

MBA⁺

**ARTIFICIAL INTELLIGENCE
& MACHINE LEARNING**


MBA⁺

PROGRAMANDO IA COM R

Prof. Elthon Manhas de Freitas

elthon@usp.br

2018

A high-angle, wide shot of a busy New York City street, likely Times Square, filled with yellow taxis and pedestrians. The scene is crowded, with cars and people filling the frame. The text 'Trabalho final da disciplina' is overlaid in white, semi-transparent font across the center of the image.

Trabalho final da disciplina

Corridas de táxi de NY

Orientações gerais

- Grupos de 4 alunos
- Vale 9 pontos na nota da disciplina
- Entrega em até 15 dias a partir do final da disciplina.
 - 11 de junho de 2018.
 - A entrega deve ser feita pelo portal do aluno.
- A entrega é composta por
 - Conjunto de dados obtidos externamente e utilizados para enriquecimento
 - Conjunto de scripts utilizados para análise
 - Conjunto de Notebooks explicando as análises
 - Em R-Markdown
 - Aplicação Shiny que permite a navegação nos notebooks
 - Pacote R para carregamento e distribuição

- O objetivo é verificar se o grupo
 - É capaz de fazer análises estatísticas utilizando o R
 - Expor os resultados de suas análises
 - Fazer regressões simples
 - Interpretação básica de uma regressão
 - Modelar problemas para aprendizado de máquina
 - Criar scripts complexos em R
- Estima-se pelo menos 40 horas de trabalho por grupo
 - 10 horas por integrante

Orientações – Parte 1 (Descrição)

- Dataset utilizado em competição para determinar o tempo de duração de uma corrida de táxi em NYC.
- Há dois datasets disponíveis
 - train.csv – com 1.458.644 registros de corrida de táxi
 - test.csv – contém 625.134 registros de corrida de táxi
- Para o trabalho de análise, é necessário utilizar apenas o primeiro dataset.
 - A utilização do segundo dataset é opcional.

● Campos do dataset:

- **id** – chave única de cada corrida
- **vendor_id** – Código do provedor da informação
- **pickup_datetime** – Hora que a corrida se iniciou
- **dropoff_datetime** – Hora que a corrida se encerrou
- **passenger_count** – Quantidade de passageiros na corrida
- **pickup_longitude** – Longitude do início da corrida
- **pickup_latitude** – Latitude do início da corrida
- **dropoff_longitude** – Longitude do final da corrida
- **dropoff_latitude** – Latitude do final da corrida
- **store_and_fwd_flag** – Indicada se o veículo armazenou o dado temporariamente antes de sincronizar com a central (ou seja, veículo sem conexão com a internet)
- **trip_duration** – Duração da corrida, em segundos

NEW YORK CITY MAP

FIAP

LEGEND

- Major Road
- Other Road
- Golf Course
- Park/Vegetation
- Point of Interest
- Shopping Center
- Museum
- Monument
- Hospital



Orientações – Parte 1 (enriquecimento)

- O grupo precisa enriquecer o dataset:
 - Calcular a distância de cada corrida em quilômetros.
 - É necessário calcular tanto a distância euclidiana quanto a distância de manhattan
 - Como latitude e longitude são dados contínuos, é preciso discretizar em quadrantes. Cada quadrante não pode ter mais do que 20 metros quadrados.
 - Cada quadrante deve estar associado a 1 ou mais pontos de interesse (estações de metrô, pontos turísticos, etc.). São pedidos pelo menos 15 pontos de interesse.
 - Opcional: Cada viagem pode ter a relação de quadrantes intermediários entre o início e o fim da corrida.

- Análise exploratória inicial:
 - Faça uma análise exploratória indicando
 - Os principais horários das corridas
 - Distinção por dia da semana? E por dia do ano? Por hora do dia?
 - As principais origens em função do horário
 - Os principais destinos em função do horário
 - O tempo médio da viagem em função do horário

- Filtrar o dataset para análise sobre um conjunto menor de dados, que contenha pelo menos 5000 observações.
- Algumas análises podem ser feitas sob o “minidataset” escolhido. Quando isso acontecer, informar na análise.
 - Sugestões:
 - Corridas saídas da Broadway em uma determinada época do ano
 - Corridas de fora da ilha para dentro da ilha nos finais de semana
 - etc.

Orientações – Parte 1 (Análises gráficas) FIAP

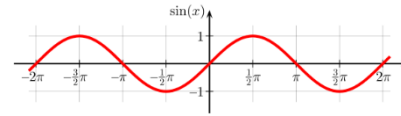
- Espera-se que a entrega possua pelo menos 1 análise gráfica de cada tipo:
- Gráfico de linha temporal por mês
- Gráfico de linha temporal por dia da semana
 - agregar por soma ou média
- Gráfico de linha temporal por hora do dia
 - Quebrar as horas do dia em 15 minutos, ou seja, 96 “quartos de hora”
- Clusterização (aprendizado não supervisionado)
- Mapa de calor

Obs.: é necessário que cada gráfico venha acompanhado de uma interpretação.

Orientações – Parte 1 (Modelagem ML)

- Definir as variáveis de entrada e as variáveis saída de um modelo.

- Considerar variáveis temporais cíclicas



- Posição geográfica

- Proximidades a pontos de interesse



- Representar graficamente a modelagem
 - pode usar qualquer ferramenta, como o powerpoint

MBA⁺

Copyright © **2018**

Prof. Elthon Manhas de Freitas

Todos direitos reservados. Reprodução ou divulgação total ou parcial deste documento é expressamente proibido sem o consentimento formal, por escrito, do Professor (autor).