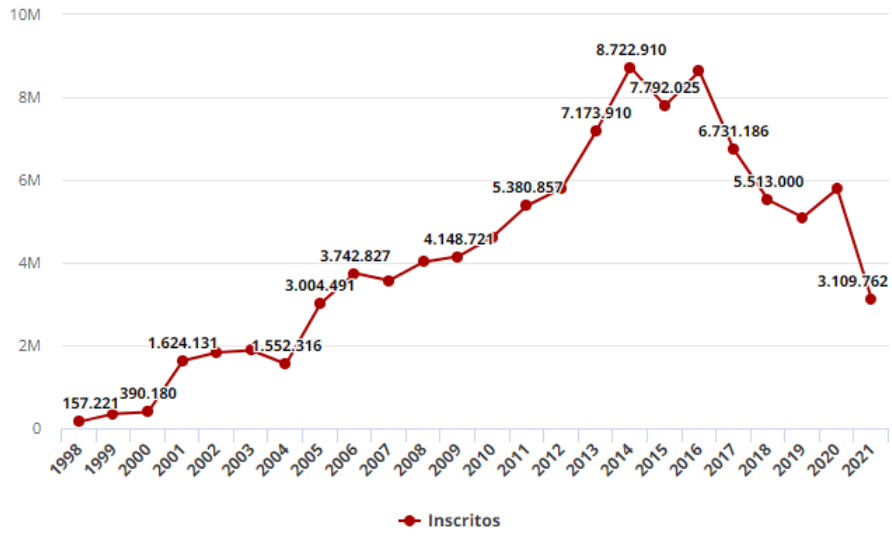


FIGURA 1.1 – Inscritos confirmados no Enem (1998-2021)
Fonte: (INEP, 2021)

1.2 Relevância do Tema

O setor educacional brasileiro é um dos maiores do mundo em termos de quantidade de estudantes. No gráfico da Figura 1.2, nota-se o declínio de matrículas totais dos anos de 2016 até 2020.

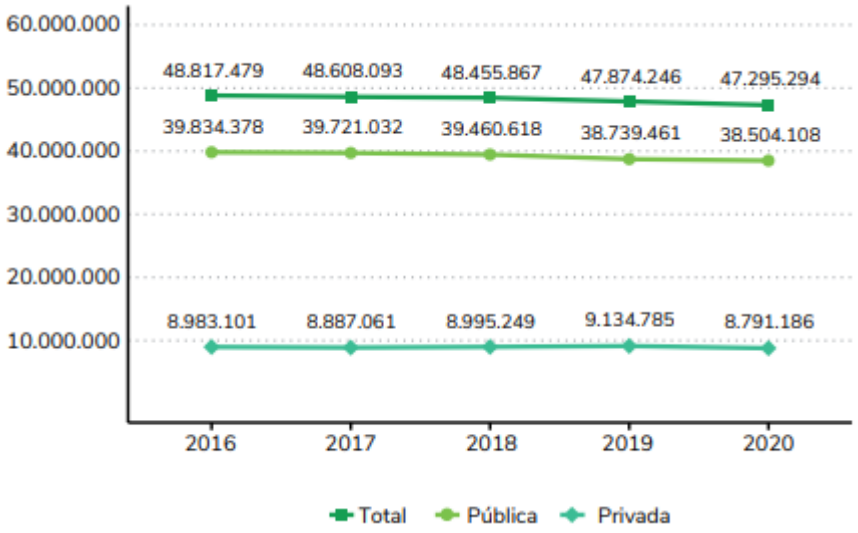


FIGURA 1.2 – Alunos matriculados em instituições de ensino básico brasileiro
Fonte: (INEP, 2020a)

No cenário acima descrito, nota-se que 79,7% das matrículas da educação básica se encontram no setor público, donde dessa parcela, grande quantidade se encontra na esfera municipal de acordo com o explicitado na Figura 1.3.

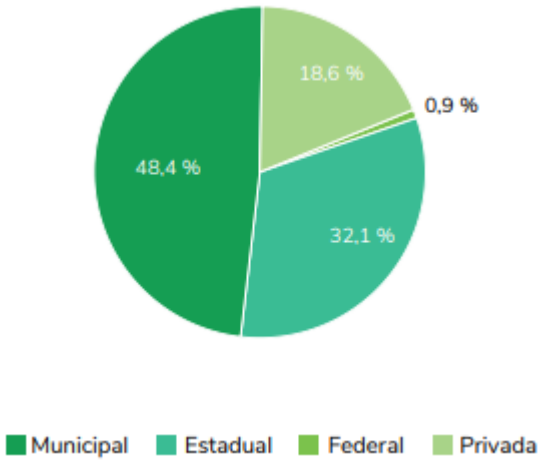


FIGURA 1.3 – Discretização de alunos nas esferas particular e privada
Fonte: (INEP, 2020a)

Uma importante característica do cenário educacional brasileiro é a distorção idade-série presente no dia a dia das escolas. Ressalta-se que é descrito aluno em distorção idade-série caso possuam a idade superior à recomendada para a série frequentada - a idade de 6 anos é considerada idade ideal para o ingresso no 1º Ano. Conforme descrito, ilustra-se o diagrama da Figura 1.4 de distorção idade-série do ensino básico brasileiro.

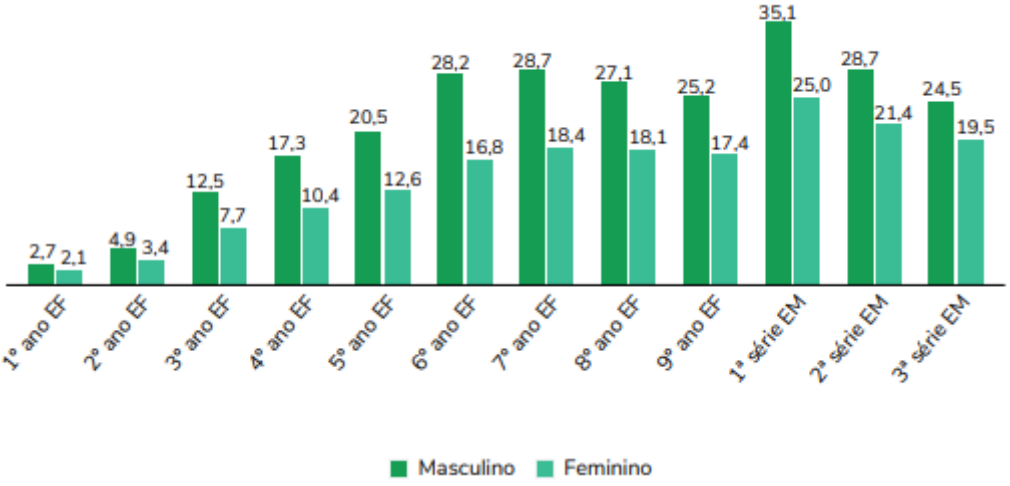


FIGURA 1.4 – Histograma de distorção idade-série no ensino básico brasileiro
Fonte: (INEP, 2020a)

De acordo com a imagem supracitada, o sistema educacional brasileiro tem-se encontrado com uma grande irregularidade de idade-série no ensino básico, acentuando-se principalmente no Ensino Fundamental Anos Finais e Ensino Médio conforme Censo Escolar de 2020.

Para a evasão escolar, com os dados do Censo Escolar de 2020 do INEP, torna-se possível comparar a evasão ou o insucesso escolar dos alunos nos anos de 2017, 2018 e 2019. Define-se insucesso escolar como o aluno que reprovou a série que estava cursando. Vale mencionar que casos de insucesso escolar contribuem diretamente com o aumento da taxa de distorção idade-série acima descrita.

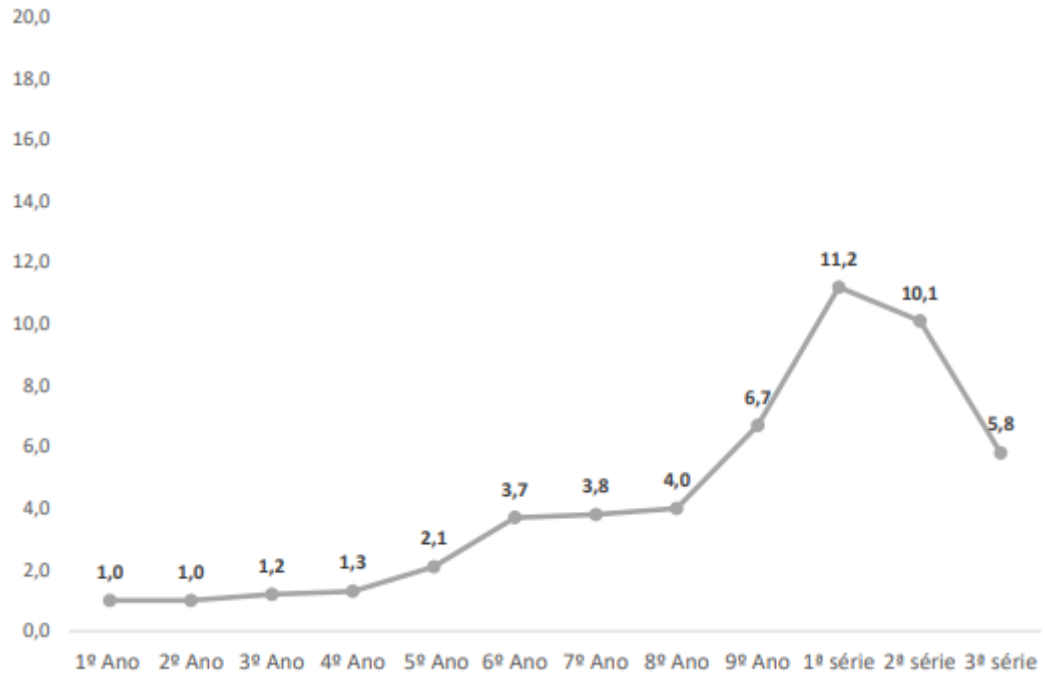


FIGURA 1.5 – Taxa de insucesso por série/ano nos ensinos fundamental e médio por série de ensino Brasil 2019
Fonte: (INEP, 2020b)

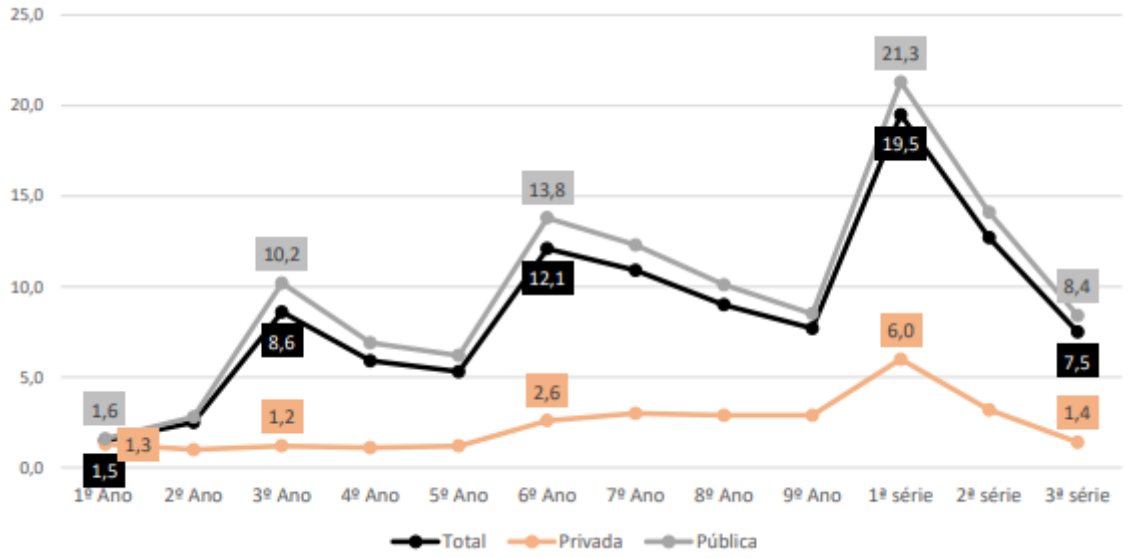


FIGURA 1.6 – Taxa de evasão na rede pública por série/ano nos ensinos fundamental e médio regular - Brasil 2017/18
Fonte: (INEP, 2020b)

Com isso, nota-se o aumento da evasão escolar em relação à 2017 em todas as séries de ensino da rede pública. Já na rede particular, esse permanece em pequenos patamares, somente reafirmando a desigualdade que ocorre nessa divisão público-privada que está presente no Brasil atual.

Diante do texto acima disposto, torna-se claro a relevância do tema de evasão escolar no contexto brasileiro atual.

1.3 Definição do problema

Dado o problema de evasão escolar e todos os parâmetros que podem influenciar positivamente ou negativamente esse processo na vida do aluno, optou-se por fazer uma análise mais quantitativa do que levaria a evasão escolar em uma escola particular.

O primeiro passo para resolução do problema foi a definição da escola modelo para o estudo de caso a seguir proposto. Assim, mantendo as informações da escola em confidencialidade, listam-se as seguintes características da escola modelo relativas aos alunos ativos na instituição durante o ano letivo de 2020:

- Idade média dos responsáveis: 44,43 anos.
- Localizada na Barra Funda - São Paulo.
- 54,09% dos responsáveis são mulheres.
- 45,91% dos responsáveis são homens
- 35,86% dos responsáveis são casados.
- 312 estão regularmente matriculados na instituição de ensino (IE).
- Ensino completamente remoto de Março/2020 até Outubro/2020.
- Período de análise dos dados: 01/01/2020 - 31/12/2020.
- Calendário escolar seguindo a segmentação de bimestres.

Em seguida, para matematizar o problema de evasão escolar foi necessário entender quais *inputs* a família e o aluno passam para escola durante o período do ano letivo. Em seguida, procurou-se entender quais desses é possível a escola fornecer com exatidão e precisão.

A partir das definições de quais dados seriam utilizados para a tese aqui disposta, optou-se por analisar somente os dados que a escola possuía referente ao ano de 2020, pois devido ao ano de pandemia, o engajamento metrificável discente e docente nas plataformas digitais aumentou drasticamente.

Com isso, optou-se por analisar as bases de dados a seguir descritas:

- Engajamento da Família na Plataforma de Comunicação da Família com a Escola: Esse banco de dados é o conjunto de informações da quantidade de vezes que o responsável de cada aluno interagiu em um canal de comunicação. Alguns exemplos de canais de comunicação são: Coordenação Fundamental I, Secretaria, Diretoria Pedagógica.

- Engajamento do aluno na Plataforma de Gestão Pedagógica: Esse banco de dados é o conjunto de informações de quantidade de conteúdos que o aluno assistiu, qual disciplina ele acessou, qual dia e duração ele fez aquele acesso.

- Aproveitamento do aluno na Plataforma de Gestão Pedagógica: Esse banco de dados é o compilado de todas as performances educacionais dos alunos. É o compilado de todas as pontuações de avaliações que ocorreram em todas as disciplinas.

1.4 Objetivo do trabalho

O objetivo do presente trabalho é identificar padrões de famílias e de alunos que existem na escola de estudo de caso. A partir dos padrões, será estudado quais destes podem levar à evasão escolar a partir de uma predição por meio de redes neurais artificiais (RNA).

Para os métodos de identificação de padrões, serão utilizados algoritmos de aprendizado de máquina não-supervisionados, pois o propósito é agrupar e prever alunos que irão ser transferidos do colégio. Assim, não há *a priori* a base de treino supervisionada para o problema em questão.

Em suma, segue abaixo a listagem dos objetivos específicos que serão abordados no trabalho:

- Analisar exploratoriamente as informações das bases de dados enunciadas e identificar *outliers* com o uso de visualizações gráficas.

- Identificar correlações entre variáveis existentes nas bases de dados enunciadas.

- Clusterizar os alunos do colégio em grupos com base nos algoritmos não supervisionados escolhidos.

- Treinar e validar uma rede neural artificial preditora de evasão escolar com os inputs das bases de dados em questão.

2 Conceitos Fundamentais

A definição de evasão escolar na educação básica não é tão clara, fato este que dificulta os estudos de causa e dos princípios básicos que podem acarretar tal evento. Mencionam-se aqui as principais ideias abordadas no cenário educacional brasileiro atual. (FILHO; ARAÚJO, 2017)

Algumas propostas aceitas e trabalhadas no cenário da educação brasileira são como as explicitadas por Riffel e Malacarne (2010) que nos indicam que entende-se evasão escolar como fuga ou abandono da escola em função da realização de outra atividade. (RIFFEL; MALACARNE, 2010)

Já Steinbach (2012) e Pelissari (2012), optam por não denominar evasão escolar e sim, abandono escolar, pois consideram que evasão é um ato de um indivíduo, já abandono, é de um grupo. (STEINBACH, 2012) (PELISSARI, 2012)

Até alguns institutos brasileiros como o IDEB em 2012 e o INEP em 1998 já entraram em indefinições, mostrando a não clareza do conceito em si. Por fim, traz-se a proposta utilizada pelo IDEB em 2012, que propõe que o abandono escolar seja o afastamento do aluno do sistema de ensino e desistência das atividades escolares sem solicitar transferência.

Para o estudo de caso em questão, optou-se pela ideia explicitada por Steinbach (2012) e Pelissari (2012) e analisar o problema da evasão escolar como um tudo, não só com o indivíduo. Para isso, na fundamentação teórica do problema utilizar-se-á a quantidade que os responsáveis dos alunos interagiram com a família. Com isso, acredita-se considerar uma parcela que não corresponde ao indivíduo aluno.

2.1 Contexto da educação no Brasil e mundo

Durante o período de pandemia, o fechamento das escolas e o forçoso *homeschooling* imposto pela situação que se encontrou em 2020 provocou muitos impactos negativos no cenário da educação básica atual.

Conforme explicitado por Khan *et al.* (2021) na sua pesquisa referente ao impacto da

pandemia no Paquistão, tem-se números assustadores no que nos diz respeito à perda de aprendizado gerada. Khan *et al.* (2021) menciona que o maior impacto tem sido no ensino primário, o qual se inicia aos 6 anos e se estende até os 12 anos do estudante. É apontado por Khan *et al.* (2021) que os percentuais de evasão chegaram a níveis altíssimos, com uma média de 6,7% antes de completar o ensino primário e 18,6% evade antes de finalizar o primário. (KHAN; AHMED, 2021)

É importante mencionar que a queda de qualidade na educação não foi um evento isolado nem afetou somente países da Ásia e da América. Regiões como a Europa também foram impactadas conforme explicitada por Aniela *et al.* no caso da Romênia. (ANIELA; BADEA, 2021)

Todavia, apesar do impacto devido à pandemia, é de ciência que os danos aos sistemas educacionais não ocorrem unicamente por isto, dado que as notas do Programa Internacional de Avaliação de Alunos não tem mostrado bons resultados. Na Romênia, teve-se um dos piores índices históricos já apresentados, com notas de 466 em matemática, 439 em ciência e 408 em leitura. O que é bem abaixo da média satisfatória preconizada pela Organização para a Cooperação e Desenvolvimento relacionada aos países que é de 487 em leitura, 489 em matemática e 489 em ciências. (ANIELA; BADEA, 2021)

De acordo com o resultado da prova de PISA de 2018, o Brasil também se encontra bastante abaixo com as notas de 413 em leitura, 384 em matemática e 404 em ciências e está em uma estagnação há 10 anos, conforme explicitado na Figura 2.1:

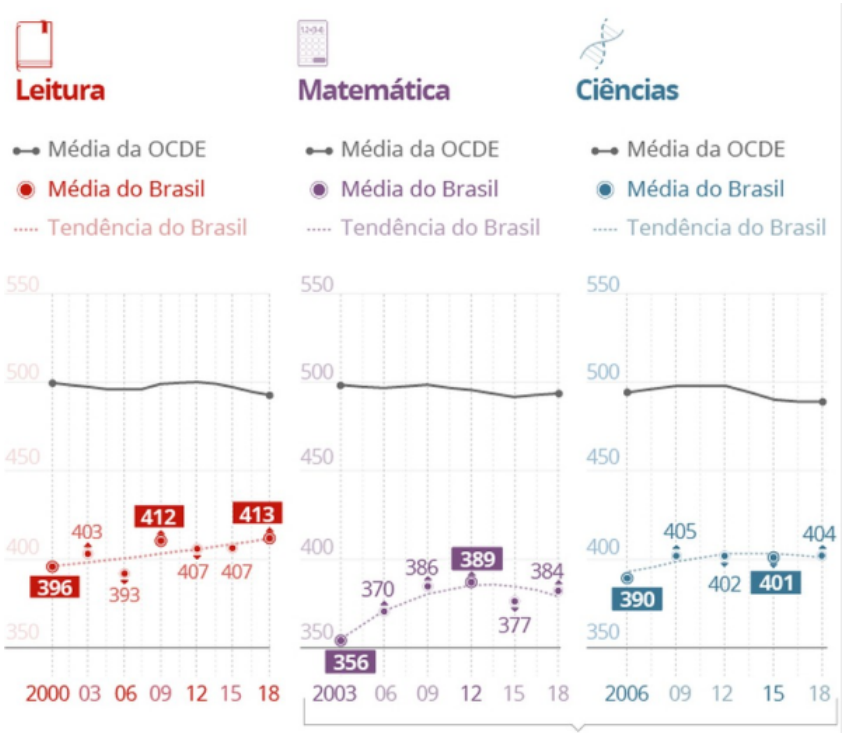


FIGURA 2.1 – Evolução do Brasil nas competências do PISA

2.2 Variáveis que permeiam a evasão escolar

A análise das variáveis que contribuem para a evasão escolar por si só já é uma pauta imprecisa, dado que existem várias abordagens para tal assunto.

Diante do apresentado por Bezerra *et al.* em seu estudo de aplicação de mineração de dados para encontrar correlações relevantes na evasão escolar, ressalta-se que é importante levar em consideração três grandes pilares para a análise desse problema (BEZERRA *et al.*, 2016):

i) o indivíduo

ii) a escola

iii) o sistema de ensino

Para a abordagem do problema aqui disposto e dos dados que são possíveis de serem minerados para o colégio em análise, tem-se:

- Dados de Engajamento na Plataforma de comunicação da Família com a Escola: Com esses dados, acredita-se entender a contribuição da escola junto ao indivíduo. Isso se dá por meio de cada um dos canais de comunicação dispostos na plataforma. O banco de dados de engajamento na plataforma de comunicação nos fornece para cada aluno, quantas vezes ambos os responsáveis interagiram em cada um dos canais de comunicação. Os canais de comunicação presentes na plataforma são:
  - Direção Administrativa
  - Secretaria
  - Financeiro
  - Coordenação Médio
  - Coordenação Fund II
  - Coordenação Fund I
  - Rematrículas
  - Direção Pedagógica
  - Saídas (Autorizações)
  - Comunicação Fund I
  - Coordenação Integral
  - Coordenação Cursos Livres
  - Coordenação Estudo Direcionado

– Cultura Inglesa

Para a análise em questão, optou-se por relacionar os canais que possuíam os mesmos fins, como por exemplo no que diz respeito às coordenações. Assim, optou-se por considerar somente um canal para as Coordenações, assim não há divergência entre qual segmento escolar o aluno se encontra. Explicita-se na Tabela 2.1 o exemplo da base de dados nesse ponto mencionada com os dados da Coordenação discretizados por segmento.

TABELA 2.1 – Dados de Engajamento na Plataforma de comunicação da Família com a Escola

Aluno	**Série**	**Diretoria Administrativa**	**Secretaria**	**Financeiro**	**Coordenação Fundamental I**	**Coordenação Fundamental II**	**Coordenação Ensino Médio**
João	3º Ano	1	0	0	4	0	0
Arthur	5º Ano	1	3	3	2	0	0
Marcela	1ª Série	1	0	0	0	0	2
Rodrigo	3ª Série	1	2	4	0	0	5

- Engajamento dos alunos na Plataforma de Gestão Pedagógica: A partir da análise dos acessos na plataforma de Gestão Pedagógica utilizada na escola, é possível entender o engajamento e a participação dos alunos nas aulas onlines que ocorreram durante o ano de 2020.
  Com essa base de dados, torna-se tangível o entendimento da contribuição do indivíduo estudante e como ele é responsável pela sua rotina escolar implementada pela escola. Esse *dataset* nos retorna importantes informações do dia a dia da rotina escolar, como qual aula foi assistida, se ela foi assistida no dia exato da aula, qual foi a primeira interação do aluno com a aula e quanto tempo ele ficou naquele conteúdo específico acessado.

TABELA 2.2 – Engajamento dos alunos na Plataforma de Gestão Pedagógica

Aluno	**Bimestre**	**Série**	**Data da Aula**	**Disciplina**	**Segmento da aula**	**Aula Assistida**	**Aula Assistida no Dia**	**Tempo de Acesso (s)**	**Data da Primeira Iteração**
Rafael	1º Bimestre	1ª Série	18/01/2021	Geografia	Conteúdo	Sim	Sim	800	18/01/2021
Walter	1º Bimesre	5º Ano	18/01/2021	História	Tarefa de Casa	Sim	Não	983	20/01/2021
Bruna	1º Bimestre	1ª Série	19/01/2021	Língua Portuguesa	Conteúdo	Não	Não	764	
Gabriel	1º Bimestre	3ª Série	19/01/2021	Matemática	Atividade Complementar	Sim	Sim	35	19/01/2021

- Aproveitamento do aluno na Plataforma de Gestão Pedagógica: Por fim, a última base de dados utilizada para estruturação de todos os *dataset* relacionados à proposta de solução para essa discussão é o compilado de todas as notas de atividades, provas e recuperações que ocorreram no ano de 2020. Ressalta-se que para a escola em análise há algumas especificidades.

  – Atividades: Cada professor tem a opção de realizar no máximo 5 Atividades e no mínimo 2 Atividades. Com isso, nota-se que na base de dados há vários campos nulos.

– Prova Bimestral: De cunho obrigatório para cada um dos bimestres da escola modelo. As provas bimestrais não ocorrem para as disciplinas eletivas, Educação Física e Artes. A composição da nota da prova bimestral dessas matérias é realizada a partir de uma relação com as atividades.

– Recuperação Bimestral: Somente obrigatória para os alunos que ficaram com a média final bimestral abaixo de 6,5. Para a escola modelo, as recuperações são realizadas bimestralmente.

TABELA 2.3 – Aproveitamento do aluno na Plataforma de Gestão Pedagógica

Aluno	**Série**	**Bimestre**	**Disciplina**	**Atividade 1**	**Atividade 2**	**Atividade 3**	**Atividade 4**	**Atividade 5**	**Prova Bimestral**	**Recuperação Bimestral**
Rafael	1ª Série	1º Bimestre	Geografia	7,6	8,0	4,6			6,5	
Walter	5º Ano	1º Bimestre	História	2,3	5,6	3,4			4,3	7,3
Bruna	1ª Série	1º Bimestre	Matemática	7,3	8,2	6,6	6,4	6,1	8,2	
Gabriel	3ª Série	1º Bimestre	Física	6,1	7,3	7,4	8,8		6,4	

3 Fundamentação Teórica do Método

Nessa seção menciona-se as ferramentas matemáticas e computacionais utilizadas para a análise exploratória da base de dados, identificação de clusters na instituição de ensino e predição de evasão escolar na instituição de ensino em questão.

Para a análise exploratória dos dados, utilizou-se as visões computacionais de *boxplot*, histograma e matriz de correlação.

Para o modelo matemático de identificação de clusters, foram utilizados os algoritmos *KMeans* e *KMeans with noisy cluster*. O segundo algoritmo foi utilizado visando o problema de ruído identificado no *dataset* proposto para o estudo de caso.

Já para o algoritmo preditivo, utilizou-se as Redes Neurais Artificiais binárias. Ao longo do disserto a seguir, também são elucidados trabalhos científicos aplicados com essas práticas tanto no contexto educacional quanto fora, as condições de aplicabilidade e as limitações do estudo.

3.1 Análise Exploratória

No tópico a seguir serão apresentados descrições mais específicas dos métodos de análise exploratória utilizados neste trabalho de graduação. Para o relatório aqui apresentado foram utilizados os seguintes algoritmos de análise abaixo listados:

- *Boxplot*.
- Histrograma.
- Matriz de correlação.
- *Scatterplot*.

3.1.1 Descrição do método

Para a visualização da base de dados em análise, optou-se pelos modelos abaixo descritos:

- *Boxplot*: De acordo com Williamson *et al.* (1989), esse tipo de visualização se baseia em métricas importantes da base de dados como: a média, o primeiro e o terceiro quartil e os valores máximos e mínimos do *dataset*. (WILLIAMSON *et al.*, 1989) Conforme explicitado pelo autor, esse tipo de visualização permite encontrar padrões que não estão claros de imediato no banco de dados analisado.

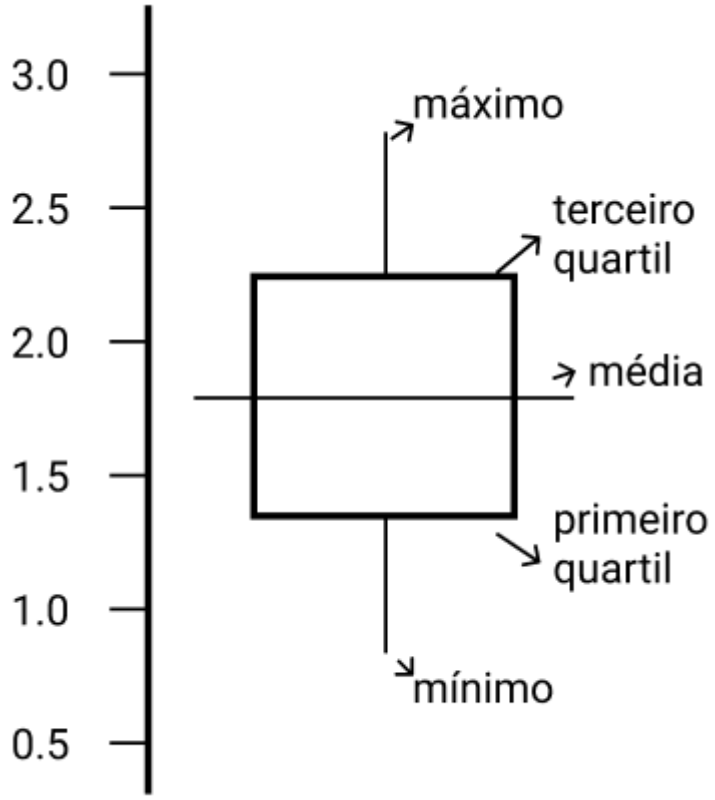


FIGURA 3.1 – Exemplo de uma visualização *boxplot*
Fonte: Adaptado de (WILLIAMSON *et al.*, 1989)

- Histograma: Esse tipo de visualização de dados é uma clássica que é utilizada desde o século 19. De acordo com Nuzzo (2019), esse tipo de análise possibilita entender a distribuição de informação que existe no *dataset* para um tipo de variável em análise. (NUZZO, 2019)

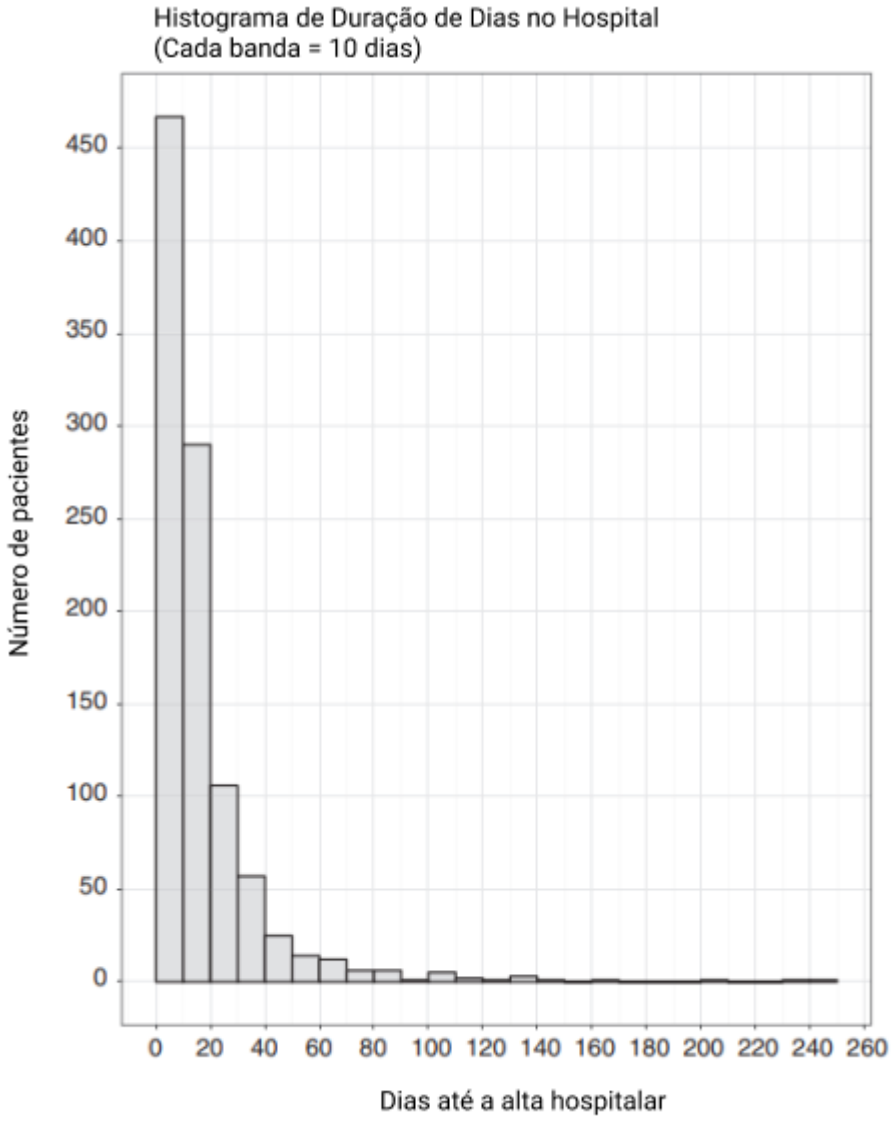


FIGURA 3.2 – Exemplo de uma visualização histograma
Fonte: Adaptado de (NUZZO, 2019)

- Matriz de correlação (*heatmap*): Esse método tem-se tornado bastante utilizado para analisar associação de variáveis de um problema. Conforme explicitado por Haarman *et al.* (2014), nota-se que esse método é de suma importância também para outros setores, como a medicina, mostrando que é possível entender correlações entre complexos sistemas biológicos a partir desta visualização. (HAARMAN *et al.*, 2015)

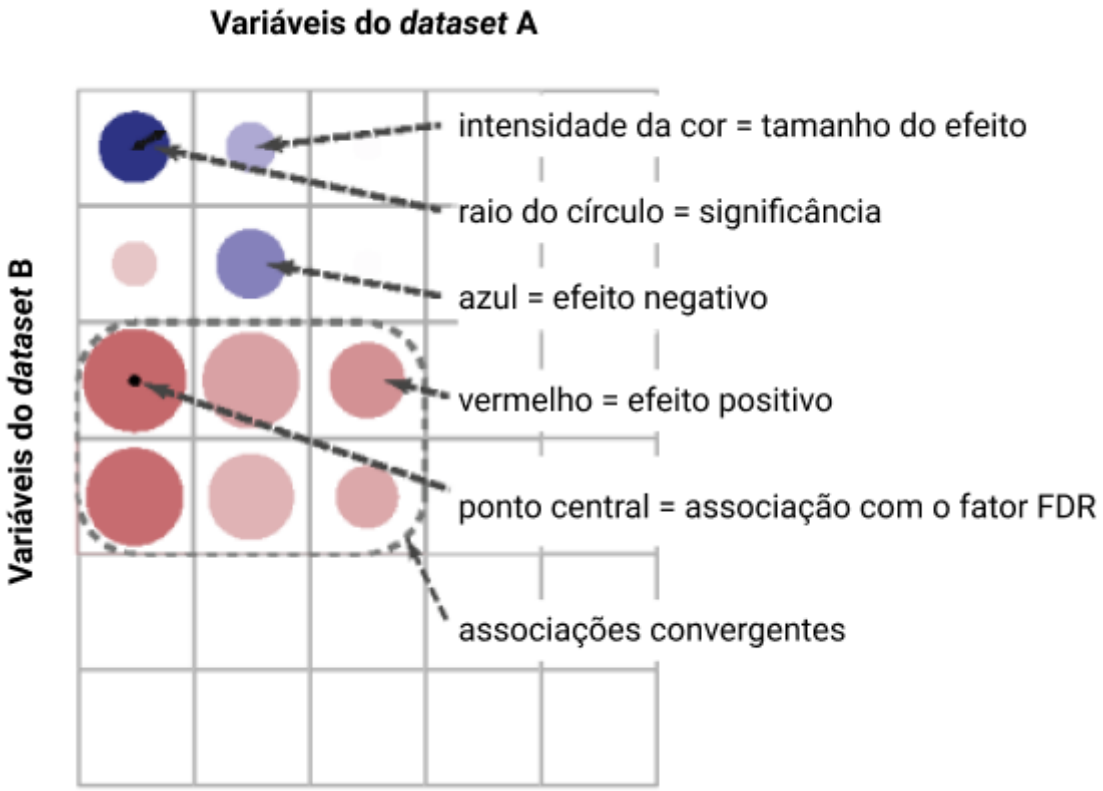


FIGURA 3.3 – Exemplo de uma visualização *heatmap*
Fonte: Adaptado de (HAARMAN *et al.*, 2015)

- *Scatterplot*: As visualizações 3D têm se tornado cada vez mais utilizadas nos dias atuais, apesar do conservadorismo existente em relação a sua utilização (MARRIOTT *et al.*, 2018). (2018) É válido mencionar importantes ganhos na visualização 3D do *dataset* conforme listado abaixo:
  - Adicionar um novo canal de visualização de dados.
  - Tornar o ambiente de análise imersivo.
  - Ir além da eficácia esperada para a análise exploratória.

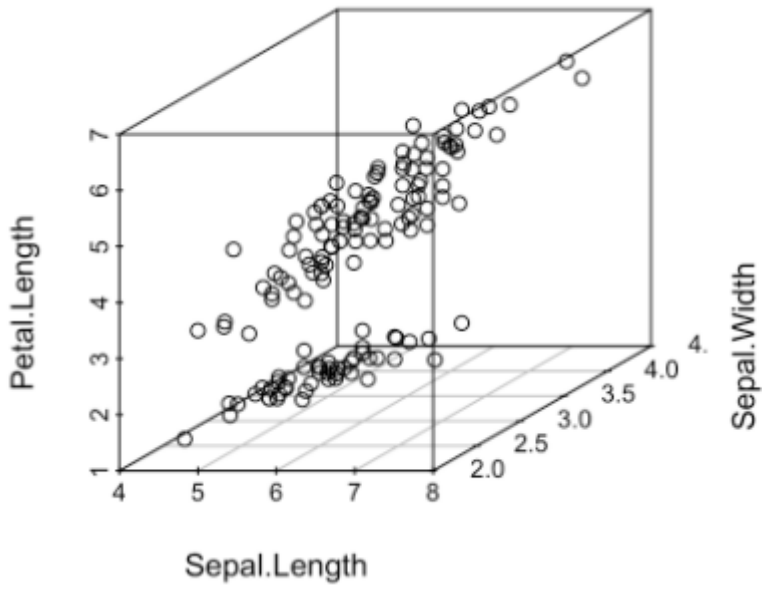


FIGURA 3.4 – Exemplo de uma visualização *scatterplot*

3.2 *KMeans*

No tópico abaixo elucidado será abordado uma discussão mais específica do algoritmo *KMeans*. O enfoque dos tópicos a seguir serão:

- Descrição do método: Aqui será abordado a origem do algoritmo como também a análise matemática envolvida na lógica de funcionamento do *KMeans*.
- Características gerais do *KMeans*: Nesse tópico serão abordados em quais situações o algoritmo escolhido é utilizado, como também a análise sobre a ordem do que foi implementado e quais os impactos de ruídos nos *datasets* utilizados.

3.2.1 Descrição do método

Apesar do termo *KMeans* ser primeiro introduzido na literatura como um artigo de MacQueen sobre alguns métodos de classificação e de análises de observações de multivariáveis em 1967, essa ideia de algoritmo já havia aparecido em escritas de Hugo Steinhaus em 1956 e de Stuart Lloyd em 1957. (MACQUEEN *et al.*, 1967)

O algoritmo proposto por Lloyd é o que mais se aproxima computacionalmente do que está implementado na biblioteca utilizada em questão. (LLOYD, 1982)

Para exemplificar a utilização do algoritmo escolhido, o *KMeans* funciona alternando entre duas etapas, a de atribuição e a de atualização abaixo descritas.

- Etapa de atribuição

  Para etapa de atribuição, define-se para cada observação x_p em qual *cluster* essa observação se encontra a partir do espaço de k centróides $(m_1, m_2, ..., m_k)$. Para cada observação P calcula-se:

$$S_i^t = \{x_p : ||x_p - m_i^{(t)}||^2 \leq ||x_p - m_j^{(t)}||^2, \forall j, 1 \leq j \leq k\} \tag{3.1}$$

  Em que x_p será assinalado para um único cluster S_i^t, mesmo que o x_p possa ser introduzido em mais de um *cluster*.

- Etapa de atualização

  Na etapa de atualização, para cada valor de *cluster* definido, calcula-se o novo centróide a partir de:

$$m_i^{i+1} = \frac{1}{|S_i^t|} \sum_{x_j \in S_i^t} x_j \tag{3.2}$$

  Assim, tem-se os rebalanceamento dos centróides dada a iteração de atribuição de cada ponto x_p ao centróide. Ao recalcular os centróides de todos os k *cluster*, realiza-se a etapa de atribuição novamente até que as variações dos valores dos centróides estejam muito pequenas. Vale mencionar que tal método não garante encontrar a solução ótima do problema de classificação.

3.2.2 Características gerais do *KMeans*

De acordo com MacQueen (1967), o algoritmo *KMeans* pode ser utilizado para alguns problemas típicos como:

- Problemas de encontrar similaridade entre grupos
- Encontrar classificações relevantes em *datasets*
- Aproximar uma distribuição geral do *datasets*

Tem-se também que o algoritmo *KMeans* é da ordem $O(nkdi)$, onde:

- n é o número de pontos no *datasets*
- d é o número de dimensões do *datasets*
- k é o número de *clusters*

- i é o número de iterações até a convergência

Por último, frisa-se que o presente algoritmo é bastante sensível a ruído devido à função de distância descrita na etapa de atribuição. Ao utilizar-se o erro quadrático médio em tal caso, uma pequena quantidade de pontos de ruído afeta fortemente a análise (DAVE, 1991).

3.3 *KMeans with noisy cluster*

No tópico abaixo descrito foram abordados os seguintes pontos abaixo elucidados do algoritmo *KMeans with noisy cluster*:

- Descrição do método: Nesse tópico foi abordado a principal diferença matemática envolvida na implementação do *KMeans with noisy cluster* em comparação ao *KMeans*.
- Características gerais do *KMeans with noisy cluster*: Nesse ponto foi explicitado como deve ser considerada o δ para a primeira iteração do algoritmo.

3.3.1 Descrição do método

A primeira literatura em que há menção a este tipo de variante do algoritmo *KMeans*, encontra-se no artigo proposto por Rajesh N. Dave. Conforme por ele mencionado, após a criação dos primeiros algoritmos de clusterização, perceberam-se que não haveria um algoritmo para solucionar todos os problemas presentes na realidade. (DAVE, 1991)

Assim, começaram a ser estudados alguns métodos para a análise de ruídos e *outliers* presentes em *datasets*.

Uma frequente abordagem utilizada, criada por Jain e Dubes em 1988, propõe identificar e remover todos os ruídos e *outliers* presentes na base de dados. Todavia, conforme posteriormente descrito, muitas vezes tal abordagem é de grande dificuldade ou até impossível.

Para exemplificação do método descrito, vale-se ressaltar que ambas as etapas dispostas no algoritmo *Kmeans* são utilizadas. Todavia, o diferencial para a detecção de ruídos e *outliers* se dá na distância que será utilizada. A partir do conceito de *Noise cluster*, é definido que este será o cluster que conterá todos os pontos de ruído. As distâncias utilizadas no problema são as apresentadas na Equação 3.3 e Equação 3.4:

$$d_{i,j} =< x_k - v_i >^T A_i < x_k - v_i > \tag{3.3}$$

$$d_{i,j} = \delta^2 \tag{3.4}$$

Em que na Equação 3.3 será utilizada a distância euclidiana para todos os pontos que não incluem o *Noise cluster* e o 3.4 somente para o *Noise cluster*. Ressalta-se que inicialmente o δ será a distância do *noise prototype* definido por Rajesh. É válido enfatizar que o *noise prototype* é a entidade em que estará a mesma distância de todos os pontos do *dataset*. Com essa definição, tem-se que todos os pontos são equiprováveis de serem ruídos ou *outlyers*.

Por último, as etapas são executadas até que as partições estejam estáveis.

3.3.2 Características gerais do *KMeans with noisy cluster*

Um grande entrave à implementação do algoritmo é a consideração inicial que deve ser adotada para o δ. Com isso, Rajesh sugere que estatisticamente se tenha a distância tal que:

$$\delta^2 = \lambda \frac{\sum_{i=1} \sum_{k=1} d_{i,j}^2}{n(c-1)} \tag{3.5}$$

3.4 Redes Neurais Artificiais (RNA)

Nos tópicos a seguir foram apresentados os apontamentos sobre o algoritmo das Redes Neurais Artificais utilizado neste trabalho. Dessa maneira, elucidam-se os seguintes pontos:

- Descrição do método: Nesse ponto foi abordado uma breve contextualização da origem dos algoritmos de redes neurais, como também os apontamentos matemáticos referentes ao modelo de rede escolhido para implementação no estudo de caso realizado..
- Características gerais das Redes Neurais Artificiais: Neste tópico abordou-se uma discussão mais específica sobre os pontos positivos e negativos de algoritmos de aprendizado supervisionado como a Rede Neural.

3.4.1 Descrição do método

As redes neurais artificiais têm sido um método computacional de implementação de inteligência artificial para resolver problemas de clusterização, reconhecimento de padrão,

predição e classificação. A utilização de redes neurais tem sido implementada nos mais diversos campos como computação, engenharia, medicina, meio ambiente e clima (ABIODUN *et al.*, 2018).

As primeiras escritas referentes às redes neurais artificiais surgiram em 1943 com Warren McCulloch e Walter Pitts, onde eles descrevem teoricamente como seria um modelo computacional para redes neurais artificiais. (MCCULLOCH; PITTS, 1943)

Somente em 1958 tem-se a primeira menção sobre a rede neural artificial do tipo *Perceptron* dada por Rosenblatt, uma RNA que é composta por somente uma única camada com um único neurônio. Esse modelo de RNA é a que está implementada dentro da biblioteca *neuralnet* utilizada no algoritmo elaborado para a resolução deste problema. (ROSENBLATT, 1958)

Para a concepção de uma RNA do tipo *Perceptron*, tem-se o seguinte esboço dos componentes da rede:

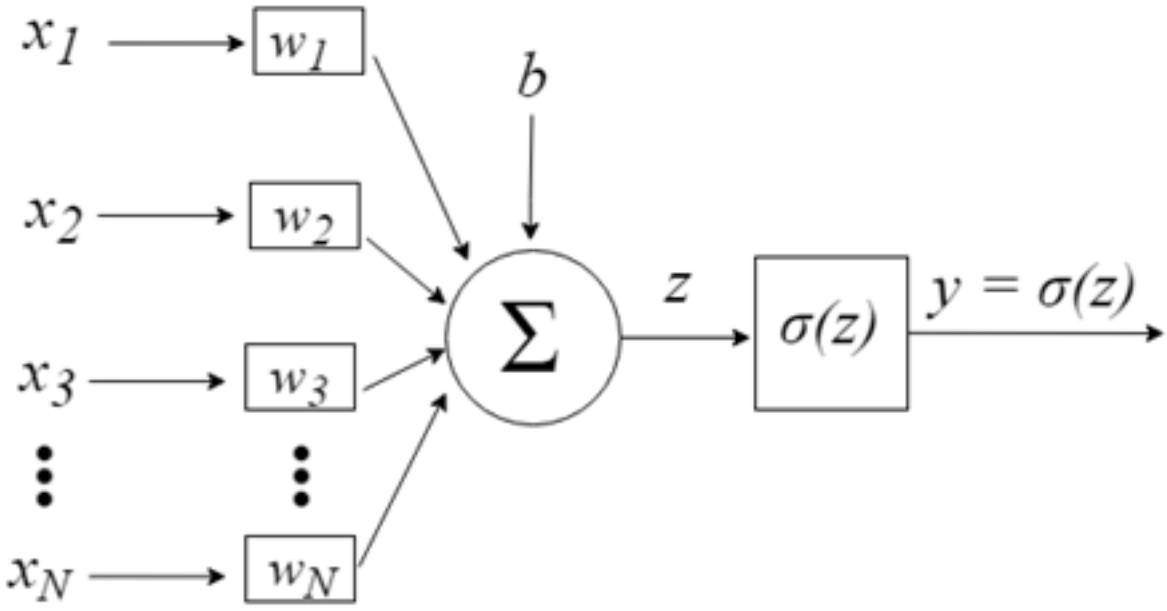


FIGURA 3.5 – Exemplo de RNA com somente uma camada (*Perceptron*)

Onde no diagrama acima, a primeira camada seria a referente aos *inputs*, a camada central seria a *hidden layer* e a última camada seria composto pelos *outputs*. Note que em cada uma das camadas há a existência de neurônios, cada um dos neurônios possuirá sua função conforme a Equação 3.6:

$$z(x) = w_n \cdot x_n \tag{3.6}$$

Tendo em vista que w_n é o peso referente aquele neurônio e x_n é o *input* dado. Ao

combinar todas as funções dos neurônios, tem-se o *input* da função de ativação da RNA. Para o caso da rede do tipo *Perceptron*, utiliza-se a função sigmóide logística como função de ativação do problema.

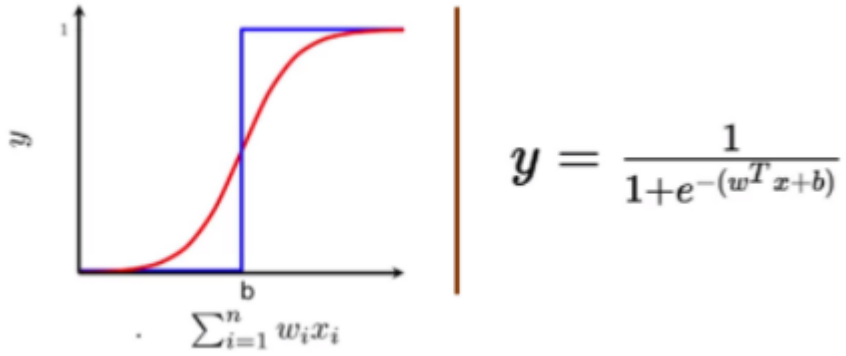


FIGURA 3.6 – Exemplificação de uma função sigmóide com a sua função de ativação

Assim, tem-se que o valor do *output* da rede será um valor obrigatoriamente entre 0 e 1. Dessa maneira, pode-se interpretar o *output* da RNA como uma probabilidade do evento em questão. No caso do referido problema, o *output* da RNA irá se referir de forma binária às predições do aluno que será transferido ou não.

3.4.2 Características gerais das Redes Neurais Artificiais

As redes neurais artificiais aqui utilizadas possuem a concepção do algoritmo de aprendizado diferente da implementada na clusterização. Nota-se que o método de aprendizado supervisionado, que é o usado na RNA, tem algumas vantagens e desvantagens em relação ao método não-supervisionado de aprendizado de máquina.

Dentro o mencionado acima, os pontos em que o algoritmo supervisionado desempenha melhor que o não-supervisionado são:

- Ganho em relação ao nível de especificidade do problema. Em algoritmos supervisionados, é necessário fornecer ao computador toda a base de treino para que ele parametrize a inteligência artificial. Com isso, essa etapa de fornecer ao computador o que de fato ele precisa, permite o algoritmo atingir maiores níveis de discretização em alguns casos comparado ao aprendizado não-supervisionado.

- Maior acurácia das predições. Isso se dá devido ao aprendizado supervisionado passado ao computador. No método não-supervisionado, ao não se definir com clareza as condições do problema, os impactos de uma base de dados menos precisa pode acarretar piores previsões.

- Os resultados encontrados no aprendizado supervisionado podem ser julgados de acordo com a realidade. Como no aprendizado não-supervisionado não são passadas informações de aprendizado já esperadas, torna-se difícil o processo de cruzar as informações encontradas de outputs com a realidade.

4 Aplicação, Resultados e Discussão

Para a abordagem numérica do problema de evasão escolar em questão, iniciou-se pela análise exploratória da base de dados em questão. Procura-se entender quais são os canais de comunicação mais suscetíveis à família interagir. Assim, optou-se por entender como seria a correlação entre os canais de comunicação da escola presentes na Base de Dados de Engajamento na Plataforma de Comunicação da Família com a Escola. Conforme elucidado sobre a matriz de correlação, trançando-se a correlação entre os canais plota-se o disposto a seguir:

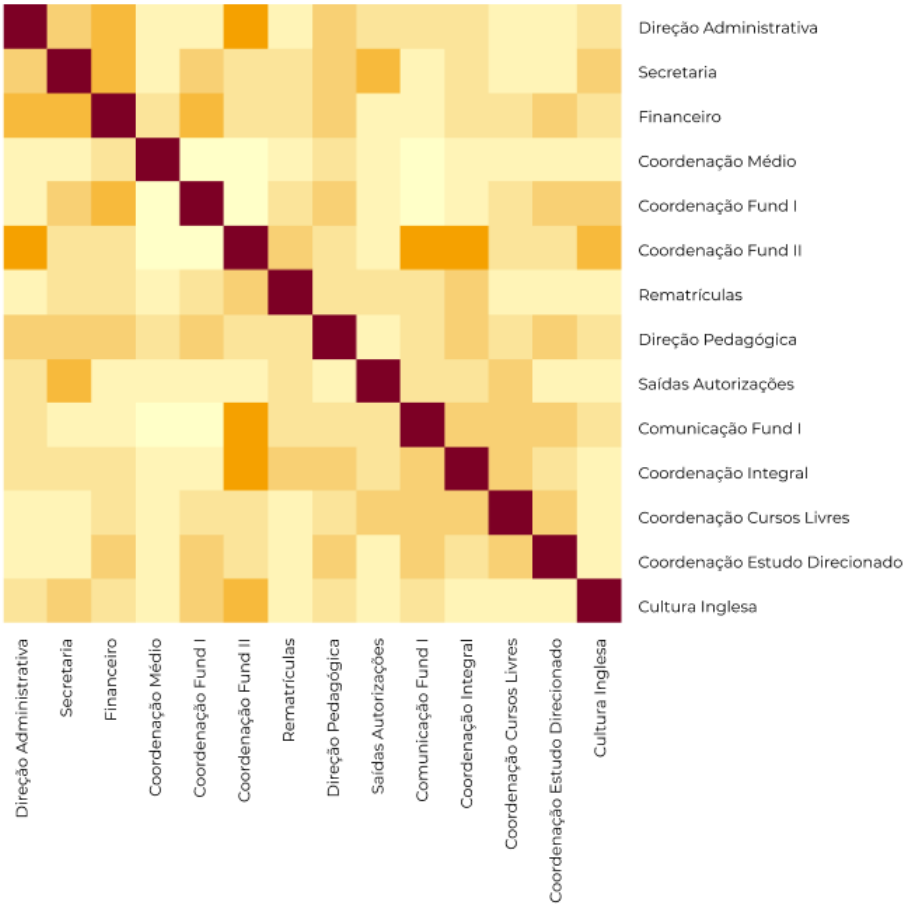


FIGURA 4.1 – *Heatmap* de correlação - Matriz de correlação

Com isso, é possível inferir alguns pontos sobre a escola modelo:

- Os responsáveis que se comunicam pela Coordenação do Fundamental I são mais propensos a se comunicarem via o canal da Direção Administrativa.
- É de se esperar uma maior presença de mensagens dos responsáveis do Fundamental I nos canais do programa de Integral da escola modelo.
- Nota-se uma menor correlação, porém ainda existente entre os responsáveis do Fundamental I e a coordenação de Inglês.

Em seguida, opta-se pela análise das frequências de mensagens por meio da visualização de histograma. Para a análise do histograma, optou-se por clusterizar os canais de comunicação do colégio em dois grandes grupos, o Administrativo e o Pedagógico, conforme explicitado no Anexo. A partir da elaboração de tais visualizações, têm-se:

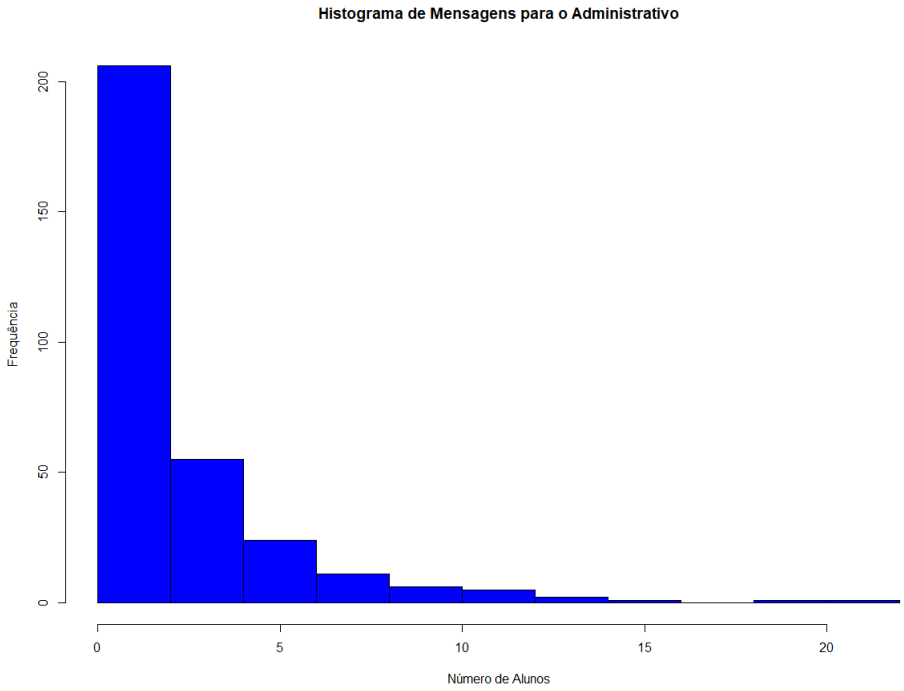


FIGURA 4.2 – Histograma da frequência de mensagens no administrativo por aluno

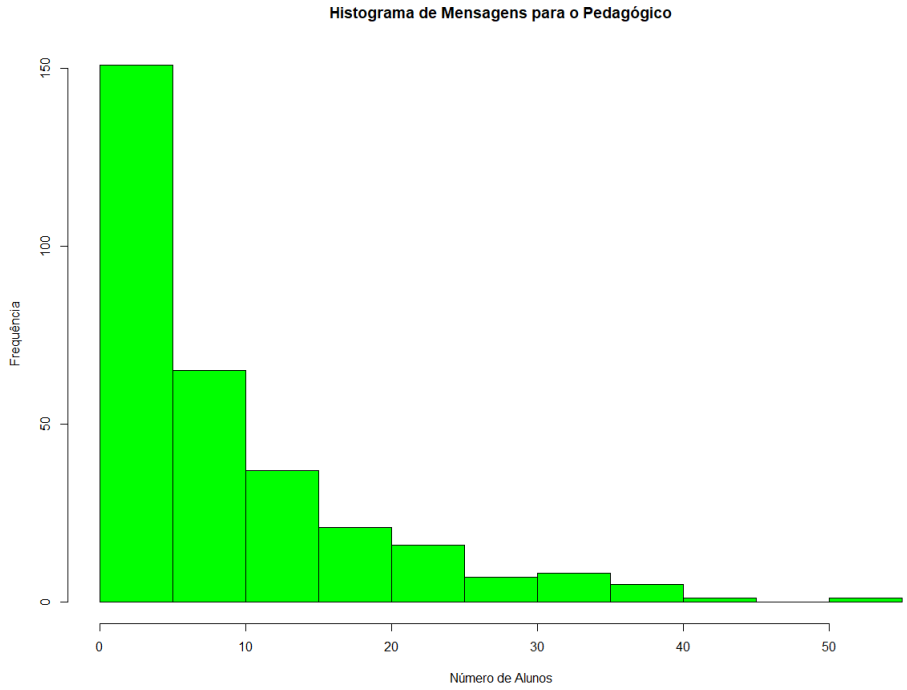


FIGURA 4.3 – Histograma da frequência de mensagens no pedagógico por aluno

Nota-se uma maior frequência máxima presente nos canais de administrativo do que nos canais pedagógicos. Isso elucida um importante ponto que é sobre como são divididas as tarefas burocráticas de atendimento às famílias via canal de comunicação online, mostrando assim que a equipe administrativa possui por vezes uma maior demanda que a pedagógica.

Por último, visando entender o que o colégio modelo enfrenta na rotina e quem são seus *outliers*, optou-se pela visualização do tipo *boxplot*. Com isso, pode-se traçar importantes métricas como os valores máximos, mínimos e média de mensagem por canal.

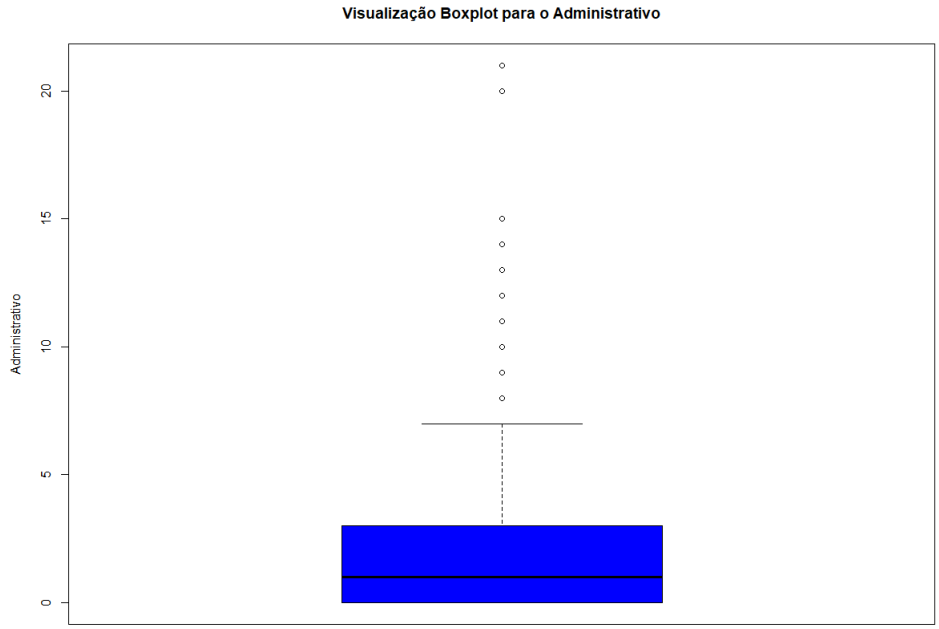


FIGURA 4.4 – *Boxplot* com a base de dados de mensagens das famílias para os canais administrativos

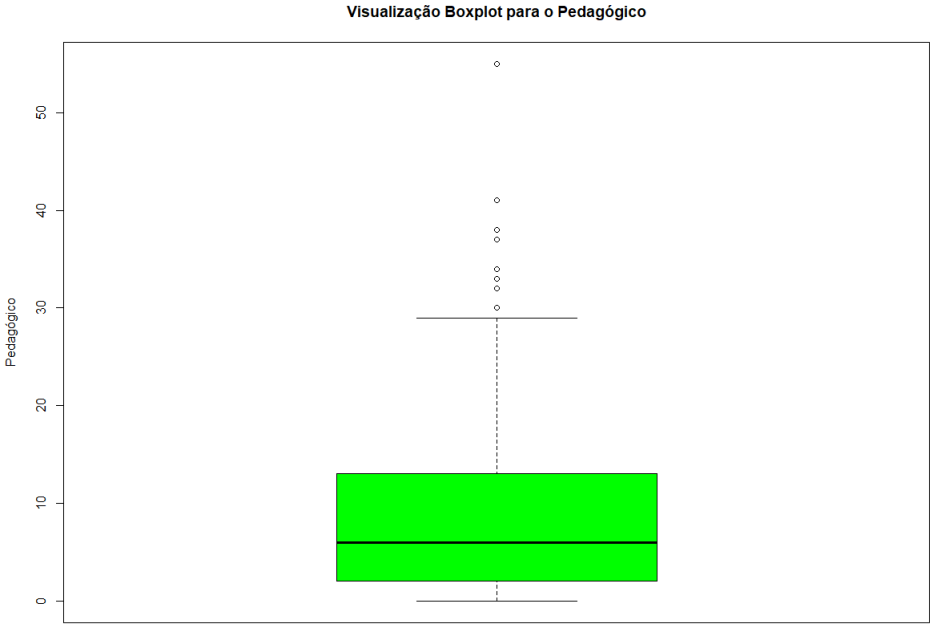


FIGURA 4.5 – *Boxplot* com a base de dados de mensagens das famílias para os canais pedagógicos

Assim, notam-se importantes pontos a partir dos gráficos ilustrados acima. Portanto, é válido mencionar que:

- Mesmo que haja maiores frequências nos canais de Administrativo, nota-se que há um valor máximo admissível maior no canal do pedagógico. Famílias que enviam até

aproximadamente 30 mensagens no canal do pedagógico não devem ser consideradas *outlyers*, diferentemente do canal administrativo que a partir de 7-8 mensagens a família já é considerada uma *outlyer*

- Ressalta-se que a média de mensagens no canal do pedagógico é maior por família. Isso faz jus ao histograma explicitado acima, pois mostrou uma curva mais distribuída ao longo da quantidade de alunos. Esse é um importante ponto de que os processos podem ser melhorados para reduzir a carga de trabalho. Um incremento pequeno nos alunos pode sobrecarregar o setor de atendimento.

Para a visualização 3d com o *scatterplot*, optou-se por transformar cada uma das bases em um único vetor coluna. As transformações propostas para cada uma das bases de dados está disposta nos Anexos.

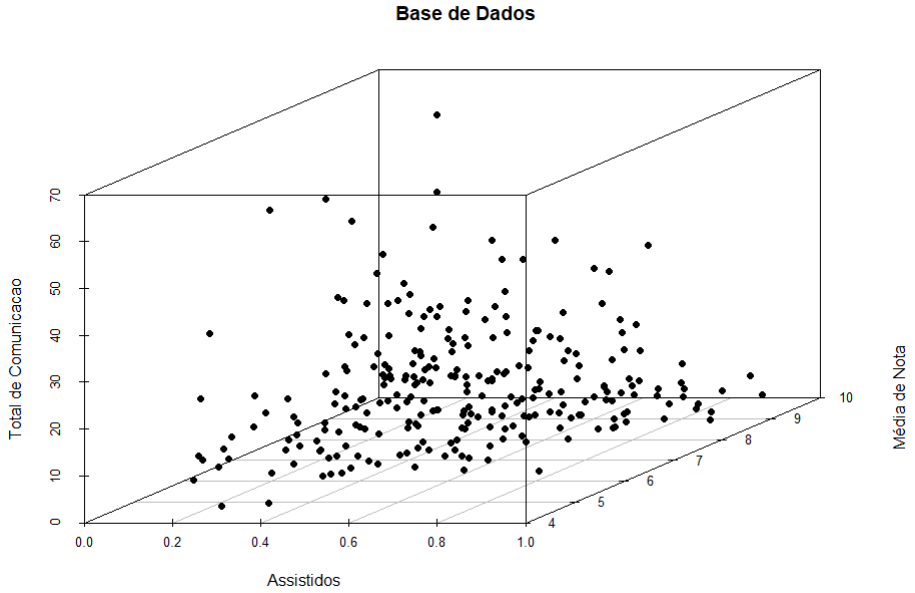


FIGURA 4.6 – *Scatterplot* dos dados combinados

A partir da visualização acima, é possível observar a identificação de ruído nos dados em questão. Para a escolha do número de clusters para o método do *Kmeans*, opta-se por utilizar o *Elbow Method* onde será escolhido o valor de k que provoca a maior variação do valor do erro quadrático médio intra-clusters.

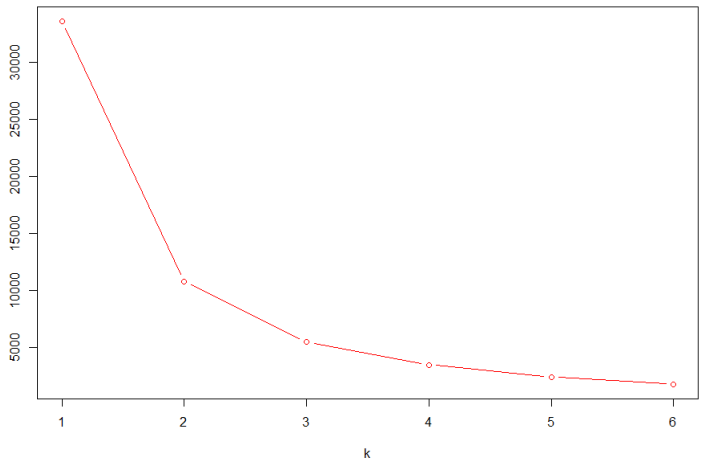


FIGURA 4.7 – Método de escolha do número de *clusters* para o *Kmeans*

Portanto, o valor de k que provoca a maior variação é k igual a 2. Assim, realizando-se a clusterização pelo método *Kmeans* com o valor de k sugerido, tem-se:

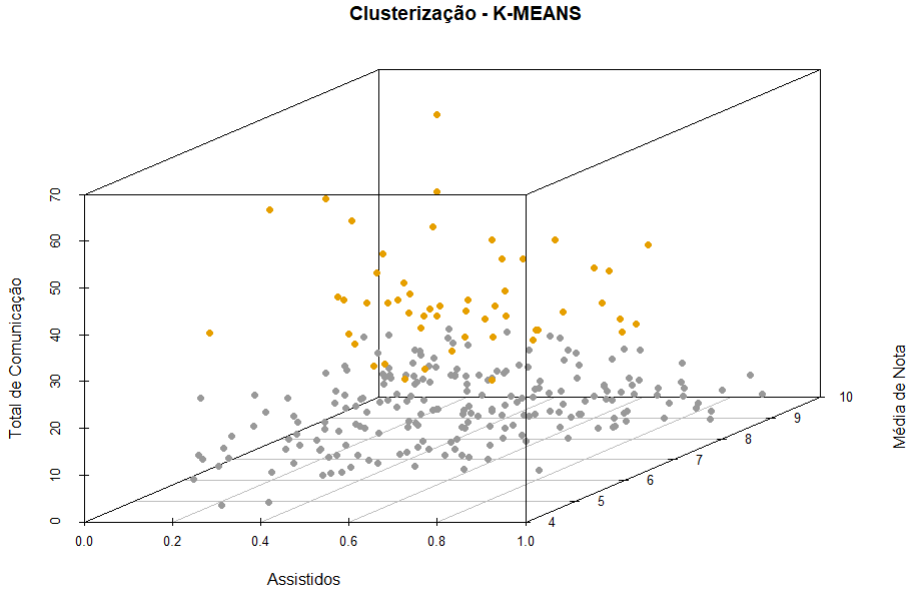


FIGURA 4.8 – Clusterização utilizando o algoritmo *Kmeans*

A partir da análise do *scatterplot* do *output* da aplicação do algoritmo *Kmeans*, torna-se nítido que houve um problema na clusterização devido ao ruído presente no *dataset*. É válido mencionar que nessa clusterização basicamente teve-se dois clusters, um seria composto por todo o ruído e o outro composto pelos alunos que não estão contidos no ruído.

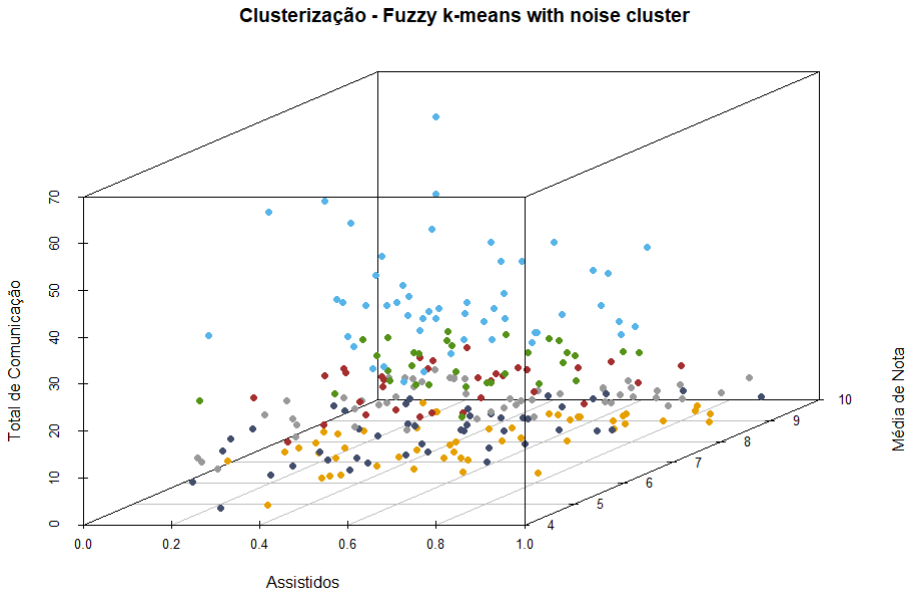


FIGURA 4.9 – Clusterização utilizando o algoritmo *Fuzzy k-means with noise cluster*

Com a aplicação do algoritmo *Fuzzy k-means with noise cluster*, percebe-se a existência de 6 clusters na base de dados analisada. Percebe-se que o cluster o em azul claro é praticamente o mesmo acusado pelo ruído no algoritmo *Kmeans*.

Todavia, é válido mencionar que foram encontrados outros 5 padrões de famílias na escola. Isso possibilita uma análise personalizada dos indivíduos que se encontram em cada um desses grupos.

Por último, implementa-se uma rede neural artificial com o intuito de predizer os alunos que serão transferidos no final do ano letivo. É de importância mencionar que para tal abordagem, utilizou-se 70% da base para treinamento e 30% para validação após calibração da RNA.

Com isso, teve-se a seguinte RNA para a análise em questão:

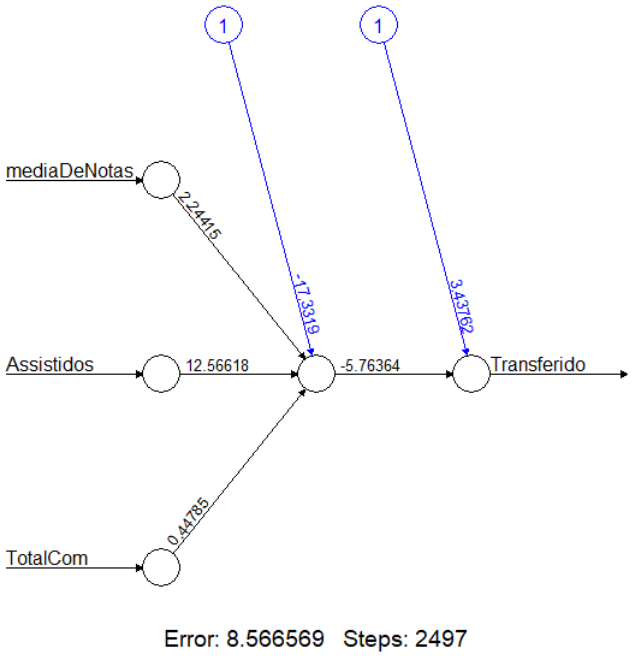


FIGURA 4.10 – Modelo de RNA implementada após calibração

Realizando-se a validação para 30% da base de dados restante, obteve-se a seguinte matriz de confusão para a predição proposta:

TABELA 4.1 – Matriz de Confusão da predição

		Referência	
		0	1
Predição	0	69	7
	1	3	0

Por fim, é notório que a rede neural estruturada anteriormente possui uma boa previsibilidade dos alunos que permaneceram no colégio modelo. Ressalta-se também que não foi identificado nenhum caso positivo de transferência com os *outputs* fornecidos.

5 Conclusões

Os objetivos centrais deste trabalho são entender a correlação dos vários *inputs* que existem no dia a dia escolar e como isso se correlaciona com a evasão escolar ou mudança de escola no final de um ano letivo e tornar possível a escola prever com antecedência as famílias que poderiam deixá-la no período letivo.

É importante mencionar que, mesmo após os algoritmos implementados, é necessário uma melhora na base de dados em questão para que seja viável para a aplicação da escola. É válida a ressalva em relação ao tamanho do *dataset* em questão, dado que foi enfrentado problemas de a base de dados ser pequena demais para o treinamento e o teste da rede neural.

A partir da análise da matriz de correlação do problema aqui disposto, nota-se também que é possível melhorar os *inputs* utilizados na análise dos canais de comunicação. Devem-se manter o máximo de informações linearmente independentes no que diz respeito às comunicações da família com a escola. É importante a identificação de tais dependências tanto de um modo supervisionado quanto não-supervisionado.

Vale salientar também que após o algoritmo com a detecção de ruído implementado, percebeu-se a presença de 6 grupos na rotina da escola. É necessário que haja a reflexão sobre o que isso quer dizer para cada uma das instituições de ensino que cheguem a tal conclusão. Isso se dá devido ao método de clusterização não supervisionado implementado, dificultando entender as conclusões retiradas pela matemática no dia a dia.

Por fim, os resultados encontrados pela rede neural mostram que é possível tal abordagem, todavia ainda é necessário uma melhor base de dados de treino e de teste de forma que se torne acessível o aprendizado por parte de uma maior quantidade de casos de transferência escolar no *dataset*. Para o problema aqui analisado, a pequena relevância dessas informações não tornou tangível prever nenhum caso positivo de transferência escolar, mostrando uma imprecisão do método implementado. Para os casos de permanência na escola, a matriz de confusão elucidou um número considerável de identificações, entretanto esse valor possui um grande viés devido à baixa densidade de transferências identificadas na base de treino da rede neural.

5.1 Sugestões para trabalhos futuros

Dado as considerações supracitadas, sugerem-se alguns importantes pontos de continuação da pesquisa aqui realizada.

5.1.1 Validação das conclusões da escola modelo

Alguns pontos importantes sobre a gestão escolar foram apontados na análise exploratória realizada para a escola modelo. Com isso, é válido ressaltar que é de suma importância a continuidade da pesquisa para escolas de maior porte para buscar entender como é o impacto no número de alunos nas premissas e indagações propostas aqui nessa pesquisa.

5.1.2 Maior assertividade nos dados escolares minerados

Entende-se que um ponto chave da pesquisa aqui realizada é entender como os dados escolares se apresentam e o que eles sugerem. Assim, acredita-se que um importante próximo passo é melhorar os dados coletados por parte da escola. Aumentando o número de medições presente nas bases de dados de forma que torne viável uma melhor implementação dos métodos de clusterização e de rede neural.

Referências

ABIODUN, O. I.; JANTAN, A.; OMOLARA, A. E.; DADA, K. V.; MOHAMED, N. A.; ARSHAD, H. State-of-the-art in artificial neural network applications: A survey. **Heliyon**, Elsevier, v. 4, n. 11, p. e00938, 2018.

ANIELA, B.; BADEA, D. N. Challenges of the transition from classical teaching to online teaching in romanian schools as a result of the covid-19 pandemic. **Annals of'Constantin Brancusi'University of Targu-Jiu. Economy Series**, n. 1, 2021.

BEZERRA, C.; SCHOLZ, R.; ADEODATO, P.; LUCAS, T.; ATAIDE, I. Evasão escolar: aplicando mineração de dados para identificar variáveis relevantes. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.: s.n.], 2016. v. 27, n. 1, p. 1096.

BRASIL. **Lei nº 10.172, de 09 de janeiro de 2001**. Aprova o Plano de Educação e dá outras providências — *Diário Oficial da União*, Brasília, 2001. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/leis_2001/l10172.htm>.

DAVE, R. N. Characterization and detection of noise in clustering. **Pattern Recognition Letters**, Elsevier, v. 12, n. 11, p. 657–664, 1991.

FILHO, R. B. S.; ARAÚJO, R. M. de L. Evasão e abandono escolar na educação básica no brasil: fatores, causas e possíveis consequências. **Educação por escrito**, v. 8, n. 1, p. 35–48, 2017.

HAARMAN, B. C. B.; LEK, R. F. Riemersma-Van der; NOLEN, W. A.; MENDES, R.; DREXHAGE, H. A.; BURGER, H. Feature-expression heat maps–a new visual method to explore complex associations between two variable sets. **Journal of biomedical informatics**, Elsevier, v. 53, p. 156–161, 2015.

IBGE. **Educação : 2019**. Brasil, Rio de Janeiro, 2020. Disponível em: <https:/-/biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhesid=2101736>. Acesso em: 26 ago. 2021.

INEP: Resumo técnico | censo escolar 2020. 2019. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores-/censo-escolar/resultados>. Acesso em: 26 ago. 2021.

INEP: Total médio de anos de estudo cresce no brasil. 2020. Disponível em: <https://agenciabrasil.ebc.com.br/educacao/noticia/2019-06/total-medio-de-anos-de-estudo-cresce-no-brasil-diz-pesquisa-do-ibge>. Acesso em: 26 ago. 2021.

INEP: Apresentação da coletiva de imprensa | censo escolar 2020. 2020. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores-/censo-escolar/resultados>. Acesso em: 26 ago. 2021.

INEP: Inscritos confirmados no enem (1998-2021). 2021. Disponível em: <https://g1.globo.com/educacao/enem/2021/noticia/2021/09/27/enem-2021-tem-280145-novos-participantes-apos-reabertura-de-inscricoes-para-isentos.ghtml>. Acesso em: 26 ago. 2021.

KHAN, M. J.; AHMED, J. Child education in the time of pandemic: Learning loss and dropout. **Children and Youth Services Review**, Elsevier, v. 127, p. 106065, 2021.

LLOYD, S. Least squares quantization in pcm. **IEEE transactions on information theory**, IEEE, v. 28, n. 2, p. 129–137, 1982.

MACQUEEN, J. *et al.* Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**. [S.l.], 1967. v. 1, n. 14, p. 281–297.

MARRIOTT, K.; CHEN, J.; HLAWATSCH, M.; ITOH, T.; NACENTA, M. A.; REINA, G.; STUERZLINGER, W. Immersive analytics: Time to reconsider the value of 3d for information visualisation. In: **Immersive analytics**. [S.l.]: Springer, 2018. p. 25–55.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, Springer, v. 5, n. 4, p. 115–133, 1943.

NUZZO, R. L. Histograms: A useful data analysis visualization. **PM&R**, Wiley Online Library, v. 11, n. 3, p. 309–312, 2019.

PELISSARI, L. B. O fetiche da tecnologia e o abandono escolar na visão de jovens que procuram a educação profissional técnica de nível médio. 2012.

RIFFEL, S. M.; MALACARNE, V. Evasão escolar no ensino médio: o caso do colégio estadual santo agostinho no município de palotina. **O professor PDE e os desafios da escola pública paranaense**, v. 1, p. 01–24, 2010.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, American Psychological Association, v. 65, n. 6, p. 386, 1958.

SCHLEICHER, A. **PISA 2018: Insights and Interpretations**. PISA 2018: Insights and Interpretations — *OECD*, 2018. Disponível em: <https://www.oecd.org/pisa/publications/>.

STEINBACH, A. A. Juventude, escola e trabalho: razões da permanência e do abandono no curso técnico em agropecuária integrado. 2012.

WILLIAMSON, D. F.; PARKER, R. A.; KENDRICK, J. S. The box plot: a simple visual method to interpret data. **Annals of internal medicine**, American College of Physicians, v. 110, n. 11, p. 916–921, 1989.

Anexo A - Transformações realizadas nos Bancos de Dados

A.1 Clusterização realizada no banco de dados de comunicação da família com a escola

Canal Macro	**Canal Micro**
Administrativo	Direção Administrativa
	Secretaria
	Rematrículas
	Financeiro
	Saídas e Autorizações
Pedagógico	Coordenação Médio
	Coordenação Fund I
	Coordenação Fund II
	Direção Pedagógica
	Comunicação Fund I
	Coordenação Integral
	Coordenação Cursos Livres
	Coordenação Estudo Direcionado
	Cultura Inglesa

A.2 Transformações realizadas nos bancos de dados para calibração dos *inputs* para rede neural artificial

Transformação Realizada	**Inputs da Rede Neural**
Média das colunas: Atividade 1, Atividade 2, Atividade 3 , Atividade 4, Atividade 5, Prova Bimestral e Recuperação Bimestral. Notas ausentes são descartadas.	Média total de notas
Porcentagem de "Sim"relativo ao total de aulas que o aluno teve. A coluna usada para cálculo foi a "Aula Assistida"	Porcentagem de Aulas Assistidas
Soma de todas as comunicações nos clusteres Pedagógicas e Administrativas	Total de Comunicação

FOLHA DE REGISTRO DO DOCUMENTO

1. CLASSIFICAÇÃO/TIPO	2. DATA	3. REGISTRO N°	4. N° DE PÁGINAS
TC	23 de novembro de 2021	DCTA/ITA/TC-129/2021	54

5. TÍTULO E SUBTÍTULO:

Perfis Familiares Presentes nos Ensinos Fundamental e Médio e suas Influências na Evasão Escolar.

6. AUTOR(ES):

Rafael Lima Gonzaga

7. INSTITUIÇÃO(ÕES)/ÓRGÃO(S) INTERNO(S)/DIVISÃO(ÕES):

Instituto Tecnológico de Aeronáutica - ITA

8. PALAVRAS-CHAVE SUGERIDAS PELO AUTOR:

Educação, Ensino Fundamental, Evasão Escolar, Ensino Médio e Algoritmos preditivos

9.PALAVRAS-CHAVE RESULTANTES DE INDEXAÇÃO:

Ensino básico; Ensino médio: Abandono; Algoritmos; Educação

10. APRESENTAÇÃO: **(X) Nacional ( ) Internacional**

ITA, São José dos Campos. Curso de Graduação em Engenharia Civil-Aeronáutica. Orientadora: Giovanna Miceli Ronzani Borille. Publicado em 2021.

11. RESUMO:

A educação básica tem sido um importante tópico referente ao desenvolvimento de um país na atualidade. No Brasil, o contexto atual tem-se mostrado frágil e preocupante, principalmente devido a como foi enfretada a realidade da pandemia e como estão sendo dados os passos da gestão federal nesse ramo. Dito isso, esse trabalho busca analisar um aspecto crítico na educação básica global que é a evasão escolar. Nesse trabalho de graduação, busca-se encontrar correlações e previsões entre como a família e o aluno atuam na rotina escolar com a possível evasão ou mudança escolar do aluno no final do período letivo. O modelo de estudo aqui projetado utilizou uma escola modelo de São Paulo cuja representatividade não é significativa no âmbito federativo. Propôs-se uma abordagem sistemática e numérica de parâmetros-chave adquiridos da mineração de dados escolares para clusterizar perfis familiares com a presença ou não de ruído e entender qual o impacto dos eventos em uma análise de transferência escolar. Nos modelos aqui implementados abordou-se algoritmos não-supervisionados e supervisionados para as análises de clusterização e de predição, respectivamente.

12. GRAU DE SIGILO:

(X) OSTENSIVO ( ) RESERVADO ( ) SECRETO