Mineração de Dados Aplicada à Predição do Desempenho de Escolas e Técnicas de Interpretabilidade dos Modelos

Milton V. da Gama Neto[1], **Germano C. Vasconcelos**[1], **Cleber Zanchettin**[1]

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)
{mvgn,cz,gcv}@cin.ufpe.br

Abstract. *This paper presents a data mining analysis of schools in the state of Sao Paulo in the SARESP exam, using data from SEDUC-SP. The methodology*[1]*, based on CRISP-DM, proposes a machine learning solution not only to predict school performance but also to extract insights using explainable AI techniques. Experiments with 7 different classifiers achieved high accuracy (93%) and AUROC metrics (0.96) when predicting performance. Features analyzed included student, school and external socio-economic data. The non-linear model together with explanations given from SHAP and Counterfactual techniques highlighted relevant factors that can impact the educational results and evidenced the usefulness of the methodology in decision making.*

Resumo. *Este trabalho analisa o desempenho com mineração de dados das escolas de São Paulo no exame SARESP, com dados da SEDUC-SP. A metodologia*[1]*, baseada no CRISP-DM, propõe uma solução de aprendizagem de máquina para prever o desempenho das escolas e extrair padrões relevantes do desempenho educacional com técnicas de IA Explicativas. Sete classificadores alcançaram alta acurácia (93%) e AUC ROC (0.97) na previsão do desempenho das escolas, com dados do perfil de alunos, escolas e valores sócio-econômicos externos. O modelo não-linear e as técnicas SHAP e Counterfactual evidenciaram fatores relevantes que podem impactar o resultado educacional e a utilidade da metodologia no apoio à decisão.*

1. Introdução

A qualidade da educação está relacionada a diversos fatores, ainda que exista certa subjetividade para definir se um plano educacional está ou não alcançando seus objetivos. Desta forma, existem diversas avaliações e métricas para medir o desempenho de alunos e das instituições. Buscando aumentar o desempenho de seus alunos nestas avaliações, os gestores educacionais necessitam de conhecimento especializado em educação e evidências dos fatores que levam as instituições ao sucesso. Por meio da Mineração de Dados, na qual são empregadas técnicas analíticas, estatística e inteligência artificial, é possível encontrar padrões escondidos nos dados [Han et al. 2011]. Segundo [Baker et al. 2011], o uso desse tipo de técnica no contexto educacional é definido como Mineração de Dados Educacionais (EDM), campo este que busca descobrir padrões ou evidências a respeito dos alunos, professores e dos processos de aprendizagem e gestão.

Os avanços dos modelos de aprendizagem de máquina (AM) nas últimas décadas conduziram a elevadas taxas de acurácia, entretanto os modelos ficaram mais complexos e menos interpretáveis. Para alguns domínios, como saúde e educação, por exemplo, o entendimento e interpretação dos sistemas inteligentes é algo fundamental. Em

[1]Disponível em https://github.com/miltongneto/Explainable-Educational-Data-Mining

DOI: 10.5753/sbie.2021.217421

[Qin et al. 2020], os autores apresentam um estudo realizado na China sobre a confiança dos usuários em sistemas educacionais baseado em inteligência artificial, concluindo que a interpretabilidade dos sistemas é um dos principais fatores.

Com intuito de tornar os resultados mais compreensíveis e prover explicações das decisões tomadas pelos modelos de AM, surgiu a área *Explainable Artificial Intelligence*, comumente denominada de *Explainable AI* ou XAI (em português, IA Explicativas) [Gunning 2017]. Uma área fundamental para aplicações reais, em que as técnicas de XAI impulsionam responsabilidade e ética das soluções de inteligência artificial.

Este trabalho emprega a metodologia CRISP-DM para construir uma solução de Mineração de Dados Educacionais para predizer o desempenho escolar e interpretar dos modelos inteligentes com técnicas de XAI. A solução foi construída a partir da base de dados pública fornecida pela Secretaria de Educação do Estado de São Paulo (SEDUC-SP) com registros das notas e detalhes das escolas. O desempenho foi baseado no SARESP (Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo) que é realizado anualmente com provas para disciplinas de Língua Portuguesa e Matemática.

As principais contribuições deste trabalho correspondem à adoção de modelos não-lineares combinado de técnicas para prover explicações (XAI), conduzindo à descoberta de conhecimento e uma avaliação crítica, que ocorrem tanto de forma geral como específica para determinada escola, sendo a primeira vez no contexto escolar brasileiro. A análise realizada gerou artefatos que podem auxiliar na tomada de decisão pelos gestores educacionais. A partir de nosso levantamento na literatura, este é o primeiro trabalho de Aprendizagem de Máquina e *Analytics* usando a base de dados fornecida pela SEDUC-SP. Por fim, o método proposto pode ser replicado em outras bases de dados educacionais para construção de novos modelos e para a avaliação do impacto destes atributos.

2. Trabalhos Relacionados

[da Silva Pinto et al. 2019] descreveram um modelo para classificar o desempenho dos alunos do 9^o ano do ensino fundamental da cidade de Maceió no IDEB nas disciplinas Matemática e Língua Portuguesa. Os autores também realizaram uma seleção dos atributos mais importantes, através das técnicas de filtro, embaralhamento e embutida. Apesar de avaliarem os principais atributos e sugerirem algumas hipóteses das possíveis razões para o desempenho dos alunos, as técnicas utilizadas servem apenas para indicar a importância do atributo na modelagem, sem sinalizar a forma que influenciou a predição.

[Lacruz et al. 2019] apresentaram uma análise do desempenho na Prova Brasil (2013) de escolas do estado de Espírito Santo dos anos finais do ensino fundamental, com objetivo de verificar os indicadores de qualidade do ensino. As escolas foram separadas entre piores e melhores através do primeiro e terceiro quartis das notas, respectivamente. A base de análise contou com 124 escolas. Os autores empregaram a técnica linear de análise discriminante para identificar as variáveis mais relevantes para explicar as diferenças entre os grupos e o que existe de comum nos elementos do mesmo grupo.

[Calixto et al. 2017] propuseram um estudo para identificar fatores que influenciam na evasão escolar, para isso, seguiram a metodologia CRISP-DM. A base de dados utilizada foi do censo educacional de 2014, 2015 e 2016 dos estados do Ceará e Sergipe. A modelagem foi realizada utilizando a técnica de Regressão Logística, a qual forneceu os coeficientes como forma de avaliar os fatores importantes. Por fim, foram construídas

árvores de decisão curtas para gerar regras de indução por meio do algoritmo o *RIPPER*.

[Annegues et al. 2020] analisaram a relação entre a quantidade de alunos em uma sala de aula e o desempenho acadêmico de diversos centros da Universidade Federal da Paraíba. O trabalho apontou um efeito negativo do aumento da quantidade de alunos na turma no desempenho dos alunos. Os autores construíram um modelo linear com regressão quantílica e analisaram seus coeficientes para chegar à essa evidência.

Através de regressão linear simples, [Silva et al. 2018] modelaram o desempenho dos estados brasileiros e suas capitais medido pelo Índice de Oportunidades da Educação Brasileira (IOEB) de 2014, em função de fatores relacionados aos gastos públicos e pelo quantitativo de alunos e docentes das instituições. Os resultados mostraram que uma quantidade alta de alunos em relação ao número de docentes pode diminuir o desempenho dos alunos, enquanto que aumentar a quantidade de docentes nas escolas e os gastos públicos per capita em relação à população podem melhorar o desempenho.

3. Método Proposto

A construção da solução proposta segue a metodologia CRISP-DM (do inglês *Cross Industry Standard Process for Data Mining*) [Chapman et al. 2000]. Este é um processo bastante utilizado para abordar problemas de Ciência de Dados, sendo formado por 6 etapas: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação. A etapa final, que corresponde a implantação da solução, não é apresentada neste trabalho. O método proposto possui a seguinte estrutura: (1) Análise da literatura e do problema; (2) Compreensão dos dados, informando as fontes de dados e suas características; (3) Preparação dos dados, com a realização do processamento dos dados e engenharia de características para construir a base de dados final com os atributos a serem utilizados como insumo do modelo preditivo; (4) Modelagem, fase que são aplicados modelos de Aprendizagem de Máquina para predizer o desempenho escolar (baseado no SARESP); e (5) Avaliação, empregando metodologias de interpretabilidade do modelo para extração de informações sobre a relevância dos atributos e o que induziu a predição do desempenho escolar.

3.1. Entendimento dos dados

Os dados foram coletados no Portal de Dados Abertos da Secretaria da Educação do Estado de São Paulo [2], o qual possui informações da rede estadual de ensino. Das informações disponíveis no portal, fizeram parte deste estudo os dados do SARESP de 2018, com informações das classes e turmas, endereço das escolas, histórico de mudança dos últimos 5 anos na gestão (diretor, vice-diretor e coordenador), servidores ativos na rede de ensino, carga horária e dados sobre a formação dos professores. Além destes, foram inseridos dados externos, como informação dos municípios [3], um compilado de informações públicas dos municípios brasileiros, e os resultados da Avaliação Nacional da Alfabetização (ANA), realizado em 2016.

3.2. Preparação dos dados

Uma das principais etapas de projetos de Ciência de Dados, a engenharia de características ou engenharia de recursos, consiste em construir atributos por meio do conhecimento

[2]Disponíveis em: https://dados.educacao.sp.gov.br/
[3]https://www.kaggle.com/crisparada/brazilian-cities

no domínio para serem utilizadas no modelo de aprendizagem de máquina. Com um processo robusto e um bom conjunto de características, isto é, atributos que representem informações relevantes para o problema em questão, é possível treinar modelos preditivos que apresentem bom desempenho.

Algumas das fontes estavam em um nível de informação mais detalhado que a nível da escola, ou seja, em uma granularidade menor, como é o caso dos dados de servidores ativos e turmas, por exemplo. Nestes casos, foi necessário transformar os dados para o grão ideal através do processo de agregação com a construção de características, foi utilizado a frequência, média e valor máximo. Além disso, outras técnicas de pré-processamento foram aplicadas para aprimorar a qualidade da representação dos dados [Han et al. 2011]. A Tabela 1 apresenta os atributos que foram construídos.

Tabela 1. Descrição dos atributos avaliados

Nome	Descrição
MUNICIPIO CAPITAL	Indica se é capital ou não (variável binária)
MUNICIPIO AREA	Área do município
MUNICIPIO POPULACAO	População do município
MUNICIPIO AREA RURAL	Indica se é área rural ou não (variável binária)
MUNICIPIO VALOR ACRESCENTADO BRUTO	Produto Acrescido Público
MUNICIPIO PIB PER CAPITA	Produto Interno Bruto (PIB) per Capita
DEPENDENCIAS	Quantidade de dependências. Uma coluna para cada informação: salas de aula, sala dos professoreres, laboratório de ciência e laboratório de informática.
FORMACAO	Distribuição da formação dos professores. Um coluna para cada formação com seu percentual (sem informação, ensino médio, bacharelado, licenciatura, especialização, mestrado, doutorado)
QTD SERVIDORES	Quantidade de servidores
QTD PROFESSORES	Quantidade de professores
MEDIA FORMACOES	Valor médio da formação dos professores (conversão das categorias para números, quanto maior a formação maior o valor)
QTD FORMACAO CONTINUADA	Quantidade de professores com pós-graduação
QTD CARGOS DISTINTOS	Quantidade de cargos distintos
QTD TOTAL ALUNOS	Quantidade total de alunos
QTD CLASSES	Quantidade de classes
MEDIA ALUNOS SALA	Média de alunos nas salas
QTD CLASSES TIPO ENSINO	Quantidade de classes de acordo com o tipo de ensino. Uma coluna para cada modalidade (fundamental e médio).
QTD ALUNOS TIPO ENSINO	Quantidade de alunos de acordo com o tipo de ensino. Uma coluna para cada modalidade (fundamental e médio).
DIRETORES QTD 2018	Quantidade de diretores em 2018
COORDENADORES QTD 2018	Quantidade de coordenadores em 2018
DIRETORES QTD 5 ANOS	Quantidade de diretores nos últimos 5 anos
COORDENADORES QTD 5 ANOS	Quantidade de coordenadores nos últimos 5 anos
DIRETOR IDADE	Idade do(a) diretor(a)
DIRETOR CARGO CLAS EXER IGUAIS	Indica se o diretor tem o cargo de contrato igual ao de exercício (informação binária)
DIRETOR ANOS TRAB CARGO C	Anos de trabalho do diretor no cargo de contrato

DIRETOR ANOS TRAB CARGO E	Anos de trabalho do diretor no cargo de exercício
JORNADA QTD DISCIPLINAS MEDIA	Quantidade média de disciplinas dos professores
JORNADA QTD DISCIPLINAS MAX	Quantidade máxima de disciplinas dos professores
JORNADA QTD TOTAL AULAS MEDIA	Quantidade média de aulas dos professores
JORNADA QTD TOTAL AULAS MAX	Quantidade máxima de aulas dos professores
SERVIDORES IDADE MEDIA	Idade média dos servidores
SERVIDORES TEMPO CARGO C MEDIA	Tempo médio de contrato dos servidores
SERVIDORES CAT FUNCIONAL	Distribuição dos servidores de acordo com a categoria funcioanl. Uma coluna com percentual para cada tipo (A, F e O)
ANA NOTAS	Notas no exame ANA em Matemática, Escrita e Leitura (uma coluna para cada)
RELACAO ALUNO POR SERVIDOR	Relação entre o número de alunos para o de servidores
RELACAO ALUNO POR PROFESSOR	Relação entre o número de alunos para o de professores

3.3. Modelagem

A partir das notas obtidas no SARESP por cada escola e o conjunto de características elaborado na etapa anterior, foi construído um modelo de aprendizagem de máquina para predizer o desempenho da escola. O modelo foi treinado de forma supervisionada. Para isto, o problema foi modelado em um processo de classificação binária, na forma clássica de representação das classes como "Bom" e "Ruim". O processo para criação da variável alvo foi definido nos seguintes passos:

1. De acordo com a série e a disciplina verificou-se em qual faixa a nota obtida se encontrava, discretizando a nota de acordo com os níveis de proficiência do SARESP: abaixo do básico, básico, adequado e avançado.
2. Com os valores discretizados, converteu-se as categorias "abaixo do básico" e "básico" em Ruim, e "adequado" e "avançado" em Bom.
3. Para cada escola, realizou-se a contagem dos indicadores Bons e Ruins obtidos. Considerou-se uma escola como Boa (com bom desempenho) se ela obteve mais de 50% das notas boas. Caso contrário, a escola foi rotulada como Ruim.

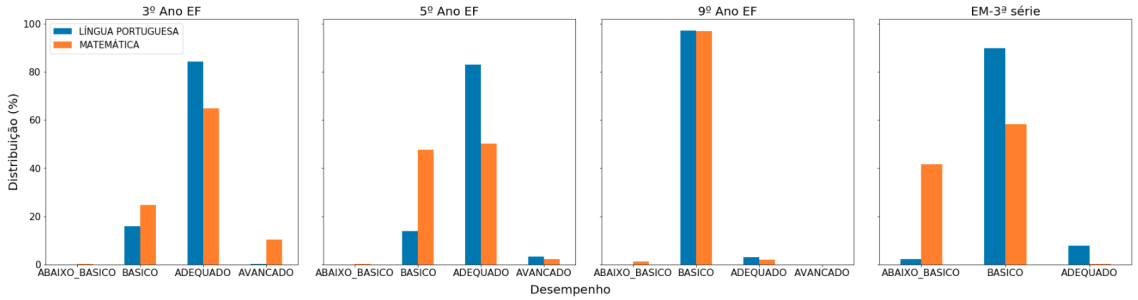

Figura 1. Distribuição das notas por níveis de proficiência do SARESP

A primeira discretização foi importante para estabelecer uma comparação justa entre as escolas, dado que os limiares numéricos são diferentes de acordo com a série e a disciplina consideradas. Assim, essa categorização segue um indicativo de qualificação adotado pelo SARESP e permite calcular de forma mais consistente o desempenho por escola. A Figura 1 apresenta a distribuição do desempenho por série e disciplina após essa transformação, onde fica evidente o forte desbalanceamento entre as classes. Após transformar o resultado das avaliações para forma binária (Bom ou Ruim), os dados são processados para o grão escola, resultando em uma distribuição de 84% das escolas classificadas como Boas e 16% das escolas classificadas como Ruins.

3.4. Avaliação

A avaliação consiste em duas etapas, a primeira para medir o desempenho do modelo na tarefa de classificação de desempenho escolar, empregando as métricas de avaliação. Enquanto a segunda está voltada para análise e interpretabilidade do modelo, através de técnicas de *Explainable AI*, para entender o que os conduziu a predizer determinado valor de saída, extrair conhecimento na identificação dos fatores de influência e guiar o processo com responsabilidade para evitar decisões precipitadas.

Para obter explicações serão aplicadas técnicas de XAI no escopo global e local, que dizem respeito a explicação do comportamento do modelo como um todo e para instâncias individuais, respectivamente [Gunning 2017]. A técnica SHAP (*SHapley Additive exPlanation*) [Lundberg and Lee 2017] é baseada na teoria dos jogos e fornece detalhes de como cada atributo impacta na predição do modelo, para este contexto, indicará as características que influenciaram o desempenho de determinada escola. Ainda no escopo local, será aplicado a técnica Counterfactual [Van Looveren and Klaise 2019], que modifica as características de uma instância e realiza uma otimização para encontrar a menor mudança para alcançar uma predição de outra classe, que neste caso, será as mudanças mínimas para que uma escola com desempenho ruim alcançasse um desempenho bom. A técnica SHAP também será utilizada no escopo global, com a combinação das contribuições individuais é inferido como os atributos estão influenciando de forma geral o modelo, indicando quais são os atributos mais relevantes e como se comportam.

As técnicas aplicadas para obter explicações são agnósticas ao modelo, ou seja, funcionam independentemente do tipo do modelo, podem ser aplicada tanto em uma árvore de decisão quanto em uma rede neural, por exemplo. Estas técnicas fazem parte do estado da arte e são capazes de capturar mecanismos para decisões de modelos não lineares complexos. Sendo assim, possibilitam a utilização de modelos mais acurados e fornecem interpretabilidade. Diferentemente dos trabalhos relacionados que foram apresentados que, em geral, utilizaram modelos lineares para obter entendimento dos fatores que influenciaram a decisão do modelo, ou em comparação com [Calixto et al. 2017] que aprende as regras diretamente através de modelos mais simples como a árvore de decisão e indução regras. Além disso, o método proposto conta com explicações locais para fornecer insumos para análises detalhadas de cada escola.

4. Experimentos

O conjunto de dados final é composto por 4.523 instâncias, em que cada uma representa uma escola diferente, e 58 atributos descritivos que foram listados na Tabela 1. Os dados foram separados em 80% para o conjunto de treinamento e 20% para o conjunto de teste. O conjunto de teste, com 904 registros, é completamente independente e o tamanho possibilita uma análise consistente de desempenho e para a fase de interpretabilidade, evitando uma análise enviesada. Para verificar a performance dos modelos de forma robusta, o conjunto de treinamento foi empregado usando a técnica de validação cruzada, com o método *k-fold* (com $k = 10$). Os resultados obtidos nas iterações são analisados e servem de insumo para determinar qual o melhor modelo para a base de dados utilizada.

Os experimentos foram realizados na linguagem de programação Python com a biblioteca *scikit-learn* [Pedregosa et al. 2011] que implementa diversos algoritmos clássicos de Aprendizagem de Máquina e a biblioteca *LightGBM* [Ke et al. 2017] com

a implementação do modelo em questão. A Tabela 2 apresenta os resultados obtidos na validação cruzada para as métricas de Acurácia, AUC ROC (do inglês, *Area Under the Receiver Operating Characteristic Curve*), e o valor macro das métricas de Precisão, *Recall* e *F1-score*, que computam a média do resultado obtido para cada classe. Foram avaliados os modelos *Decision Tree*, *Random Forest*, *Gradient Boosting*, *LightGBM*, *k-NN*, *MLP* e *Linear SVM*, todos utilizando os hiperparâmetros padrões. Na análise dos resultados observou-se o altíssimo desempenho obtido, alcançando $0,96$ na AUC ROC (quando comparado ao máximo de 1 para a métrica), indicando uma alta capacidade do modelo de discriminar as classes Bom e Ruim. O bom desempenho foi alcançado por alguns classificadores, o que serve como indicador de qualidade dos atributos utilizados, tendo boa relação com o atributo alvo. Entretanto, modelos como *Decision Tree* e *k-NN* apresentaram desempenhos inferiores comparado aos outros modelos mais complexos.

Tabela 2. Resultados com diferentes classificadores

Classificador	AUC ROC	Acurácia	Precisão	*Recall*	*F1-score*	*Tempo*
Decision Tree	$0,80$	$89,0\%$	$0,79$	$0,80$	$0,79$	**1s**
Random Forest	**0,96**	$92,3\%$	**0,85**	$0,87$	**0,86**	$5s$
Gradient Boosting	$0,95$	$91,9\%$	$0,84$	$0,86$	$0,85$	$17s$
k-NN	$0,91$	$90,5\%$	$0,83$	$0,81$	$0,82$	**1s**
LightGBM	**0,96**	$91,5\%$	$0,84$	$0,85$	$0,84$	**1s**
MLP	**0,96**	$91,5\%$	$0,84$	$0,85$	$0,84$	$21s$
Linear SVM	**0,96**	**92,4%**	**0,85**	**0,88**	**0,86**	$7s$

Após executar os experimentos foi realizado o teste estatístico não paramétrico Friedman para comparar os resultados dos diferentes classificadores. Com nível de significância $0,05$, a hipótese nula que as amostras são iguais foi rejeitada, pois o *p-value* obtido foi aproximadamente 0, sendo forte evidência de que as distribuições são diferentes. Com exceção dos classificadores *Decision Tree* e *k-NN*, os resultados foram bem próximos. O resultado do teste Kruskal-Wallis para estes classificadores foi um *p-value* de $0,15$, que falha em rejeitar a hipótese nula, então assumimos que são da mesma distribuição. O modelo selecionado foi o *LightGBM* devido ao tempo de execução mais baixo em relação aos outros de resultados similares. Esta escolha também facilitará a etapa de interpretação, ainda que as abordagens sejam agnósticas ao modelo. O algoritmo SHAP será mais rápido por ser baseado em árvores, segundo os autores do algoritmo, e o fato de fornecer as probabilidades das predições melhor que o algoritmo *Random Forest*, que apenas computa uma média das árvores, torna mais fácil para o Counterfactual. O resultado no conjunto de teste foi $93,3\%$ de acurácia e AUC ROC de $0,97$.

5. Resultados e Discussão

5.1. Interpretação Global

A técnica SHAP fornece a importância dos atributos que o modelo aprendeu. Os resultados são apresentados com magnitude, sentido e distribuição dessa importância ao longo das variações da entrada. A Figura 2 apresenta os 15 atributos mais importantes baseado nesta técnica. Os dois atributos mais importantes estão associados a turmas do ensino médio, através das informações da quantidade de alunos e de classes. Considerando que

nem toda escola na base de dados possui ensino médio, o resultado indica que as escolas com essa modalidade de ensino tem grande probabilidade de ter desempenho ruim. Essa informação é extraída do gráfico, onde tons vermelhos e azuis representam, respectivamente, valores altos e baixos nos atributos. No eixo x, valores positivos impactam na predição do modelo para a classe positiva (desempenho bom), e os negativos, o inverso. Nota-se que escolas com número alto de alunos no ensino médio possuem baixo desempenho.

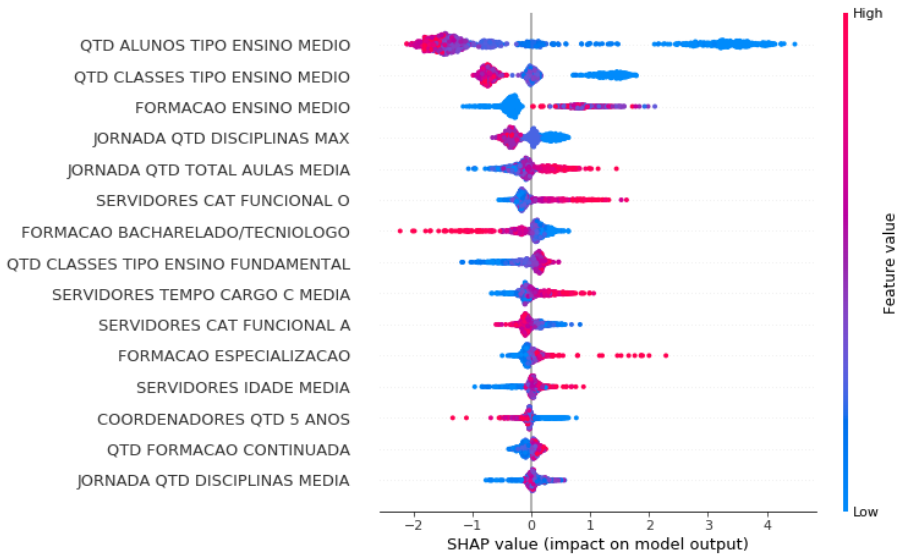

Figura 2. Resultado do SHAP global para os 15 atributos mais relevantes

Retomando a Figura 1, na Seção 3.3, é notório que o ensino médio não teve um destaque positivo em sua na maioria, pois as notas estão concentradas nas classes "abaixo do básico" e, na maior parte, em "básico". A transformação do alvo para binário considera ambos valores pertencentes a classe "Ruim". O modelo consegue perceber esse comportamento e por isso atribui um peso alto a atributos associadas ao tipo de ensino.

5.2. Interpretação Local

A interpretação local é realizada para uma instância, dessa forma, serão selecionados dois casos para obter explicações, um com desempenho bom e outro ruim.

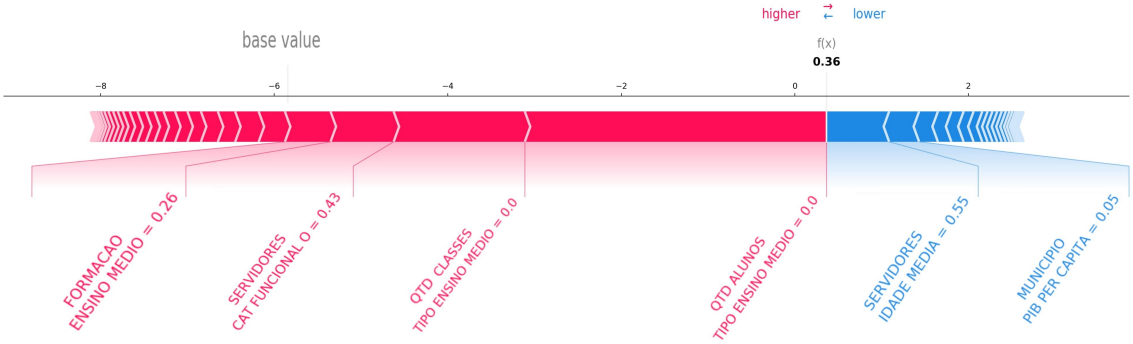

Figura 3. Resultado do SHAP local para importância dos atributos da instância

A Figura 3 apresenta as contribuições dos atributos calculados com a técnica SHAP. No gráfico, é possível notar que o resultado final ($f(x)$) foi positivo, o que indica

que a escola em análise teve um bom desempenho. As barras mostram a contribuição, em que o sentido indica se foi positivo ou negativo e a magnitude da barra indica o impacto. Na parte inferior estão os valores dos atributos de entrada. O gráfico só exibe as maiores contribuições, porém a técnica fornece o valor exato de cada atributo. No exemplo em questão, é possível notar que não ter turmas do ensino médio influenciou o modelo para classe positiva, junto com o percentual de profissionais com o tipo do contrato temporário e a formação de ensino superior. Enquanto os principais fatores negativos foram a idade média dos servidores o Produto Interno Bruto do município.

Com a aplicação do algoritmo Counterfactual em uma instância do conjunto de teste que obteve probabilidade de $0,76$ de apresentar um desempenho ruim, em que o modelo acertou, foi gerada uma nova instância artificial com probabilidade $0,64$ de alcançar um bom desempenho. Esse exemplo construído pelo algoritmo realizou pequenas modificações em apenas duas características, aumentando a quantidade média de formações dos profissionais e a quantidade de profissionais que tem especialização.

5.3. Principais descobertas

As técnicas de XAI possibilitam ter um entendimento melhor do modelo. Neste caso, foi possível identificar que o desempenho ruim do ensino médio influenciou bastante no resultado das escolas. Dessa forma, uma predição do desempenho escolar acaba sofrendo grande influência deste fator, sendo um forte viés para o desempenho negativo.

Além do tipo de ensino, vale destacar outros fatores importantes para o modelo, como a influência positiva no desempenho com aumento dos valores da especialização dos servidores, tempo exercendo o cargo, valor médio da quantidade de aulas dos professores na escola (associada a maior dedicação do professor na mesma), PIB do município. Enquanto algumas características impactam de forma negativa a medida que seus valores crescem, por exemplo, o número máximo de disciplinas que um professor ministra em uma escola, a média de alunos por classe e mudanças de coordenação.

6. Considerações finais

Foi apresentada uma solução para predição do desempenho das escolas no SARESP. O modelo *LightGBM* atingiu $93,3\%$ de acurácia e uma AUC ROC de $0,97$ no conjunto de teste. Este é o primeiro trabalho de Aprendizagem de Máquina com dados da Secretaria de Educação do Estado de São Paulo. A metodologia proposta é capaz de integrar dados heterogêneos para criar um conjunto de características das escolas. Após o ótimo desempenho obtido pelo modelo, foi realizada a interpretabilidade do modelo, aplicando a técnica SHAP que permite identificar o impacto dos atributos de modelos não-lineares de maneira global e local. Também foi empregado a técnica Counterfactual que indica as modificações mínimas para que uma escola com desempenho ruim alcance um desempenho bom. Esta é primeira vez que as técnicas de *Explainable AI* agnósticas ao modelo são utilizadas em dados educacionais brasileiros. A interpretabilidade do modelo junto com um aprofundamento na análise dos dados permitiu identificar um conjunto de características que impactam no desempenho de forma positiva e negativa. As descobertas obtidas podem servir de apoio à tomada de decisão de gestores educacionais.

Para trabalhos futuros, pretende-se adicionar novas características escolares, incluindo fontes distintas, dado que a metodologia é escalável e novos atributos podem

melhorar o modelo e apresentarem impacto relevante. Além disso, explorar modelos para cada tipo de ensino separadamente, verificando a maneira que as características impactam no desempenho dos alunos.

Referências

Annegues, A. C., Porto Júnior, S., and Figueiredo, E. (2020). Tamanho da turma e desempenho acadêmico dos universitários: evidência para a UFPB. *Estudos Econômicos (São Paulo)*, 50(1):99–124.

Baker, R., Isotani, S., and Carvalho, A. (2011). Mineraçao de dados educacionais: Oportunidades para o Brasil. *Revista Brasileira de Informática na Educação*, 19(02):03.

Calixto, K., Segundo, C., and de Gusmão, R. P. (2017). Mineração de dados aplicada a educação: um estudo comparativo acerca das características que influenciam a evasão escolar. In *Simp. Brasileiro de Inf. na Educação-SBIE*, volume 28, page 1447.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). Crisp-dm 1.0 step-by-step data mining guide.

da Silva Pinto, G., Júnior, O. F., Costa, E., Barbirato, J. C. C., and Rodrigues, W. R. M. (2019). Identificação dos fatores de melhorias no IDEB pelo uso de mineração de dados: Um estudo de caso em escolas municipais de maceió. In *Simpósio Brasileiro de Informática na Educação-SBIE*, volume 30, page 1828.

Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2(2).

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pages 3146–3154.

Lacruz, A. J., Américo, B. L., and Carniel, F. (2019). Indicadores de qualidade na educação: análise discriminante dos desempenhos na prova brasil. *Revista Brasileira de Educação*, 24.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Qin, F., Li, K., and Yan, J. (2020). Understanding user trust in artificial intelligence-based educational systems: Evidence from china. *British Journal of Educational Technology*, 51(5):1693–1710.

Silva, M. C. d., Souza, F., Tavares, A., and Silva, J. D. (2018). Índice de oportunidades da educação brasileira: Variáveis explicativas de rendimento dos alunos das capitais estaduais e dos estados brasileiros. *Revista Científica Hermes*, 20:20.

Van Looveren, A. and Klaise, J. (2019). Interpretable counterfactual explanations guided by prototypes. *arXiv preprint arXiv:1907.02584*.