Projeto semestral Disciplina ACH-2016 – INTELIGÊNCIA ARTIFICIAL

Previsão de Desempenho no ENEM por escolas:

Uma abordagem comparativa com Regressão Linear, Random Forest e MLP

Autores: Antonia Bandeira de Melo Coimbra – Nº USP.10875951;
Eduardo Rodrigues de Oliveira – Nº USP.13671921;
Pedro Henrique Lima de Andrade – Nº USP.15427077;
Reinaldo Pedro Bom Mendes – Nº USP.14605932

EACH USP – Escola de Artes, Ciências e Humanidades da Universidade do Estado de São Paulo

Abstract (Resumo)

O Exame Nacional do Ensino Médio (ENEM) é um indicador fundamental para a avaliação de políticas educacionais no Brasil. Este trabalho desenvolve e avalia modelos de aprendizado de máquina para prever as notas médias por área de conhecimento em escolas brasileiras, utilizando dados de infraestrutura, corpo docente e localização. Foram comparados três modelos: Regressão Linear, Random Forest e um Perceptron de Múltiplas Camadas (MLP). O modelo Random Forest apresentou o desempenho superior, alcançando um Coeficiente de Determinação (R^2) médio de 0.725. A análise de importância das features revelou que a dependência administrativa da escola (pública vs. privada), a taxa de participação no exame e a qualificação docente são os fatores preditivos mais relevantes. O modelo proposto constitui uma ferramenta com potencial para auxiliar na identificação de fatores críticos e na gestão de políticas educacionais.

1. Introdução

O Exame Nacional do Ensino Médio (ENEM) consolidou-se como o principal instrumento de avaliação da educação básica no Brasil e como principal via de acesso ao ensino superior. Seus resultados, geridos pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), são cruciais para a formulação de políticas públicas e a alocação de investimentos.

Contudo, a análise desses dados evidencia o desafio persistente da desigualdade educacional, onde o desempenho escolar é influenciado por uma complexa gama de fatores estruturais e socioeconômicos. Nesse contexto, a capacidade de estimar o desempenho escolar a partir de dados observáveis emerge como uma ferramenta de gestão estratégica, permitindo a identificação de escolas que necessitam de intervenção e a compreensão dos fatores de maior impacto.

Este estudo visa, portanto, desenvolver e comparar modelos preditivos para estimar as notas médias das escolas em cada uma das cinco competências do ENEM. Para isso, comparamos três algoritmos de aprendizado de máquina: uma Regressão Linear como baseline, um ensemble de árvores de decisão Random Forest, e uma rede neural MLP Regressor, avaliando suas capacidades de predição e extraindo insights a partir do modelo de melhor performance.

2. Descrição do Problema

O problema principal deste trabalho está relacionado com o Objetivo de Desenvolvimento Sustentável (ODS) 4 da ONU, que busca garantir uma educação de qualidade e equitativa para todos. As grandes diferenças de desempenho entre as escolas brasileiras, representam um obstáculo claro ao cumprimento dessa meta global no país.

Dessa forma, o objetivo deste trabalho é transformar dados em uma ferramenta de diagnóstico e previsão que ajude a entender as causas da desigualdade educacional no Brasil. Sem uma ferramenta que permita analisar e ponderar a relevância dos principais fatores e características apontados nas pesquisas e dados coletados, as políticas públicas correm o risco de serem ineficazes.

3. Trabalhos Relacionados

Este trabalho se insere no campo da Mineração de Dados Educacionais (Educational Data Mining - EDM), uma área interdisciplinar que aplica métodos computacionais para analisar dados gerados pelo ambiente de aprendizagem e, assim, compreender e otimizar os processos educacionais (Baker & Yacef, 2009). A previsão de desempenho acadêmico, seja de alunos ou de instituições, é uma das tarefas mais clássicas e de maior impacto dentro da EDM.

No cenário internacional, diversos estudos têm explorado a previsão de sucesso acadêmico. Por exemplo, Aulck et al. (2016) utilizaram modelos de regressão logística e árvores de decisão para identificar os fatores que levam à graduação de estudantes universitários, demonstrando a eficácia desses modelos em dados contextuais. Da mesma forma, pesquisas frequentemente empregam variáveis socioeconômicas e de engajamento para prever notas e taxas de evasão, validando a abordagem de usar dados não-pedagógicos para entender resultados educacionais.

No contexto brasileiro, a vasta quantidade de dados públicos gerados pelo ENEM tem fomentado diversas análises. Trabalhos como o de Costa e de Mello (2021) aplicaram técnicas de machine learning para identificar a importância de fatores socioeconômicos e da infraestrutura escolar no desempenho dos alunos, corroborando que a desigualdade de recursos é um forte preditor das notas. Outros estudos focam em modelos específicos; por exemplo, o uso de algoritmos de ensemble como Random Forest e Gradient Boosting é recorrente na literatura por sua capacidade de capturar as interações complexas e não-lineares presentes em dados tabulares heterogêneos, como os de contexto escolar (Kotsiantis, 2007).

O diferencial deste estudo reside na sua abordagem específica: a comparação direta de múltiplos modelos de regressão, incluindo um baseline linear (Regressão Linear), um robusto modelo de ensemble (Random Forest) e uma rede neural complexa (MLP Regressor), com o objetivo de prever, simultaneamente, as cinco notas do ENEM por escola. Ao focar em um conjunto de dados que combina características de infraestrutura, gestão e corpo docente, nosso trabalho busca preencher um nicho importante, fornecendo uma análise comparativa que contribue para a literatura de EDM no Brasil.

4. Metodologia

4.1. Dados e Tratamento Inicial

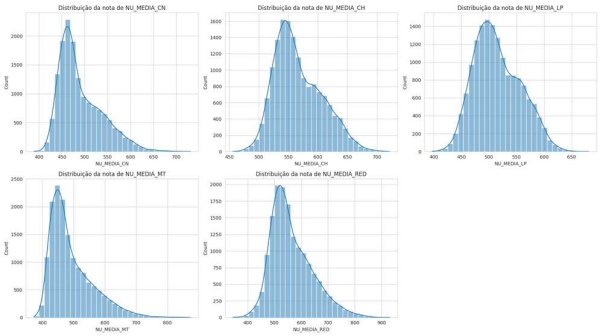
O conjunto de dados foi construído a partir dos Microdados do ENEM por Escola e do Censo Escolar, disponibilizados pelo INEP.

Os Microdados do Enem por Escola foram publicados, de forma inédita, no Portal do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Os Microdados do Enem por Escola contemplam resultados das 11 edições do Exame Nacional do Ensino Médio (Enem) até 2015. Os dados são calculados para estabelecimentos de ensino que tenham, matriculados, no mínimo, dez estudantes da terceira ou da quarta série do ensino médio regular seriado e 50% de estudantes dessas mesmas séries como participantes do Enem.

Foi feita a integração de duas bases de dados: Microdados ENEM por Escolas e os resultados do Censo Escolar de 2014, visto que a base de dados ENEM por Escolas foi descontinuada pelo INEP a partir de 2015.

O tratamento inicial incluiu a imputação de valores ausentes em PC_FORMACAO_DOCENTE e NU_TAXA_APROVACAO utilizando a mediana, e o tratamento de outliers em variáveis numéricas através do método do Intervalo Interquartil (IQR).

Gráfico de Distribuição das Notas Médias do ENEM por Áreas



4.2. Pré Processamento e Seleção de Features

Para análise de Significância e Colinearidade das features foi desenvolvido um script ‘verifica_significancia’ para uma
análise rigorosa onde variáveis com p-valor superior a 0,10 ou com alto Fator de Inflação da Variância (VIF) foram removidas. Este passo eliminou diversos preditores e diversas variáveis de infraestrutura que não apresentaram impacto estatístico significativo.

Variáveis categóricas nominais (PORTE_ESCOLA, TP_LOCALIZACAO, TP_DEPENDENCIA) foram transformadas em variáveis binárias através da técnica de One-Hot Encoding. A seleção final de features removeu identificadores (CO_ENTIDADE, SG_UF) e a variável MEDIA_TOTAL para evitar vazamento de dados, resultando em um conjunto final de preditores. As cinco notas médias foram definidas como as variáveis-alvo (multi-saída).

O conjunto de dados final para modelagem foi, portanto, composto por todas as colunas restantes após estas

remoções, sendo elas: CO_MUNICIPIO, TP_DEPENDENCIA, TP_LOCALIZACAO, NU_TAXA_PARTICIPACAO, PC_FORMACAO_DOCENTE, NU_TAXA_PERMANENCIA, NU_TAXA_APROVACAO, PORTE_ESCOLA, IN_LABORATORIO_CIENCIAS, QT_COMP_ALUNO , QT_FUNCIONARIOS, IN_FORMACAO_ALTERNANCIA, TP_ATIVIDADE_COMPLEMENTAR, mais as cinco variáveis alvo: NU_MEDIA_CN, NU_MEDIA_CH, NU_MEDIA_MAT, NU_MEDIA_LP e NU_MEDIA_RED.

4.3. Pipeline de Modelagem e Avaliação

4.3.1. Modelos:

Os modelos avaliados neste trabalho são Regressão Linear, Random Forest Regressor e MLP Regressor. A seguir descrevemos, de forma resumida, como cada algoritmo foi configurado no código, quais hiperparâmetros foram explorados e qual papel ele desempenha na comparação.

a) Regressão Linear: Implementada com LinearRegression() do scikit-learn usando todos os parâmetros-padrão. Atua como baseline por sua capacidade de capturar apenas relações lineares entre as variáveis independentes e cada uma das cinco saídas. Implementada com LinearRegression() do scikit-learn usando todos os parâmetros-padrão.

b) Random Forest Regressor: Ensemble de 100–200 árvores de decisão (RandomForestRegressor(random_state = 42)). A otimização de hiperparâmetros foi conduzida por GridSearchCV com cv = 3, buscando as melhores combinações de n_estimators ∈ {100, 200} e max_depth ∈ {20}. O algoritmo foi escolhido por lidar bem com não-linearidades e interação entre atributos sem exigir escalonamento rigoroso.

c) MLP Regressor: Rede neural feed-forward multissaída (MLPRegressor(random_state = 42, max_iter = 1000, early_stopping = True)). O grid considerou hidden_layer_sizes = (50, 50) e alpha = 0,001 para regularização L2. O early stopping interrompe o treinamento se o erro de validação não melhorar em 10 iterações, evitando overfitting.

4.3.2. Treinamento:

O treinamento de cada algoritmo foi organizado em duas etapas sucessivas que compartilham a mesma lógica de validação: primeiro estimamos um desempenho-base pela técnica de validação cruzada e, em seguida, refinamos o modelo mais promissor ajustando seus hiperparâmetros. Para preservar comparabilidade, todas as operações foram implementadas com classes do scikit-learn, que tratam automaticamente problemas de regressão multi-saída.

Na etapa de linha de base, aplicamos uma validação cruzada 5-fold sobre o subconjunto de treino (X_train, y_train). Esse procedimento divide os dados em cinco partições estratificadas, treina o algoritmo em quatro partes e valida na quinta, repetindo o ciclo até que cada partição assuma o papel de conjunto de validação. O resultado é um valor médio de R^2 que nos dá referência rápida de quão bem o método se comporta antes de qualquer ajuste fino.

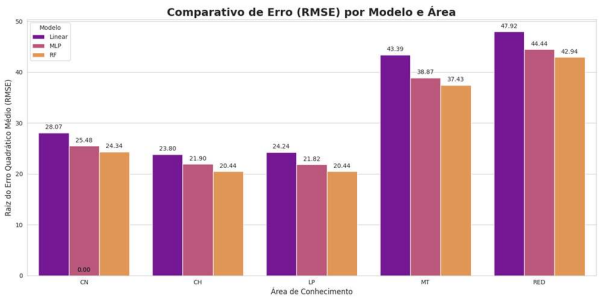
Esse fluxo garante isolamento entre treino e teste e fornece estimativas confiáveis de desempenho fora da amostra.

4.3.3. Otimização e Teste:

A fase seguinte foi dedicada à busca em grade (GridSearchCV, cv = 3), mas somente para o Random Forest e o MLP, cujos desempenhos-base indicaram espaço para melhoria via ajuste de hiperparâmetros. O grid percorreu combinações discretas de profundidade e número de árvores no Random Forest e de tamanho de camada oculta e regularização no MLP. O parâmetro n_jobs = -1 habilitou o uso de todos os threads da CPU, reduzindo o custo computacional total para algo em torno de quatro minutos em uma máquina de oito núcleos. A métrica-alvo do GridSearchCV permaneceu sendo o R^2 de validação, calculado para a média uniformemente ponderada das cinco saídas.

Após identificado o conjunto ótimo de hiperparâmetros, cada estimador foi re-treinado integralmente sobre todo o conjunto de treino para aproveitar a máxima quantidade de dados disponível. É nesse ponto que as particularidades de cada algoritmo lidam de forma nativa com a regressão multivariada: o LinearRegression() calcula um conjunto de coeficientes por saída; o RandomForestRegressor() gera um vetor de previsões em cada folha de suas árvores; e o MLPRegressor() projeta cinco neurônios de saída, um para cada área do ENEM, aprendendo seus pesos por retropropagação.

Finalmente, os modelos prontos foram avaliados sobre X_test, completamente inédito para eles, gerando predições (y_pred) que alimentaram o cálculo de R^2 e RMSE por disciplina, conforme apresentado na Tabela 1.



Área	R^2 Linear	R^2 MLP	R^2 RF	RMSE Linear	RMSE MLP	RMSE RF
CN	0,645	0,708	**0,733**	28,07	25,48	**24,34**
CH	0,643	0,697	**0,736**	23,80	21,90	**20,44**
LP	0,660	0,724	**0,758**	24,24	21,82	**20,44**
MT	0,603	0,681	**0,704**	43,39	38,87	**37,43**
RED	0,616	0,670	**0,692**	47,92	44,44	**42,94**
Média	0,633	0,696	**0,725**	33,88	30,10	**29,10**

TABELA 1

5. Resultados e Discussão

5.1. Resultados de Perfomance

Após o treinamento e otimização, os modelos foram avaliados no conjunto de teste.

5.2. Discussão e Análise de Features

Os resultados indicam a clara superioridade do Random Forest, que obteve o maior R^2 em todas as áreas, seguido pelo MLP Regressor. A baixa performance da Regressão Linear sugere que as relações entre as características escolares e as notas são predominantemente não-lineares.
O melhor desempenho do Random Forest foi em Linguagens e Códigos (R^2 = 0,758), enquanto o menor foi em Redação (R^2 = 0,692). É possível que as escolas privadas, variável mais relevante, implique num maior incentivo a prática da leitura. O pior desempenho foi em matemática (0,704) e redação (0,692), Motivos possíveis disso levantados são o de que em Matemática os aspectos individuais de cada aluno, não incluído no modelo, têm grande peso e, em Redação, possivelmente a subjetividade do tema e da banca e o grande foco que esse aspecto recebe nas diferentes instituições adicione ruído à variável-alvo.

A análise de importância de features do modelo Random Forest (Figura 2) revela os principais preditores de desempenho.

As variaveis mais importantes foram TP_DEPENDENCIA_2, mostrando grande impacto positivo caso uma escola seja privada, destacando a vantagem que as instituições privadas tem sobre a publica, mostrando uma possível tendencia de menor qualidade de preparação e ensino em escolas públicas. Também NU_TAXA_PARTICIPACAO, % de alunos que participaram do ENEM, possivelmente mostrando que escolas que incentivam os alunos a fazerem as provas atingem melhores resultados, ou, opostamente, que escolas melhores acabam tendo mais alunos motivados a fazer as provas.

As features CO_MUNICIPIO e TP_LOCALIZACAO_2 destacam diferenças regionais, com escolas rurais obtendo pior resultado, e por fim, QT_FUNCIONARIOS e PC_FORMACAO_DOCENTE, mostra que quantidade e qualificação dos profissionais têm impacto considerável nos resultados.

No geral, a variável TP_DEPENDENCIA_2 (indicando se a escola é privada) foi, de longe, a mais importante, destacando a desigualdade entre os sistemas de ensino. Em seguida, a NU_TAXA_PARTICIPACAO e a qualificação docente (PC_FORMACAO_DOCENTE) emergem como fatores cruciais, reforçando que o engajamento dos alunos e a qualidade do corpo docente são mais determinantes que a mera infraestrutura.

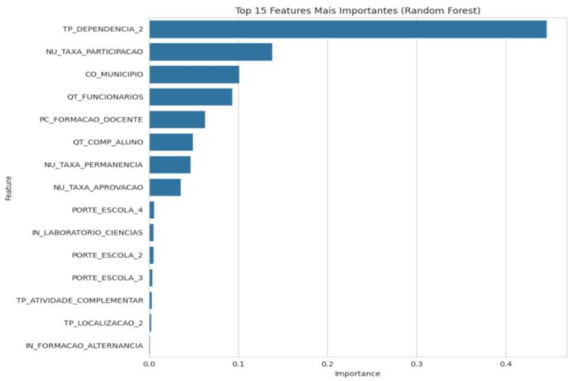


FIGURA 2

6. Conclusão

Este estudo demonstra a viabilidade de prever as notas das escolas no ENEM com boa precisão usando aprendizado de máquina. O modelo Random Forest foi o grande destaque, conseguindo explicar cerca de 72,5% da variação das notas e superando com clareza tanto a Regressão Linear quanto a rede neural MLP.

Um ponto crucial da análise é a margem de erro: um erro médio de 29 pontos, em uma escala que vai até 1000, pode parecer grande à primeira vista. No entanto, isso representa um desvio de menos de 3% do total, o que sugere que as previsões do modelo são, na prática, bastante precisas e confiáveis.

Os resultados indicam que a dependência administrativa (se a escola é pública ou privada) figura como o fator de maior impacto, o que expõe numericamente a desigualdade do sistema educacional. Além disso, o engajamento dos alunos, medido pela taxa de participação no exame, e a qualificação dos professores se mostraram mais determinantes para o sucesso do que apenas a infraestrutura.

Embora o projeto apresentou bons resultados, é fundamental, contudo, reconhecer suas limitações. A principal

delas é a ausência de dados socioeconômicos detalhados sobre os alunos e suas famílias, que probabilisticamente têm um peso enorme no desempenho. Além disso, é importante lembrar que nosso modelo encontra correlações fortes, mas não consegue, sozinho, provar uma relação de causa e efeito, funcionando apenas como um possível direcionador para melhores políticas educacionais, especialmente nas escolas públicas que apresentaram as piores médias quando comparadas às privadas.

No futuro, a pesquisa pode ser aprofundada ao incorporar novas fontes de dados para criar um modelo mais completo, testar algoritmos ainda mais avançados (como XGBoost ou LightGBM) e, como um passo prático, desenvolver um painel interativo (dashboard). Uma ferramenta assim poderia permitir que gestores simulassem o impacto de diferentes investimentos, transformando os resultados desta análise em decisões reais e bem-informadas.

6. Referências

1. Aulck, L., Aras, A., Li, L., & Luan, S. (2016). Predicting Student Success: A Comparison of Two-Year and Four-Year College Students. Proceedings of the Sixth International Conference on Learning Analytics & Knowledge.

2. Baker, R. S. J. d., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1), 3-17.

3. Costa, E. B., & de Mello, R. F. (2021). A Machine Learning Approach to Predict Student Performance in a Brazilian Public University. Anais do XXXII Simpósio Brasileiro de Informática na Educação.

4. Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. Informatica, 31(3), 249-268.

5. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). (Ano). Relatório do ENEM. Brasília, DF: INEP.

6. Ministério da Educação. Portal do Governo Brasileiro: microdados do ENEM por Escolas é divulgado pela primeira vez. Disponível em: https://portal.mec.gov.br/component/tags/tag/enem-por-escola

7. Stack Overflow, “Alternate different models in Pipeline for GridSearchCV" Disponivel em: https://stackoverflow.com/questions/50265993/alternate-different-models-in-pipeline-for-gridsearchcv (codigo adaptado para o loop de otimização)

8. Scikit-learn User Documentation, “sklearn.metrics.r2_score”. Disponivel em: https://scikitearn.org/stable/modules/generated/sklearn.metrics.r2_score.html