1. Introdução

Este documento descreve a etapa de preparação dos dados para o modelo de previsão de preços de corridas por aplicativo (Uber, 99), conforme os princípios do CRISP-DM.

1.1 Objetivo

O objetivo é consolidar, limpar, transformar e integrar os dados fornecidos pela empresa Khipo em uma tabela final derivada com alta qualidade para alimentar o modelo de Machine Learning.

2. Tabelas Utilizadas

- **Ride_v2:**
  Informações gerais da corrida (data, status, etc.).
- **Rideaddress_v1:**
  Endereços e coordenadas de origem e destino.
- **Rideestimative_v3:**
  Estimativas de preço de produtos para a corrida.
- **Product.:**
  Informações de cada tipo de produto ofertado.

3. Etapas de Preparação

3.1. Seleção dos Dados:

- Escolha das colunas relevantes de cada base: horário, coordenadas, endereços e estimativas de produto.

3.2. Limpeza e Uniformização:

- Conversão de **RideID** para string
- Tratamento de coordenadas (vírgula → ponto, tipo float)
- Remoção de 2 linhas com dados nulos essenciais

3.3 Derivação de Dados

- Colunas temporais a partir de **'Schedule': Dia, Hora, Minuto, Hora Decimal, Faixa15min**

- Cálculo de **'Distancia_km' com geopy (distância geodésica entre origem e destino)**

3.4 Integração dos Dados

- Concatenação de **dfTempo, dfCoords e dfEstimadaSelecionada** com base em **RideID comuns**.
- Garantia de consistência dos dados por interseção de chaves

3.5. Formatação Final:

- Arredondamento de coordenadas
- Conversão de tipos e reordenação de colunas
- Base final nomeada **dfDerivado**

4. Resultado

A tabela dfDerivado contém:

- RideID (identificador da corrida)
- Dados temporais derivados
- -Coordenadas e endereços de origem e destino
- Informações do produto selecionado
- Distância geodésica estimada da corrida

Conclusão

A etapa de preparação dos dados foi conduzida com sucesso, garantindo que todas as informações relevantes estivessem limpas, consistentes e organizadas para análises e modelagens futuras. A base final derivada (dfDerivado) reflete uma integração completa entre os dados de corrida, localização e estimativas de preço, com enriquecimento por variáveis temporais e cálculo de distância geográfica.

Esse processo permitiu transformar dados brutos e dispersos em um conjunto único, confiável e pronto para alimentar algoritmos de aprendizado de máquina com o objetivo de prever valores de corrida com maior precisão.

A consistência dos dados foi assegurada por meio da filtragem por RideID, conversão e padronização de formatos, além da derivação de variáveis estratégicas como HoraDecimal, Faixa15min e Distância_km. A base está apta para suportar análises exploratórias, construção de modelos preditivos e visualizações que darão suporte à tomada de decisões orientadas por dados.