Introdução

Com a crescente popularidade da Internet, um grande número de s dados faz parte da vida cotidiana na WWW (World Wide Web). Big data é o termo usado para nomear essa enorme quantidade de dados. Diante de uma quantidade tão grande de dados que são produzidos todos os dias, abre-se uma janela para explorá-los, as informações úteis escondidas por trás de todo esse volume de dados. Esse estudo de dados é chamado de ciência de dados.

“O Exame Nacional do Ensino Médio (Enem) foi instituído em 1998, com o objetivo de avaliar o desempenho escolar dos estudantes ao término da educação básica” (INEP, 2021). O Enem significa para muitos alunos a porta de entrada para acesso ao ensino superior gratuito de qualidade em instituições públicas selecionado pelo Sistema Unificado (SISU). Após a inscrição, os participantes devem preencher um formulário socioeconômico. Além dos dados socioeconômicos, outras informações são fornecidas sobre os participantes, por exemplo, idade, sexo, cidade de residência, tipo de escola que terminou o ensino médio, etc. Os dados deste formulário foram fornecidos pelo Inep e são a matéria-prima para realização desse projeto.

Ciência de dados

Como outros campos técnicos, a ciência de dados tem um ciclo de vida relacionados ao seu projeto. O ciclo de vida da ciência de dados não segue um único padrão para todos os projetos, cada trabalho tem suas necessidades e requisitos específicos. Por conta disso, é comum que diferentes trabalhos utilizem diferentes manifestações deste ciclo de vida.

Neste trabalho, algumas atividades e fases do ciclo de vida da ciência de dados são essenciais para a conclusão do todo projeto de dados. Estas etapas são:

- Problema de Negócio: pode ser considerada uma das etapas mais importantes do ciclo. O entendimento do problema a ser resolvido a partir dos dados permite a seguir métodos que auxiliam as organizações a tomarem as melhores decisões e alcançar os resultados desejados.
- Coleta de dados: É aqui que a extração de dados ocorre após a definição do problema. Os dados são divididos em estruturados e não estruturados e podem vir de diferentes fontes, como por exemplo, planilhas, arquivos de texto, áudio, vídeo, API independente, etc.
- Análise Exploratória: A partir da análise dos dados é possível identificar padrões e relações assim como também levantar hipóteses a respeito deles. Para isso, é fundamental que se tenha uma "soft skill" analítica

bem desenvolvida com objetivo de gerar “insights” e agregar valor para empresa.

- Análise profunda de dados: É nesta fase que os modelos preditivos, estatísticas e técnicas de aprendizado de máquina são aplicados para testar as hipóteses propostas.
- Comunicação de resultados e Feedback: Pode-se agora encerrar o estudo que foi realizado graças aos resultados que estão sendo divulgados nessa etapa.

Estatística descritiva

O uso da estatística descritiva facilita a compreensão e a descrição dos dados.

- Média: o valor médio dos dados.
- Mediana: Os dados devem estar em rol, ou seja, ordenados de forma crescente, esse valor divide os 50% dos dados menores dos 50% dos dados maiores.
- Moda: o valor que mais se repete em um conjunto de dados.

Distribuições

Embora valiosas, as estatísticas descritivas podem ocultar detalhes cruciais sobre a amostra que está sendo analisada. A média de um conjunto de dados será distorcida se incluir valores significativamente maiores do que outros e não pode ser considerada uma representação precisa dos dados.

Um histograma pode ser usado para representar uma distribuição. É Um tipo de gráfico de barras que exibe uma distribuição de frequência. O boxplot é outro um tipo de gráfico que pode ser usado para examinar a distribuição de dados. É necessário compreender os quartis para descrever o que é um boxplot.

Os quartis 25, 50 e 75, que correspondem ao primeiro, segundo e terceiro quartis, respectivamente, são usados para representar os dados mostrados no boxplot. O segundo quartil compreende 50% da amostra, também chamado de mediana. Podemos ver outliers do conjunto no boxplot.

Materiais e Métodos

Esta seção descreve as ferramentas que foram usadas para coletar dados do Enem e lidar com esses dados para concluir as investigações sugeridas.

Linguagem de programação

A linguagem de programação Python foi utilizada neste projeto. O Python possui muitas bibliotecas que auxiliam muito na realização das pesquisas necessárias, por isso sua praticidade foi o principal fator na escolha dessa linguagem de programação. O desenvolvimento do estudo fez uso das seguintes bibliotecas:

- **Pandas**: é um pacote Python focado em análise de dados. Com essa biblioteca é possível fazer a leitura, tratamento e processamento dos dados.
- **Matplotlib:** é uma biblioteca para criar gráficos.
- **Seaborn:** Outra biblioteca que ajuda na criação gráfica é a Seaborn, normalmente têm um layout mais atraente do que os gráficos produzidos pelo matplotlib.
- **Numpy:** Para realizar cálculos matemáticos rapidamente, use essa biblioteca. É utilizado principalmente para cálculos de matrizes multidimensionais.
- **Sklearn:** Essa biblioteca auxilia na aplicação de algoritmos de “Machine Learning”.

O Colab, também conhecido como Google Collaboratory, serviu como ambiente de execução. É um ambiente de execução virtual que pode ser acessado por meio de seu navegador da Web que usa poder de processamento e recursos dos servidores do Google para lidar com grandes conjuntos de dados em PCs que não possuem hardware tão poderoso.

Obtenção dos dados

Há um repositório de dados de todas as versões anteriores do Enem no site do Instituto Nacional de Estudos e Pesquisas Educacionais Ansio Teixeira (INEP), que pode ser consultado em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem . Para este estudo, utilizamos os dados do exame de 2021.

Múltiplas colunas no arquivo csv servem para especificar várias características administrativas do exame, como dependência administrativa da escola, cor do teste, exigência de ajustes de acessibilidade, etc. Usando análise descritiva, as colunas mais pertinentes do DataFrame, com objetivo de identificar o perfil dos participantes sob os aspecto socioeconômico e

geográfico foram filtrados. As colunas consideradas para este estudo estão listadas tabela abaixo.

Tabela 1 – Colunas utilizadas no estudo

Nome da coluna	Descrição
NU_INSCRICAO	Número de inscrição
TP_FAIXA_ETARIA	Idade
TP_SEXO	Sexo
TP_COR_RACA	Cor/raça autodeclarada
TP_ESCOLA	Tipo de escola do Ensino Médio
NU_NOTA_CN	Nota da prova de Ciências da Natureza
NU_NOTA_CH	Nota da prova de Ciências Humanas
NU_NOTA_LC	Nota da prova de Linguagens e Códigos
NU_NOTA_MT	Nota da prova de Matemática
NU_NOTA_REDACAO	Nota da prova de redação
Q001	Até que série seu pai, ou o homem responsável por você, estudou?
Q002	Até que série sua mãe, ou a mulher responsável por você, estudou?
Q006	Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)
Q022	Na sua residência tem telefone celular?
Q024	Na sua residência tem computador?
Q025	Na sua residência tem acesso à Internet?

A retirada de entradas vazias do DataFrame, ou seja, de inscritos que não participaram de uma ou mais fases do exame, é necessária porque o objetivo deste trabalho é validar conexões entre as características dos participantes e o desempenho.

Conclusão

A aplicação de conceitos de ciência de dados a dados educacionais do Brasil serviu como estrutura para este trabalho. Seu objetivo foi usar métodos estatísticos e ciência de dados para saber mais sobre os candidatos ao Exame Nacional do Ensino Médio de 2021 a partir do “dataset” que o Inep cedeu. Os resultados dos dois primeiros experimentos nos permitiram concluir que nem a idade e nem o sexo dos participantes são determinantes para seu desempenho. O tipo de escola em que o participante concluiu o ensino médio foi objeto da análise seguinte. Foi possível confirmar que os alunos que frequentaram escolas particulares tiveram melhor desempenho em todas as provas do Enem. Por fim, a partir de dados históricos dos inscritos no Enem, foram utilizadas técnicas de aprendizado de máquina para criar uma linha de tendência relativa ao número de participantes ao longo do tempo. Apesar de usar um pequeno conjunto de dados, foi capaz de obter uma visão geral do padrão de matrículas para os próximos anos. Essas análises são cruciais porque podem prever dados que ajudarão no sucesso dos próximos testes, como preocupações sobre a adequação da infraestrutura para todos os inscritos.