Detalhes sobre o Projeto Final

Saul Leite

02 dezembro, 2024

Objetivo

O objetivo do projeto final é fazer síntese do que foi apresentado durante a disciplina em um problema com dados reais. O projeto será realizado por um grupo de até **6 integrantes**.

Os integrantes do grupo devem buscar um conjunto do *tidytuesday* que seja de seu interesse. O problema pode ser de regressão ou classificação. Os dados podem ser encontrados no link abaixo:

https://github.com/rfordatascience/tidytuesday.

Os dados acima geralmente não possuem suas colunas classificadas como variáveis preditoras e de saída. Pense em um problema de predição que seja interessante para os dados que escolheu. Também serão aceitos trabalhos usando os bancos de dados disponibilizados pelo Kaggle: https://www.kaggle.com/datasets.

Atenção: Escolha o problema e os dados que vão utilizar com cuidado, discuta no grupo os problemas que vocês podem resolver. Dê preferência para problemas interessantes com várias variáveis e de **aprendizado supervisionado**. Além disso, vocês devem fazer suas **próprias análises**, e não se basearem *em análise feitas por outras pessoas na internet.*

Observação: O número máximo de grupos disponíveis é 14, caso contrário, não teremos tempo para todas as apresentações.

Métodos

Os passos esperados para o desenvolvimento do projeto são os seguintes:

1. Fazer limpeza dos dados, agregar as tabelas (em alguns casos, os dados estão dispostos em tabelas diferentes, use comandos como `inner_join` para a junção, que funcionam como para bancos de dados, veja detalhes em https://dplyr.tidyverse.org/reference/join.html). Muitas vezes os dados do tidytuesday já possuem código de limpeza e carregamento dos dados, vocês podem aproveitar esse material.
2. Explicar os dados que estão utilizando, identificando do que se tratam e fazendo uma **breve** análise exploratória. Por exemplo, identificando os tipos de variáveis, contínuas, qualitativas nominais, e ordinais. Identifique dados faltantes e remova-os de sua base. Boas ferramentas para este estágio são: `skimr` e `ggpairs`.
3. Prepare os dados com receitas para o treinamento. Vocês podem fazer seleção de características ou redução de dimensionalidade se achar necessário.
4. Comparar **no mínimo 3 algoritmos** de regressão ou classificação (dependendo do problema) e justificar suas escolhas. Por exemplo, dizer qual foi a razão da escolha destes algoritmos e porque acreditam que eles iram funcionar bem para o seus dados. Vocês podem usar os algoritmos apresentados em aula ou outros que acharem interessantes.
5. Fazer ajuste de hiper-parâmetros para ajustar os algoritmos aos dados da melhor forma possível, justificando a abordagem escolhida.

6. Comparar os algoritmos e definir o melhor algoritmo para o problema escolhido. Justificar a métrica utilizada na comparação dos algoritmos.

O projeto deve ser desenvolvido em repositório associado à atividade no **Github Classroom**. O link para atividade está abaixo:

https://classroom.github.com/a/OlpaHjkf

Material para Entregar

Os integrantes do grupo devem preparar:

1. Um relatório em RMarkdown com a análise e o texto explicativo dos passos utilizados durante a análise. É importante a explicação dos passos tomados em conjunto com a justificativa e contextualização. Relatórios contendo praticamente *código puro* ***não serão bem avaliados***. O relatório final deve ser entregue em formato RMarkdown e em PDF.
2. Arquivo contendo a apresentação que será apresentada no final da disciplina. Vocês podem optar por apresentar o própria saída do relatório em RMarkdown.

Data para Entrega

O relatório deve estar pronto até o dia:

Domingo, 15 de dezembro de 2024.

A submissão será feita com o último *commit* anterior a data de entrega no **github classroom**.

Boas Sorte para Todos e Divirtam-se!