1. Introdução

Este documento tem como objetivo descrever a etapa de **Entendimento dos Dados**, conforme a metodologia **CRISP-DM**, abordando todas as atividades necessárias para compreender os dados disponíveis, garantir sua qualidade e prepará-los para a modelagem.

1.1 Objetivo

Este projeto, tem como objetivo desenvolver um modelo de Machine Learning capaz de prever o preço estimado de corridas dos serviços Uber e 99taxi entre cada categoria de cada serviço. Além disso, o projeto visa incluir dados de outras empresas (como **99** e **transporte público**) para estimar a forma mais barata de chegar ao destino desejado.

2. Coleta e Compreensão dos Dados

2.1 Fontes de Dados

Os dados foram fornecidos pela empresa Khipo conforme os documentos fornecidos pela Fecap para o desenvolvimento do nosso PI conforme descrito nos documentos do projeto e estão distribuídos em quatro tabelas principais:

- **ride.csv** – Contém informações detalhadas sobre as corridas.
- **rideestimative.csv** – Apresenta estimativas de preços para as corridas.
- **product.csv** – Armazena dados sobre as categorias de serviço disponíveis.
- **rideaddress_v1.csv** – Inclui informações sobre os endereços associados às corridas.

2.2 Carregamento dos Dados

Cada uma dessas tabelas será carregada e tratada individualmente para garantir que os dados sejam analisados de forma isolada, permitindo uma compreensão melhor e mais detalhada de cada tabela. Esse processo visa manter a consistência e a integridade das informações, evitando sobreposição de dados e nos entregando segurança e consequentemente garantindo a qualidade sobre os dados manipulados.

3. Descrição e Exploração dos Dados

Após a etapa de carregamento de dados, serão realizadas análises exploratórias para compreender a estrutura e características de cada tabela. Essa etapa inclui:

✔ Listagem das colunas disponíveis e seus respectivos tipos de dados para conseguirmos entender as informações.
✔ Contagem e identificação de valores nulos para determinarmos a qualidade dos dados e necessidade de tratamento de cada coluna ou linha.
✔ Preenchimento de colunas vazias com valores apropriados ou através da média com base em dados da coluna.
✔ Verificação dos dados para identificar possíveis inconsistências.

3.1 Descrição e Exploração dos Dados

Abaixo vamos descrever a mineração de dados que realizamos em uma das tabelas (**rideaddress_v1.csv**) e que vamos manter como um modelo para as demais que recebemos da Khipo, com as devidas alterações para cada uma das tabelas.

Descrição das Colunas

A tabela **rideaddress_v1.csv** contém informações sobre endereços associados às corridas, incluindo colunas como:

- **Neighborhood** – Bairro associado ao endereço.
- **City** – Cidade associada ao endereço.
- **State** – Estado associado ao endereço.
- **Street** – Rua associada ao endereço.
- **Number** – Número do imóvel
- **Lat e Lng** – Coordenadas geográficas (latitude e longitude).
- **RideAddressID e RideID** – Identificadores para referenciar as corridas.

Análise de Valores Nulos

A análise revelou uma quantidade significativa de valores nulos nas colunas Neighborhood, City, State, Street e Number. Após uma análise mais detalhada, foi possível constatar que:

- A coluna **Number** apresentava um alto índice de nulos e como não representação uma informação relevante para o objetivo final do projeto decidimos remover a coluna.
- As colunas **Neighborhood**, **City**, **State** e **Street** foram tratadas usando uma combinação de técnicas de preenchimento, conforme descrito mais abaixo na etapa de tratamento e preparação inicial dos Dados.

4. Tratamento e Preparação Inicial dos Dados

4.1 Estratégia de Tratamento Utilizada

A estratégia de tratamentos que utilizamos foram basicamente para tratar os valores nulos e preparar os dados para modelagem, conforme demonstramos abaixo:

1° Preenchimento de valores nulos com "Desconhecido"

- Utilizado para as colunas Neighborhood, City, State e Street, como alternativa segura para evitar a perda excessiva de dados.

2° Remoção de colunas irrelevantes ou com excesso de dados ausentes

- A coluna **Number** foi removida, pois não contribui para a análise e apresentava um alto índice de nulos.

4° **Remoção de linhas com valores nulos restantes**

- Após as tentativas de preenchimento e inferência, as linhas que ainda continham valores ausentes foram removidas.

4.2 Resultados Após o Tratamento

- Todas as colunas relevantes foram preservadas.
- O número de valores nulos foi significativamente reduzido.
- As colunas preenchidas por inferência agora apresentam dados consistentes e prontos para serem utilizadas no processo de modelagem.

5. Verificação da Qualidade dos Dados

Após o tratamento, realizamos uma verificação completa para garantir que:

Os dados foram carregados corretamente e estão completos.
As colunas possuem valores coerentes e consistentes.
A remoção e preenchimento de dados nulos foi conduzida de forma segura e precisa.
As colunas irrelevantes foram removidas para evitar ruído nos modelos de previsão.

6. Conclusão

A primeira etapa de compreensão e tratamento dos dados foi concluída com êxito. A tabela rideaddress_v1.csv foi limpa e preparada para integração com os demais conjuntos de dados, e tomaremos as mesmas regras para as demais tabelas fornecidas.

Com todas as tabelas limpas podemos gerar apenas uma com todos os dados relevantes para o treinamento do machine learning com dados seguros e com uma boa qualidade para que o estudo tenha êxito.