Machine Learning e a influência das características infraestruturais e pedagógicas no desempenho escolar para o programa Milênio do MPBA

Leandro Soriano Ferreira
Salvador-BA, Brasil
leandro.ferreira@aln.senaicimatec.edu.br

Gessé Pinto da Silva
Salvador-BA, Brasil
gesse.silva@aln.senaicimatec.edu.br

Davi Lourenço Oliveira dos Santos
Salvador-BA, Brasil
davi.l.santos@aln.senaicimatec.edu.br

Éldman de Oliveira Nunes
Salvador-BA, Brasil
eldman.nunes@fieb.org.br

***Resumo*—O acesso a educação de qualidade é de suma importância para a população, e diversos aspectos podem influenciar no desempenho escolar dos alunos. Através do censo escolar é possível conhecer os recursos disponibilizados pelas unidades de ensino. Na etapa final da educação básica, através de um sistema de avaliação, como o ENEM, é possível, para além de outros objetivos, medir a qualidade do ensino das escolas a partir do desempenho dos seus respectivos alunos. Dentro deste cenário, conhecer quais as características infraestruturais e pedagógicas mais influenciam no desempenho escolar se apresenta como uma ferramenta importante para atuar de forma mais assertiva nos pontos que podem melhorar a qualidade do ensino. Sendo assim, este trabalho tem como objetivo selecionar as características infraestruturais e pedagógicas mais relevantes para o desempenho dos alunos nas unidades de ensino públicas e privadas do país. Para isto, foram utilizados os microdados do censo escolar, bem como os micro dados do ENEM por escola, com o recorte temporal de 2009 a 2015. Para realizar a seleção de características, foram criados dois grupos de atributos relacionados a infraestrutura e pedagogia, na sequência, foram utilizados métodos de seleção baseados em filtros, *wrapper* e *embbeded*. As escolas foram classificadas em conceitos que vão de A a D. Das 389 colunas presentes no *dataset* foram selecionadas 44 e seus respectivos graus de influência. O modelo de *machine learning* resultante da seleção de atributos alcançou 98.9% do desempenho do modelo sem seleção de atributos.**

***Palavras Chaves*—Machine Learning, educação, infraestrutura, pedagógico, desempenho, ENEM**

I. Introdução

Notadamente, a educação possui impacto em todas as áreas da vida do cidadão e o acesso a ela permite que o indivíduo forme sua opinião, aprenda a proteger os interesses sociais e exerça seus direitos e deveres [1]. No Brasil, a partir de mudanças constitucionais (EC 14/1996, §VII, art. 206), iniciativas públicas deram forma a um robusto e eficiente sistema de avaliação em todos os níveis e modalidades de ensino, consolidando uma efetiva polıtica de avaliação educacional [2].

Considerada hoje uma das mais abrangentes e eficientes do mundo, a política de avaliação brasileira engloba diferentes programas que, em conjunto, configuram um macrossistema de avaliação da qualidade da educação brasileira [2]. Nesse ínterim, o ENEM figura como uma valiosa ferramenta de avaliação e medição de desempenho discente, como porta de entrada para o ensino superior [3], e como principal fonte para a formulação de políticas públicas eficazes que promovam a equidade e a eficiência educacional. Adicionalmente, foram instituídos um sistema de micro dados, com detalhes relacionados às escolas, e o Censo Escolar, um instrumento de coleta de dados e pesquisa estatística da educação básica do país [3], onde, juntos, ampliam significativamente o campo de estudo.

No âmbito do estado da Bahia, o programa Saúde + Educação: futuro para o Milênio, criado pelo MPBA em 2008, tem como objetivo fiscalizar os setores de educação e saúde, visando efetivar os direitos de cidadania e contribuindo para a melhoria da prestação de serviços públicos nessas áreas [4]. No que concerne à educação, o programa consiste em observar as condições estruturais, sanitárias e de prestação do serviço, além da correta aplicação de verbas públicas com equipes multi-institucionais *in loco*, aplicando formulários, identificando e documentando os problemas encontrados, através de fotos e vídeos, registrando também os aspectos positivos [4]. Após a fiscalização, os dados são analisados, onde é produzido um diagnóstico em caráter de devolutiva. Com esse documento, é feita uma notificação à unidade de ensino fiscalizada e, a partir dela, são aguardadas as devidas justificativas e ajustes por parte dos gestores e administradores. Os diagnósticos são produzidos em função de algumas características elencadas nos formulários e avaliadas pelos membros das equipes.

Tanto o Milênio quanto o Censo Escolar/Micro Dados do ENEM apontam semelhanças, porém, esses últimos, complementam o primeiro com resultados de desempenho no ENEM dos estabelecimentos de ensino. Assim, tornando-se factível associar temas das questões do formulário as variáveis do Censo, mas sem permitir avaliar quais temas/variáveis mais impactam nos resultados e assim possibilitar devolutivas mais assertivas. Essas associações, análises e demais ações são feitas de forma manual, por equipes, sem uso de recursos computacionais sofisticados, operacionalmente custosa e propícia a falhas. Em ambos os sistemas, os dados disponibilizados necessitam de análises mais elaboradas, envolvendo estatística inferencial

e/ou mineração de dados, que produzam informações relevantes [5] e que permitam estabelecer quais os critérios impactam no desempenho discente/escolar e embase decisões político-administrativas nas áreas que envolvem a educação básica do país.

Estudos recentes, como [6], destacam variáveis que avaliam um conjunto de informações sobre as características da escola, da infraestrutura, do corpo docente, da escolaridade dos pais, do tamanho das turmas e indicadores de fluxo para alunos do ensino fundamental em seu trabalho intitulado "Os determinantes do aprendizado com dados de um painel de escolas do SAEB".

Considerando o problema, os estudos mencionados e suas conclusões, surge uma lacuna no que tange a aspectos mais específicos das características infraestruturais e pedagógicas. Nesse cenário, o uso de técnicas de Ciência de Dados e aplicação de IA se fazem importantes. O presente artigo busca explorar as características infraestruturais e pedagógicas, elencando-as por seu grau de importância e impacto nos resultados de desempenho, com o uso técnicas de *feature selection*, para que, através do mapeamento dessas características, com temas tratados nas questões do formulário de fiscalização do Milênio, permita-lhe melhorá-lo, criando ou removendo eventuais perguntas, além de construir ao final um modelo de *machine learning* com capacidade de predizer o desempenho da escola com base nas variações dessas características.

Este artigo está organizado da seguinte forma: a seção II traz informações sobre os algoritmos de *machine learning* utilizados; a seção III aborda conceitos sobre os métodos de seleção de atributos; a seção IV define o método de consolidação, pré-processamento e redução da dimensionalidade dos dados; a seção V apresenta os experimentos e resultados obtidos; e por fim, a seção VI traz conclusões, incluindo contribuições, limitações e sugestões de pesquisas futuras.

II. Algoritmos de Machine Learning

Dentre as técnicas e métodos de *machine learning* utilizados na solução, estão os algoritmos de aprendizado supervisionado (*supervised learning*). Por aprendizado supervisionado, pode-se entender o processo de fornecer a um algoritmo um conjunto de dados (*dataset*) de entrada, associados a valores desejados para a saída, de modo que o algoritmo encontre um padrão capaz de produzir tais valores de saída a partir dos valores de entrada fornecidos [7], [8]. Os valores de entrada podem ser chamados como *variáveis preditoras* ou *independentes*, e os valores de saída, podem ser chamados como *variáveis alvo* ou *dependentes*. Após o aprendizado, o algoritmo é capaz de produzir um valor de saída a partir de dados de entrada que nunca lhe foram fornecidos, sem auxílio ou intervenção humana [7].

Dentre os algoritmos de aprendizado supervisionado, existem aqueles cujo objetivo é produzir um valor numérico (contínuo ou discreto) como saída, tal como a renda anual de uma pessoa, ou a previsão de vendas de uma empresa para o próximo ano. Estes algoritmos são tipificados como modelos de *regressão*. Também existem aqueles cujo objetivo é produzir um valor nominal, tal qual o gênero de um livro ou filme, ou ainda rotular um e-mail como *spam* ou não. Estes algoritmos são categorizados como modelos de *classificação* [7], [8].

O k-NN (k-*Nearest Neighbor* ou k-Vizinhos mais próximos) utiliza uma métrica de distância para encontrar as k instâncias mais próximas [7], [8]. Já modelos lineares de regressão (e suas derivações, como *Lasso*, *Ridge*, regressão logística, dentre outros) se baseiam no uso de uma função linear das variáveis preditoras [7], [8]. O algoritmo Naïve Bayes se baseia na teoria de probabiliade de Bayes [8], fornecendo a probabilidade de uma dada instância pertencer a uma determinada classe da variável alvo. Árvores de decisão e suas variantes (como *Random Forest* e *Gradient Boosting*, por exemplo) se baseiam na segmentação do espaço de busca da variável alvo [7]–[9], onde o resultado da predição é a média dos valores no nó folha, para problemas de regressão ou a mediana, para problemas de classificação. Máquinas de Vetores de Suporte (*Support Vector Machines*) são algoritmos que se baseiam em encontrar o melhor hiperplano que separe as instâncias das diferentes classes da variável alvo [8].

III. Seleção de Atributos

Seleção de atributos (ou *feature selection*), segundo [10], tem o objetivo de reduzir a dimensionalidade do *dataset*, chegando-se ao número de variáveis de entrada que acredita-se ser o mais útil para o modelo, com o objetivo de predizer a variável alvo. Alguns problemas de modelagem preditiva podem ter um número alto de variáveis de entrada, o que torna lento o desenvolvimento e o treinamento dos modelos, além de exigirem uma quantidade grande de memória do sistema [10]. Segundo [11], muitos pesquisadores concordam que não existe um método de seleção de atributos que seja chamado de "o melhor", sendo seus esforços focados em encontrar um bom método para um problema específico.

De acordo com [9], [11], os métodos de seleção de atributos podem ser divididos em três grupos: métodos de filtro, *wrapper* e *embbeded*. Os métodos de filtro selecionam os atributos baseados nas características intrínsecas dos dados, através de testes estatísticos, ignorando sua interação com o modelo de *machine learning*. Os métodos *embbeded* incorporam a seleção de atributos no treinamento do modelo preditivo, desde que o processo de otimização deste modelo tenha alguma forma de discernir a importância de cada um dos atributos. Já os métodos *wrapper* envolvem a seleção de atributos em torno de um modelo preditivo, gerando múltiplos subconjuntos de atributos e avaliando o seu desempenho baseado no modelo de classificação ou regressão.

Para os métodos de seleção de atributos baseados em filtro, [9], [11] destacam os testes estatísticos ANOVA, o qui-quadrado (χ^2), o método Mutual Information e correlações como Pearson, Spearman e Kendall, dentre outros. O teste ANOVA, acrônimo para *Analysis of Variance* (análise de variância), tem o seu uso recomendado quando as variáveis preditoras/independentes são numéricas e a variável alvo/dependente é categórica, ou vice versa, assim como a correlação de Kendall. O teste qui-quadrado é indicado quando ambas as variáveis preditoras e

alvo são categóricas. As correlações de Pearson e Spearman são apropriadas para quando ambas as variáveis preditoras e alvo são numéricas. Apesar de o método *Mutual Information* ser diretamente aplicável quando ambas as variáveis preditoras e alvo são categóricas [11], ele pode ser adaptado para todos os tipos de variável preditora e alvo [9], [11].

Dentre os métodos de seleção de atributos *embbeded*, algoritmos baseados em árvores de decisão como *Random Forest* e *Gradient Boosting* podem ser utilizados. Para estes algoritmos, a importância de cada atributo pode ser medida de acordo com a redução no critério utilizado para particionar o *dataset* em cada uma das ramificações (ou nós) da árvore, durante a sua construção. Para problemas de classificação, este critério pode ser, por exemplo, o Gini ou a entropia, enquanto que, para problemas de regressão, este critério pode ser a raiz do erro quadrático médio (RSME) [9], [11]. Algoritmos como Regressão Linear e Regressão Logística, dentre outros, são construídos com base na soma ponderada dos valores das variáveis de entrada e, para estes algoritmos, a importância de cada atributo pode ser obtida utilizando-se os coeficientes atribuídos a cada variável [9], [11].

Dentre os métodos de seleção de atributos do tipo *wrapper*, o método *Recursive Feature Elimination* (RFE) é um método que utiliza como núcleo um modelo preditivo capaz de determinar a importância dos atributos de um *dataset*. O RFE funciona buscando um subconjunto de atributos, partindo do conjunto completo dos atributos do *dataset* de treino. O modelo preditivo usado como núcleo é treinado/ajustado, os atributos do *dataset* de treino são ordenados de acordo com a sua importância, removendo-se os n menos importantes. O modelo preditivo núcleo é, então, reajustado, utilizando-se o *dataset* de treino com o novo conjunto reduzido de atributos [9], [11].

IV. Método

Nesta seção, descreve-se todo o procedimento realizado para desenvolver a solução proposta neste artigo, desde a obtenção dos dados até a avaliação dos modelos preditivos selecionados.

A. Arquitetura da Solução Proposta

A figura 1 ilustra a arquitetura para a solução proposta por este trabalho. As fases da proposta de solução passam pela obtenção e consolidação dos microdados do Censo Escolar e do ENEM por Escola, bem como a divisão deste *dataset* consolidado em treino e teste, o pré-processamento dos *datasets* de forma individualizada (a fim de evitar vazamento de dados).

Em um primeiro momento, realizou-se o treinamento dos modelos de classificação *Extra Trees Classifier, XGBoost, Random Forest, Naive Bayes, SVM - Suport Vector Machine* e *Ada Boost Classifier* e ajuste dos hipermarâmetros, utilizando o *dataset* de treino. Os modelos resultantes foram avaliados de acordo com a métrica de avaliação escolhida e o modelo mais eficiente foi selecionado.

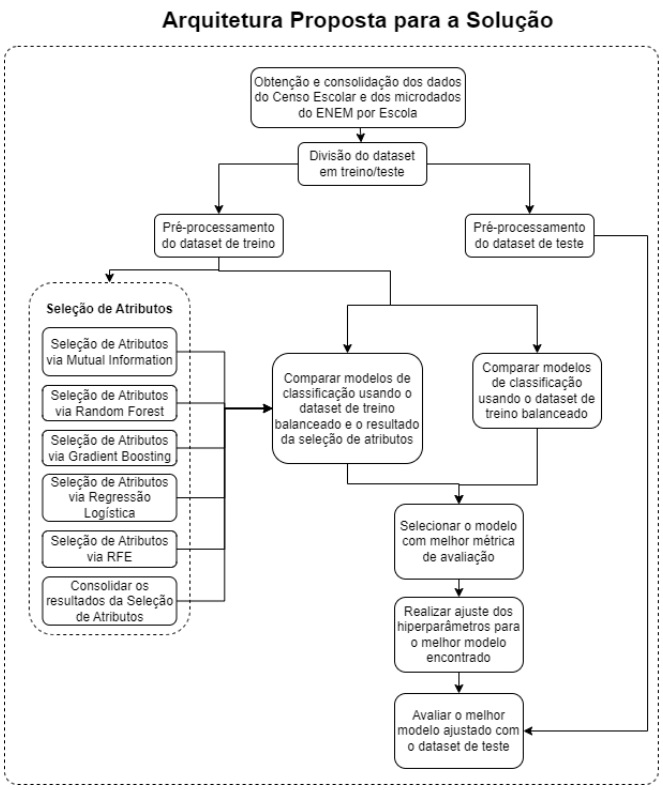

Figura 1. Arquitetura da Solução Proposta

Em um segundo momento, realizou-se a seleção de atributos para o *dataset* de treino, utilizando diferentes métodos. O resultado destes diferentes métodos foi consolidado numa lista única de atributos, a ser utilizada para reduzir a dimensionalidade do *dataset* de treino. Feito isso, repetiu-se o treinamento dos modelos de classificação e ajuste dos hipermarâmetros com este novo *dataset* de treino, selecionando-se o melhor modelo encontrado. Este segundo modelo resultante foi avaliado de acordo com a mesma métrica de avaliação.

B. Descrição dos dados

Os dados do Censo Escolar foram obtidos para os anos de 2007 a 2015 [12]. Para este *dataset*, o período anterior a 2007 foi descartado, devido a uma mudança de estrutura nos arquivos fornecidos, o que poderia dificultar a extração e análise dos dados. Sendo cada ano fornecido em um *dataset* individualizado, todos estes dados foram consolidados em um único *dataset*, totalizando 2.016.308 registros e 370 colunas. Para este *dataset*, as colunas `NU_ANO_CENSO` e `CO_ENTIDADE` foram elencadas como chave. Não foram encontradas linhas duplicadas, nem valores negativos para este dataset. O dicionário de dados menciona o preenchimento de valores extremos (indicados com o código/valor "88888") para o período a partir de 2019, valores estes não identificados no período de interesse (2007 a 2015). No entanto, foram identificadas 299 colunas com dados faltantes, a serem tratadas posteriormente.

Os microdados do ENEM por escola [13], relativos ao período de 2005 a 2015, também foram obtidos diretamente

da base de dados do INEP. Este *dataset* contém 172.305 registros e 27 colunas. Para este *dataset*, as colunas `NU_ANO` e `CO_ESCOLA_EDUCACENSO` foram elencadas como chave. Para o período posterior a 2015, o INEP realizou uma reestruturação dos microdados, com o objetivo de suprimir a possibilidade de identificação de pessoas, em atendimento às normas previstas na Lei Geral de Proteção de Dados (LGPD). Deste modo, após este período, não foi possível obter as médias das notas do ENEM por escola, cruciais para este estudo. Também não foram identificadas linhas duplicadas ou valores negativos para este *dataset*. No entanto, foram identificadas 15 colunas com dados faltantes, a serem tratadas em etapa posterior.

Os *datasets* de ambas as fontes foram consolidados em um único *dataset*, utilizando as colunas elencadas como chave. O *dataset* resultante apresenta 140.601 registros e 397 colunas. Ao remover colunas em duplicidade, contidas tanto no Censo Escolar quanto nos microdados do ENEM por Escola, o número de atributos do *dataset* resultante foi reduzido para 389.

C. Divisão em Treino e Teste

O *dataset* resultante da união do Censo Escolar e dos microdados do ENEM por Escola foi submetido a uma divisão onde 95% dos dados foram reservados para o treino/validação (totalizando 133.570 registros), e os outros 5%, para o teste dos modelos de classificação (totalizando 7.031 registros), conforme mostrado na figura 2. A divisão treino/teste e treino/validação foi realizada de forma extratificada a partir da região geográfica, estado e mesorregião.

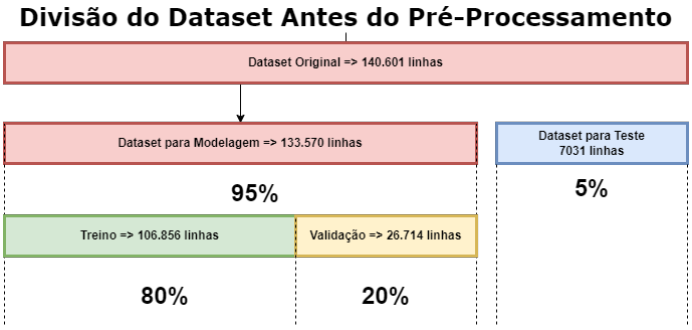

Figura 2. Divisão do dataset em treino, validação e teste, antes do pré-processamento

D. Pré-processamento dos dados

As fases do pré-processamento realizado no *dataset* consolidado a partir do Censo Escolar e dos microdados do ENEM por Escola são ilustradas na figura 3, e detalhadas nas subseções a seguir.

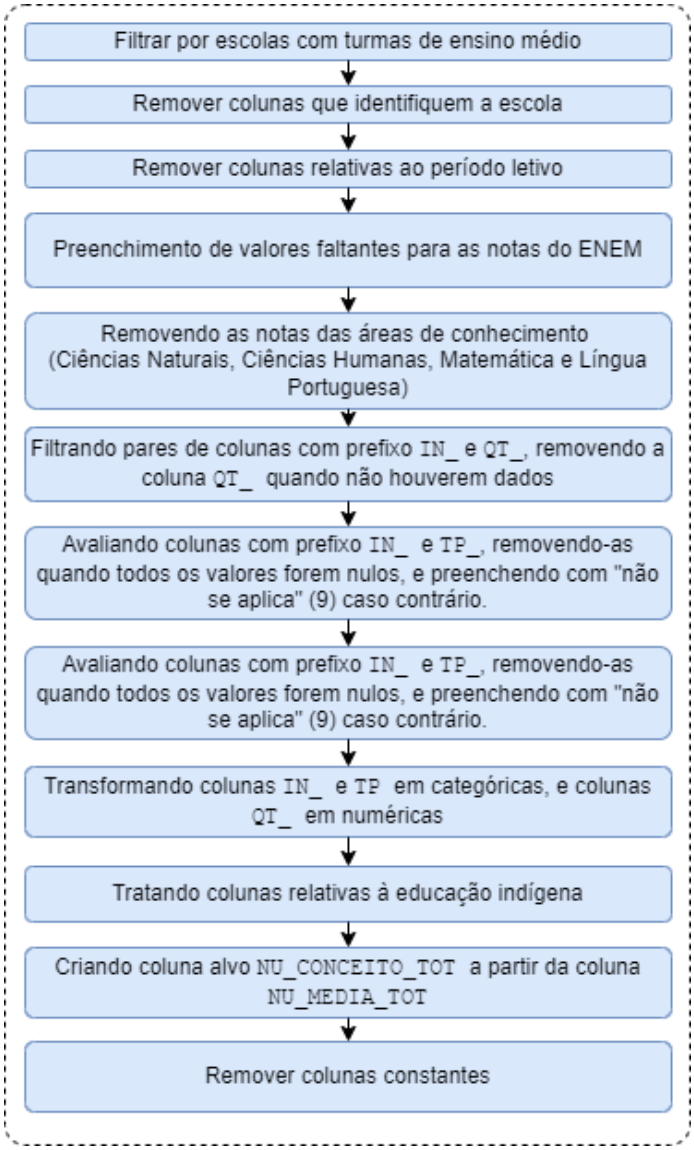

Figura 3. *Pipeline* do Pré-Processamento

1) Filtrando escolas com turmas do ensino médio: Apenas as escolas participantes do Censo Escolar que possuem turmas do ensino médio foram mantidas no *dataset*. Essa escolha se deve ao fato de que a influência do ensino dessas escolas no desempenho dos seus alunos no ENEM se dá de forma direta, ao invés das escolas que possivelmente possuem apenas turmas do ensino fundamental. No *dataset*, a coluna que representa este dado é `IN_MED`.

2) Remover colunas que identificam a escola: Colunas que, de alguma forma, identifiquem a escola também foram removidas do *dataset*. Essa decisão tem como objetivo forçar o modelo de *machine learning* a analisar apenas as características estruturais/pedagógicas fornecidas no *dataset*. As colunas removidas com esse intuito foram:

- `DS_ENDERECO`
- `NU_ENDERECO`
- `DS_COMPLEMENTO`

- NO_BAIRRO
- CO_CEP
- NU_DDD
- NU_TELEFONE
- CO_ESCOLA_SEDE_VINCULADA
- CO_IES_OFERTANTE
- CO_ORGAO_REGIONAL
- NU_CNPJ_ESCOLA_PRIVADA
- NU_CNPJ_MANTENEDORA
- NO_ENTIDADE
- CO_ENTIDADE

3) Remover colunas relativas ao período letivo: As colunas relativas ao período letivo da escola (DT_ANO_LETIVO_INICIO e DT_ANO_LETIVO_TERMINO) foram removidas do *dataset*. Entende-se que atrasos no início ou término do período letivo por motivos diversos (incluindo eventuais greves, reformas, dentre outros) afetam diretamente o desempenho dos alunos de uma escola. No entanto, a magnitude deste atraso não poderia ser determinada, já que o Censo Escolar não determina o início/término previsto e o início/término real do ano letivo.

4) Preenchimento de valores faltantes para as notas do ENEM: Durante a análise da quantidade de dados faltantes para as notas do ENEM no *dataset*, foi encontrada a situação relatada na tabela I. Para as notas das áreas de conhecimento (NU_MEDIA_CN, NU_MEDIA_CH, NU_MEDIA_MT e NU_MEDIA_LP), um percentual de aproximadamente 25,55% de dados faltantes foi encontrado. O dicionário de dados fornecido junto ao *dataset* dos microdados do ENEM por Escola informa que tais colunas só foram preenchidas para os anos entre 2009 e 2015. A coluna NU_MEDIA_OBJ teve o preenchimento feito apenas para o ano de 2008, a coluna NU_MEDIA_RED teve o preenchimento feito para os anos de 2008 a 2015, enquanto que a coluna NU_MEDIA_TOT teve o seu preenchimento feito apenas para o ano de 2008.

Tabela I
DADOS FALTANTES PARA AS NOTAS DO ENEM

Coluna	Dados Faltantes	Percentual (%)
NU_MEDIA_CN	35914	25.543204
NU_MEDIA_CH	35914	25.543204
NU_MEDIA_LP	35914	25.543204
NU_MEDIA_MT	35914	25.543204
NU_MEDIA_RED	17150	12.197637
NU_MEDIA_OBJ	121457	86.384165
NU_MEDIA_TOT	123831	88.072631

Ao avaliar as notas técnicas fornecidas junto com o *dataset* dos microdados do ENEM por Escola [13], observou-se que seria possível preencher os dados faltantes para a coluna NU_MEDIA_OBJ, de maneira correta, apenas para o intervalo de tempo em que as notas das áreas de conhecimento estivessem preenchidas. Deste modo, decidiu-se então filtrar as instâncias do *dataset*, mantendo apenas aquelas referentes aos anos de 2009 a 2015. Com isso, as instâncias com dados preenchidos para a coluna NU_MEDIA_TOT (preenchida para os anos de 2005 a 2007) e para a coluna NU_MEDIA_OBJ (preenchida apenas para o ano de 2008) também foram eliminadas.

O preenchimento adotado para a coluna NU_MEDIA_OBJ seguiu o disposto nas notas técnicas dos microdados do ENEM por Escola [13], utilizando a média aritmética simples das notas das áreas de conhecimento, como mostrado a equação 1:

$$MO = \frac{M_{CN} + M_{CH} + M_{MT} + M_{LP}}{4} \tag{1}$$

Onde:

- MO é a coluna NU_MEDIA_OBJ;
- M_{CN} é a coluna NU_MEDIA_CN;
- M_{CH} é a coluna NU_MEDIA_CH;
- M_{MT} é a coluna NU_MEDIA_MT;
- M_{LP} é a coluna NU_MEDIA_LP.

Uma vez tendo preenchido a coluna NU_MEDIA_OBJ, analisou-se a forma como a coluna NU_MEDIA_TOT foi preenchida nas notas técnicas, ilustrada pela equação 2:

$$MT = \frac{NO \cdot MO + NR \cdot MR}{NO + NR} \tag{2}$$

Onde:

- MT é a coluna NU_MEDIA_TOT, representando a nota média total dos alunos do ENEM daquela escola, naquele ano;
- NO é o número de alunos da escola que participaram da prova objetiva do ENEM naquele ano;
- MO é a coluna NU_MEDIA_OBJ, representando a nota média na prova objetiva;
- NR é o número de alunos da escola que participaram da prova da redação do ENEM naquele ano;
- MR é a coluna NU_MEDIA_RED, representando a nota média da redação.

Observou-se que, no *dataset* fornecido para os microdados do ENEM por Escola, não constam os números de alunos participantes de quaisquer provas, tornando difícil o preenchimento da coluna NU_MEDIA_TOT usando a equação 2. No entanto, ainda analisando as mesmas notas técnicas já citadas, pôde-se observar que as notas das áreas de conhecimento foram preenchidas levando em consideração o número de participantes da escola nas respectivas provas, segundo a equação 3:

$$M_{AC} = \frac{\sum_i M_i}{NO} \tag{3}$$

Onde:

- M_{AC} é a média da área de conhecimento;
- M_i é o desempenho do i-ésimo aluno da escola naquela Área de Conhecimento;
- NO é o número de alunos daquela escola que fizeram as provas objetivas.

Deste modo, neste trabalho, decidiu-se preencher a coluna NU_MEDIA_TOT usando a média aritmética simples da média da prova objetiva (NU_MEDIA_OBJ) e da prova de redação (NU_MEDIA_RED), conforme a equação 4:

$$MT = \frac{MO + MR}{2} \tag{4}$$

Onde:

- MT é a coluna NU_MEDIA_TOT, representando a nota média total dos alunos do ENEM daquela escola, naquele ano;
- MO é a coluna NU_MEDIA_OBJ, representando a nota média na prova objetiva;
- MR é a coluna NU_MEDIA_RED, representando a nota média da redação.

Após esta etapa, a situação referente aos dados faltantes observada para as notas do ENEM é descrita pela tabela II. Ainda restaram 380 instâncias sem a média da prova de redação e a média total. Tais instâncias foram removidas do *dataset*, por não haver como preenchê-las e por representar um percentual pequeno, em relação ao total.

Tabela II
DADOS FALTANTES PARA AS NOTAS DO ENEM APÓS O PREENCHIMENTO

Coluna	Dados Faltantes	Percentual (%)
NU_MEDIA_CN	0	0
NU_MEDIA_CH	0	0
NU_MEDIA_LP	0	0
NU_MEDIA_MT	0	0
NU_MEDIA_RED	380	0.362987
NU_MEDIA_OBJ	0	0
NU_MEDIA_TOT	380	0.362987

5) Removendo as notas das áreas de conhecimento: As notas das áreas de conhecimento contribuem diretamente para a composição do valor da nota média total do ENEM, conforme equações 1, 3 e 4, fazendo com que a sua correlação com a média total do ENEM seja alta. Sendo o objetivo principal deste trabalho avaliar como as características estruturais e pedagógicas de uma dada escola afetam o desempenho de seus alunos na média total do ENEM, optou-se por removê-las do *dataset*, pois tais colunas poderiam assumir o papel de *preditoras* em relação à variável alvo, e esse não é o objetivo.

6) Filtrando pares de colunas com prefixo IN_ e QT_: Analisando o dicionário de dados fornecido para o *dataset* do Censo Escolar [12], observa-se que algumas colunas surgem aos pares, diferenciando-se pelo prefixo IN_ e QT_, como exemplificado na figura 4. Em alguns casos, o preenchimento da coluna com prefixo IN_ se faz presente em todo o período a ser analisado (2009 a 2015), mas a sua correspondente com prefixo QT_ apresenta alguma defasagem em relação ao ano de início do seu preenchimento. Em situações como essas, não foi encontrada forma razoável para o devido preenchimento das colunas com prefixo QT_ no período faltante. Nesse caso, optou-se por manter apenas a coluna com prefixo IN_, excluindo do *dataset* a sua correspondente com prefixo QT_.

Também foram encontradas situações em que ambas as colunas com prefixo IN_ e QT_ não foram preenchidas no período de interesse, como exemplificado na figura 5. Em casos como esses, optou-se por remover ambas as colunas.

7) Avaliando colunas com prefixo IN_ e TP_: Ainda durante a avaliação do dicionário de dados fornecido para o *dataset* do Censo Escolar [12], observou-se que existem colunas com prefixo IN_ que não possuem correspondente com prefixo QT_. Tais colunas se assemelham às com prefixo TP_, em relação ao tipo de valores preenchidos. Neste caso, estas colunas foram avaliadas da mesma forma.

Nos casos em que não houve preenchimento em todo o período de interesse (2009 a 2015), tais colunas foram removidas do *dataset*. Para os casos em que foi detectada a presença de valores nulos para colunas deste tipo no período de interesse, optou-se por preencher os valores faltantes com o código "9", que para algumas destas colunas corresponde a "não se aplica", e para outras corresponde a "não informado".

8) Transformando colunas IN_ e TP_ em categóricas, e colunas QT_ em numéricas: Para garantir que o modelo de *machine learning* interprete as colunas tratadas nas subseções IV-D6 e IV-D7 de maneira correta, definiu-se que as colunas com prefixo IN_ e TP_ são do tipo "categórico", enquanto que as colunas com prefixo QT_ são do tipo "numérico", devido à natureza dos dados preenchidos.

9) Tratando colunas relativas à educação indígena: De acordo com o dicionário de dados fornecido para o *dataset* do Censo Escolar [12], a coluna IN_EDUCACAO_INDIGENA indica a presença de educação escolar indígena para aquela escola/instituição de ensino. Quando este tipo de educação está presente, pode-se avaliar o preenchimento da coluna TP_INDIGENA_LINGUA, correspondente ao idioma indígena em que as aulas são ministradas, podendo ter os valores:

- 1: quando as aulas são ministradas apenas em Língua Indígena;
- 2: quando as aulas são ministradas apenas em Língua Portuguesa;
- 3: quando as aulas são ministradas tanto em Língua Indígena quanto em Língua Portuguesa;
- vazio: não aplicável para escolas sem Educação Escolar Indígena.

Como descrito na subseção IV-D7, os valores nulos/vazios para a coluna TP_INDIGENA_LINGUA foram preenchidos com o código "9". Para efeito prático, devido ao número reduzido de instâncias no *dataset* relativas à presença de Educação Escolar Indígena, optou-se por remover as colunas CO_LINGUA_INDIGENA_1, CO_LINGUA_INDIGENA_2 e CO_LINGUA_INDIGENA_3, que representam em qual língua indígena as aulas são ministradas.

10) Criando coluna NU_CONCEITO_TOT: Na subseção IV-D4, a coluna NU_MEDIA_TOT (correspondente à média total da escola, naquele ano, para o ENEM) foi tratada. Com o objetivo de facilitar o entendimento dos resultados do modelo de *machine learning* gerado, a coluna chamada NU_CONCEITO_TOT foi criada, representando conceitos de A a D.

N	Nome da Variável	Descrição da Variável	Tipo	Tam.(1)	Categoria	Coleta por ano ("s"=sim;"-"=não) 07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	Notas sobre diferenças entre os anos
157	IN_EQUIP_DVD	Equipamentos existentes na escola para o processo ensino aprendizagem - DVD/Blu-ray	Num	1	0 - Não 1 - Sim	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	
158	QT_EQUIP_DVD	Quantidade de Aparelhos de DVD/Blu-ray	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo de 4 equipamentos para cada 3 salas existentes - foram marcados apenas valores>3)	-	-	-	-	-	-	s	s	s	s	s	s	s	s	s	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
159	IN_EQUIP_SOM	Equipamentos existentes na escola para o processo ensino aprendizagem - Aparelho de som	Num	1	0 - Não 1 - Sim	-	-	-	-	-	s	s	s	s	s	s	s	s	s	s	
160	QT_EQUIP_SOM	Quantidade de Aparelhos de som	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo de 4 equipamentos para cada 3 salas existentes - foram marcados apenas valores>3)	-	-	-	-	-	-	s	s	s	s	s	s	s	s	s	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.
161	IN_EQUIP_TV	Equipamentos existentes na escola para o processo ensino aprendizagem - Aparelho de televisão	Num	1	0 - Não 1 - Sim	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	
162	QT_EQUIP_TV	Quantidade de Aparelhos de televisão	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo de 4 equipamentos para cada 3 salas existentes - foram marcados apenas valores>3)	-	-	-	-	-	-	s	s	s	s	s	s	s	s	s	Em 2019: primeiro ano em que foi aplicado tratamento de valores extremos, que são marcados com o código 88888.

Figura 4. Dicionário de dados do Censo Escolar e a defasagem no preenchimento de colunas com prefixo IN_ e QT_

N	Nome da Variável	Descrição da Variável	Tipo	Tam.(1)	Categoria	Coleta por ano ("s"=sim;"-"=não) 07	08	09	10	11	12	13	14	15	16	17	18	19	20	21
181	QT_DESKTOP_ALUNO	Quantidade de computadores em uso pelos alunos - Computador de mesa (desktop)	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo de 4 equipamentos para cada 3 matrículas - foram marcados apenas valores>3)	-	-	-	-	-	-	-	-	-	-	-	s	s	s
182	IN_COMP_PORTATIL_ALUNO	Computadores em uso pelos alunos - Computador portátil	Num	1	0 - Não 1 - Sim	-	-	-	-	-	-	-	-	-	-	-	s	s	s
183	QT_COMP_PORTATIL_ALUNO	Quantidade de computadores em uso pelos alunos - Computador portátil	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo de 4 equipamentos para cada 3 matrículas - foram marcados apenas valores>3)	-	-	-	-	-	-	-	-	-	-	-	s	s	s
184	IN_TABLET_ALUNO	Computadores em uso pelos alunos - Tablet	Num	1	0 - Não 1 - Sim	-	-	-	-	-	-	-	-	-	-	-	s	s	s
185	QT_TABLET_ALUNO	Quantidade de computadores em uso pelos alunos - Tablet	Num	5	88888 - registro com marcação de valor extremo (valor superior ao limite máximo de 4 equipamentos para cada 3 matrículas - foram marcados apenas valores>3)	-	-	-	-	-	-	-	-	-	-	-	s	s	s

Figura 5. Dicionário de dados do Censo Escolar e o não preenchimento de colunas com prefixo IN_ e QT_

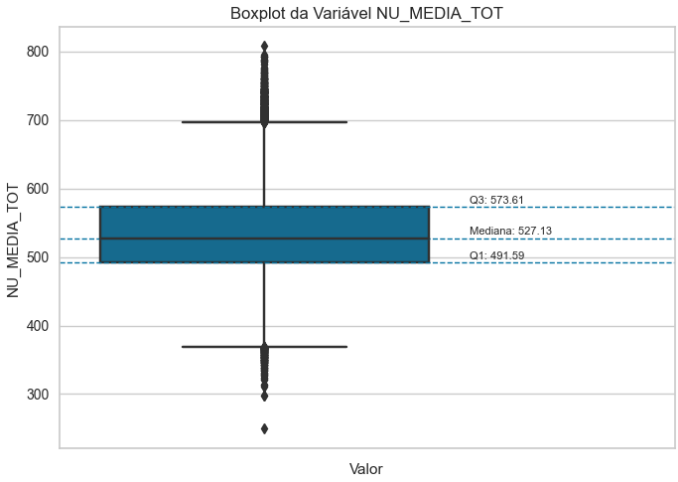

Figura 6. Boxplot para a variável NU_MEDIA_TOT

Até o momento da escrita deste artigo, não foi encontrada uma tabela oficial de equivalência entre notas e conceitos, utilizada pelo Ministério da Educação do Brasil ou outra instituição de relevância similar. Então, para traduzir a média total num conceito global, inicialmente tentou-se utilizar os quartis relativos à coluna NU_MEDIA_TOT. O boxplot representado pela figura 6 ilustra a distribuição dos valores da coluna NU_MEDIA_TOT. A partir deste boxplot, foi criada uma tabela de equivalência inicial (representada pela tabela III).

Ao analisar a tabela III, observou-se que o intervalo correspondente ao conceito "A" era muito extenso, o que não representa uma realidade plausível. Os intervalos para os conceitos "B" e "C" também eram muito curtos, o que também representa uma distorção. Para chegar num meio termo aceitável, foram alteradas as faixas de correspondência entre a média total e o conceito global, resultando na tabela IV. Para

Tabela III
EQUIVALÊNCIA ENTRE CONCEITOS E NOTAS A PARTIR DOS QUARTIS PARA A VARIÁVEL NU_MEDIA_TOT

Conceito	Intervalo
A	$[573.61, 1000]$
B	$[527.13, 573.61]$
C	$[491.59, 527.13]$
D	$[0, 491.59]$

efeito de treinamento e teste dos modelos de *machine learning*, utilizou-se esta tabela para gerar os valores para a coluna NU_CONCEITO_TOT a partir da coluna NU_MEDIA_TOT, sendo a coluna NU_CONCEITO_TOT a variável alvo dos modelos a partir de então.

Tabela IV
EQUIVALÊNCIA ENTRE CONCEITOS E NOTAS

Conceito	Intervalo
A	$[700, 1000]$
B	$[600, 699]$
C	$[400, 599]$
D	$[0, 399]$

11) Removendo Colunas Constantes: Todas as colunas com variância abaixo de 0.01 foram removidas do *dataset* nesta última fase do pré-processamento. Colunas que possuem apenas um único valor têm variância 0.00, enquanto que colunas com poucos valores únicos podem ter uma variância baixa [10]. Colunas constantes ou com baixa variância podem ser seguramente removidas do *dataset* [9]. Como resultado, 157 colunas foram removidas.

E. Seleção de Atributos

A fim de realizar a seleção de atributos no *dataset* pré-processado, seus atributos foram divididos em dois macrogru-

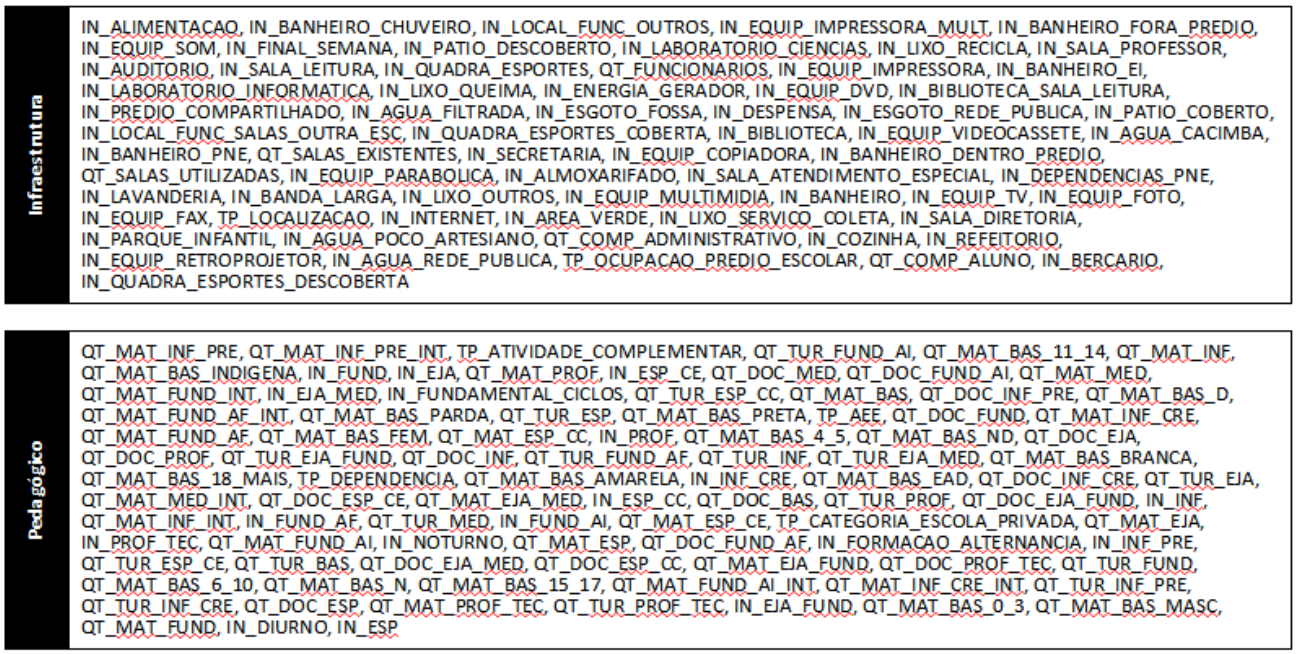

Figura 7. Divisão dos atributos nos grupos Infraestrutura e Pedagógico

pos, denominados “Infraestrutura“ e “Pedagógico“. Tal divisão foi realizada a partir do nome e da descrição dos atributos, presentes no dicionário de dados fornecido para o *dataset* do Censo Escolar [12]. O grupo “Infraestrutura“ foi criado com 67 colunas, enquanto o grupo “Pedagógico“ foi criado com 94 colunas. A figura 7 ilustra esta divisão dos atributos em grupos. Atributos que não foram enquadrados em um destes grupos foram removidos do *dataset*.

Considerando o *dataset* de treino e a união dos dois conjuntos de atributos mencionados anteriormente, foram executadas as técnicas de seleção de atributos a seguir: *Mutual Information*, *Random Forest*, *Gradient Boosting*, Regressão Logística e *Recursive Feature Elimination* (RFE). Ao executar cada uma dessas 5 técnicas, foi associado um *score* de importância a cada atributo do *dataset*. Tal *score* foi normalizado utilizando o algoritmo `MinMaxScaler` [10], [14]. Também foi associado um indexador percentílico a cada atributo, a partir deste *score* normalizado. Adicionalmente, a fim de unificar as técncias de seleção mencionadas, um *score* totalizador foi criado para cada atributo do *dataset*, a partir dos *scores* normalizados obtidos em cada uma das 5 técnicas de seleção mencionadas. Também foi criado um indexador percentílico para este *score* totalizador.

Tendo sido calculado o percentual de importância de cada atributo (tanto para as 5 técnicas de seleção mencionadas quanto para a unificação destas com base no somatório dos *scores*), adotou-se a linha de corte de 0.5%. Para cada uma das técnicas mencionadas (e a unificação delas), criou-se um conjunto de atributos cujo percentual de importância é igual ou maior que esta linha de corte, descartando os demais.

F. Seleção de Modelos

Para fins de comparativo, utilizou-se 6 modelos preditivos para problemas de classificação, disponíveis no *framework* PyCaret [15], listados na tabela V. Em todos os comparativos, tais modelos foram treinados e tiveram os seus hiperparâmetros ajustados, de forma que as métricas utilizadas sejam oriundas dos melhores resultados encontrados para estes modelos.

Tabela V
ALGORITMOS DE MACHINE LEARNING UTILIZADOS NO COMPARATIVO

Nome
Naive Bayes
SVM - Linear Kernel
Random Forest Classifier
Ada Boost Classifier
Extreme Gradient Boosting
Extra Trees Classifier

Inicialmente, comparou-se os modelos destacados utilizando o *dataset* de treino já pré-processado, sem remoção adicional de atributos. Posteriormente, comparou-se os modelos destacados utilizando os conjuntos de atributos criados para cada uma das 5 técnicas de seleção mencionadas na subseção IV-E, além do conjunto de atributos oriundo da unificação destas técnicas. Todos os comparativos foram realizados em um *laptop* ThinkPad, com processador Intel(R) Core(TM) i7-1165G7 (11ª Geração) com 2.80GHz de velocidade, e 32GB de memória RAM.

G. Avaliação dos Modelos

De modo geral, a fim de contextualizar a definição das métricas de avaliação mencionadas neste artigo, algumas terminologias foram adotadas. Em se tratando de problemas

de classificação multi-classes, pode-se definir o conjunto de classes $C = \{c_1, c_2, ..., c_n\}$, onde cada c_i (para $i = 1, 2, ..., n$) representa uma classe distinta para o problema em questão. Deste modo, tem-se:

- Verdadeiro Positivo (*True Positive* - TP): o total de instâncias do *dataset* originalmente pertencentes à classe c_i, e corretamente classificadas pelo modelo como c_i.
- Falso Positivo (*False Positive* - FP): o total de instâncias do *dataset* que não pertencem à classe c_i, mas erroneamemnte classificadas pelo modelo como c_i.
- Verdadeiro Negativo (*True Negative* - TN): o total de instâncias do *dataset* que não pertencem à classe c_i, e corretamente classificadas pelo modelo como não pertencentes à classe c_i.
- Falso Negativo (*False Negative* - FN): o total de instâncias do *dataset* que originalmente pertencem à classe c_i, mas foram erroneamente classificadas pelo modelo como pertencentes a uma outra classe c_j, tal que $i \neq j$.

Deste modo, a *precisão* pode ser definida pela equação 5 como a razão entre o total de verdadeiros positivos (TP) pelo total de instâncias classificadas pelo modelo como positivas [16]:

$$Precisão = \frac{TP}{TP + FP} \tag{5}$$

A revocação (ou *recall*), pode ser definida pela equação 6 como a razão entre o total de verdadeiros positivos (TP) pelo total de instâncias originalmente classificadas como positivas [16]:

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

O F_1-*score* [17], [18] foi utilizado como métrica de avaliação para os modelos tratados neste estudo. Tal métrica foi originalmente proposta no domínio de recuperação de informação [19], para avaliar a qualidade dos documentos obtidos como resultado de consultas em motores de busca, mas pode ser utilizada para avaliar modelos de classificação [19]. Pode-se definir o F_1-*score* como na equação 7:

$$F_1 = 2 \cdot \left(\frac{Precisão \cdot Recall}{Precisão + Recall} \right) \tag{7}$$

O F_1-*score* pode ser interpretado como a média harmônica entre a precisão e o *recall* [17], onde alcança seu melhor valor em 1 e o seu pior valor em 0 [16]. Ainda de acordo com [16], a contribuição relativa da precisão e do *recall* pode ser considerada igual para o F_1-*score*, e a média harmônica é útil para encontrar o melhor balanceamento entre essas duas medidas de avaliação que o compõem.

V. Experimentos

Nesta seção, serão abordados os resultados dos experimentos realizados neste trabalho, de acordo com a solução proposta na seção IV-A.

A. Pré-processamento dos datasets de treino e teste

Partindo da divisão do *dataset* consolidado, descrito pela figura 2, realizou-se o pré-processamento dos *datasets* de treino/validação e teste, com o objetivo de prepará-los para o uso em conjunto com os modelos de *machine learning* avaliados neste trabalho. A figura 8 ilustra a nova dimensionalidade destes *datasets*. O *dataset* reservado para treino/validação agora possui 99.419 registros e 343 colunas/atributos, enquanto que o *dataset* reservado para teste possui 5.268 registros e as mesmas 343 colunas.

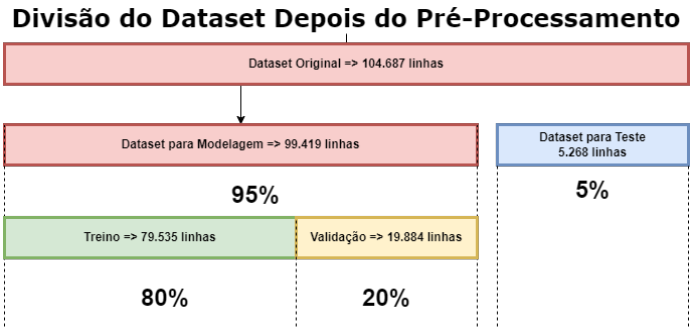

Figura 8. Divisão do dataset em treino, validação e teste, depois do pré-processamento

B. Avaliando modelos antes da Seleção de Atributos

Dentre os modelos avaliados (vide tabela V), após o treinamento e ajuste de hiperparâmetros, o *Random Forest* [20], [21] obteve o melhor F_1-*score* com valor 0.9152. Vale destacar que os modelos *Extra Trees Classifier* e *Extreme Gradient Boosting* obtiveram valores para F_1-*score* muito próximos ao obtido para o *Random Forest*, com 0.9148 e 0.9142 respectivamente. Também é importante destacar que estes 3 modelos de *machine learning* são baseados em árvores de decisão.

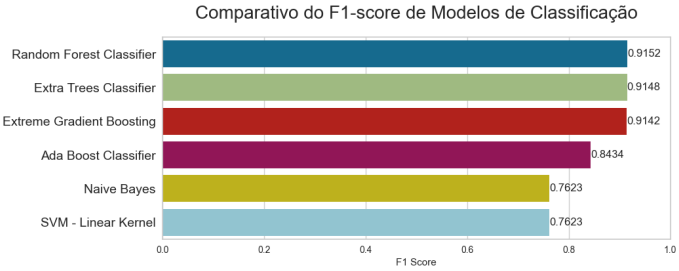

Figura 9. Resultado do Comparativo de Modelos Antes da Seleção de Atributos

C. Seleção de Atributos

Ao realizar a seleção de atributos discutida na subseção IV-E, utilizando a linha de corte de 0.5% no percentual de importância, obtiveram-se ao todo 6 conjuntos com quantidade variada de atributos. A figura 10 ilustra a quantidade de atributos associada à interseção entre tais conjuntos.

Na figura 10, podemos observar as 5 técnicas de seleção de atributos e a unificação delas pelo somatório dos *scores*. À

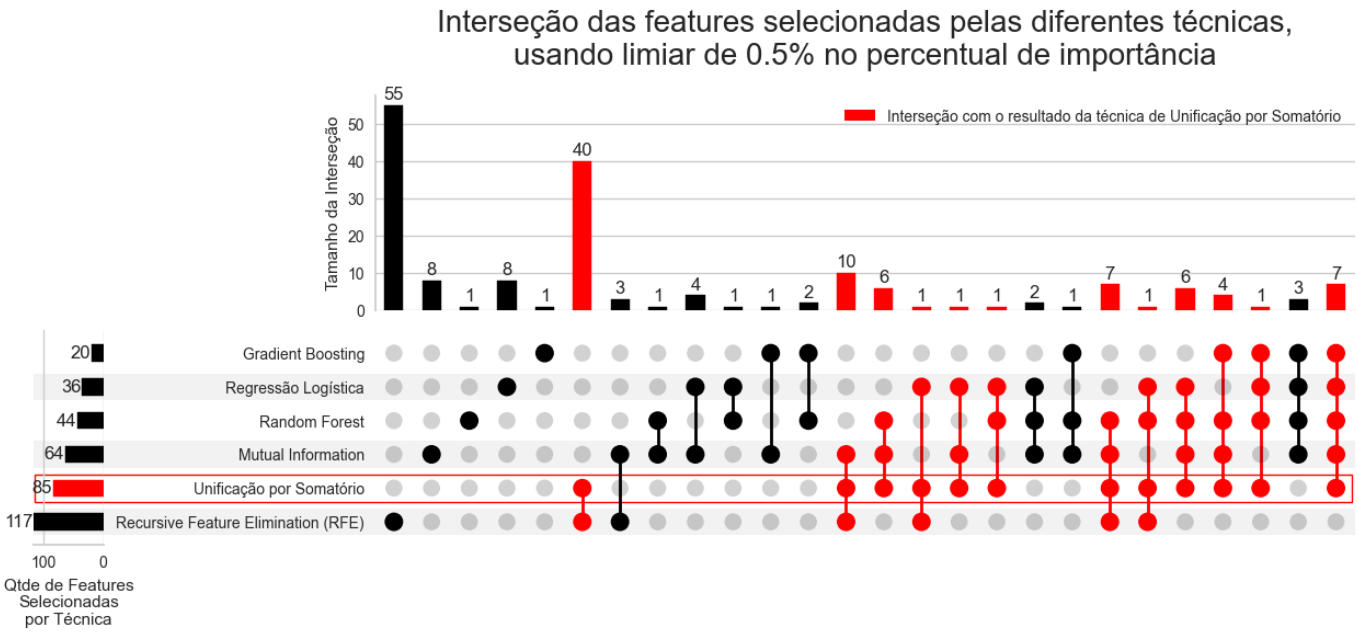

Figura 10. Analisando a Interseção dos diferentes conjuntos de atributos

esquerda do nome das técnicas de seleção, está a quantidade de atributos obtida para cada uma das técnicas de seleção. Há círculos preenchidos em preto ou vermelho para cada uma das técnicas de seleção. Quando tais círculos estão isolados, ou seja, sem ligação vertical a outros círculos, a barra vertical associada a este círculo representa a quantidade de atributos presente **apenas** no conjunto de atributos obtido através da respectiva técnica de seleção. Quando tais círculos estão ligados verticalmente a outros, a barra vertical representa a quantidade de atributos presente na **interseção** dos conjuntos de atributos obtido para as respectivas técnicas de seleção. Nesta mesma figura, destaca-se a unificação das técnicas de seleção utilizando a cor vermelha. Neste caso, quando dois ou mais círculos são preenchidos com a cor vermelha, indica que há uma interseção de um ou mais conjuntos de atributos com aquele obtido para a unificação das técnicas de seleção.

Analisando a figura 10, nota-se que não há informação referente a quantos atributos estão presentes **apenas** no conjunto de atributos referente à unificação das técnicas de seleção. Também nota-se que 55 atributos estão presentes **apenas** no conjunto de atributos obtidos para a técnica *Recursive Feature Elimination* (RFE). Também observa-se que a interseção entre os conjuntos de atributos obtidos para a técnica *Recursive Feature Elimination* e a Unificação por Somatório contém a maior quantidade (40 atributos).

D. Avaliando modelos depois da Seleção de Atributos

Como descrito na subseção IV-F, reduziu-se a dimensionalidade do *dataset* de treino pré-processado utilizando os conjuntos de atributos obtidos a partir de 5 técnicas de seleção e da unificação destas pelo somatório dos *scores* de importância. Ao todo, foram criados 6 *datasets* diferentes, um para cada técnica de seleção. Cada *dataset* com dimensionalidade reduzida foi utilizado no comparativo dos modelos de *machine learning* destacados na tabela V. Todos os modelos foram treinados e tiveram os seus hiperparâmetros ajustados.

A figura 11 ilustra os resultados obtidos para cada comparativo de modelos de *machine learning*. Em todos os comparativos utilizando *datasets* com dimensionalidade reduzida, observa-se que os modelos *Random Forest*, *Extra Trees Classifier* e *Extreme Gradient Boosting* obtiveram valores muito próximos, mesmo quando analisamos o comparativo que utiliza o *dataset* oriundo da unificação das 5 técnicas de seleção. A maior variação de valores para a métrica F_1-*score* é observada apenas para o modelo *Naive Bayes*, com valores entre 0.5945 e 0.7618.

Dentre todos os comparativos e conjuntos de atributos utilizados, destaca-se o conjunto de atributos obtido a partir da técnica de seleção *Random Forest*. Para este conjunto, o modelo de *machine learning Extra Trees Classifier* obteve o melhor resultado, com valor 0.9051 para a métrica F_1-*score*, muito próximo do valor obtido para o modelo *Random Forest* sem seleção de atributos (0.9152). A figura 12 exibe a lista de atributos obtida através da técnica de seleção de atributos *Random Forest*. Com isso, pode-se concluir que houve sucesso ao reduzir a dimensionalidade do *dataset* de treino pré-processado, pois houve pouca variação na métrica F_1-*score* dos modelos de *machine learning* utilizados.

No período de desenvolvimento deste artigo, o formulário de inspeção a unidades de educação presente no **Milênio**, apresentava, no total, 80 itens para inspeção. Dessas, 39 (49.37%) não contém relação com nenhuma das características selecionadas pelo modelo. Das 44 características selecionadas, 41 (93.18%) contém relação com algum item de inspeção do formulário do Milênio. Esse números evidenciam que há

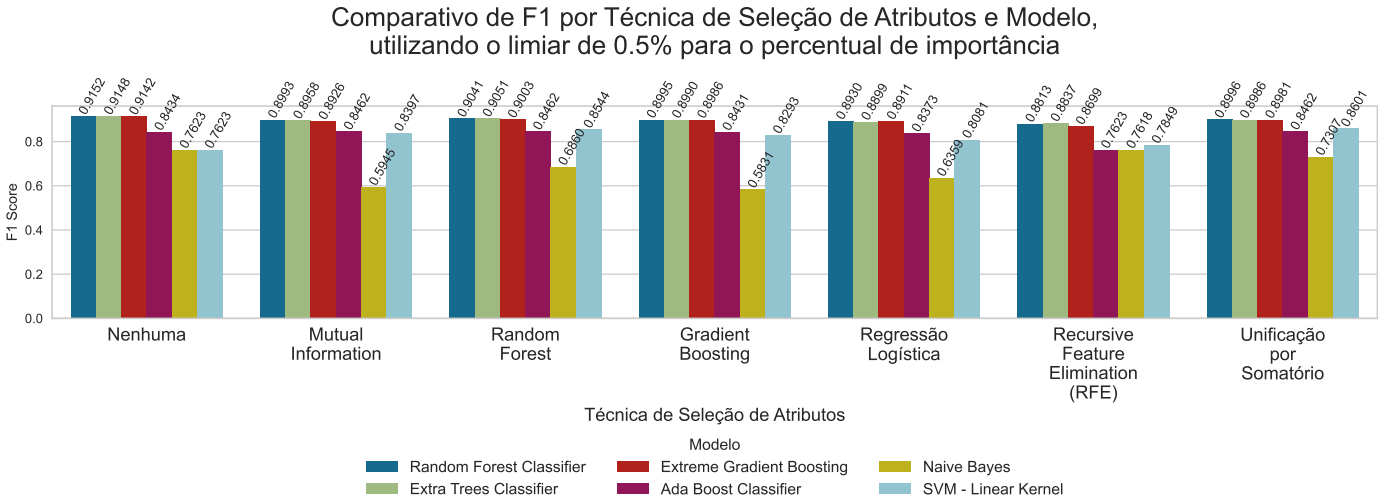

Figura 11. Analisando a Interseção dos diferentes conjuntos de atributos

uma proximidade entre as características mais importantes apresentadas neste trabalho e os itens inspecionados pelo MPBA.

VI. Conclusão

É pacífico que a educação impacta todas as áreas da vida do cidadão, permitindo que ele forme suas opiniões, proteja interesses sociais e exerça direitos e deveres. No Brasil, marcos legais, reforçaram o ensino fundamental gratuito e obrigatório como direito público subjetivo, exigindo qualidade e responsabilizando governantes por falhas. Isso levou gradativamente à criação de um robusto sistema de avaliação educacional, com destaque para o ENEM, que avalia o ensino médio, facilita o acesso ao ensino superior e é essencial para a formulação de políticas públicas eficazes.

Na Bahia, o programa Saúde + Educação: Futuro para o Milênio, iniciado pelo MPBA em 2008, fiscaliza a educação e saúde, observando condições estruturais e aplicação de verbas públicas através de visitas *in loco* e preenchimento de formulários. Todavia a análise manual dos dados é operacionalmente custosa e propensa a falhas, necessitando recursos computacionais mais sofisticados para melhor precisão.

Após os experimentos observou-se que a técnica de seleção de atributos *Random Forest* associada ao modelo de classificação *Extra Trees Classifier* apresentou o melhor resultado baseado na métrica *f1-score* (0.9051) dentre todas as outras técnicas apresentas neste trabalho, implicando na redução do custo computacional, sem perdas substancias em relação ao melhor modelo sem uso de seleção alguma (0.9152). Reforçando o resultado obtido, 93.18% das características selecionadas contém relação com algum item de inspeção do formulário do Milênio, evidenciando uma proximidade entre as características mais importantes obtidas pela técnica de seleção e os itens inspecionados pelo MPBA.

Como contribuição, este trabalho traz um método bem definido para avaliar, a partir de um conjunto de questionamentos sobre o perfil das escolas brasileiras, quais destes tem real relevância em termos da avaliação educacional. Com isso, a atuação dos órgãos fiscalizadores pode ser direcionada para os pontos que, de fato, possam melhorar a qualidade da educação no país.

Como limitações deste trabalho, pode-se destacar a quantidade reduzida de informações referente a escolas com turmas do ensino médio, presentes no **Milênio**. O número pequeno de informações não permitiu fazer um estudo direto a partir das questões presentes nos formulários do Milênio, e como alternativa, foi realizada a busca de informações semelhantes em outras fontes de dados. Também pode-se destacar o período relativo aos microdados do ENEM por Escola (de 2005 a 2015) fornecido pelo INEP. A descontinuidade do fornecimento dos dados para o período pós 2015, nos mesmos moldes, teve como objetivo inibir o seu uso para fins de avaliação da qualidade do ensino pela mídia e por gestores educacionais. A pandemia de Corona Vírus (Sars-Cov-2) também pode ser considerada como uma linha de corte relevante em termos de qualidade de educação, devido ao isolamento social, às dificuldades relativas à educação básica à distância, readaptação dos alunos, dentre outros. A indisponibilidade do fornecimento de dados relativo ao período pós 2020, nos mesmos moldes dos microdados do ENEM por Escola, impôs uma limitação ao estudo num recorte temporal mais atualizado.

Como sugestão de trabalhos futuros, pode-se destacar a realização deste estudo com dados oriundos dos formulários do sistema Milênio do MPBA. Ao obter um volume de dados consistente com a aplicabilidade deste estudo, pode-se direcionar a atuação fiscalizadora do órgão de maneira mais efetiva em termos da qualidade da educação no estado da Bahia. Também pode-se destacar a integração de um modelo de *machine learning* ao Milênio, de modo que seja possível se ter uma avaliação direta da qualidade da gestão escolar a partir destes formulários.

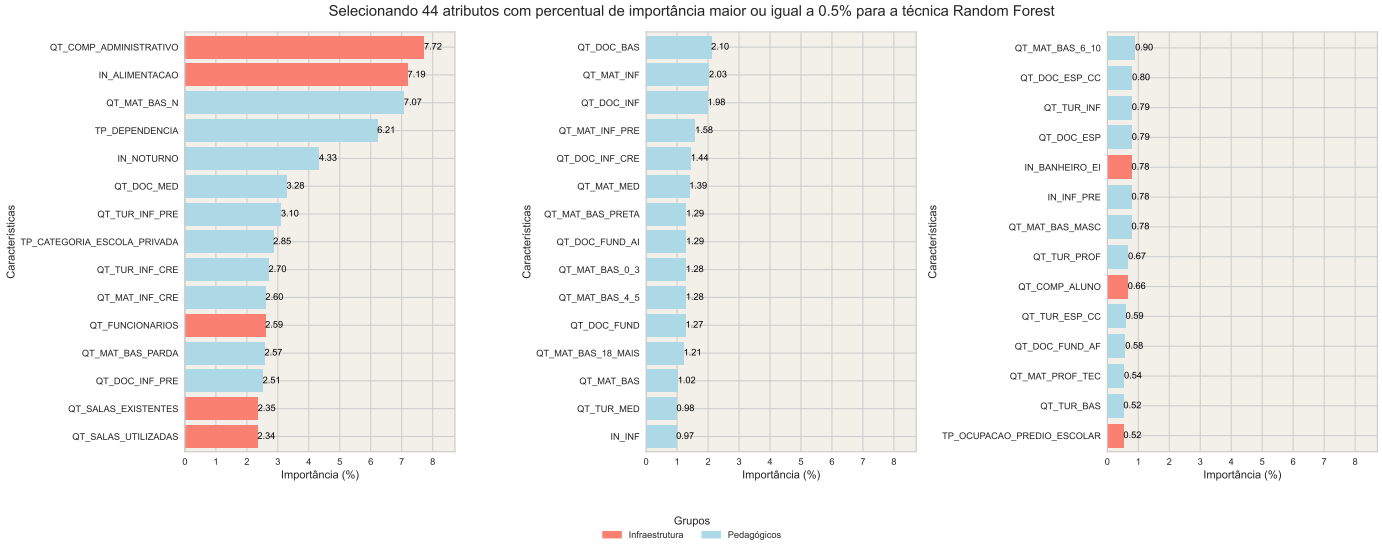

Figura 12. Analisando a Interseção dos diferentes conjuntos de atributos

REFERÊNCIAS

[1] H. F. SANDES, "O papel da educação na formação do cidadão brasileiro," *egov.ufsc.br*, 2012.

[2] M. H. G. Castro, "Sistemas de avaliação da educação no brasil: avanços e novos desafios," *educa.fcc.org.br*, 2009.

[3] INEP, "Enem." [Online]. Available: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem

[4] Ministério Público do Estado da Bahia, "Portal milênio," 2024. [Online]. Available: https://milenio.mpba.mp.br/

[5] S. KULKARNI, G. RAMPURE, and B. YADAV, "Understanding educational data mining (edm)," *https://citeseerx.ist.psu.edu*, 2013.

[6] A. M. P. Franco and N. A. Menezes-Filho, "Os determinantes do aprendizado com dados de um painel de escolas do saeb. economia aplicada," *Economia Aplicada*, 2017.

[7] A. C. Müller and S. Guido, *Introduction to Machine Learning with Python: A Guide for Data Scientists*, 1st ed. 1005 Gravenstein Highway North, Sebastopol: O'Reilly Media, Inc., Sep 2016.

[8] P. B. Harrington, *Machine Learning in Action*, Jan 2012.

[9] G. S. Galli, *Feature Selection in Machine Learning with Python*. Lulu. com, 2022.

[10] J. Brownlee, *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python*. Machine Learning Mastery, 2020.

[11] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowledge and information systems*, vol. 34, pp. 483–519, 2013.

[12] INEP, "Censo escolar." [Online]. Available: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-escolar

[13] ——, "Inep republica microdados do enem por escola." [Online]. Available: https://www.gov.br/inep/pt-br/assuntos/noticias/enem/inep-republica-microdados-do-enem-por-escola

[14] S. learn Developers, "MinMaxScaler - sklearn preprocessing," https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html, 2024, accessed: 2024-07-04.

[15] M. Ali, *PyCaret: An open source, low-code machine learning library in Python*, 2024, pyCaret version 3.0.4. [Online]. Available: https://www.pycaret.org

[16] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview," *arXiv preprint arXiv:2008.05756*, 2020.

[17] Y. Sasaki *et al.*, "The truth of the f-measure," *Teach tutor mater*, vol. 1, no. 5, pp. 1–5, 2007.

[18] P. Christen, D. J. Hand, and N. Kirielle, "A review of the f-measure: Its history, properties, criticism, and alternatives," *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–24, 2023.

[19] C. v. Rijsbergen, *Information retrieval*. Butterworth-Heinemann, 1979.

[20] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, pp. 3–42, 2006.

[21] scikit-learn developers, "Randomforestclassifier," Ago 2024, scikit-Learn version 3.2. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Sistema FIEB | SENAI CIMATEC
PELO FUTURO DA INOVAÇÃO

CENTRO UNIVERSITÁRIO SENAI CIMATEC
ESPECIALIZAÇÃO EM DATA SCIENCE & ANALYTICS

ATA DE APRESENTAÇÃO DE PROJETO FINAL DE CURSO

Ata de apresentação do Projeto Final de Curso, "**Machine Learning na Análise das Características Infraestruturais e Pedagógicas no Desempenho Escolar: um Estudo de Caso para Contribuição com o Programa Milênio do MPBA"**, submetido pelo aluno **Gessé Pinto da Silva**, como parte dos requisitos para obtenção do Certificado de **Especialista em Data Science & Analytics** pelo Centro Universitário SENAI CIMATEC, às 18h20 do dia 19 de Julho de 2024. Reuniu-se remotamente pela plataforma Google Meet, a Banca Examinadora designada pelo Prof Dr. Éldman de Oliveira Nunes – Orientador, constituída pelo Prof Dr. Éldman de Oliveira Nunes e Prof MSc Braian Varjão Gama Bispo. O Orientador deu início aos trabalhos com as devidas orientações, e a exposição foi realizada pelo estudante dentro do prazo de tempo estabelecido. Ao final da apresentação a banca reuniu-se atribuindo a seguinte nota: **9,2** (nove pontos e dois décimos).

A banca de avaliadores decidiu pela:

(X) Aprovação do trabalho
Caberá ao aluno apresentar em no máximo em 30 (trinta) dias a contar da data de assinatura desta Ata, uma cópia do trabalho em PDF, constando as considerações pontuadas pela banca. A Ata de Apresentação do Projeto Final de Curso deve ser digitalizada e inserida na terceira página do TCC ou como anexo do artigo.

( ) Reprovação do trabalho
O aluno terá que se matricular novamente no TCC – Trabalho de Conclusão de Curso e ser submetido a uma banca avaliadora no semestre seguinte**.**

As ações consequentes ao status de Aprovação deverão obedecer ao prazo proposto acima sob pena do parecer final ser modificado para o status de Reprovado automaticamente e sem possibilidade de recurso.

Para constar, lavrou-se a presente ata que vai assinada por todos os membros da Banca. Por estarem cientes de suas obrigações estão de acordo com os termos desse documento:

Salvador, 19 de Julho de 2024.

gov.br
Documento assinado digitalmente
ELDMAN DE OLIVEIRA NUNES
Data: 20/07/2024 12:38:50-0300
Verifique em https://validar.iti.gov.br

Éldman de Oliveira Nunes
Professor Orientador

Electronically signed by:
Braian Varjão Gama Bispo
CPF: ***.291.145-**
Date: 8/2/2024 8:09:57 PM -03:00

Braian Varjão Gama Bispo
Membro da banca

Assinado eletronicamente por:
Patricia Freitas Tourinho
CPF: ***.733.265-**
Data: 30/07/2024 18:56:31 -03:00

Patricia Freitas Tourinho
Coordenadora de Pós-Graduação *Lato Sensu*