Insper

Aprendizagem Estatística de Máquinas II
Projeto Final - Análise de dados

A análise de dados será dividida em duas partes

1. Seleção de modelos para um problema que envolva métodos de aprendizagem supervisionada (regressão ou classificação).
2. Análise de um problema não supervisionado.

Requisitos:

- Toda análise deve ser feita em R.
- A análise supervisionada deverá ser feita utilizando as funções do pacote tidymodels, com o objetivo de comparar pelo menos 3 modelos preditivos distintos. Será considerado ponto extra na entrega para quem incluir na análise a modelagem de uma rede neural (aqui pode-se usar o pacote keras ao invés do tidymodels).
- A análise não supervisionada pode envolver um problema de agrupamento, redução de dimensionalidade, análise de texto ou análise de sentimentos.

Considere escolher um base de dados em que podem ser aplicados tanto modelos supervisionados quanto não supervisionado.

Datas importantes:

Entrega 1 (21/11): definição dos integrantes do grupo (até 3 pessoas), da escolha dos dados que serão usados nas análises. Apresentação, de até 15 minutos, da(s) base(s) escolhida(s) e da abordagem utilizada para implementação da solução. Submeter arquivo Excel "Projeto Final - Entrega 1.xlsx" devidamente preenchido.

Entrega 2 (04/12): relatório final contendo descrição dos dados utilizados, objetivos de análise, decisões tomadas no processo de modelagem, as comparações entre os modelos considerados e o modelo final. Apresentação de até 20 minutos.

Critérios de avaliação:

Entrega 1:

- Arquivo Excel template preenchido corretamente.
- Apresentação (será avaliada a clareza na definição dos objetivos e a metodologia a ser implementada).

Entrega 2:

- Um relatório final bem estruturado incluindo uma contextualização dos dados utilizados e objetivos da análise, análise exploratória dos dados e gráficos que

fazem sentido, explicação das decisões importantes que foram tomadas (remoção de dados aberrantes, input de dados faltantes, etc), estudo comparativo dos modelos preditivos e apresentação da conclusão sobre o modelo mais adequado e interpretação da performance do modelo escolhido. Além da apresentação da análise não supervisionada e interpretação dos resultados.

Escolha das bases de dados:

Os grupos poderão utilizar os conjuntos indicados no fim desse arquivo ou considerar algum outro conjunto de interesse. Além disso, é permitido enriquecer os dados apresentados com fontes externas de informações.

Para a análise de dados supervisionada, a sugestão é utilizar o conjunto de dados **bisnode** que foi usado na atividade integradora do primeiro trimestre.

Dica: escolha uma base de dados em que seja possível aplicar tanto uma solução para um problema supervisionado quanto um não supervisionado.

Anexo fontes de dados sugeridas:

- https://github.com/rfordatascience/tidytuesday
- https://www.kaggle.com/
- http://basedosdados.org/
- https://github.com/rfordatascience/tidytuesday
- https://painel.seade.gov.br/repositorio-de-dados/
- http://www.ipeadata.gov.br/

Algumas bases de dados selecionadas:

Airbnb
Dados de valores de diárias em imóveis do Airbnb em várias cidades.
Fonte: http://insideairbnb.com/get-the-data.html

Coffe ratings
Dados sobre avaliações de café com informações como país de origem, sabor, aroma, acidez etc. O objetivo dessa análise é predizer a pontuação (total_cup_points) ou agrupar cafés semelhantes. Essa base conta com 1330 avaliações e 42 preditoras.
Fonte: https://github.com/rfordatascience/tidytuesday/tree/master/data/2020/2020-07-07

Bank marketing
Dados sobre uma campanha de marketing direto de uma instituição portuguesa. A campanha é baseada em ligações telefônicas. Essa base conta com aproximadamente

45.000 observações e 16 preditoras. O objetivo é predizer se os clientes subscreveram depósitos a prazo.
Fonte: https://www.kaggle.com/aakashverma8900/portuguese-bank-marketing

Injury prediction for competitive runners
Dados sobre treinamentos de corredores no período de 2012 a 2019. Essa base de dados conta com informações de aproximadamente 42.800 treinamentos e 72 preditoras. O objetivo é predizer lesões de acordo com o perfil do treinamento.
Fonte: https://www.kaggle.com/shashwatwork/injury-prediction-for-competitive-runners

Viena – hotel Booking
Dados sobre preços e características de hotéis em Viena coletados em novembro de 2017. Essa base de dados conta com informações de 430 hotéis. O objetivo é predizer o preço de uma diária ou fazer um sistema de recomendação.
Fonte: https://gabors-data-analysis.com/

Loan data
Dados sobre 9.578 empréstimos e pagamento. O objetivo é obter um bom classificador para pagamento/não pagamento do empréstimo. Esse conjunto de dados conta com 13 preditoras.
Fonte: https://www.kaggle.com/itssuru/loan-data

Hospital Sirio-Libanês
Dados sobre 1.925 indivíduos com informações demográficas, exames de sangue, sinais vitais e doenças prévias. O objetivo dessa análise é criar um classificador para indicar se o paciente necessitará de cuidados de unidade de tratamento intensiva.
Fonte: https://www.kaggle.com/S%C3%ADrio-Libanes/covid19

Municípios de São Paulo
Dados sobre renda, IDH, número de domicílios, população (masculina, feminina, urbana, rural e grau de urbanização dos municípios), PIB e saúde de São Paulo. O objetivo é agrupar municípios similares em relação às características que você julgar interessante. Nessa análise você pode adicionar informações de interesse sobre esses municípios.
Fonte: https://repositorio.seade.gov.br/group/seade-municipios