



Disciplina:	Introdução a Ciência de Dados
Estudantes:	Angela Favretto Pastorello - 1196243 Bruna Oenning Amador - 2241463 Daniel de Viveiros Inácio - 1923820 Matheus Biscaya Gutierrez - 1657470
Equipe:	Dados Educação

PROJETO DE ANÁLISE

1. Tema

O tema do projeto é 'Analisar o impacto da pandemia na educação através da análise das desistências do ENEM', sendo o objetivo principal avaliar o quanto a pandemia impactou nas desistências do ENEM e buscar a relação dos resultados com os aspectos socioeconômicos, identificando quais obtiveram maior influência.

2. Perguntas de pesquisa

Com relação ao Questionário Socioeconômico apresentado nos Dados do ENEM, foi questionada 'Qual a influência do acesso à internet, computadores, geladeira, carro, entre outros fatores socioeconômicos sobre as desistências?'.

Com relação aos aspectos socioeconômicos foi disposta a pergunta inicial 'Qual a influência dos aspectos socioeconômicos do município/estado nas desistências?'.

Por fim, para a análise com dados do COVID, as perguntas iniciais foram 'A situação da pandemia na semana antes da prova está relacionada com as desistências no município?' e 'O quão relacionada estava a situação da pandemia durante o período antes da prova com os aspectos socioeconômicos dos municípios?'.

3. Hipóteses

Dadas às perguntas mencionadas na Seção 2, foram descritas uma ou mais hipóteses a serem exploradas:

- Qual a influência do acesso à internet, computadores, geladeira, carro, entre outros fatores socioeconômicos sobre as desistências?



- Hipótese: Aspectos que refletem a capacidade do estudante de ter acesso a informação e locomoção são os fatores que mais pesam na taxa de desistência.
- Qual a influência dos aspectos socioeconômicos do município/estado nas desistências?
 - Hipótese: Os aspectos socioeconômicos do município/estado onde a pessoa estudou tiveram maior influência nas desistências do que o município/estado em que a pessoa realizou a prova;
 - Hipótese: Os aspectos socioeconômicos do município/estado tiveram uma influência significativa nas desistências.
- A situação da pandemia na semana antes das provas está relacionada com as desistências no município?
 - Hipótese: Municípios que tiveram uma grande quantidade de casos na semana antes da primeira prova apresentaram maior desistência.
- O quão relacionada estava a situação da pandemia durante o período antes da prova com os aspectos socioeconômicos dos municípios?
 - Hipótese: Os aspectos socioeconômicos dos municípios estão relacionados de forma significativa com a quantidade de casos e mortes per capita.

4. Dados e modelos

Serão utilizados os seguintes dados para responder as perguntas e/ou testar as hipóteses:

- Dados do Enem¹: Serão utilizados os dados da escola e do local de aplicação da prova de cada participante, dados da presença nas provas e, por fim, dados do questionário socioeconômico.
- Produto Interno Bruto (PIB)²: Será utilizado o PIB per capita por municípios/ estados.
- Índice de Desenvolvimento Humano Municipal (IDHM)³: Serão utilizados os dados do IDHM para análises por municípios e estados.

¹ <https://www.gov.br/inep/pt-br/acao-a-informacao/dados-abertos/microdados/enem>

² <https://www.ibge.gov.br/estatisticas/economicas/contas-nacionais/9088-produto-interno-bruto-dos-municipios.htm>

³ <https://www.br.undp.org/content/brazil/pt/home/idh0.html>



- Dados sobre o COVID-19⁴: Será utilizada a quantidade de casos e mortes per capita para cada município.
- Desemprego⁵: Serão utilizados os dados de desocupação e subutilização da força de trabalho por estados.

Os modelos escolhidos pela equipe para a realização do projeto foram gráficos de mapa e distribuição dos dados para a visualização, regressão linear e logística para a análise estatística e, por fim, o classificador para a aplicação de conceitos de *machine learning*.

Dessa forma, para a etapa de visualização, vão ser utilizados gráficos de mapa, para auxiliar nas análises geográficas, e a distribuição dos dados, para auxiliar nos testes estatísticos.

O modelo de regressão logística será utilizado para analisar os fatores socioeconômicos que influenciaram os indivíduos a desistirem do ENEM no ano de 2020. Por outro lado, a regressão linear será feita em nível municipal e estadual e visa verificar a relação entre os fatores econômicos e os dados da pandemia com o aumento das desistências.

Por fim, o classificador foi escolhido para avaliar o risco de desistência de estudantes e, a partir disso, viabilizar medidas para diminuir estes casos, como entrar em contato com estudantes que obtiverem altos riscos de desistência.

5. Cronograma

Os próximos passos a serem executados pela equipe são:

- Distribuição dos dados com base na análise exploratória (Todos) - 5 horas;
- Aplicar regressão linear para os aspectos socioeconômicos (Bruna e Angela) dos municípios/estados e para os dados do COVID (Matheus) com a quantidade de desistências dos municípios/estados - 10 horas;
- Aplicar regressão logística para dados do questionário socioeconômico do ENEM (Daniel) - 5 horas;

⁴ <https://www.gov.br/saude/pt-br>

⁵

<https://www.ibge.gov.br/estatisticas/sociais/populacao/9221-sintese-de-indicadores-sociais.html?edicao=29143&t=resultados>



-
- Realizar todas as etapas necessárias para utilização do classificador (Todos) - 20 horas;
 - Relatório final (Todos) - 5 horas;
 - Montar a apresentação final (Todos) - 3 horas.