Disciplina INF2979 – Aprendizado de Máquina

**Proposta de Projeto**

**Análise da relação entre fatores socioeconômicos e performance no ENEM**

Alunos: Gustavo Coelho, Andre Sobrenome, Gerardo Sobrenome

1. Descrição

Segundo o Ministério da Educação, O ENEM (Exame Nacional do Ensino Médio) tem o objetivo de avaliar o desempenho do estudante ao fim da escolaridade básica. Porém, o ENEM também pode ser uma fonte rica de informações a respeito da desigualdade no ensino médio brasileiro. Um dos motivos para isso é a alta adesão de alunos de variados espectros sociais e econômicos. Em 2019 (sua última edição) o exame contou com a inscrição de mais de 5 milhões de alunos, o que fornece um excelente espaço amostral da população de estudantes que desejam ingressar no ensino superior. Além disso, cada candidato é convidado a responder uma ampla gama de perguntas em um questionário apresentado no momento da inscrição. A aplicação deste questionário resulta em mais de 130 informações diferentes para cada um dos inscritos, incluindo informações básicas como idade, sexo e município de residência até informações detalhadas como a escolaridade dos pais, o número de banheiros na residência e o acesso de internet sem fio em casa.

A ideia do trabalho proposto é fazer uso deste valioso conjunto de dados para tentar compreender como fatores socioeconômicos se correlacionam com o desempenho de alunos no ENEM através de ferramentas de aprendizado de máquina. Para isso, propõe-se primeiramente a criação de um modelo preditivo que seja capaz de prever a nota de um aluno a partir das informações disponíveis. Com o modelo devidamente implementado e otimizado, podemos então extrair apenas os atributos mais significativos usados pelo modelo e realizar a segmentação dos alunos em clusters a partir destes atributos. Dessa forma, garantimos que a segmentação será feita com o uso das características mais importantes para a performance na prova, o que deve impactar na nota média de cada cluster.

A partir dos resultados obtidos, poderemos compreender o perfil socioeconômico dos diferentes segmento de alunos, entendendo as diferenças entre os clusters com maiores e menores notas e talvez auxiliar especialistas na área da educação com “insights” para formulação de políticas públicas que visem a diminuição na desigualdade do ensino médio.

1. Dados

Os dados a serem utilizados neste projeto são abertos e disponibilizados pelo Inep em seu site oficial. As informações dos inscritos são disponibilizados em uma única tabela com mais de cinco milhões de linhas e 136 colunas no formato csv. Uma rápida análise exploratória confirma uma clara relação entre alguns parâmetros e nossa variável alvo:

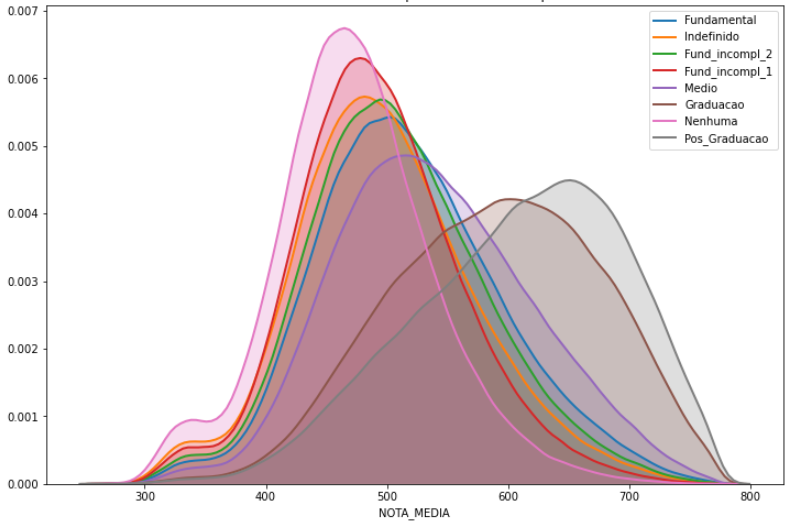


Figura 1: Densidade de notas por escolaridade do pai

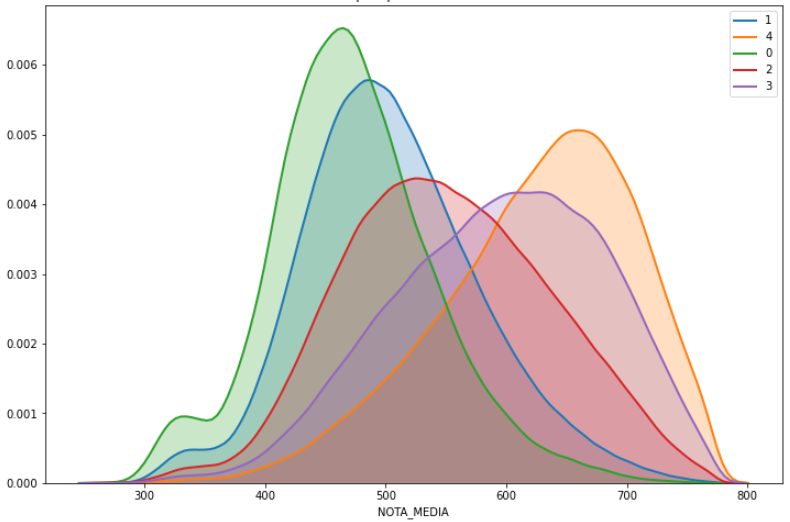


Figura 2: Densidade de notas por quantidade de banheiros na residência

1. Desafios

Após uma análise primária nos dados, é possível identificar alguns desafios que poderão ser abordados ao longo do projeto:

* Alguns importantes parâmetros possuem uma grande quantidade de valores ausentes. Como tratar estes valores sem alterar significativamente a representatividade dos dados?
* Como lidar com o conceito de “data fairness” dos dados? Devemos usar dados como Sexo e Raça? Devemos tratar dados categóricos como a formação dos pais como hierárquicos?
* Como definir as variáveis mais relevantes para a criação do modelo?
* Como encontrar a melhor combinação entre modelos de aprendizado de máquina e seus hiperparâmetros?
* Qual o melhor método de aprendizado não-supervisionado para a clusterização dos dados?
* Os modelos são interpretáveis? Isso é especialmente importante se for considerado como uma ferramenta para fomentar políticas públicas.