05 - Resultados Finais

Tema: Análises dos dados do ENEM em anos pré pandemia e durante a pandemia.

Equipe:

Gabriel Vieira Rodrigues, 1759620, gabrielvr97, EC, UTFPR

Rodrigo William Chavoni, 1996169, rodrigochavoni, BSI, UTFPR

Introdução:

Com as restrições de distanciamento social impostas pela pandemia de COVID-19, muitas áreas tiveram que se adaptar para continuar em funcionamento. O ensino foi uma delas e o ensino remoto foi a solução adotada para seguir com as aulas.
O presente projeto tem como pergunta de pesquisa: "quais grupos socioeconômicos e regionais que realizam o Exame Nacional do Ensino Médio (ENEM) foram os mais impactados durante o período da pandemia?".
Para isso foi aplicado um modelo de regressão múltipla com os dados fornecidos pelo INEP de 2019, mais especificamente as questões socioeconômicas para compreender como esses dados afetam a média final de notas. Após isso, utilizar esse modelo para prever as notas com os dados socioeconômicos dos candidatos de 2020 e fazer uma análise dos *under/over perfomers* na prova, ou seja, candidatos que tiveram resultados acima e abaixo do esperado para o seu perfil socioeconômico.
As hipóteses levantadas foram:

- Candidatos com pai/mãe com ensino qualificado foram menos afetados.
- Candidatos com pai/mãe com profissões "elitizadas" foram menos afetados.
- Candidatos sem internet e computador foram mais afetados.
- Candidatos com baixa renda familiar foram mais afetados.

- **Processamento de dados:**

Os dados utilizados são disponibilizados pelo INEP (Instituto Nacional de Estudos e Pesquisas) referentes aos exames do ENEM dos anos de 2019 e 2020, que podem ser acessados através do link: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem.

Após analisar o dicionário de dados do *dataset*, foram descartadas algumas colunas como número de inscrição, códigos de município, estado e colégio e gabaritos das questões. Após isso foram renomeadas algumas colunas referentes ao questionário socioeconômico, com o objetivo de facilitar as futuras análises. Por exemplo: a coluna “Q006” foi renomeada para “Renda_Familiar”. O próximo passo foi filtrar somente os alunos presentes, já que o *dataset* possui os dados de todos os alunos que foram inscritos, inclusive os que não compareceram no dia da prova.
Foi realizado também o processo de criação de variáveis *dummy* para análise de variáveis categóricas como profissão e nível de escolaridade dos pais.

Resultados:

Para o modelo de regressão múltipla foi utilizado as seguintes variáveis:
Variável resposta: Nota do exame.
Variáveis explicativas: Questões socioeconômicas das hipóteses.

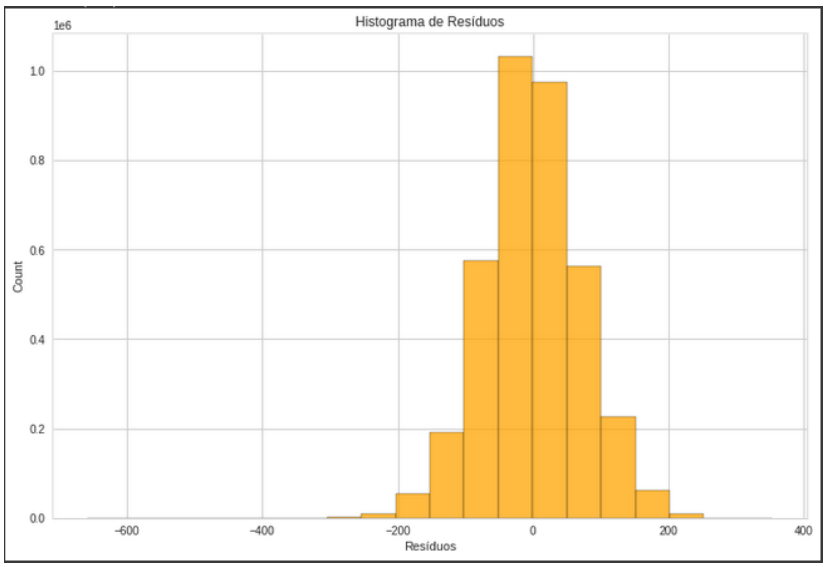
Utilizando os dados de 2019, foi possível obter os seguintes resultados a partir das análises de:

- Teste F (Valor obtido: 0,00): A primeira variável analisada foi a do resultado do teste F, que responde se o modelo utilizado é útil para prever as notas. Como o valor obtido foi menor que 0,05, indica-se que há evidências de que pelo menos uma variável está relacionada com a nota.
- R^2 e R^2 ajustado (Valor obtido 28% para ambos): O valor de R^2 representa qual a porcentagem que as variáveis explicativas explicam na variabilidade da nota. Apesar de ser considerado baixo, o problema mapeado é um tanto quanto complexo, sendo assim foi aceito um R^2 abaixo de 50%.
- p-value das variáveis explicativas: todos os valores obtidos foram de 0, portanto todas as variáveis influenciam no modelo.

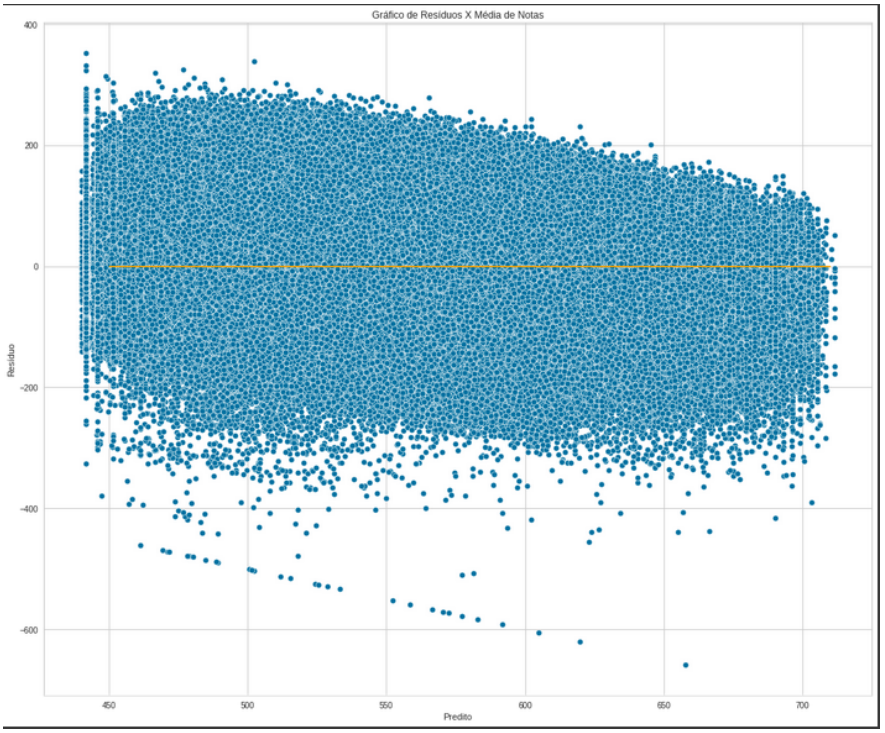
Agora analisando cada constante das variáveis explicativas, foi possível observar que os valores que mais aumentam o valor da nota em cada categoria são:

- Mães que completaram o ensino superior (32 pontos).
- Pais que completaram o ensino superior (27 pontos).
- Pais que trabalham com: Professor (de ensino fundamental ou médio, idioma, música, artes etc.), técnico (de enfermagem, contabilidade, eletrônica etc.), policial, militar de baixa patente (soldado, cabo, sargento), corretor de imóveis, supervisor, gerente, mestre de obras, pastor, microempresário (proprietário de empresa com menos de 10 empregados), pequeno comerciante, pequeno proprietário de terras, trabalhador autônomo ou por conta própria (23 pontos)
- Ter computador (16 pontos)
- Mães que trabalham como Professora (de ensino fundamental ou médio, idioma, música, artes etc.), técnica (de enfermagem, contabilidade, eletrônica etc.), policial, militar de baixa patente (soldado, cabo, sargento), corretora de imóveis, supervisora, gerente, mestre de obras, pastora, microempresária (proprietária de empresa com menos de 10 empregados), pequena comerciante, pequena proprietária de terras, trabalhadora autônoma ou por conta própria (13 pontos).

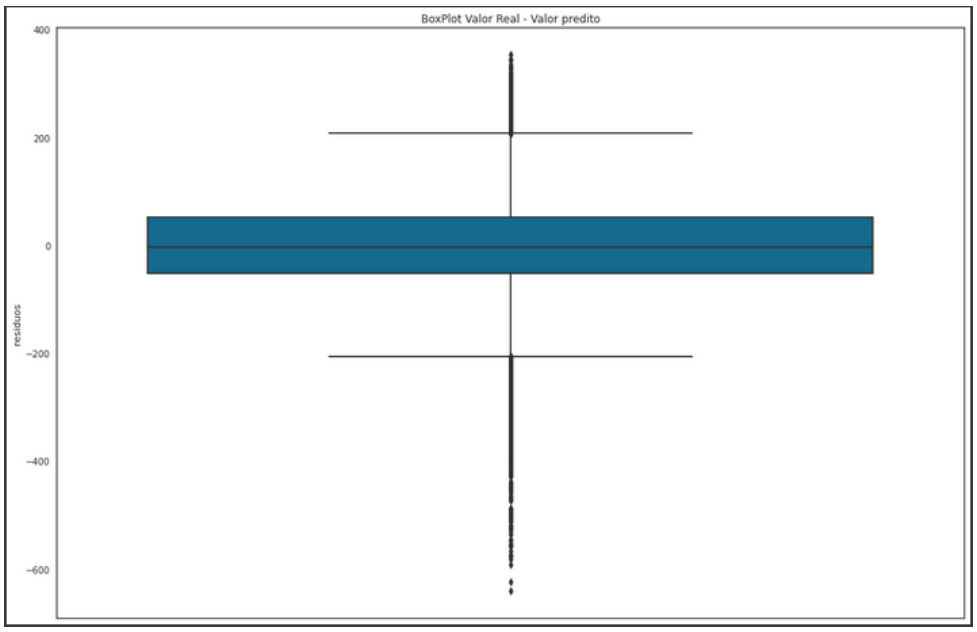
Após essas análises, foram gerados alguns gráficos para verificar se o modelo estava bom e ajustado para as próximas etapas, para isso foi analisado os resíduos do modelo.
O primeiro gráfico é um histograma, no qual espera-se uma distribuição normal com média 0. É o que o gráfico mostra.



O segundo gráfico é de dispersão dos resíduos x média de notas, para esse gráfico espera-se que não haja tendências ou muita dispersão. O que o gráfico nos mostra é uma tendência de queda nas notas mais altas, mas isso acontece também nos gráficos entre os resíduos x variáveis de pesquisa, o que pode explicar essa tendência de queda. E indica também que o modelo tende a errar em notas maiores

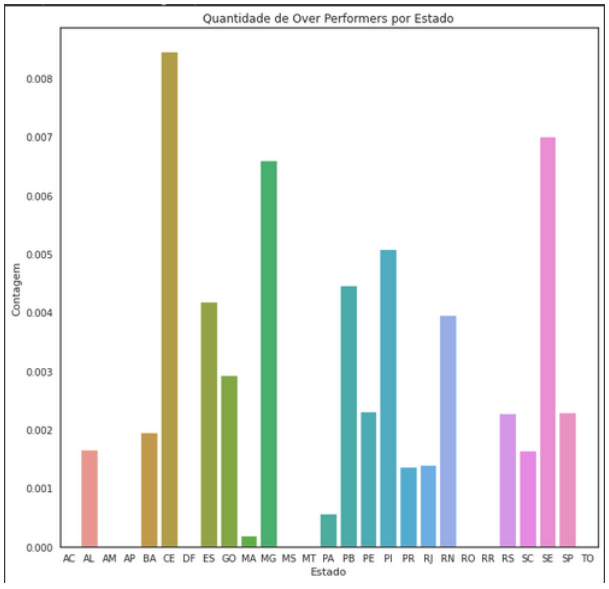


Após a análise do modelo, foi feita a previsão de notas para os alunos do ano 2020 a partir dos dados socioeconômicos. Com isso feito, subtraiu-se das notas reais de 2020, os valores previstos pelo modelo. Baseando-se no valor da subtração, foi plotado um gráfico boxplot para analisar os valores que estão fora do intervalo.

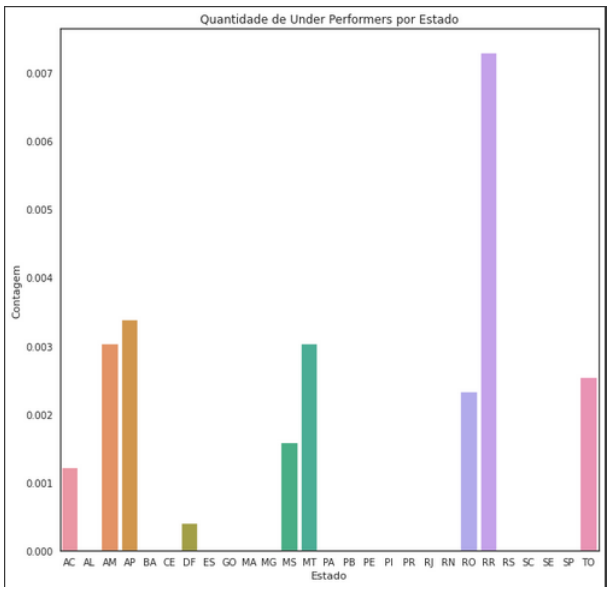


Montou-se um data frame com os dados fora dos intervalos do boxplot, que foi separado em underperformers e over performers. De primeira impressão foi possível perceber que over performers são o dobro de underperformers, um indicativo na melhora das notas do que se esperava para o ano de 2020.

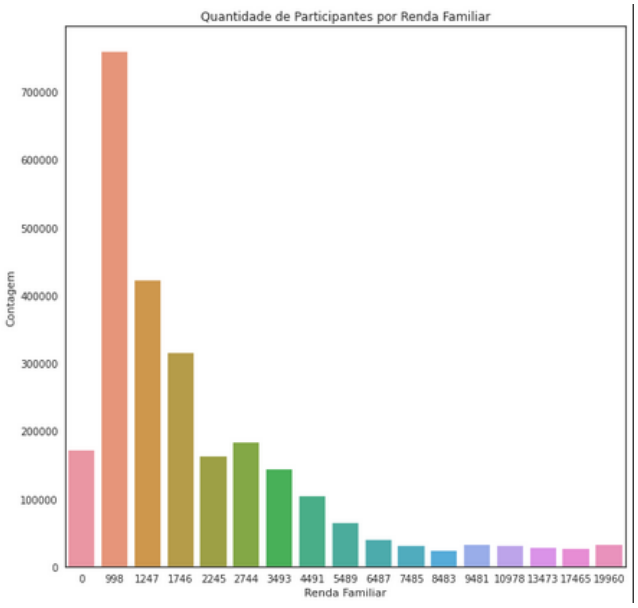
Embora o número de participantes de São Paulo seja dominante, ao normalizar os dados é possível perceber que o estado que teve mais over performers foi Ceára, outro estado que surpreendeu foi Sergipe.



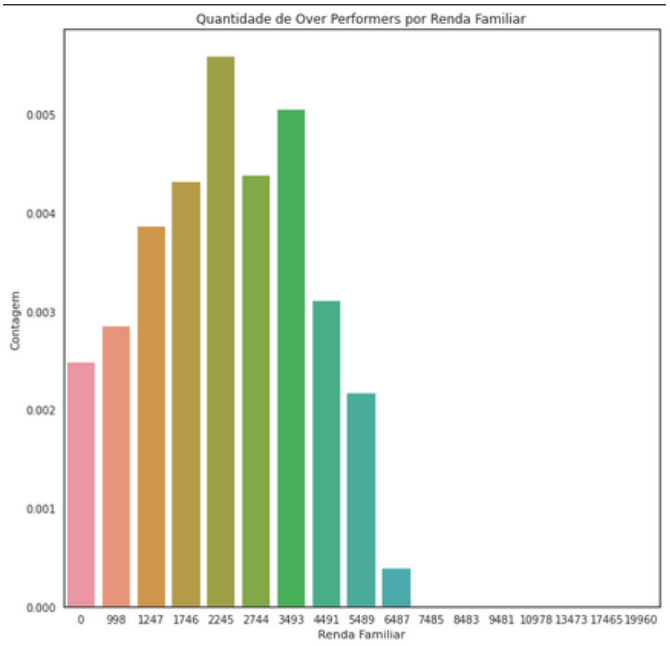
Para os participantes underperformers, o estado de Roraima ficou na frente.



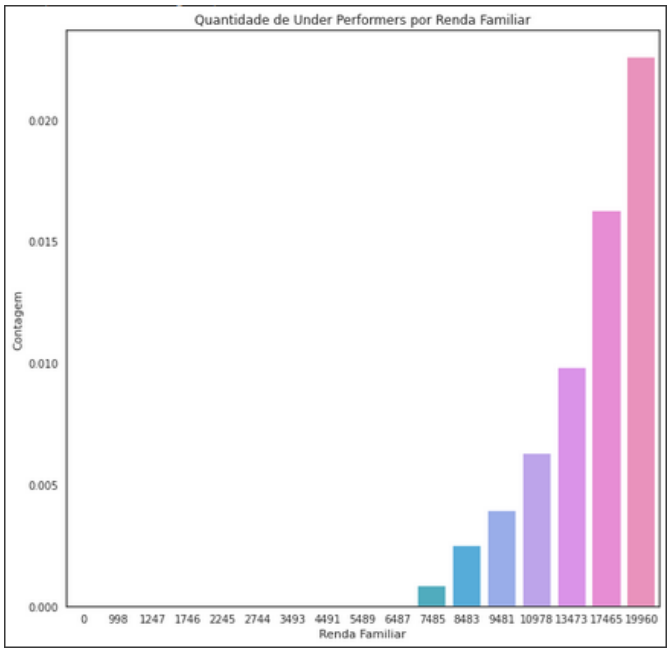
Na categoria renda, a maior parte dos participantes são estão na categoria B, como é possível visualizar no gráfico abaixo:



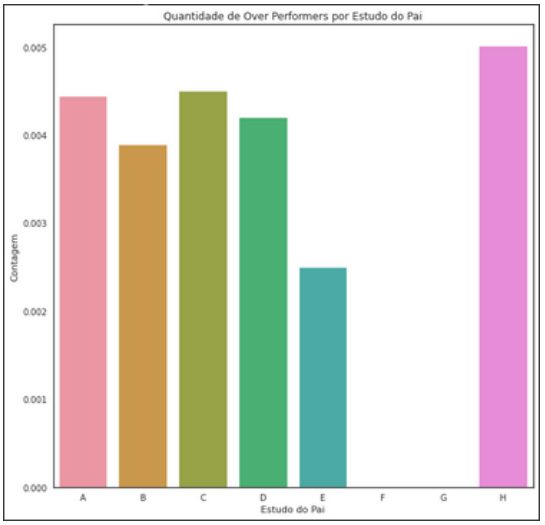
Embora esperasse que a categoria B tenha a maior quantidade de participantes, são as categorias E e G que se superaram nas notas.

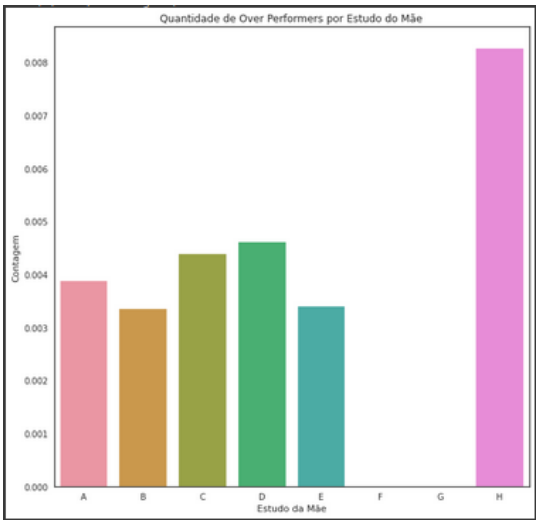


Entre os underperformers a hipótese se confirma, são os alunos da categoria mais alta de renda que têm notas abaixo do esperado.

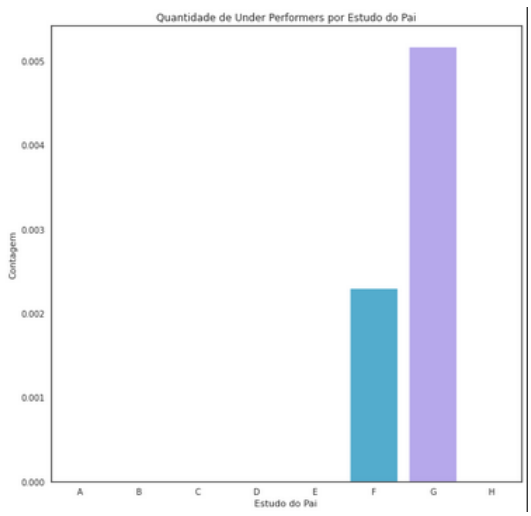


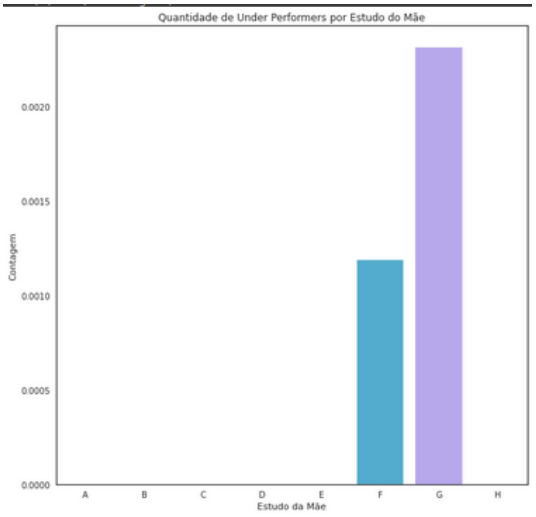
A partir do estudo dos pais, em que a maior parte deles fazia parte da categoria E. Normalizando os dados conseguimos perceber que a maior parte dos over performers se concentra na categoria H, que diz não saber qual o nível de estudo dos pais. Outros indicativos desses gráficos são que participantes que têm os pais nas categorias com níveis mais baixos de escolaridade, tem uma performance melhor do que esperado. Isso pode indicar que os pais têm influenciado e apontado para seus filhos o quanto é importante estudar.



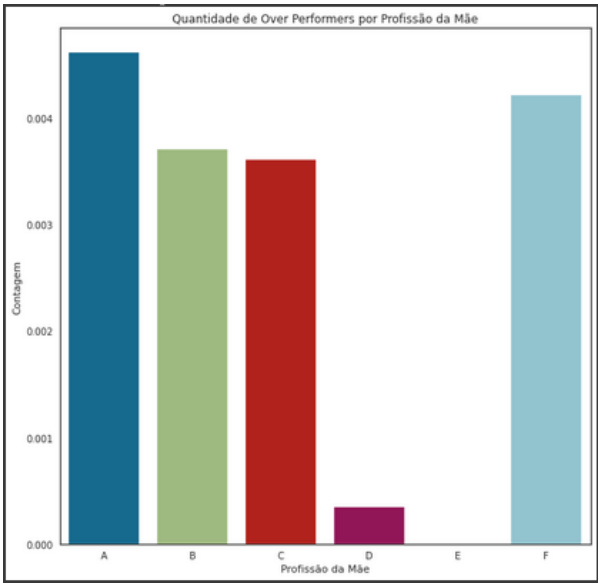


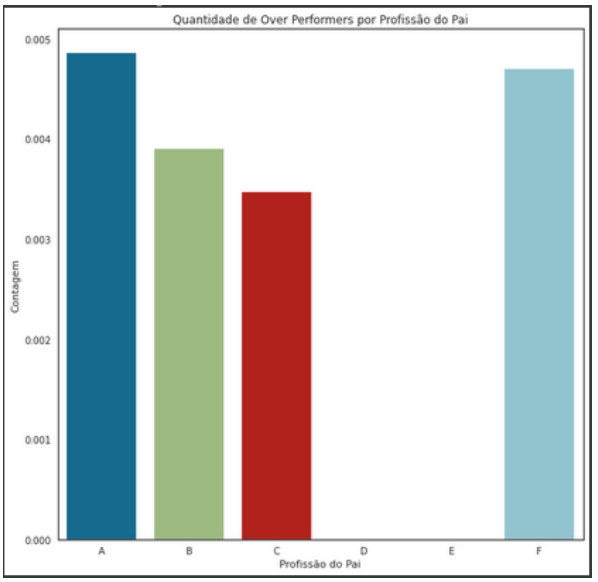
Entre os under performances se mantém as categorias onde se esperava mais, que são participantes que têm pais com níveis de escolaridade mais altos. Algumas das hipóteses para isso é que esses alunos não se importam tanto com a prova do ENEM, que para alguns é a única possibilidade de cursar uma universidade.



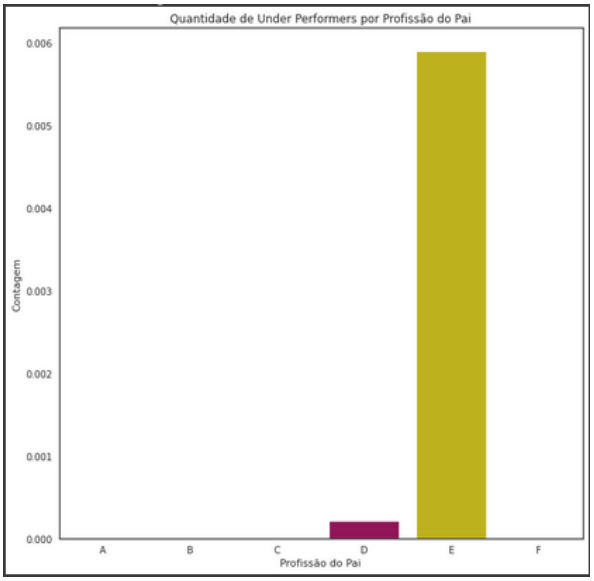


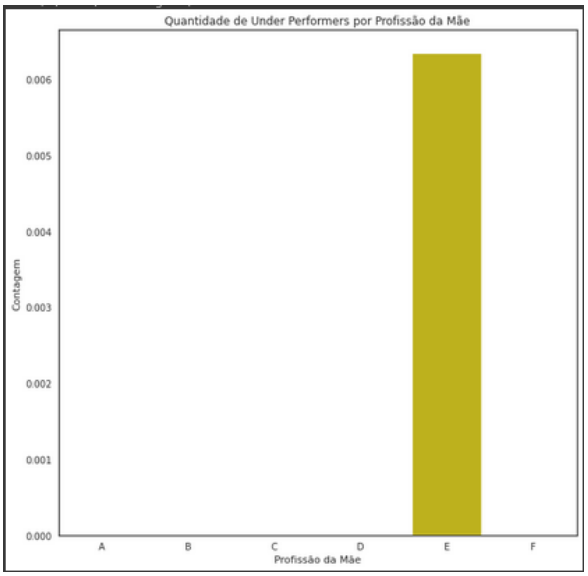
As categorias de Profissões dos pais apresentam a mesma característica das análises de ensino. Para os over performers, todas de profissão menos renomadas apresentam uma porcentagem de over performers, com destaque para as categorias A e F, sendo que F não sabiam responder qual era a profissão dos pais.



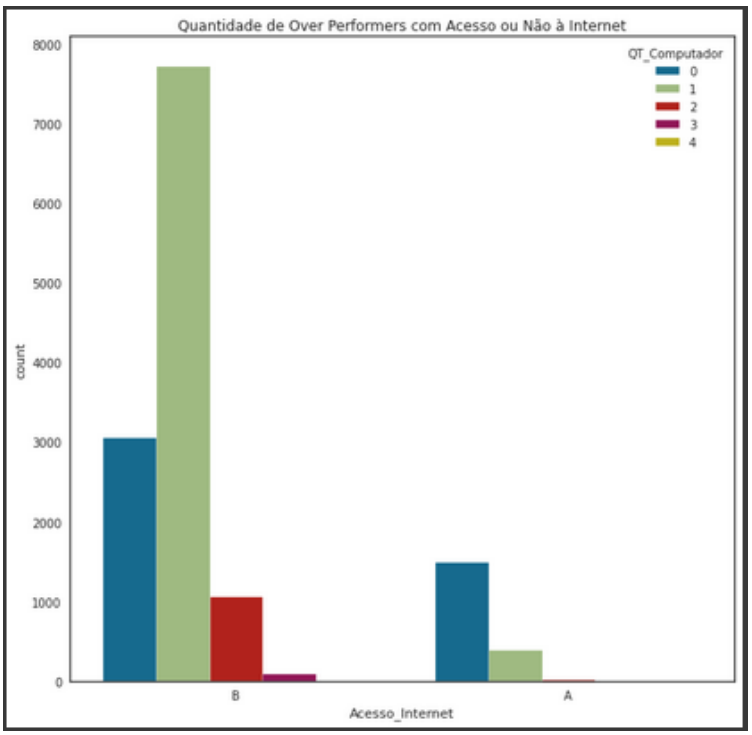


- Os underperformers dominam na categoria mais alta de profissões.





Outra análise feita foi para verificar alunos que não tinham acesso à internet e também não tinham computadores, lembrando que estávamos em período de pandemia e o ensino estava totalmente remoto. Quase 2 mil alunos performaram melhor do que a expectativa para essas condições.



Com essas análises foi possível entender um pouco como as questões socioeconômicas influenciam nas notas e quais as classes mais afetadas. Embora sejam pontos muito importantes para que os participantes estudem com qualidade e façam uma prova tranquila, foi possível concluir que não são pontos decisivos para a nota final.
Tendo em vista que 2020 era o primeiro ano de pandemia e que o ensino teve que ser totalmente transformado, os participantes foram melhores que no ano de 2019. Com destaque para os over performers, participantes que se superaram. Dentre os outliers eles foram mais que o dobro dos underperformers.

Limitações/ Trabalhos Futuros:
Um dos pontos limitantes foi imaginar que as variáveis socioeconômicas poderiam explicar bem mais sobre a nota dos alunos, pelas hipóteses criadas esperava-se que o R2 fosse maior, mas já era sabido que as questões socioeconômicas são apenas uma parte do que pode compor a nota final do participante. Seria interessante ter mais algumas variáveis para poder explicar o desenvolvimento do aluno até chegar na prova. Um dos pontos que não foi possível utilizar no trabalho foi o tipo de ensino cursado (particular/privado), o que poderia agregar muito para a análise e também para o modelo.

Além disso, surgiram algumas dúvidas quanto a veracidade dos dados, e se realmente os participantes responderam com sinceridade o questionário. Por exemplo, participantes que têm uma renda significativa, mas não possuem computador e nem internet, isso com certeza pode ser uma opção da pessoa, ou talvez morar em algum lugar remoto. Mas fica a dúvida quanto a sinceridade dos participantes no questionário.