DEP–DM: Uma Abordagem baseada em *Ensemble Regression* para o Diagnóstico de Problemas Educacionais

Paulo Mello da Silva - CIN /UFPE – PE – pms3@cin.ufpe.br
Marilia N. C. A. Lima – ECOMP/ UPE – mncal@ecomp.poli.br
Roberta A. A. Fagundes – ECOMP/UPE – roberta.fagundes@upe.br
Fernando da F. de Souza – CIN/UFPE – fdfd@cin.ufpe.br

Resumo. *Este artigo apresenta a abordagem DEP-DM, baseada em modelos Ensemble Regression para o diagnóstico do problema do desempenho escolar dos estudantes das escolas públicas de Pernambuco. O estudo baseou-se em informações do SAEB do ano de 2013. O conhecimento foi extraído por meio da abordagem proposta, sendo possível identificar os fatores associados e estabelecer as relações de causa e efeito com o problema do desempenho escolar. Por fim, foram aplicados modelos de regressão paramétricos e não paramétricos para a previsão desse desempenho. Os resultados e diagnóstico apresentam que os fatores relacionados a infraestrutura da escola, ensino e familiares, exercem forte influência sobre o desempenho escolar.*

Palavras-chaves: *Mineração de dados educacionais, modelos de regressão combinados, SAEB, diagrama causa e efeito, diagnóstico educacional.*

DEP –DM: An Ensemble Regression-Based Approach to Diagnosing Educational Problems

Abstract. *This article presents the DEP-DM approach based on Ensemble Regression models for the diagnosis of the problem and the performance of the student public school of the Pernambuco. The study used information from the SAEB in 2013. The knowledge was extracted through the proposed approach, making it possible to identify the associated factors and defined as cause and effect relationships with school performance problems. Finally, parametric and nonparametric regression models were used to predict this performance. The results and diagnosis show the factors related to the infrastructure of the school, education, and family, a strong influence on school performance.*

Keywords: *Education data mining, ensemble regression models, SAEB, cause and effect diagram, educational investigation.*

1. Introdução

A compreensão dos desafios da Educação em busca de um ensino de qualidade perpassa por estudos que sejam baseados em evidências, a partir de estatísticas e análise de dados oriundos de bases de dados educacionais. No Brasil, o Instituto Nacional de Pesquisas Educacionais Anisio Teixeira (INEP), detém a mais completa base de dados sobre educação do país. Os dados educacionais são de origem censitária ou de avaliações em larga escala, como por exemplo: o Sistema de Avaliação da Educação Básica – SAEB (INEP, 2018).

Os dados contidos nas bases do SAEB, caracterizam-se pelas proficiências nas avaliações de Língua Portuguesa (LP) e Matemática (MT), e das respostas dos

questionários contextuais respondidos pelos estudantes, diretores, professores e pela escola. As evidências da realidade da educação básica brasileira contidas nesses dados, nos permite extrair informações relevantes que serão importantes para a identificação dos fatores associados, que afetam desempenho escolar dos estudantes.

Estudar os fatores associados que afetam o desempenho dos estudantes torna-se relevante para a pesquisa educacional, porque permite destacar, com mais acurácia, a importância relativa dos fatores, provenientes de três grupos: a família, o aluno e a escola. O primeiro, influencia com sua própria estrutura e seu envolvimento no processo de aprendizagem. O segundo, com suas características pessoais e atitudes em relação a escola e o terceiro com equipes de profissionais competentes, metodologia de ensino, recursos físicos e pedagógicos (SOARES, 2004).

Na busca por soluções computacionais que fossem capazes de facilitar e potencializar o processo de ensino aprendizagem, utilizou-se a Mineração de Dados Educacionais, como uma possibilidade de investigar e encontrar padrões entre o desempenho escolar e os fatores associados a aprendizagem. Nesse contexto, (Baker, Isotani e Carvalho, 2011), conceituam a Mineração de Dados Educacionais (MDE) como a área de pesquisa que tem como principal foco o desenvolvimento de métodos para explorar um conjunto de dados coletados em cenários educacionais.

No trabalho de Adeodato *et al*. (2014), os autores analisaram dados do ENEM, e do Censo Escolar, baseado no uso da Regressão Logística, com o objetivo de produzir um classificador capaz de gerar uma pontuação de propensão de sucesso da escola, a partir das suas características e daquelas dos seus docentes, discentes e famílias. Como resultados os coeficientes de pontuação de qualidade medida pela AUC_ROC= 0,987 e Máx_KS2=0,632, mostraram que o fator econômico e a infraestrutura dos laboratórios são fatores que influenciam a boa qualidade da escola.

Em Bezerra *et al.* (2016) os autores abordaram a evasão escolar no último ano do ensino fundamental nas escolas públicas estaduais do estado de Pernambuco, baseado nos dados do Censo Escolar dos anos de 2011 e 2012. Utilizou-se técnicas de mineração de dados para identificar o perfil do aluno evadido e estimar a propensão de evasão. Foram utilizados algoritmos de Árvore de Decisão, Indução de Regras e Regressão Logística. Os resultados mostraram que os fatores como idade, turno das aulas e região geográfica das escolas influenciam fortemente na evasão.

Em Pinto *et al*. (2019), os autores identificaram os fatores que afetam o desempenho escolar (IDEB) dos alunos do ensino fundamental do Município de Teotônio Vilela através dos resultados obtidos na prova Brasil. Foram utilizados técnicas de seleção de atributos e os algoritmos de classificação J48, *OneR, LibSVM*, *Random Forest, IBK*, *NaiveBayes* e o *REPTree*. Como resultados os algoritmos *OneR, LibSVM* e J48, apresentaram os melhores resultados com mais de 98% de acurácia para o conjunto de dados de português e matemática.

Este trabalho propõe uma abordagem denominada DEP–DM (*Diagnosis of Educational Problems using Data Mining)* que utiliza modelos de *Ensemble Regression* (Montgomery et al. 2012) para a previsão do desempenho escolar. Para isso investiga as relações de causa e efeito entre os fatores associados ao desempenho em Língua Portuguesa (LP) e Matemática (MT), dos alunos do ensino fundamental das escolas públicas do estado de Pernambuco. O objetivo é trazer informações que contribuam para o desenvolvimento de políticas públicas visando a melhoria da aprendizagem e

consequentemente do desempenho nessa fase escolar utilizando o diagrama de causa e efeito (SLACK et al. 2009).

Nesta perspectiva, este trabalho se diferencia dos demais por propor a aplicação de modelos de *Ensemble Regression* utilizando métodos paramétricos e não-paramétricos para identificação dos fatores associados que afetam o desempenho dos estudantes em LP e MT no ano de 2013. De acordo com os seguintes aspectos: i) realizar análise do problema, investigando os fatores mais preponderantes que afetam o desempenho escolar dos estudantes do 5º ano ensino fundamental das escolas públicas do Estado de Pernambuco; ii) análise das bases de dados do SAEB, para determinar quais fatores exercem influência significativa no desempenho dos estudantes em LP e MT iii) a utilização de modelos de Regressão proporciona a previsão de condições futuras, dando suporte quantitativo a decisão, apontando falhas, e fornecendo *insights* que podem ajudar os gestores educacionais e professores a tomar decisões estratégicas em relação a aprendizagem de seus alunos (iv) contribuir com o fortalecimento do campo da MDE, na resolução de problemas relacionados a previsão do desempenho no contexto educacional.

O trabalho está organizado da seguinte forma: Na Seção 2 apresentamos a abordagem proposta, os experimentos realizados e suas respectivas análises. A Seção 3 apresenta os resultados e discussões. Por fim, na Seção 4 foram são apresentadas as conclusões da abordagem proposta.

2. Abordagem Proposta (DEP-DM)

Em função da natureza dos objetivos propostos neste trabalho direcionarem para o uso e aplicação de técnicas de MDE, como forma de determinação dos modelos preditivos, para o diagnóstico dos problemas educacionais, e para adequação ao contexto dos dados educacionais, foi proposta uma abordagem específica, com base na metodologia CRISP-DM, denominada DEP-DM (*Diagnosis of Educational Problems using Data Mining*). A Figura 1 apresenta a abordagem DEP-DM, a qual consiste em 5 (cinco) fases. A sequência de fases não é obrigatória, podendo ocorrer a transição para diferentes fases dependendo do resultado apresentado em cada fase. Os fluxos indicam as mais importantes e mais frequentes dependência entre as fases.

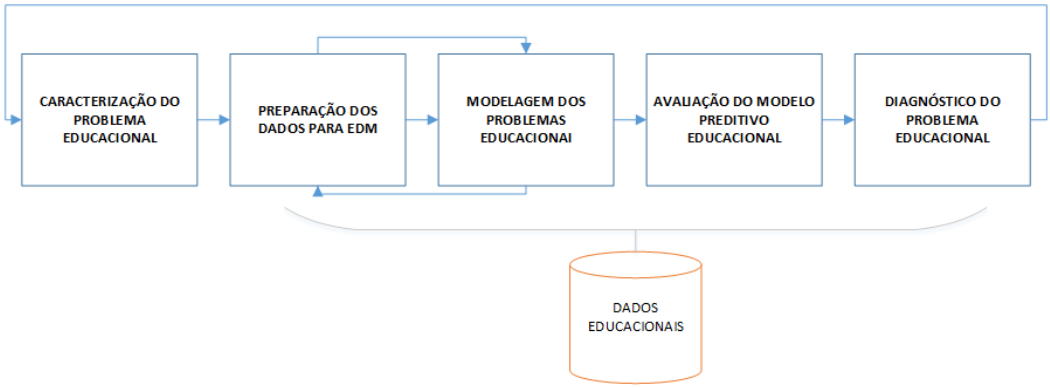

Figura 1. Abordagem DEP-DM

2.1. Caracterização do Problema Educacional

O sucesso escolar nem sempre é alcançado por todos os alunos ao longo dos ciclos escolares. Muitas são as dificuldades enfrentadas pelos estudantes ao longo do percurso escolar que se manifestam por meio do fenômeno do fracasso escolar que se caracteriza

por problemas relacionados a aprendizagem (Dazzano et al. 2016). Dentre os problemas relacionados a aprendizagem destaca-se o baixo desempenho escolar que se caracteriza quando um estudante apresenta em tarefas de sala de aula ou em avaliações internas ou externas, habilidades e competências abaixo do nível esperado para sua idade/série (D´ABREU et al. 2010).

2.2. Preparação dos dados para MDE

Para este contexto de aplicação, utilizou-se a base de dados do SAEB (Sistema de Avaliação da Educação Básica), realizado no ano de 2013 (INEP, 2018). Essa base foi considerada para verificar os fatores associados que influenciam o baixo desempenho escolar dos estudantes do 5º ano das escolas públicas do Estado de Pernambuco. O exame é composto de alguns instrumentos. Nesta pesquisa foi utilizado os questionários contextuais da escola e do aluno, considerando informações que caracterizam as condições de infraestrutura, de ensino e o perfil sócio econômico dos estudantes.

Os conjuntos de dados devem ser preparados adequadamente. Assim, foram realizadas as seguintes atividades: (i) Filtrar os dados para atender ao foco principal deste trabalho; (ii) Verificar os valores ausentes ou em branco; (iii) Normalização dos dados; e (iv) Seleção de variáveis.

O processo de filtragem de dados tem como objetivo extrair os dados de interesse deste trabalho da base original. Neste caso, serão considerados os dados referentes ao Estado de Pernambuco. A base de dados do SAEB neste caso, possui dimensão inicial de 105.451 instâncias com 214 atributos referentes a informações dos estudantes e das escolas públicas e privadas do Estado.

Buscando identificar de forma mais ampla os fatores associados que afetam de forma expressiva o desempenho dos estudantes pernambucanos, realizou-se a junção das bases de dados aluno e escola. O objeto deste trabalho restringiu-se a análise apenas dos estudantes e das escolas públicas (estaduais e municipais) do Estado de Pernambuco. A razão desta restrição deve-se a esses estudantes e estas escolas apresentarem os maiores índices de baixo desempenho escolar conforme dados do INEP (2018). Os dados mostram que apenas 19% dos alunos dessas escolas atingiram o aprendizado esperado em LP e 16% em MT no referido ano. Os números ainda mostram que 37% dos estudantes pertencentes as escolas estaduais apresentam níveis baixos de aprendizado em MT e LP. A partir desse cenário identificou-se os diversos fatores associados a aprendizagem que compõem as dimensões propostas no modelo teórico proposto por Soares (2004) que são: Aluno, Família, Escola e Sociedade.

Foi realizada uma análise das variáveis e as transformações destas de acordo com os objetivos e das necessidades identificadas neste estudo. O processo foi iniciado com a seleção dos dados, sendo excluídas as instâncias que possuíam o valor 0 (zero) para o desempenho, ou seja, foram excluídas as escolas que não responderam o questionário contextual. A caracterização do desempenho (variável dependente) foi definida pelas proficiências dos alunos nas disciplinas de LP e MT. Já os fatores associados (variáveis independentes) foram definidos pelas informações presentes nos questionários contextuais respondidos pelos alunos e pelas escolas. As variáveis redundantes ou irrelevantes foram excluídas, e os registros com valores não preenchidos (*missing values*), foram preenchidos utilizando a mediana entre os atributos. Os dados do questionário contextual, para efeitos da aplicação das técnicas de regressão, foram transformados em dados numéricos e em alguns casos dicotômizados.

Para a seleção das variáveis foram utilizados dois métodos: *Stepwise* e a Correlação de *Pearson*. Assim, a quantidade de instâncias depois do pré-processamento realizado foi 48.609 com 147 atributos referentes as informações dos estudantes e das escolas públicas e privadas do Estado de Pernambuco.

A descrição das variáveis selecionadas pelos métodos *Stepwise* e Correlação de *Pearson* são apresentadas na Tabela 1, a partir da seleção das variáveis relacionadas aos cenários de LP e MT, foi possível a construção dos modelos preditivos. Ainda nessa Tabela 1 foi possível visualizar as variáveis associadas com maior influência no desempenho dos estudantes em LP e MT. Além disso, para a correlação dessas variáveis com o desempenho escolar foi tomado como base o modelo teórico proposto por Soares (2004). Tal modelo classifica os fatores associados ao desempenho como pertencendo a quatro grupos: relacionados ao aluno, relacionados a família, relacionados a escola e relacionados a sociedade.

Tabela 1. Variáveis Selecionadas pelo Método Stepwise e Pearson

Variável	Cenário/Método	Coef LP	Coef MT
Participação nas Avaliações	LP/PEARSON	0,69	
Estado de Conservação da Escola	LP/MT/STEPWISE	0,51	0,44
Iluminação das salas de aula	MT/STEPWISE		0,66
Escola possui sala de estudo	LP/PEARSON	0,42	
Uso de mídias educativas	LP/MT/STEPWISE	0,33	0,45
Formação inicila do Professor	MT/STEPWISE		0,48
Segurança em horário de funcionamento	LP/MT/STEPWISE	0,41	0,43
Sistema anti-furto	LP/MT/STEPWISE	0,49	0,36
Mecanismo de proteção dos equipamentos	LP/MT/STEPWISE	0,59	0,52
Hábito de leitura da mãe	LP/MT/STEPWISE	0,85	0,51
Icentivo para realização das tarefas escolares	LP/MT/STEPWISE	0,4	0,44
País conversam com os filhos sobre o seu andamento na escola	LP/MT/STEPWISE	0,45	0,65

Dos fatores associados ao desempenho escolar identificados, os que exercem influência mais significativa no desempenho em LP e MT são: a participação do aluno na avaliação, a existência de sala de estudo na escola, o hábito de leitura da mãe, a formação docente inicial, o estado de conservação da escola, a existência de segurança durante o período de funcionamento da escola, a existência de sistema antifurto na escola e mecanismos de proteção dos equipamentos.

2.3. Modelagem dos Problemas Educacionais

Nesta etapa, utilizou-se os seguintes algoritmos: RL (Regressão Linear), RLR (Regressão Linear Robusta), RD (Regressão Ridge), SVR (*Support Vector Regression*), *Bagging* (B-RL, B-RR, B-RD, ML-SVR) e o modelo proposto nesse estudo (MP1, MP2, MP3 e LM4). Os modelos propostos foram combinados da seguinte forma: (i) MP1: *Ensemble Bagging* com Regressão Linear; (ii) MP2: *Ensemble Bagging* com Regressão Robusta; (iii) MP3: *Ensemble Bagging* com Regressão Ridge; e (iv) ML4 (Modelo da Literatura), representa o modelo proposto por Nascimento (2018) para a construção do modelo baseado em *Stacking* como meta preditor a RD e as regressões para compor o *Ensemble* são: Regressão Linear, Regressão Lasso, *Bagging, Boosting*, *Randon Forest, Support Vector Machine, Knearest Neighbors.*

2.4. Avaliação do Modelo Preditivo Educacional

Nesta etapa foram avaliados os resultados obtidos verificando sua relação com os objetivos propostos neste trabalho para o diagnóstico e predição do desempenho educacional. Para avaliar com precisão o desempenho dos modelos preditivos utilizou-se os índices MAE (*Mean Absolute Error*) e GR (Ganho Relativo). Estes índices são muito utilizados para o cálculo da previsão baseado no erro de previsão e o ganho em relação a minimização do erro de predição. O cálculo do erro absoluto médio (MAE) é mostrado na Equação 1.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i | \qquad (1)$$

Onde, n é o tamanho do conjunto de dados, i representa cada modelo, y_i é o valor real da variável e $\hat{y}_i$ é a variável estimada pelo modelo. Outra forma de medição de desempenho é através do ganho relativo (GR). O GR é aplicado para mensurar o ganho em relação a minimização do erro de predição, dado em porcentagem conforme Equação 2.

$$GR = 100 \left(\frac{MAE_a - MAE_b}{MAE_a}\right) \qquad (2)$$

Onde, MAE_a representa o MAE de um modelo em relação ao MAE_b que é o MAE de outro modelo.

A análise gráfica na Figura 2 apresenta os *boxplots* gerados pelas 500 interações. Com a métrica MAE destaca-se nos gráficos (a) e (b) que não houve diferença significativa na mediana dos erros entre MP1, MP2 e ML4 para as variáveis selecionadas com a Correlação de *Pearson* e *Stepwise*. Também foi identificada a presença de *outliers,* para os modelos utilizando a técnica *Stepwise* para a escolha das variáveis independentes para a construção do modelo. Além disso, o modelo MP2 proposto apresenta menor erro de previsão.

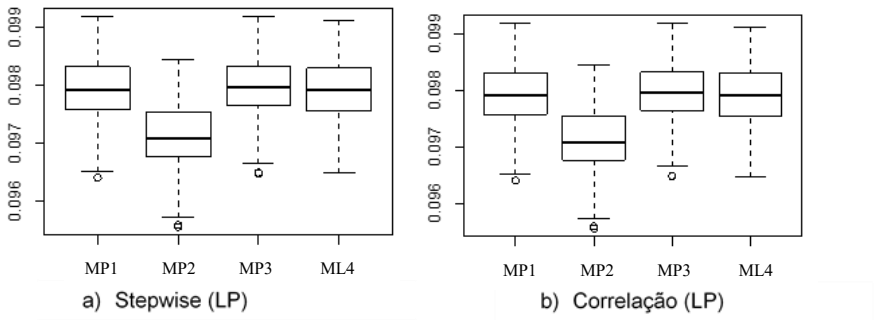

Figura 2. Boxplot dos modelos do conjunto (LP)

A Tabela 2 apresenta o RG do MP2 em relação aos outros modelos utilizados neste trabalho. Ela mostra que MP2 é mais eficiente que os demais modelos propostos (MP1 e MP3), bem como ao modelo da literatura (ML4), ratificando os valores médios obtidos na Figura 2. Portanto, destaca-se que o modelo proposto baseado em *Ensemble Bagging* (MP2) obteve resultado superior ao modelo da literatura baseado em *Ensemble Stacking* (ML4) para o problema do desempenho dos estudantes pernambucanos em LP.

Tabela 2. Resultados do RG

Técnica	MP2 x MP1	MP2 x MP3	MP2 x ML4
Stepwise	0.305%	5.79%	9.86%
Correlação	0.1%	5.69%	0%

A análise gráfica na Figura 3 apresenta os *boxplots* gerados pelas 500 interações. Com a métrica MAE observa-se nos gráficos (a) e (b) que não houve diferença significativa na mediana dos erros entre o MP1, MP3 e ML4 para as variáveis selecionadas com Correlação/*Stepwise*. Também foi identificada a presença de *outliers* para os modelos utilizando a técnica *Stepwise* para a escolha das variáveis independentes para a construção do modelo. Além disso, o modelo MP2 proposto apresenta menor mediana em relação ao erro de previsão dos demais modelos apresentados.

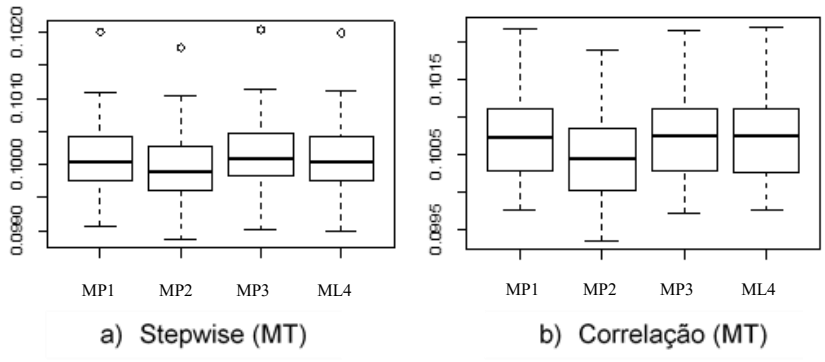

Figura 3. Boxplot dos modelos do conjunto (MT)

A Tabela 3, apresenta o RG do MP2 em relação aos outros modelos utilizados neste trabalho, comprovando que MP2 é mais eficiente que os demais modelos propostos (MP1 e MP3), bem como ao modelo da literatura (ML4). Isto ratifica os valores médios obtidos na Figura 3. Portanto, destaca-se que o modelo proposto baseado em *Ensemble Bagging* (MP2) obteve resultado superior ao modelo da literatura baseado em *Ensemble Stacking* (ML4) para o problema do desempenho da proficiência de MT dos estudantes pernambucanos. Pois, apresenta uma estimação mais robusta através de menores erros de predição para o modelo desenvolvido em relação ao modelo da literatura, tornando-se importante por obter um desempenho em relação a dados que agreguem o modelo final.

Tabela 3. Resultados do RG

Técnica	MP2 x MP1	MP2 x MP3	MP2 x ML4
Stepwise	0%	0%	8%
Correlação	0%	1%	1.3%

Realizou-se o teste de *Kolmogorov – Smirnov* para verificar se o vetor de erros (MAE: MT e LP) referentes às 500 interações seguia uma distribuição normal. Essa hipótese foi rejeitada. Desta forma foi utilizado o teste de *Wilcoxon* para realizar o teste de hipótese com 5% de significância. A hipótese alternativa elaborada é que o modelo MP2 apresenta melhor desempenho em relação ao MP1, MP3 e ML4. Assim foi comprovado estatisticamente, com um nível de confiança de 95% que o modelo MP2 para a proficiência de MT apresenta menores erros de predição em relação a todos os modelos utilizados para o cenário das variáveis selecionadas com os métodos *Stepwise/* Correlação. Utilizando a Correlação o modelo PM2 obteve erros menores que o MP1, MP3 e o ML4 já que os valores de *p-value* obtidos foram: 2,23 $x10^{-16}$; 2,2 x 10^{-16} e 1,11 x 10^{-15} , respectivamente.

2.5. Diagnóstico do Problema Educacional

Esta etapa apresenta o diagnóstico do problema educacional no contexto estudado. O processo de diagnóstico baseia-se no modelo teórico do desempenho escolar proposto por Soares (2004). O diagnóstico tem como objetivo identificar os fatores associados ao desempenho escolar a partir dos resultados dos modelos preditivos e estabelecer uma relação de causa e efeito entre os fatores associados (causas) com o problema educacional (efeito). Para isto foi proposto a aplicação do diagrama de *Ishikawa* (1993). Esta ferramenta busca simplificar processos considerados complexos dividindo-os em processos mais simples. Portanto, mais controláveis, sendo uma ferramenta bastante efetiva na busca das raízes do problema (Slack et al. 2009). O diagnóstico proposto neste trabalho apresentou a relação entre a causa e seus efeitos, bem como a proposição de soluções para minimizar a ocorrência do problema estudado. Contribuindo para intervenções a serem realizadas pelos agentes educacionais (governo, gestores, professores). A Figura 4, apresenta o diagrama de causa e efeito para o problema do desempenho educacional dos estudantes das escolas públicas de Pernambuco.

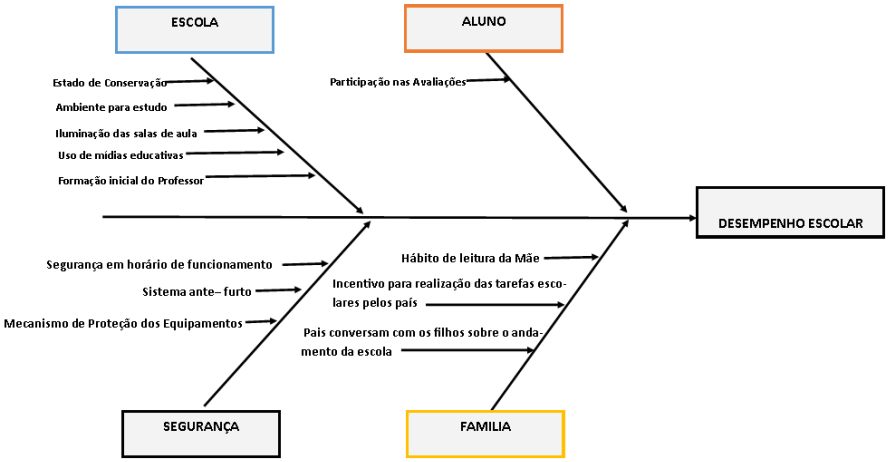

Figura 4. Diagrama de causa e efeito do desempenho escolar

O diagrama de causa e efeito apresentado na Figura 4, destaca as principais causas relacionadas ao desempenho escolar, identificadas neste trabalho, conforme as dimensões do modelo teórico proposto por Soares (2004). Contudo, devido a mudanças sociais e educacionais ocorridas nos últimos anos, novas dimensões para problemas educacionais surgiram e veem interferindo diretamente nas questões relacionadas ao ensino-aprendizagem, o comportamento e a vida social dos estudantes. Dentre essas novas dimensões, destacam-se aspectos relacionados à tecnologia, saúde, finanças e segurança. Observa-se que a exposição do problema pelo diagrama mostra as possíveis causas de forma estruturada, permitindo uma melhor visualização e compreensão do cenário que envolve o problema do desempenho escolar.

Os resultados mostram que as causas relacionadas à dimensão escola apresentam uma influência mais significativa no problema pelo quantitativo de fatores associados relacionados. Dentre as causas associadas, destacam-se os fatores: relacionados a infraestrutura da escola (iluminação das salas, ambientes para estudo); ao ensino aprendizagem (a existência e o uso de mídias educativas e de laser no cotidiano escolar); à formação inicial do professor (poucos professores com formação específica para as

disciplinas que lecionam). As demais causas também têm sua contribuição ao problema mencionado. Em relação ao aluno: o fato do mesmo não comparecer às avaliações escolares; já com relação a família: à formação educacional e cultural dos pais aparece como um fator preocupante, além do conhecimento dos pais sobre o andamento escolar dos filhos. Por fim, um fator associado que está em evidência na educação brasileira "a violência". Os dados deste estudo comprovam que fatores associados a violência contribuem com o baixo desempenho escolar dos estudantes. Destaca-se ainda neste estudo, a inexistência de segurança seja física (policiamento) ou eletrônica (monitoramento) nas imediações da escola durante o período de funcionamento.

Como sugestões de ações de intervenção: desenvolvimento de políticas públicas e adoção de ações por parte dos governos estaduais e municipais visando a melhoria da infraestrutura e da segurança pública e patrimonial de suas escolas. Além do desenvolvimento e implementação de estratégias para estreitar a relação dos pais dos alunos com a escola.

3. Discussões dos Resultados

Este trabalho utilizou modelos de *Ensemble Bagging* para a predição do desempenho escolar dos alunos das escolas púbicas (estaduais e municipais) do estado de Pernambuco. Além disso foram utilizados o método *Stepwise* e a Correlação de *Pearson* para a seleção das variáveis que serviram como base para a construção dos modelos preditivos. A vantagem de combinar regressores com o método *Ensemble Bagging,* em uma única previsão, é otimizar os resultados das estimativas dos modelos.

Os experimentos mostraram que o modelo MP2, o qual combina com *Ensemble Bagging* com Regressão Robusta, obteve os melhores resultados em relação ao modelo ML4 que utiliza o método *Ensemble Stacking*, como também em relação ao *Ensemble Bagging* com Regressão Linear (MP1) e Ensemble Bagging com regressão Ridge (MP3). As métricas com o Método *Stepwise* e a correlação alcançaram 95% na previsão do erro em ambos os cenários. Destaca-se ainda, a possibilidade de diagnóstico das causas do baixo desempenho escolar a partir da identificação dos fatores associados, e a sua relação de causa e efeito em relação ao problema do desempenho escolar.

4. Conclusão

Os resultados comprovam que os modelos que combinam *Ensemble Bagging* com métodos de Regressão, possuem menor erro absoluto médio, em relação ao modelo da literatura, ou seja, o *Bagging*, pode melhorar o desempenho de preditores instáveis com alta variância. De acordo com os resultados apresentados os modelos *Ensemble Regression*, trazem ganhos para a área educacional em direção a predição dos problemas educacionais com maior precisão e menor erro.

Neste estudo utiliza-se um conjunto de variáveis independentes, as quais foram selecionadas por cenários de aplicação. Estes cenários correspondem a predizer o valor da proficiência dos alunos em LP e MT utilizando dados do SAEB do ano de 2013.

Além disso, foi possível propor uma abordagem para diagnosticar o problema do desempenho educacional a partir das relações de causa e efeito entre os fatores associados e suas relações com o problema de forma sistemática. A utilização dessa abordagem visa minimizar a falta de compreensão dos resultados relacionados a predição no contexto

educacional servindo como um instrumento para que os agentes educacionais possam intervir nos problemas, adotando estratégias e ações pontuais.

5. Referências

ADEODATO, P, J, L.; FILHO, M, M, S.; RODRIGUES, R. L. **Predição de desempenho de escolas privadas usando o ENEM como indicador de qualidade escolar**. In: III Congresso Brasileiro de Informática na Educação (CBIE, 2014)**.** Anais do XXV Simpósio Brasileiro de Informática na Educação (SBIE 2014). 2014.

BAKER, R., S., J.; CARVALHO, M., J., D.; ISOTANI. **Mineração de Dados Educacionais Oportunidades para o Brasil.** Revista Brasileira de Informática na Educação, 2011.

BEZERRA, C.; SCHOLZ, R.; ADEODATO, P.; PONTES, T.; SILVA, I. **Evasão Escolar: Aplicando Mineração de Dados para identificar Variáveis Relevantes.** V Congresso Brasileiro de Informática na Educação (CIBIE, 2016). Anais do XXVII Simpósio Brasileiro de Informática na Educação (SBIE, 2016). 2016.

DAZZANI, M., V., M.; CUNHA, E., O.; LUTTGARDS, P., M.; ZUCOLOTO, P., C., S., V.; SANTOS, G., L. **Queixa escolar: Uma Revisão Crítica da Produção Científica Nacional.** Revista Quadrimestral da Associação Brasileira de Psicologia Escolar e Educacional (ABRAPEE). 421-428. ISSN 2175-3539, 2014.

D´ABREU, L., C., F.; MATURANO, E., M. **Associação entre comportamentos externalizantes e baixo desempenho escolar: Uma revisão de estudos prospectivos e longitudinais.** Revista Estudos de Psicologia (Natal). 43-51. ISSN 1678-4669, 2010.

INEP/MEC. **Sistema de Avaliação da Educação Básica (SAEB).** Acessado em 20 de maio de 2018. Disponível em: http:// www.inep.gov.br/SAEB.

ISHIKAWA, K. **Controle da Qualidade Total a Maneira Japonesa.** São Paulo: Editora Campus, 1993.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to Linear Regression Analysis**, volume 821. John Wiley & Sons, 2012.

NASCIMENTO, R.L.S.; **Combinação de Regressores no Contexto da Mineração de Dados: Uma aplicação *Stacking***. Dissertação (Dissertação em Engenharia da Computação), Universidade de Pernambuco, p.98. 2018.

PINTO, G., S.; JUNIOR, O., G., F.; COSTA, E., B. **Identificação dos Fatores de Melhorias do IDEB pelo uso de Mineração de Dados: Um Estudo de Caso em escolas municipais de Teotônio Vilela Alagoas.** Revista Novas Tecnologias em Educação, RENOTE, CINTED, UFRGS, 2019.

SOARES, J., F. **O efeito da escola no desempenho cognitivo de seus alunos.** Revista Eletrônica Iberoamericana sobre Calidad, Eficacia Y Cambio em Educacion, Madrid, v2, n.2, p 83-104, 2004.

SLACK, N.; CHAMBERS, S.; HARLAND, C.; HARRISON, A.; JOHNSTON, R. **Administração da Produção;** São Paulo: Editora Atlas, 2009.