Relatório – Trabalho Final (Parte 1)

Disciplina: Ciência de Dados

Especialização: Ciência de Dados

Equipe: Lays de Freitas Melo

Base de dados utilizada: Subconjunto do Censo Escolar 2024 (MEC/Inep) focado em infraestrutura tecnológica.

1. Business Understanding

1.1 Negócio por trás dos dados

O conjunto de dados representa o setor de **Educação**, especificamente o levantamento da infraestrutura de tecnologia (desktops, notebooks, tablets) disponível para alunos em escolas de educação básica no Brasil.

1.2 Objetivos estratégicos da organização

A organização detentora dos dados (ex: Ministério da Educação) possui os seguintes objetivos:

- **Promover a Inclusão Digital:** Garantir que estudantes de todas as regiões e classes sociais tenham acesso a ferramentas tecnológicas para potencializar o aprendizado.
- **Otimizar a Alocação de Recursos:** Identificar as escolas e regiões com maior carência de equipamentos para direcionar investimentos públicos de forma mais eficiente e reduzir desigualdades.

1.3 Problema de negócio

O problema de negócio central é a **grande desigualdade na distribuição de recursos tecnológicos** entre as escolas brasileiras. É fundamental entender quais fatores (como localização, tipo de administração, etc.) estão associados à disponibilidade de equipamentos para os alunos.

1.4 Tradução para Ciência de Dados

O problema pode ser abordado como um desafio de **Regressão**, onde buscaríamos prever a quantidade de dispositivos (ex: `QT_TABLET_ALUNO`) com base em características da escola, como sua região e dependência administrativa. Alternativamente, poderia ser um problema de **Clusterização** para agrupar escolas em categorias de "bem equipada", "moderadamente equipada" и "mal equipada".

1.5 Hipóteses iniciais

- **H1:** Escolas de dependência Federal possuem, em média, mais recursos tecnológicos por aluno do que escolas Estaduais ou Municipais.
- **H2:** Escolas localizadas nas regiões Sul e Sudeste são mais bem equipadas tecnologicamente do que as escolas das regiões Norte e Nordeste.

1.6 Restrições

- **Qualidade dos Dados:** O censo representa a quantidade declarada de equipamentos, mas não informa sobre a qualidade, o estado de conservação ou a taxa de uso real desses dispositivos. Os dados são um retrato de um ano específico e podem estar desatualizados.
- **Tempo:** O projeto possui um cronograma acadêmico definido para sua conclusão.

1.7 Critério de sucesso

Para um modelo de regressão, o critério de sucesso será a obtenção de um modelo com baixo erro de predição (RMSE) e, crucialmente, que seja **interpretável**, permitindo identificar os fatores de maior impacto na quantidade de equipamentos.

1.8 Métricas de avaliação

As métricas utilizadas para avaliar o modelo de regressão seriam o **RMSE (Root Mean Squared Error)** para medir o erro médio das previsões e o **R^2 (Coeficiente de Determinação)** para medir o quanto o modelo explica a variabilidade dos dados.

2. Data Understanding

2.1 Coleta

- **Dados disponíveis:** O conjunto de dados trabalhado (`df_infra_tec`) possui 215.545 registros e 6 colunas.
- **Origem dos dados:** A base de dados é pública, derivada do Censo Escolar, disponibilizado pelo Inep/MEC.
- **Principais variáveis:** As variáveis categóricas são `TP_DEPENDENCIA`, `NO_REGIAO`, `TP_LOCALIZACAO`. As numéricas são `QT_DESKTOP_ALUNO`, `QT_COMP_PORTATIL_ALUNO`, `QT_TABLET_ALUNO`. A análise com a biblioteca `Fitter` mostrou que a distribuição de `QT_TABLET_ALUNO` é altamente assimétrica, com forte concentração em zero, semelhante a uma distribuição Exponencial ou `halflogistic`.

2.2 Exploração

- **Valores faltantes:** Sim, foi identificada a necessidade de tratar valores faltantes (NaN), sendo aplicada a remoção de linhas onde a variável de interesse era nula.
- **Outliers:** A análise da distribuição de `QT_TABLET_ALUNO` revelou uma cauda longa à direita, indicando a presença de outliers (escolas com um número muito elevado de tablets) que precisam ser analisados.

- **Correlações:** Uma análise de correlação entre as variáveis numéricas e entre as numéricas e a variável-alvo seria um próximo passo importante para entender as relações entre elas.
- **Sugestões preliminares:** A análise exploratória e os testes de amostragem sugerem que há uma desigualdade significativa na distribuição de equipamentos. A grande maioria das escolas parece ter poucos ou nenhum tablet.
- **Limpeza necessária:** Os dados que precisam de tratamento são os registros com valores faltantes e os outliers identificados.

3. Data Preparation (Parte 1)

3.1 Seleção de atributos

Inicialmente, todas as 6 variáveis do `df_infra_tec` são consideradas relevantes para a modelagem.

3.2 Tratamento de registros

Registros com dados faltantes na variável-alvo foram descartados. Uma verificação de dados duplicados também seria necessária.

3.3 Subconjuntos de dados

Sim, o desenvolvimento de um modelo de Machine Learning, será essencial dividir os dados em conjuntos de **treino, validação e teste**.

3.4 Outliers

Os outliers de quantidade de equipamentos poderiam ser tratados com **transformação logarítmica** para reduzir sua influência ou com a técnica de **clipping** (limitar os valores a um teto máximo).

3.5 Criação de novas variáveis (feature engineering)

Uma variável útil a ser criada seria a `QT_TOTAL_DISPOSITIVOS`, somando as colunas `QT_DESKTOP_ALUNO`, `QT_COMP_PORTATIL_ALUNO` e `QT_TABLET_ALUNO` para ter uma visão geral da infraestrutura.

3.6 Codificação de variáveis categóricas

Sim, será necessário aplicar **One-Hot Encoding** nas variáveis `TP_DEPENDENCIA`, `NO_REGIAO` и `TP_LOCALIZACAO` para que possam ser utilizadas por algoritmos de Machine Learning.

3.7 Normalização/padronização de variáveis numéricas

Sim, as variáveis numéricas (quantidades de equipamentos) estão em escalas diferentes e precisarão ser padronizadas, por exemplo, com o **StandardScaler**.

3.10 Conjunto final

Após a execução das etapas acima (limpeza, feature engineering, codificação e normalização), o conjunto de dados estará preparado e refletirá o problema de negócio, pronto para ser usado no treinamento de modelos de Machine Learning.