Educational Data Mining: A Study of the Factors That Cause School Dropout in Higher Education Institutions in Brazil

Marília N. C. A. Lima[1], **Roberta A. de A. Fagundes**[1]

[1]Departamento de engenharia da Computação – Universidade de Pernambuco
Recife – Pernambuco – Brasil

mncal@ecomp.poli.br,roberta.fagundes@upe.br

Abstract. ***Context:****In Brazil, there is a high dropout rate in higher education institutions. Thus, it is clear that evasion is a frequent problem and that it is necessary to analyze the factors that cause it to enable solutions that can mitigate/reduce this problem.* ***Objetive:*** *(1)perform a correlation analysis (Pearson and Spearman) of the educational factores of the School Census; (2)propose school dropout prediction models taking into account educational and economic factors using regression methods (linear, robust, ridge, lasso, clusterwise regression).* ***Methodology:*** *used the phases of the CRISP-DM methodology.* ***Results:*** *the factors related to not allowing financial assistance are related to as evasion, namely: food, permanence, didactic material, transportation. There are also factors related to the study period. The regression robust and linear regression show fewer errors.* ***Conclusion:*** *the correlations used present the selection of factors in a similar way, thus following a linear distribution. This study can help to create more investment in public policies, as it ratifies factors are related to this dropout problem.*

Mineração de Dados Educacionais: Um Estudo dos Fatores que causam Evasão Escolar em Instituições do Ensino Superior no Brasil

Resumo. ***Contexto:*** *No Brasil, há altos índices de evasão escolar nas instituições do ensino superior. Dessa forma, percebe-se que a evasão é um problema frequente. Por isso, é preciso analisar os fatores que o influenciam, para possibilitar soluções que possam mitigar ou diminuir esse problema.* ***Objetivo:*** *(1) realizar uma análise correlacional (Pearson e Spearman) dos fatores educacionais e econômicos presentes no Censo Escolar; (2) propor modelos de predição da evasão escolar levando em consideração os fatores educacionais e econômicos baseado na análise correlacional para a utilização dos métodos de regressão (linear, robusta, ridge, lasso e clusterwise).* ***Metodologia:*** *utilizou-se as fases da metodologia CRSIP-DM.* ***Resultados:*** *os fatores relacionados a evasão escolar que não possibilitam auxílio financeiros são: alimentação, permanência, material didático e transporte. Há também fatores relacionados ao período de estudo. A regressão robusta e a regressão linear apresentam menores erros.* ***Conclusão:*** *as correlações utilizadas apresentam a seleção dos fatores de forma semelhante, assim seguem uma distribuição linear. Esse estudo pode ser utilizado como auxílio para criação de mais investimento de políticas públicas, ratificando que esses fatores estão relacionados ao problema de evasão escolar.*

1. Introdução

A educação brasileira apresenta mudanças que buscam melhorias para as instituições de ensino. Entretanto, a evasão escolar ainda é um problema frequente e complexo

nos diferentes níveis de ensino da educação no Brasil (COLPANI, 2018). Para (STEARNS; GLENNIE, 2006) a evasão escolar acarreta altos custos sociais sob a forma de custos de encarceramento, programas de transferência de renda, e receita tributária perdida. Uma vez que, a economia e a educação influenciam fortemente a remuneração dos trabalhadores, por exemplo, os alunos que abandonam a escola no ensino médio mantém uma posição desfavorecida para empregabilidade na sociedade.

Segundo o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) com o censo realizado em 2015, mostra que o Brasil apresenta altos índices de desistência dos alunos de graduação, uma vez que em 2010, 11,4% dos alunos abandonaram o curso e em 2014, esse número chegou a 49% (INEP, acessado 3 de Abril , 2020). Dessa forma, o uso de mineração de dados é fundamental para auxílio a análise dos dados educacionais para monitorar e explorar possíveis causas e soluções para esse problema (Mineração de Dados Educacionais - MDE). A MDE faz o uso de técnicas de mineração de dados para a geração de informação que possibilitem soluções no âmbito educacional, proporcionando melhorias de infra-estrutura, na previsão de desempenho dos alunos, e no contexto geral do processo de ensino e aprendizagem (NASCIMENTO; JUNIOR; FAGUNDES, 2018).

A MDE faz uso de diferentes técnicas entre elas classificação, regressão para extrair conhecimentos das bases de dados, em virtude disso no trabalho de (MARQUES et al., 2019) é realizado um mapeamento sistemático da literatura sobre o uso da mineração de dados auxiliando a descoberta de técnicas que são usadas no contexto da evasão escolar. Neste trabalho, mostra-se o percentual de uso de de técnicas de mineração nesse problema. Pode-se perceber que técnicas de classificação (27%) e equações estruturais (20%) são as mais comumente utilizadas para estimativa desse problema, seguidas de técnicas de regressão (13%).

De acordo com (JUNIOR et al., 2019), o estudo dos fatores educacionais é frequente e objetivo de várias pesquisas, pois existe uma grande disponibilidade de dados educacionais. Entretanto, categorizar esse fatores não é uma tarefa trivial, visto que o abandono escolar pode acontecer na esfera pública e privada de ensino. Diante disso, existe algumas teorias nesse sentido, a saber: *pull-out* e *push-out*. O *pull-out* está relacionado com fatores externo às escolas, por exemplo, família e emprego, por sua vez o *push-out* está relacionado aos elementos internos da escola, como estrutura, docentes e políticas escolares (STEARNS; GLENNIE, 2006).

Com base no exposto, o presente trabalho tem como objetivos (1) realizar uma análise correlacional (Pearson e Spearman) dos indicadores educacionais do Censo Escolar, fornecidos publicamente pelo INEP, com o intuito de determinar quais variáveis educacionais estão relacionadas com a evasão escolar em instituições do ensino superior no Brasil. Além disso, realizar a categorização (*pull-out* e *push-out*); (2) propor cinco modelos preditivos, usando os métodos de regressão linear, regressão robusta, regressão ridge, regressão lasso e regressão clusterwise, para determinar qual dos modelos melhor representa o problema da evasão escolar e seus respectivos fatores.

Este trabalho está estruturado em cinco seções. A segunda seção apresenta uma breve contextualização sobre evasão escolar, bem como, os trabalhos relacionados a área de MDE. A terceira seção descreve a metodologia empregada para o desenvolvimento desse trabalho. A quarta seção apresenta os resultados e discussões obtidos. Por fim, a quinta seção aborda as considerações finais da pesquisa.

2. Fundamentação Teórica

Nesta seção, serão apresentadas algumas considerações sobre o problema da evasão escolar no Brasil e trabalhos relacionados ao tema de MDE.

2.1. Problema da evasão Escolar

A evasão escolar caracteriza-se como o fenômeno de desistência dos alunos no ciclo dos estudos. No Brasil trata-se de um fenômeno constante nos diferentes níveis de ensino (fundamental, médio e superior). Segundo (MARQUES et al., 2019) desde de 1970 a evasão escolar é considerado um problema, já que é de extrema dificuldade encontrar soluções viáveis para mitiga-lo.

Desde de muitos tempo esse problema vem sendo estudado, por exemplo, no ano de 1975 (TINTO, 1975) apresentou um modelo que aponta alguns fatores que causam o fenômeno da evasão escolar. Desse modo, investigar possíveis fatores que causam esse problema é de extrema importância, uma vez que, atinge não somente fatores educacionais, como também, afeta o desenvolvimento social (SILVA; SILVA; ALBUQUERQUE, 2020).

2.2. Trabalhos Relacionados

Alguns trabalhos na literatura apresentam contribuições para mitigar o problema na evasão escolar utilizando a MDE. Em (LIMA MARíLIA, 2019) aplicou-se a técnica de Randon Forest para seleção das variáveis para analisar o desempenho dos estudantes utilizando dados do ENADE, também usaram a técnica de regressão robusta e propuseram um modelo combinado que usa: *clustering* e regressão robusta. Além disso, a métrica erro médio absoluto foi aplicada para avaliar os modelos. Os resultados apresentados foram satisfatórios para o problema e o modelo combinado obteve melhor desempenho que o modelos de regressão robusta para previsão do desempenho dos estudantes.

Já em (SILVA et al., 2019a) tem o objetivo de verificar a aplicabilidade de modelos de regressão para a previsão desempenho de alunos pertencentes as escolas públicas do estado de Pernambuco, utilizando de regressão linear, regressão robusta, regressão quantílica e regressão vetorial de suporte, também usaram o método *Stepwise* para seleção das variáveis e apresentam que fatores como: a quantidade de pessoas que moram na residência, o incentivo dos pais as tarefas escolares e a área onde o estudante reside estão relacionadas ao desempenho dos alunos.

No trabalho de (COLPANI, 2018) os autores estimam a evasão escolar no ensino médio com base nos indicadores do censo escolar, para isso usou a regressão linear, além do coeficiente de correlação de Pearson para seleção das variáveis. Nesse trabalho, também aplicou-se a metodologia CRISP-DM, porém o modelo de estimação da regressão linear apresentava apenas duas variáveis, e a taxa de explicação desta variáveis ao modelo é de 33% (R^2 - o quanto as variáveis independentes explicam a variável dependente), podendo a pesquisa servir como auxílio para soluções de problemas e suporte para o desenvolvimento de mecanismos em apoio ao processo de ensino.

Em (PINTO; JÚNIOR; COSTA, 2019) explorou-se técnicas de seleção, identificando quais fatores impactam no IDEB das escolas municipais de Alagoas, para isso os métodos de seleção de atributos foram usados. Além disso, utilizou-se técnicas de classificação com os algoritmos NaiveBayes, J48, JRip, LinSVM, RandonForest, IBK, OneR,

REPTree, em todos esses algoritmos a taxa de precisão foi superior a 92%. Dessa forma, o estudo possibilita melhorias dos índices nas escolas públicas do estado de Alagoas.

Em (SILVA et al., 2019b) aplicou-se técnicas de *ensemble* (modelos combinados) para estimar a taxa de evasão escolar no ensino superior, para isso o método de seleção automática de variáveis (*stepwise*) e correlação de Pearson foram utilizados. As variáveis selecionadas por esses métodos são: a quantidade de alunos matriculados, o turno de estudo (matutino, noturno), participam de atividades de extensão. Além disso, utilizou-se os modelos bagging com regressão linear, bagging com regressão robusta e bagging com regressão ridge para estimar a taxa de evasão. Os autores concluíram que as técnicas de *ensemble* apresentam resultados satisfatórios para o problema me estudo.

Diante do exposto, o presente trabalho se difere dos demais citados, pois no contexto da evasão escolar será utilizado a correlação de Pearson e correlação de Spearman para a seleção das variáveis. Como também, o uso de modelos de regressão: ridge, lasso e clusterwise. Vale mencionar que a abordagem clusterwise utiliza o processo de formar grupos e modelar regressões dentro de cada grupo para obter um modelo final.

3. Metodologia

Nesta seção, serão apresentadas as etapas do processo seguido para a condução do presente estudo que fez uso da metodologia *Cross-Industry Standard for Data Mining* (CRISP-DM) (CHAPMAN et al., 2000) que é composta de seis etapas. O uso dessa metodologia é para tornar o processo de mineração mais rígido, além de ser uma das metodologias mais populares no contexto da mineração de dados.

3.1. Compreensão do Negócio

Nesta etapa, são definidos os objetivos do projeto. Nesse sentido, busca-se com este projeto identificar os fatores responsáveis pelo fenômeno da evasão escolar em instituições de ensino superior no Brasil em 2015. Para que o objetivo seja atingido duas técnicas de seleção de variáveis serão utilizadas nas bases do censo da educação superior do Brasil(DADOS. . . , acessado 3 de Fevereiro, 2020).

3.2. Compreensão dos Dados

Nesta etapa, os dados educacionais sobre o ensino superior no Brasil foram coletados no portal do INEP. De acordo com INEP, através dos dados disponibilizados verifica-se um amplo panorama da educação no Brasil, sendo portanto, uma ferramenta de transparência sobre a educação superior no Brasil. Nesse contexto, coletou-se os dados de duas bases de dados, são elas: a Base de Dados do Censo da Educação Superior e a dos Indicadores de Fluxo da Educação Superior. Utilizou-se duas bases, pois a taxa de evasão dos estudantes está em uma base distinta dos demais fatores coletados pelo INEP. Dessa forma, realizou-se uma junção das bases de dados através do código do curso, já que é um identificador/ atributo presente nas duas bases de dados. Nesse contexto, os dados coletados foram estruturados em formato CSV e depois feitos as devidas preparações.

3.3. Preparação dos Dados

Na preparação dos dados realizou-se a verificação de dados faltosos e observou-se que a variável referente a taxa de abandono (TDA) tem muitos dados ausentes. Esses dados foram excluídos e as demais variáveis que também existiam dados faltantes foram

substituídos por sua mediana, ou seja, em cada coluna do CSV que tinha dados inexistentes foi calculado a mediana e feito o preenchimento. Toda a análise de dados foi utilizando a linguagem de programação *python* através do uso do *COLAB* (ferramenta do *google* para uso dessa linguagem de forma compartilhada e utilizando GPU).

Após essas atividades, a base de dados contém 19431 instâncias. Com isso, analisou-se descritivamente a evasão escolar no conjunto de dados resultante, percebendo que a média é de 51.36, desvio padrão de 21.08 e mediana de 52.5, verificou-se que a média e mediana estão próximas, demostrando a existe uma grande possibilidade de não ter $outlier$ (dados atípicos) na base dados em estudo. Percebe-se que a taxa de evasão chegou a 100% em alguns cursos (43 cursos). Quando analisou-se a variável referente a apoio financeiro de permanência do aluno em relação as instituições que obtiveram 100% de evasão, observou-se que nenhum aluno recebeu esse auxílio, já o auxilio alimentação apenas 1 não recebeu, o auxílio material didático 40 não receberam, e o auxílio transporte 39 não tiveram acesso. Também observou-se que alguns cursos tiveram a taxa de 0% de evasão (81 cursos), desses 80 não recebem auxílio permanência, 1 não recebeu o auxílio alimentação, e 79 não receberam o auxílio material didático. Isso demostra que fatores financeiros dos alunos podem está intimamente ligados ao abandono escolar.

Ainda é analisado o contexto geral da taxa de evasão escolar em relação aos fatores financeiros considerando as 19431 instâncias. Desses 18454 não receberam auxílio material didático (94.97%), 505 não receberam auxílio alimentação (2.59%), 18541 não receberam bolsa permanência (95.41%), 19317 não receberam bolsa trabalho (99.41%), 18356 não receberam apoio ao transporte (94.46%).

Além disso, os dados ainda foram padronizados em 0.15 e 0.85 para realizar o processo de modelagem (NASCIMENTO et al., 2018). Usou-se essa padronização, pois muitos dados eram categorizados em 0 ou 1, e nesse sentido existiam muitos valores 0, dessa forma optou-se por essa padronização para que não existisse um sobre-ajuste nos modelos de regressão.

Para uma análise mais precisa sobre os fatores que causam a evasão escolar foram usadas duas técnicas de análise de correlação (MONTGOMERY; PECK; VINING, 2012) entre as variáveis, a saber: coeficiente de correlação de postos de Spearman e o coeficiente de correlação de Pearson. As Equações 1 e 2 apresentam os coeficientes de Spearman e Pearson, respectivamente. Onde, y é a variável dependente e x representa a variável independente.

$$Spearman = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}(\sum y^2 - \frac{(\sum y)^2}{n})} \tag{1}$$

$$Pearson = \frac{n\sum xy - (\sum x)(y)}{\sqrt{\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} \tag{2}$$

O coeficiente de correlação de postos de *Spearman* foi usado porque é uma métrica não paramétrica de correlação. A qual analisa a relação entre as variáveis seja o relacionamento linear ou não. Já a correlação de *Pearson* apenas analisa as relações lineares entre as variáveis. Foi utilizado o critério de apenas selecionar as variáveis que obtém uma correlação superior a 0.015 e inferior a -0.015, para que possa existir um

melhor ajuste do modelo de regressão. Neste sentido, a Tabela 1 apresenta as variáveis selecionadas e o valor obtido de correção para os dois tipos de correlação utilizados no trabalho.

Tabela 1. Variáveis Selecionadas pelos dois tipos de correlação

Correlação Spearman		Correlação Pearson	
Variáveis	Valor	Variáveis	Valor
'CO_ALUNO_SITUACAO'	-0.02476,	,'CO_ALUNO_SITUACAO'	-0.0170,
,'IN_ING_VESTIBULAR'	0.03777,	,'IN_ING_VESTIBULAR'	0.0400,
,'IN_ING_ENEM'	-0.04389,	,'IN_ING_ENEM'	-0.04437,
,'IN_ING_DECISAO_JUDICIAL'	0.01513,	,'IN_ING_DECISAO_JUDICIAL'	0.01509,
,'IN_FINANC_ESTUDANTIL'	0.0298,	,'IN_FINANC_ESTUDANTIL'	0.03248,
,'IN_APOIO_ALIMENTACAO'	-0.01511,	,'IN_APOIO_BOLSA_PERMANENCIA'	0.0188,
,'IN_APOIO_BOLSA_PERMANENCIA'	0.01793,	,'IN_APOIO_BOLSA_TRABALHO'	0.0171,
,'IN_APOIO_BOLSA_TRABALHO'	0.018576,	,'IN_APOIO_MATERIAL_DIDATICO'	0.02179,
,'IN_APOIO_MATERIAL_DIDATICO'	0.018483,	,'IN_APOIO_TRANSPORTE'	0.01954,
,'IN_APOIO_TRANSPORTE'	0.015424,	,'IN_BOLSA_ESTAGIO'	0.0179,
,'IN_BOLSA_ESTAGIO'	0.01964,	,'IN_MATRICULA'	0.02525,
,'IN_MATRICULA'	0.02473,	,'IN_INGRESSO_TOTAL'	0.01619,
,'IN_INGRESSO_TOTAL'	0.016688,	,'QT_INGRESSANTE'	0.05162,
,'QT_INGRESSANTE'	0.102085,	,'QT_PERMANENCIA'	-0.1779,
,'QT_PERMANENCIA'	-0.259926,	,'QT_CONCLUINTE'	-0.11920,
,'QT_CONCLUINTE'	-0.211299,	,'QT_DESISTENCIA'	0.08157,
,'QT_DESISTENCIA'	0.189484,	,'QT_FALECIDO'	-0.01577,
,'QT_FALECIDO'	-0.015681,	,'IN_MATUTINO_CURSO'	0.0753,
,'IN_MATUTINO_CURSO'	0.079355,	,'IN_VESPERTINO_CURSO'	-0.04652,
,'IN_VESPERTINO_CURSO'	-0.047381,	,'IN_NOTURNO_CURSO'	0.17525,
,'IN_NOTURNO_CURSO'	0.170466,	,'IN_POSSUI_LABORATORIO'	0.06910,
,'IN_POSSUI_LABORATORIO'	0.071315,	,'QT_MATRICULA_CURSO'	-0.0163
,'QT_MATRICULA_CURSO'	-0.114335,	,'QT_CONCLUINTE_CURSO'	-0.0300,
,'QT_CONCLUINTE_CURSO'	-0.187931,	,'QT_INGRESSO_CURSO'	0.02612
,'QT_INGRESSO_CURSO'	0.031481,	,'TDA'	1
,'TDA'	1		

Uma característica fundamental a ser analisada nesses dois tipos de correlação é a influência da forma da curva (linear ou não) em relação a evasão dos estudantes. Nesse sentido, percebe-se que existe uma variável selecionada, quando usou-se a correlação de *Spearman* ('IN_APOIO_ALIMENTACAO'), comprovando que esta variável tem um formato não linear, já que não é selecionada na correlação de *Pearson*. Isso demonstra um fator importante para o uso da correlação de *Spearman* em dados de evasão escolar em instituições de ensino superior no Brasil.

Analisando os resultado das correlações percebe-se que aulas realizadas no período noturno (0.17) tem grande influência na taxa de evasão, isso demostra que compreender a relação entre a escola e a realidade do aluno é fundamental, pois os alunos precisam, muitas vezes, trabalhar para se manter e não conseguem conciliar o trabalho com a graduação.

A analise dos fatores selecionados, tanto pela correlação de *Spearman* quanto *Pearson*, são categorizados como *push-out*, pois os fatores selecionados tem ligação ao financeiro, quantidade de alunos ingressos, quantidade de alunos matriculados e o período do estudo. Mas vale lembrar que os alunos que não recebem auxílios financeiros são mais propensos a evasão escolar.

3.4. Modelagem

Nesta fase são definidos os modelos utilizados para estimar a taxa de evasão escolar. Dessa forma, propomos cinco modelos de regressão (MONTGOMERY; PECK; VINING, 2012) descritos a seguir. Para entendimento das equações, y_i é o valor real $\hat{y}_i$ é o valor estimado pelos modelos, sendo i o valor de cada instância até n (tamanho da amostra). As estimativas β são determinados minimizando uma função objetivo para todos β, w é o coeficiente da regressão escolhido para minimizar a função custo, ϵ o erro associado ao modelo.

- $PM1$: nesse modelo aplicou a regressão linear múltipla, estimada por $y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i = x_i'\beta + \epsilon_i$.

DOI:

- $PM2$: nesse modelo foi usado a regressão robusta, estimada por $y_i = \sum_{i=1}^{n} \rho(e_i) = \sum_{i=1}^{n} \rho(y_i - x_i'\beta)$. Onde a função ρ fornece a contribuição de cada resíduo para a função objetivo. Esse tipo de regressão é mais adequada para dados que tem $outlier$, pois diferente da regressão linear, a regressão robusta é sensível na presença de dados atípicos, obtendo melhor desempenho, ou seja, apresenta uma modelos mais satisfatório para o conjunto de dados.
- $PM3$: nesse modelo utilizou a regressão ridge, estimada por $y_i = \sum_{i=1}^{n}[y_i - (w * x_i + \beta)]^2, + \alpha \sum_{j=1}^{p} w_j^2$
. Onde, $\alpha \geq 0$ é um parâmetro de ajuste da penalidade, que é determinado separadamente, j é a quantidade de coeficientes até p número máximo. Esse tipo de regressão penaliza quadrados dos coeficientes para que variáveis correlacionadas fortemente tenham coeficientes parecidos.
- $PM4$: nesse modelo usou a regressão lasso, estimada por $y_i = \sum_{i=1}^{n}[y_i - (w * x_i + \beta)]^2, + \alpha \sum_{j=1}^{p} |w_j|$. Onde α é igual 1, variando entre 0 e 1, j é a quantidade de coeficientes até p (número máximo). Quando existe uma correlação muito alta entre as variáveis independentes essa regressão seleciona apenas uma dessas variáveis e zera os coeficientes das outras, de forma a minimizar o erro do modelo.
- $PM5$: nesse modelo aplicou a regressão *clusterwise* que forma modelos de regressão minimizando o erro médio quadrado, utiliza-se o processo de *clustering* de criação de modelos de regressão linear. a função custo a ser minimizada é $k = argmin_{1 \leq h \leq k}\left(y_i - \hat{\beta}_{0h} - \sum_{j=1}^{p} x_{ij}\hat{\beta}_{ih}\right)$ para cada grupo (k).Onde h pode variar de 1 até k, e j é uma lista de modelos proposto até o máximo modelo P.

3.5. Avaliação

Nesta fase é realizada a avaliação dos modelos usados segundo uma métrica. Neste sentido, utilizou-se o erro médio absoluto do inglês *Mean Absolute Error* (MAE).Que é expresso por $MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|$. Onde n é o tamanho do conjunto de dados, y_i é a valor real da variável e $\hat{y}_i$ é a valor estimado/predito pelo modelo.

3.6. Aplicação

Nesta etapa do processo da metodologia as informações e conhecimentos são disponibilizados a indústria, sendo assim nesta pesquisa esta etapa não foi realizada.

4. Resultados

Nesta seção serão apresentados os resultados dos experimentos realizados de acordo com a metodologia CRISP-DM apresentada na seção 3.

A Tabela 2 apresenta a média dos erros e os desvios padrão das 30 iterações realizadas no estudo. Como pode ser analisado no cenário utilizando a correlação de Spearman, o erro do $PM1$, $PM2$ são similares, entretanto os modelos $PM3$, $PM4$ e $PM5$ possui maiores erros que os demais modelos. Ao analisar o cenário das variáveis seleciona por correlação de Pearson, o mesmo aconteceu, o $PM4$ obteve maiores erros. Não houve diferença significativa entre entre os desvios padrão dos modelos $PM1$, $PM2$, $PM3$ e $PM5$ nos dois cenários analisados.

Tabela 2. Média e Desvio Padrão dos Modelos Propostos

Tipo de Correlação	Erro (Desvio Padrão)				
	PM1	PM2	PM3	PM4	PM5
Spearman	16.073 (0.16318)	16.0683 (0.16245)	16.26708 (0.17018)	17.24815 (0.17705)	16.41804 (0.16096)
Pearson	16.12050 (0.13176)	16.06588 (0.14887)	16.31205 (0.13513)	17.28139 (0.14958)	16.44725 (0.13127)

Outra maneira de verificar os resultados é por meio do uso do gráfico $boxplot$, este gráfico mostra a variação/dispersão dos dados contidos em uma amostra, no trabalho essa amostra refere-se aos erros dos valores preditos em relação ao valor real relacionados a taxa de evasão escolar. A Figura 1 apresenta a variação dos erros dos modelos nos dois cenários de estudo. Observa-se que o $PM4$ tem erros maiores que os outros modelos, e que os modelo $PM1$ e $PM2$ apresentam mediana (representada pela linha na caixa) similares. Ainda verifica-se que para o cenário com correlação de *Spearman* todos os modelos não tiveram a presença de $outlier$ (representado pelas bolas fora da caixa), já para o cenário com a correlação de *Pearson* apenas o $PM2$.

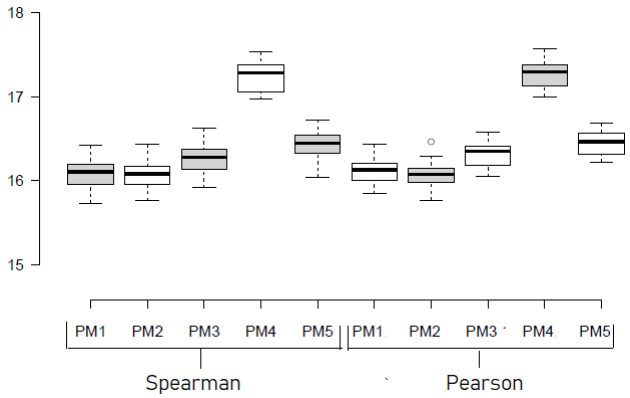

Figura 1. Boxplot dos erros da 30 iterações realizadas nos experimentos

Além disso, testes estatísticos foram realizados baseados nos erros dos modelos. Primeiro utilizou-se o teste de normalidade e foi verificado que todos os erros dos modelos não seguem uma distribuição normal. Dessa forma, usou-se o teste de *Wilcoxon* com 5% de significância, comparando os cinco modelos propostos, verifica-se o $PM1$ apresenta menores erros que os demais modelos nos dois cenários (*Pearson* e *Spearman*). Os resultados obtidos mostram que apenas para o $PM2$ não há evidências estatísticas de menores erros.

Nesse sentido, conclui-se que os modelos $PM1$ e $PM2$ são similares visualmente (box-plot) e apresentam um melhor desempenho para a predição da taxa de evasão escolar. Como os erros desses modelos são semelhantes, comprova-se que não existe presença de $outlier$ na base dados. Assim como já analisado na seção 3 pela mediana em relação a média. Dessa forma, obteve-se dois modelos para estimar com precisão o relacionamento das variáveis do estudo, pois esses modelos minimizaram o erro de predição com maior precisão que os outros modelos estudados.

Alguns artigos explicam o abandono escolar, mas poucos tentaram prever o abandono escolar aplicando técnicas de regressão. No trabalho mostra-se que uso dessas técnicas é importante para auxílio ao processo de monitoramento da evasão no Brasil, pois os modelos propostos alcançaram resultados satisfatórios.

5. Considerações Finais

Diante do exposto no estudo, os fatores selecionados pelos dois diferentes tipos de correlação (*Spearman* e *Pearson*) tiveram muitas variáveis semelhantes, isso demonstra que para esse conjunto de dados a seleção de variáveis utilizando técnica paramétrica (*Pearson*) ou não paramétrica (*Spearman*) apresentaram bons resultados. A utilização de cinco modelos de regressão foram aplicados de forma a garantir uma boa capacidade de generalização (fatores selecionados que tem dependência com a taxa de evasão escolar).

As principais variáveis selecionadas tem relação com o turno de estudo (noturno, matutino), e auxílio para os estudantes (permanência, alimentação, transporte e material didático), demonstrando que fatores *push-out* tem muita relação com a taxa de evasão dos estudantes. Ao analisar os resultados, percebe-se que os melhores resultados são da regressão linear($PM1$) e da regressão robusta($PM2$) para a previsão da evasão escolar. Esses modelos obtiveram menores erros apresentados pelos gráficos $boxplot$, médias e desvio padrão, ratificando esse desempenho através do teste estatístico (teste de hipótese).

Nesse contexto, as principais contribuições do estudo estão em aplicação de dois tipos de correlação e dois modelos de regressores que minimizam o erro de predição no contexto da evasão escolar baseado nos fatores selecionados. Desse modo, o uso de MDE nesse estudo identificou fatores que precisam de um investimento mais adequado para minimizar fenômeno o evasão de escolar. Assim, o presente estudo possibilita um ganho na literatura e aos interessados no processo de ensino e aprendizado. Além disso, sugere-se a utilização de politicas e investimentos para reduzir o efeito da evasão escolar em instituições de ensino superior no Brasil.

Como trabalhos futuros pretende-se utilizar outros métodos de seleção de variáveis como o *Random Forest*, além do uso de outros modelos de regressão (SVR, regressão quantílica). Também pretende-se utilizar fatores externo (pull-out) das universidades para analisar o impacto na evasão dos estudantes.

Agradecimento

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

CHAPMAN, P. et al. Crisp-dm 1.0: Step-by-step data mining guide. **SPSS inc**, v. 9, p. 13, 2000.

COLPANI, R. Mineraçao de dados educacionais: um estudo da evasao no ensino médio com base nos indicadores do censo escolar. **Informática na educação: teoria & prática**, v. 21, n. 3, 2018.

DADOS - INEP. acessado 3 de Fevereiro, 2020. <http://portal.inep.gov.br/web/guest/microdados>.

INEP. acessado 3 de Abril , 2020. <http://portal.mec.gov.br/component/tags/tag/32044-censo-da-educacao-superior>.

JUNIOR, R. N. et al. Estimação de índices de aprovação e reprovação escolar do ensino médio. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.: s.n.], 2019. v. 30, n. 1, p. 339.

LIMA MARíLIA, A. G. S. W. F. R. Educational data mining: A hybrid approach to predicting academic performance of students. In: **International Conference on Machine Learning and Data Mining in Pattern Recognition - MLDM**. [S.l.: s.n.], 2019.

MARQUES, L. T. et al. Mineração de dados auxiliando na descoberta das causas da evasão escolar: Um mapeamento sistemático da literatura. **RENOTE-Revista Novas Tecnologias na Educação**, v. 17, n. 3, p. 194–203, 2019.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. [S.l.]: John Wiley & Sons, 2012. v. 821.

NASCIMENTO, R. L. S. do; JUNIOR, G. G. da C.; FAGUNDES, R. A. de A. Mineração de dados educacionais: Um estudo sobre indicadores da educação em bases de dados do inep. **RENOTE-Revista Novas Tecnologias na Educação**, v. 16, n. 1, 2018.

NASCIMENTO, R. L. S. do et al. Educational data mining: An application of regressors in predicting school dropout. In: SPRINGER. **International Conference on Machine Learning and Data Mining in Pattern Recognition**. [S.l.], 2018. p. 246–257.

PINTO, G. da S.; JÚNIOR, O. d. G. F.; COSTA, E. de B. Identificação dos fatores de melhorias no ideb pelo uso de mineração de dados: Um estudo de caso em escolas municipais de teotônio vilela-alagoas. **RENOTE-Revista Novas Tecnologias na Educação**, v. 17, n. 3, p. 183–193, 2019.

SILVA, E.; SILVA, J.; ALBUQUERQUE, C. de. Uma análise da evasão escolar nos cursos de tecnologia da informação: Um estudo de caso em floresta/pe. In: SBC. **Anais do XXIV Workshop sobre Educação em Computação**. [S.l.], 2020. p. 408–416.

SILVA, P. et al. Modelos de regressão aplicados a predição do desempenho escolar de estudantes do ensino fundamental. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.: s.n.], 2019. v. 30, n. 1, p. 1621.

SILVA, P. M. da et al. Ensemble regression models applied to dropout in higher education. In: IEEE. **2019 8th Brazilian Conference on Intelligent Systems (BRACIS)**. [S.l.], 2019. p. 120–125.

STEARNS, E.; GLENNIE, E. J. When and why dropouts leave high school. **Youth & Society**, Sage Publications Sage CA: Thousand Oaks, CA, v. 38, n. 1, p. 29–57, 2006.

TINTO, V. Dropout from higher education: A theoretical synthesis of recent research. **Review of educational research**, Sage Publications Sage CA: Thousand Oaks, CA, v. 45, n. 1, p. 89–125, 1975.