ESTUDO DA CORRELAÇÃO ENTRE FATORES SOCIOECONOMICOS E A NOTA DO ENEM NO ASPECTO GERAL DA PROVA

Igor de S. Mendes
Discente do curso de Engenharia Mecatrônica
Campus Alto Paraopeba (CAP)
Universidade Federal de São João del-Rei (UFSJ)
igordesouzamendes@outlook.com

Michel C. R. Leles
Departamento de Tecnologia (DTECH)
Campus Alto Paraopeba (CAP)
Universidade Federal de São João del-Rei (UFSJ)
mleles@ufsj.edu.br

RESUMO:

O Enem é um marco na vida de todo aluno que busca uma chance de entrar numa universidade seja ela pública ou privada, juntamente a isso há uma adesão de diversos candidatos de todo o brasil, Nesse trabalho iremos avaliar e com o uso de Partial Dependence Plot quais são os impactos das diversas características geográficas e socioeconômicas na nota do candidato de forma geral para os anos de 2017,2018 e 2019. Como o esperado, a influência da renda familiar do candidato é o fator mais influente estando em uma escala totalmente diferente dos demais. Há também pouca diferença entre os três anos apresentados mostrando que não é um problema recente. Essas análises têm como objetivo subsidiar e fundamentar discussões em prol do melhoramento da educação brasileira seja ela através de políticas públicas ou pesquisas científicas na área

***Index Terms* – Ciência de Dados, Aprendizado de Máquina, Estudo Microdados do ENEM, Partial Dependence Plot**

1. Introdução

Em 1998, com o objetivo de avaliar o desempenho escolar dos estudantes ao término da educação básica, o Exame Nacional do Ensino Médio, comumente chamado de ENEM, se tornou a passagem de muitas pessoas para o ensino superior sendo uma porta de entrada para os alunos em uma universidade públicas particulares por meio de políticas públicas como Sistema de Seleção Unificada (Sisu) e ao Programa Universidade para Todos do ProUni. [1]
Juntamente da prova o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) implementou um questionário socioeconômico que não somente avaliava o ensino que o candidato teve, como também sua situação socioeconômica. Transformando o ENEM em uma poderosa ferramenta de coleta de dados de alunos que estariam ingressando no ensino superior.

Essa quantidade de dados coletados permitiu a Emanuel M. Lima (2021), fornece uma análise bem aprofundada sobre o comportamento das Notas do Enem sobre um viés socioeconômico nos Anos de 2017 até 2019. As Médias de Nota analisadas é uma prova quantificada de que há uma divergência de resultados dependendo do perfil ao qual o candidato está inserido, expondo um problema socioeconômico e educacional. [2]

Mas mesmo indicando que há uma diferença, não se dá para dizer apenas por média o quanto essa diferença impacta numa análise geral da prova. O trabalho então torna como objetivo primário demonstrar e medir o tamanho do impacto que as características socioeconômicas podem influenciar na nota do Enem

2. Fundamentação Teórica:

2.1 Ciência de Dados:

Com a tecnologia crescendo em velocidade exponencial, cada vez mais o mundo se torna um lugar menor e a cada segundo tem-se diversas interações entre diversas pessoas, plataformas e serviços. Essa quantidade de interações, diferente de tempos mais antigos, são registrados e computados com uma facilidade que não se via a tempos, a esses registros e informações damos o nome de Dado.

Devido a essa abundância de produção, tem se todos os tipos diversos de dados: organizados, caóticos, rápidos, descartáveis e inúteis também. Isso gerou uma necessidade da criação de uma nova área responsável por trabalhar, tratar, analisar, compreender, visualizar e manter os dados seguros e íntegros. [3]

Essa nova ciência ganhou o nome de Ciência de Dados, a fim de estudar esse elemento que se mostra ser a peça central de uma nova Era Tecnológica.

2.2 Machine Learning:

Segundo o site da IBM: “Machine Learning é um ramo da inteligência artificial e de ciência da computação, que se concentra no uso de dados e algoritmos para imitar a maneira como os humanos aprendem, melhorando gradualmente sua precisão”

Machine learning é uma das peças mais importantes do ramo de Ciência de dados, sua utilização não só auxilia como torna possível a análise de grandes quantidades de dados por meios estatísticos. Os algoritmos são treinados para fazer diversos tipos de cálculos estatísticos. Que vão se tornar informações importantes para as diversas tomadas de decisão em diversos aspectos. [4]

2.3 Linear Regression:

Linear Regression ou Regressão linear em português é uma análise numérica que baseado no valor de outras variáveis, chamadas de variáveis independente, conseguimos prever qual o comportamento de uma variável principal, chamada de variável dependente.

Essa forma de análise estima o coeficiente de uma equação linear, envolvendo 1 ou mais variáveis independentes procurando a melhor forma de prever o valor da variável dependente. De uma forma mais gráfica, o algoritmo tenta traçar uma linha reta que ligue o máximo de pontos possível ao qual a distância entre eles e a linha seja o menor possível. [5]

2.4 Decision Tree:

Uma árvore em programação é uma estrutura de dados que herda as características das topologias em árvore. Conceptualmente diferente das listas encadeadas, em que os dados se encontram numa sequência, nas árvores os dados estão dispostos de forma hierárquica, dessa maneira cada nó nos dá acesso a uma subárvore menor.

Numa Decision Tree, ou em português Árvore de Decisão, o dado é classificado começando do nó principal que irá possuir um teste, que de maneira subsequente vai passando para outros ramos com outros testes das subárvore até finalmente chegar em uma das folhas, nessa folha e que vai se encontrar a previsão da árvore, indicando assim sua classificação. [6]

2.5 Random Forest:

Random Forest, ou em português Floresta Aleatória, assim como seu nome sugere, é uma algoritmo ao qual você possui uma série de processos em árvores, tal como Decision Tree. Para cada árvore dessa floresta nós teremos uma combinação de todos os nossos valores de entrada, nos resultando em cada uma árvore nos dando uma previsão. As previsões de modelos arvores unitários não relacionados formam um portfólios de respostas ao qual podem com mais acurácia prever o comportamento do grupo que estamos estudando.

O motivo para tal algoritmo ser forte é a capacidade de autocorreção das árvores (contando que elas não estejam tendendo ao mesmo erro). Enquanto algumas arvores possam estar erradas, outras arvores vão estar certas, então olhando de forma geral o conjunto estará todas movendo para uma resposta em comum [7]

3. Abordagem Proposta:

O Processo de ciência de dados normalmente seguimos uma ordem de etapas, e para o caso apresentado aqui também vamos nos basear nesse padrão.

1) Análise e Compreensão do Problema;
2) Obtenção de Dados e Tratamento;
3) Análise Exploratória dos Dados;
4) Aplicação de Machine Learning;
5) Observação dos Resultados.

3.1 Obtenção dos Dados:

O Primeiro passo, e talvez um dos mais importantes depois de entender o problema, é a obtenção de dados, uma vez que sem dados confiáveis e seguros todo o processo está condenado pelo seu começo. Felizmente, por estarmos lidando com um banco de dados públicos fornecido pelo próprio INEP, o que valida sua veracidade, e está disponível a qualquer pessoa para fazer o download direto do site do ENEM.

Vale a pena ressaltar que os dados pessoais tais como documentação, informações pessoais ou qualquer forma de identificar o por meios externos não são disponibilizados. Encontramos por fim nesses dados as respostas do questionário socioeconômico, as respostas do Enem, a Prova, e outras informações estatísticas pertinentes para a confecção de pesquisas e estudos.

Para este trabalho, estaremos utilizando os dados referentes aos anos de 2017, 2018 e 2019 do Enem. Ressaltando que os anos anteriores a isso demonstram problemas de download. Para cada ano temos: 120 características (colunas) para cada Inscrito do Enem, se levarmos em consideração que para cada ano tivemos em torno de 5 a 6 milhões de inscritos (linhas) podemos facilmente concluir que teremos uma quantidade exorbitante de dados a nossa disposição.

3.2 Análise e limpeza de Dados:

Num primeiro momento é comum pensar que se um banco de dados com esse tem 120 características (colunas) teremos que utilizá-las, porém, além de isso ser uma mentira é muito mais comum ser o contrário, quanto maior o Banco de Dados mais comum é a presença de problemas diversos, como: Dados faltando, Dados inúteis, Dados desorganizados, entre outros.

Nesse caso, felizmente o Inep fornece os dados organizados e categorizados, juntamente com um dicionário responsável por explicar o nome e a função de cada uma das 120 categorias do Enem dividindo-os em 9 subcategorias:

1) Dados do Participante;
2) Dados da Escola;
3) Dados dos pedidos de atendimento especializado;
4) Dados de pedidos de atendimento específico;
5) Dados do Pedidos de Recurso Especializados e Específicos para realização das Provas;
6) Dados do Local de Aplicação da Prova;
7) Dados da Prova Objetiva;
8) Dados da Redação;

9) Dados do Questionário Socioeconômicos;

Podemos fazer uma primeira limpeza no banco de dados tirando todas as categorias que não abrange o nosso campo de pesquisa, ficando assim com 3 categorias: Dados do Participante, Dados da Prova Objetiva e Dados do Questionário Socioeconômico reduzindo assim boa parte do nosso escopo de dados. Usando de referência do trabalho já citado de Emanuel M. Lima (2021), somado a uma análise do que cada variável era responsável por representar, resultou na seleção das seguintes variáveis:

a) **NU_INSCRIÇÃO:** o número de inscrição do candidato, responsável por identificar os candidato ele é uma chave única e sem repetição, uma vez que cada candidato possui apenas um número de inscrição, o tornando nosso melhor candidato para o Index de uma tabela
b) **TP_COR_RACA:** assim como mostrado pela pesquisa de Emanuel Lima, a raça/cor é um elemento muito importante para a nossa análise, e é um dos principais influenciadores da nota.
c) **TP_SEXO:** responsável por definir se o candidato é do sexo feminino ou masculino
d) **SG_UF_RESIDENCIA:** o brasil é um país de tamanho continental, sendo o 5º maior país do mundo, as diferenças de um estado para o outro são extremamente altas não só culturalmente como financeiramente portanto não pode-se deixar de fora essa variável.
e) **Q001,Q002:** parte do questionário socioeconômico, as questões Q001 e Q002 nos dão respectivamente o nível de escolaridade que o pai e a mãe do candidato possui

f) **Q006:** outra questão vinda diretamente do questionário socioeconômico, dessa vez tratando da renda familiar que cada candidato possui, numa comparação com o salário mínimo da época.
g) **TP_PRESENCA_CN, TP_PRESENCA_LC,TP_PRESENCA_CH,TP_PRESENCA_MT**: essas quatros variáveis trata exclusivamente do registro de presença do candidato nas provas de Ciências da Natureza, Linguagens e Códigos, Ciências Humanas, e Matemática respectivamente, e por esse motivo já se torna uma variável extremamente importante como será mostrado mais a frente no estudo.
h) **NU_NOTA_CN,NU_NOTA_LC,NU_NOTA_CH,NU_NOTA_MT:** sendo a nossa principal fonte de análise, essas variáveis são responsáveis por armazenar os valores numéricos de cada candidato para as provas de Ciências da Natureza, Linguagens e Códigos, Ciências Humanas, e Matemática respectivamente.

Vale a pena citar não levaremos em consideração a nota referente a redação do Enem, devido a sua forma única de correção.

3.3 Tratamento de Dados:

Com as variáveis mais importantes separadas iniciou-se a filtragem e tratamento dos dados. Todos os processos aqui listados foram aplicados a todos os 3 bancos de dados de cada ano, respectivamente.

Primeiro passo, foi a checagem de dados faltantes e a remoção de todas as linhas que estavam com dados corrompidos ou faltando para evitar algum tipo de erro. Seguindo por uma filtragem dos candidatos, utilizando as variáveis: TP_PRESENCA_CN, TP_PRESENCA_LC, TP_PRESENCA_CH, TP_PRESNCA_MT.

De acordo com o dicionário apresentado pelo próprio Inep, essas variáveis podem receber três valores que identificam o candidato, sendo elas: 0 para os candidatos ausentes, 1 para os candidatos presentes e 2 para os candidatos eliminados. Os dois grupos formados por essa filtragem foram: os candidatos presentes, aqueles que tiveram todas as suas variáveis iguais a 1, e os Ausentes foram aqueles tiveram todas as suas variáveis iguais a zero.

Como mostra a tabela 1, os candidatos eliminados provaram ser um grupo de análise muito pequeno, abaixo de 1% do total de dados da prova, por isso foram desconsiderados.

Tabela 1 - Número de Participantes e Distribuição

Ano	Total	Presentes	%	Ausentes	%	Eliminados	%
2017	6733122	4426755	65.75	2303092	34.20	3275	0.05
2018	5515332	3893743	70.60	1619005	29.35	2584	0.05
2019	5098956	3702008	72.60	1390968	27.28	5980	0.11

Na utilização de algoritmos de Machine Learning existe a necessidade de ajustar os dados para que eles possam ser processados, muitas das vezes por lidarem com valores matemáticos utilizamos dicionários correspondentes para trocar os valores sem alterar a informação. Um exemplo é a variável TP_SEXO que consegue ter dois valores F, para feminino, e M, para masculino, sendo assim foi alterado para uma variável numérica: 1 para feminino e 0 para masculino, mantendo assim sua informação mas permitindo as análises de algoritmos de Machine Learning.

A Eliminação de dados de uma tabela são decisões que devem ter cuidado ao serem feitas, pois isso pode comprometer não só aquela variável como a análise inteira. Nesse caso, porém os dados que foram eliminados eram valores vazios que não contribuíram para o nosso problema de forma alguma.

TP_COR_RACA possuem uma categoria ao qual o candidato não precisa declarar a qual cor/raça ele pertence, para o estudo desse artigo esse valor de “Não declarado” não tem valor para nós por isso foi removido.

As Q001 e Q002 possuem algo similar, com a opção “Não Sei”. Esse valor não fornece a precisão necessária para ser relevante nessa pesquisa, por isso também foi removido

3.4 Machine Learning - Decisões e Modificações:

Antes de começar as aplicações de Machine Learning, há uma necessidade de esclarecer duas alterações nos dados que não foram mencionadas no tópico passado: A média das Notas e o Agrupamento de Estados.

3.4.1 Agrupamento de Estados:

Dados podem ser agrupados em 2 Grande grupos: os Dados Quantitativos, ou numéricos, e dados Qualitativos, ou categóricos [8]:
Dados Numéricos: Responsáveis por armazenar valores numéricos, seu crescimento é proporcional a sua representação. Quanto maior a variável numérica maior o valor que ela representa

Dados Categóricos: Responsável por representar uma certa divisão, categoria ou escolha, Seus valores não possuem valores numéricos e sim representativos sobre qual grupo aquele dado pertence

Ambos estão presentes no banco de dados do Enem, porém para a análise de dados categóricos temos que aplicar um método especial e utilizar as Dummy Variables. As Dummy Variables são colunas representativas daquela categoria onde só podemos ter valores de 1 ou 0, e quanto maior a variável categórica mais colunas serão necessárias para esse processo [9], um exemplo: a variável TP_RACA_COR possui ao todo 5 categorias (Amarela, Branca, Indígena, Parda, Preta) para cada uma das categorias teremos 1 dummy variable, ou seja, teremos 5 dummy variable, somente para TP_RACA_COR.

Separando então as variáveis em numéricas e categóricas temos:

Tabela 2 - Relações de Categoria x Variáveis

Nome da Variável	Tipo	Categorias
TP_SEXO	Categórica	2
TP_COR_RACA	Categórica	5
TP_UF_RESIDENCIA	Categórica	27
Q001	Categórica	6
Q002	Categórica	6
Q006	Numérica	0
NU_NOTA_CN	Numérica	0
NU_NOTA_CH	Numérica	0
NU_NOTA_LC	Numérica	0
NU_NOTA_MT	Numérica	0

Aplicando a ideia apresentada anteriormente de dummy variables sob a tabela 2, resultará então em uma tabela de 46 Dummy Variables. Um número grande de colunas abre espaço para problemas de Multicolinearidades, que acontece quando temos relações muito fortes entre 2 variáveis independentes. Isso pode fazer variáveis que seriam uma vez insignificantes ganharem peso prejudicando o resultado do nosso algoritmo de Machine Learning. [10]

Sendo assim temos a variável **TP_UF_RESIDENCIA** que contêm dentro do seu conteúdo o nome de cada estado do Brasil mais o distrito Federal, gerando assim 27 Dummy Variables. Felizmente podemos agrupar os estados em regiões, reduzindo assim o número de Dummy Variables para 6.

3.4.2 Média das Notas:

As notas, porém, foram unidas em uma nova variável chamada de Media_Nota, através da Média Aritmética de todas elas. Isso foi feito pois em todas as Análises de Machine Learning precisamos de um valor para ser nosso valor de Saída. Além do mais, queremos analisar o desempenho do candidato para todo o Enem.

3.5 Machine Learning - Escolhas e Aplicações:

Para as ferramentas de Machine Learning, houve algumas experimentações de alguns algoritmos que foram utilizados: Linear Regression, Decision Tree, e Random Forest Regression

A primeira tentativa foi dada a Linear Regression, por ser a mais comum dentre as diversas operações de Machine Learning, depois da sequência de tratamento de dados aplicados foi possível a geração de um resultado, porém esse resultado não possuía uma precisão agradável. Depois de uma pesquisa foi concluído que essa não seria a melhor ferramenta uma vez que os dados em sua grande maioria eram categóricos enquanto regressão linear se comporta melhor com a presença de variáveis majoritariamente numéricas.

Foi utilizado também o algoritmo de Decision Tree, uma vez que a maioria dos dados apresentados eram categóricos. Houve uma modificação nos dados para poder encaixar nesse tipo de Análise uma vez que o resultado de uma Decision Tree não se dar por valor numérico e sim por uma categoria. Portanto agrupamos as notas dos candidatos em categorias
entretanto seus resultados de precisão foram inferiores ao de Linear Regression

Depois de algumas experimentações chegamos a Random Forest, um algoritmo que mesmo sendo pesado, e utiliza diversos processos simultaneamente, foi o que mostrou o menor dos índices de erros em comparação a Linear Regression, e subsequentemente Decision Tree

4. Resultados Experimentais e Discussão:

Para a nossa análise, utilizamos uma metodologia de análise de Machine Learning chamada: Partial Dependence Plot (PDP): Exibe o efeito de uma ou duas variáveis no resultado previsto de um modelo de Machine learning. O PDP é um método global, dessa forma ele analisa a relação da variável com o Machine Learning levando em consideração todos as interações entre os estados, a variável analisada, e o resultado da saída. No nosso caso que temos o uso predominante de variáveis categóricas o PDP força cada instância a ter a mesma Categoria gerando assim um PDP estimado valor [11]

Essas ferramentas nos auxiliam a exibir uma relação de o quão importante tal categoria é responsável e influente no comportamento de predição que um Algoritmo possui. Dessa forma quanto mais influente aquele valor possui maior será o resultado que subsequentemente nos informar quais das variáveis são responsáveis por impactar na nota do Enem

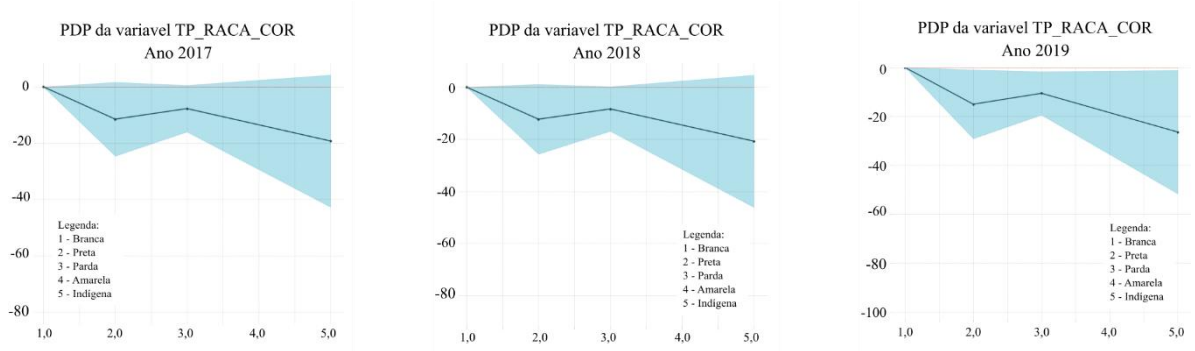

Figura 1 - PDP variável TP Raça Cor

Os resultados em Machine Learning mesmo que categóricos tomam como referência o primeiro valor e usam ele de base para então construir toda uma equação em volta. Por isso, de acordo com a Figura [1] o elemento 1, no caso a raça branca, ela é dada como elemento neutro, mas é bem perceptível que todas as outras etnias tendem a abaixar o resultado, com uma clara ênfase nos Indígenas. Isso mostra que ainda temos, o comportamento se mantém com os passar dos anos visto mesmo que um leve melhoramento quanto ao ano de 2019

.

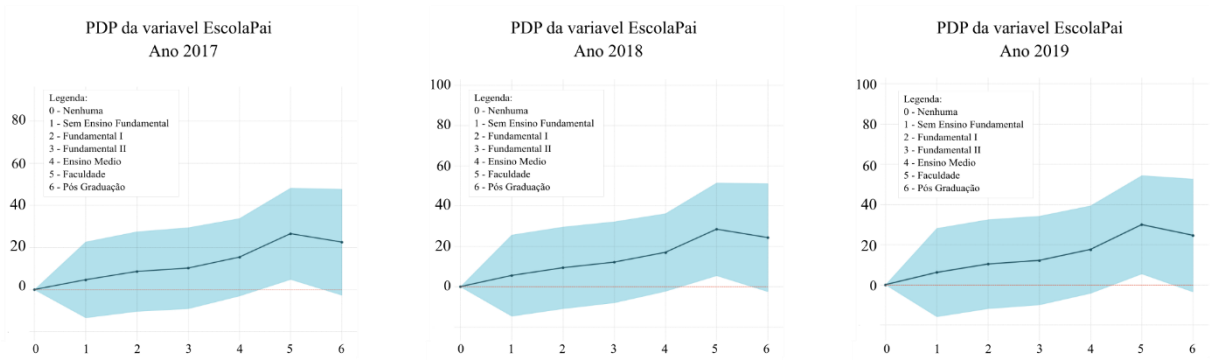

Figura 2 - PDP Variável EscolaPai

Seguindo disso, outro comportamento que pode nos mostrar são as escolaridade da figura paterna do candidato, com exceção de 2018, figura 05, que mostra um pico para os candidatos para o qual os pais concluíram a faculdade, Temos uma linearidade no resultado da nota com a escolaridade do pai

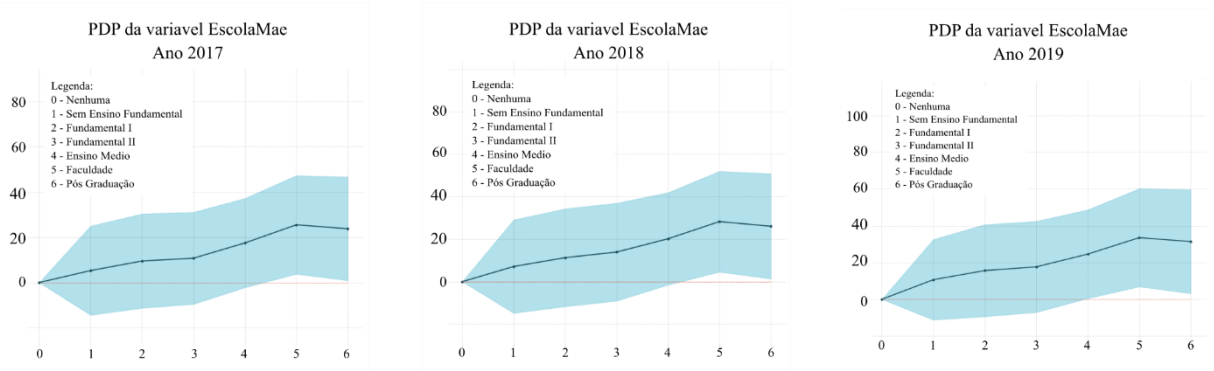

Figura 3 - PDP Vatiável EscolaMae

No caso materno dos nossos candidatos, filhos de mãe com Faculdade possuem maior tendência a girar maiores notas. Uma observação válida de se fazer é que que o comportamento do gráfico da figura materna não é tão linear quando da figura paterna

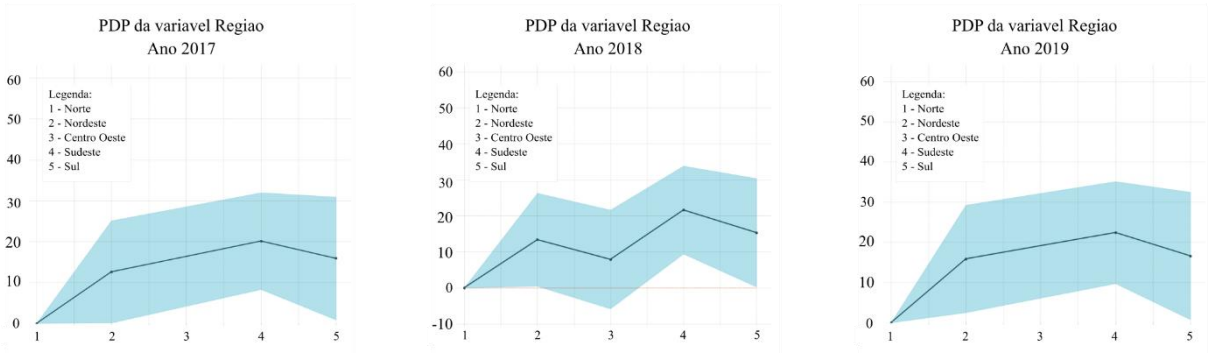

Figura 4 - PDP Variavel Região

Quanto às Diversas regiões conseguimos ver uma crescente nos valores ao passar dos anos, mas não perdemos de forma alguma a liderança do Sudeste em comparação com as demais regiões do brasil. è uma conclusão esperada visto que os 3 maiores PIBs brasileiros estão localizados no Sudeste.

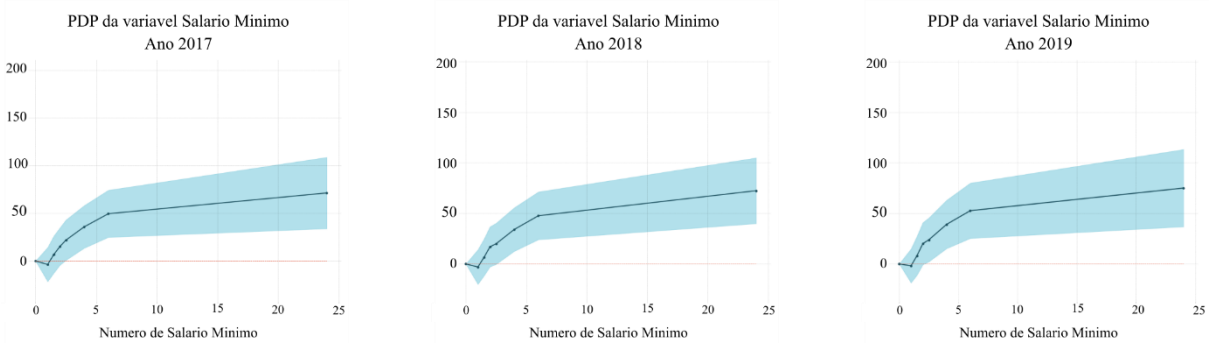

Figura 5 - PDP Variavel Salario Minimo

Sem dúvida a variável mais importante de todas as selecionadas até aqui, só olhando a escala de importância [Eixo Y] já conseguimos ver a diferença de escala que estamos trabalhando. Não é errôneo pensar que ao ter condições financeiras melhores o candidato também teria acesso a escolas melhores ou então até cursos preparatórios exclusivos em fazer a prova do Enem.

5. Considerações Finais e Trabalhos Futuros:

Com as discussões promovidas nesse trabalhos e também os dados apresentados, podemos inferir numericamente que alguns quais os tópicos que mais influenciam na nota do Enem, ressaltando a Renda Salarial consegue se destacar na previsão do Machine Learning.

É perceptível o grande abismo que alguns candidatos têm sobre os demais, baseado simplesmente em valores geográficos, econômicos e familiares. Infelizmente o padrão se manteve entre os anos de 2017 a 2019, tanto que quase não vemos diferença gritante na forma como cada variável impacta ao longo dos anos. Acentuando assim a desigualdade que atinge os brasileiros a tempos.

Como propostas para trabalhos futuros há a possível análise para os anos posteriores, de tal forma a estudar se essa diferença se mantém ou não com o passar dos anos. Um outro viés a se seguir é a tentativa de encontrar uma exatidão seja por implementação de outros algoritmos, ferramentas mais avançadas ou então o equacionamento desses atributos podendo então chegar a uma fórmula matemática que poderia prever esses resultados com exatidão

6. Agradecimento

O autor agradece o suporte financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Universidade Federal de São João del-Rei (Edital 005/2021/PROPE/UFSJ)

REFERNCIAS:

[1] INEP. Inep: Exame Nacional do Ensino Médio. Página Inicial. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem>. Acesso em: 25/09/2022

[2] LIMA, Emanuel. Aplicação de Ferramentas de Data Science na Avaliação de Desempenho de Alunos do Ensino Médio por meio do Enem, 2021, Disponível em: https://github.com/e-moncao-lima/IC_DataScience_ENEM/blob/main/IC_Emanuel%20(1).pdf

[3] AMARAL, Fernando. Introdução à ciência de dados: mineração de dados e big data. Alta Books Editora, 2016

[4] Machine Learning. IBM, 15 de Julho de 2020, O que é machine learning?. Disponível em: <https://www.ibm.com/br-pt/cloud/learn/machine-learning>. Acesso em: 27/09/2022.

[5] Linear Regression. IBM, What is linear regression?. Disponível em: <https://www.ibm.com/topics/linear-regression>. Acesso em: 26/09/2022.>

[6] QUINLAN, J.. Ross . Learning decision tree classifiers. ACM Computing Surveys (CSUR), v. 28, n. 1, p. 71-72, 1996.

[7] YUI,Tony. Understanding Random Forest. Medium,12 de Junho de 2019. Disponível em: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[8] LUCENA,Willianm. Tipos de atributos e dados. Medium,20 de Agosto de 2019. Disponível em: https://medium.com/@will.lucena/tipos-de-atributos-e-dados-7d89f47b4c8d

[9] HARDY, Melissa A. Regression with dummy variables. Sage, 1993.

[10] SULEIMAN, Ahmad A. Analysis of multicollinearity in multiple regressions. International Journal of Advanced Technology in Engineering and Science, v. 3, n. 1, p. 575-576, 2015.

[11] Interpretable Machine Learning - A Guide for Making Black Box Models Explainable: pag 117 - 118