PUCPR
GRUPO MARISTA

Orientação - Atividade Somativa 1

Ao andar em grupos do Facebook e realizar algumas pesquisas no LinkedIn, você percebe que o mercado de TI está aquecido para a contratação de pessoas com conhecimento e experiência em temas ligados à inteligência artificial. Isto inclui termos como ciência de dados, *advanced analytics*, *predictive analytics*, *machine learning*, entre outros. Em um lado, existem várias pessoas e empresas que estão animadas por implementar projetos que possam utilizar essas técnicas – muitas vezes partindo de uma premissa de que *machine learning* (ML) é uma superinteligência que geralmente vemos nos filmes, gerando uma expectativa que pode estar fora da realidade. Por outro lado, ao pesquisar pelas bibliotecas e ferramentas de ML, você se depara com uma infinidade de termos, por exemplo, scikit-learn, statsmodels, pandas, koalas, Tensorflow, Keras, AdaBoost, XGBoost, SVM, LightGBM, Fairlearn, InterpretML, Watson, spaCy, Random Forest, NLTK, Seaborn, Matplotlib, Numpy, TPOT, Anaconda, JupyterLab, SHAP, LIME, MLlib, prophet, imbalanced-learn, category-encoders, MXNet, PyTorch, GAN, KNN e assim por diante. A lista é interminável!

É humanamente impossível dominarmos todas as técnicas e bibliotecas do mundo. Por outro lado, é possível entender como as técnicas de ML operam de uma forma geral para, em um segundo momento e dependendo das nossas necessidades, aprofundarmo-nos em uma ou outra aplicação em específico. Um *chef* francês certamente dominaria um conjunto de técnicas, temperos e ferramentas da cozinha bem diferentes de um *chef* japonês, ainda que ambos sejam igualmente *chefs*, não? No entanto, ambos dominariam e entenderiam, em linhas gerais, como uma cozinha e o método de preparação de pratos funcionam. A mesma lógica aplica-se aqui: não precisamos dominar todas as técnicas do mundo, mas entender em linhas gerais como as diferentes técnicas de ML funcionam.

Nesse contexto, você entrou para o processo seletivo de um programa de estágio de uma grande consultoria. Essa consultoria presta serviços para grandes empresas do mundo inteiro, o que inclui empresas de alimentação, de *streaming* de jogos, da área da saúde, do mercado financeiro, entre outros. Sendo assim, eles lhe enviaram algumas bases de dados (também chamadas de *datasets*) para que você pudesse demonstrar os seus conhecimentos ao criar algoritmos de ML destinados para elas.

Você recebeu um *e-mail* explicando que o objetivo disso é o de entender o seu domínio em python aplicado a técnicas de ML – logo, a empresa avaliará a forma pela qual você resolveu o problema, e não apenas o resultado do seu algoritmo. Consequentemente, é uma oportunidade para demonstrar o seu raciocínio, criatividade e qualidade no processo de desenvolvimento. Esse mesmo *e-mail* possui os seguintes detalhes – leia-os **atentamente**:

1. O processo seletivo é dividido em **duas partes**. O trabalho desta semana refere-se à **primeira parte**.
2. O **mesmo *dataset*** que foi utilizado na primeira parte também deverá ser usado para a segunda parte. Imagine que a segunda parte será uma continuidade da primeira.
3. Como todo processo seletivo, a avaliação não considera somente o **resultado**, mas principalmente a **forma que se chegou ao resultado**. Nesse sentido, a organização e a legibilidade do código são partes igualmente importantes do seu trabalho. Também não são toleradas cópias ou quaisquer situações que possam ser qualificadas como plágio.

Dito isso, vamos à parte I do trabalho em si. Você deverá fazer o seguinte:

1. Crie **um** *notebook* em Jupyter utilizando python. Você executará todo o trabalho dentro dele. Existem plataformas de versionamento de código como o Github e o GitLab que podem ajudá-lo nesse sentido. Por outro lado, se fizer isso, certifique-se que o seu código não está disponível publicamente.
2. Dentro desse *notebook*, carregue um dos *datasets* disponíveis abaixo.
3. Depois de carregar esse *dataset*, execute o passo de **preparação dos dados**.
    1. Você precisará **obrigatoriamente** aplicar uma técnica de seleção de atributos ou de extração de atributos.
    2. Você poderá também aplicar de forma **opcional** outras técnicas em conjunto (ex.: remoção de *outliers* e normalização) se quiser, desde que ao menos utilize também uma técnica de seleção de atributos ou de extração de atributos.
4. Depois de realizar a preparação dos dados, faça a **divisão** do seu *dataset* entre uma base de treinamento e outra de teste. A base de treinamento deverá ter 75% da base original e a base de teste terá os 25% restantes.
5. Para o *dataset* que escolheu, defina se o problema é de **classificação**, **regressão** ou previsão de **séries temporais**. Justifique a sua escolha dentro do *notebook*.
6. Depois de definir o problema, faça o **treinamento** de um algoritmo de aprendizagem supervisionada utilizando **uma técnica de uma das bibliotecas** vistas anteriormente (ou seja, scikit-learn, LightGBM, XGBoost e Prophet). Para

fazer o treinamento, utilize a base de treino (isto é, os 75% divididos anteriormente).
7. Após finalizar o treinamento, mostre a **predição** para a base de teste (isto é, os 25% restantes da sua base de dados).
8. Não se esqueça de deixar o seu *notebook* apresentável – isto é, divida o seu código em **células,** como mostramos anteriormente, de uma maneira que os diferentes blocos do seu código sejam compreensíveis. Utilize também células do tipo ***markdown*** (isto é, que permitam códigos em HTML) para explicar às pessoas que não entendam de python o que você fez durante o trabalho.
9. Entregue **dois** arquivos: o arquivo original do *notebook* (com extensão .ipynb) e o mesmo *notebook* em HTML (no Jupyter clique em "*File > Export Notebook As... > Export Notebook as HTML*" ou "*File > Download as > HTML*").

Os *datasets* que estão disponíveis para a sua escolha estão listados abaixo. Escolha apenas **um**. Dê preferência ao *dataset* com o qual tem mais afinidade/conhecimento prévio. Isso é bem importante por dois motivos: o primeiro é para entender se os dados e as previsões fazem sentido ou não de acordo com o seu bom senso. O segundo é pelo próprio estímulo: é muito melhor trabalharmos em algo de que gostamos e que conhecemos, não é? Enfim, vamos lá:

- *Dataset*: ***CS:GO round winner***
  - **Nome do arquivo**: csgo_round_snapshots.xlsx
  - **Contexto**: esse *dataset* contém o resultado de 122 mil momentos que aconteceram durante mais de 700 partidas de campeonatos de alto nível do jogo Counter-Strike: Global Offensive. Temos dados como o arsenal de ambas as equipes (CT e T), o dinheiro disponível para as duas equipes, o mapa, entre outros. A ideia é **prever** o vencedor da partida (*round_winner*).
  - **Importante**: o *dataset* foi ligeiramente alterado em relação à versão disponível na fonte original. Considere para o trabalho somente a versão do *dataset* disponibilizada pela universidade.
- *Dataset*: **NBA *players stats***
  - **Nome do arquivo**: nba_stats.csv
  - **Contexto**: o *dataset* contém dados de milhares de jogadores que passaram pela NBA desde 1950 até 2017. A ideia é **prever** o *win share* (ou seja, uma estimativa da quantidade de vitórias nas quais o jogador contribuiu – coluna *WinShares*).
  - **Importante**: o *dataset* foi ligeiramente alterado em relação à versão disponível na fonte original. Considere para este trabalho somente a versão do dataset disponibilizada pela Universidade.
- Dataset: **Seoul Bike Data**
  - **Nome do arquivo**: seoul_bike_data.xlsx
  - **Contexto**: este dataset contém a quantidade de bicicletas alugadas na cidade de Seoul (Coreia do Sul) entre 2017 e 2018. Neste sentido, existem duas opções:
    - **Prever** a quantidade de bicicletas que serão alugadas a partir de fatores como o dia da semana, temperatura, chuva e demais condições;
    - **Prever** quantas bibliotecas seriam alugadas no futuro (isto é, durante todo o mês de janeiro de 2019 – cujas informações não existem neste dataset).
    - Para ambos os exemplos a coluna que estamos buscando prever é a "*Rented Bike Count*".
  - **Importante**: o dataset foi ligeiramente alterado em relação à versão disponível na fonte original. Considere para o trabalho somente a versão do *dataset* disponibilizada pela universidade.
- *Dataset*: **NASA *airfoil self-noise***
  - **Nome do arquivo**: nasa.csv
  - **Contexto**: o *dataset* contém os resultados de diferentes testes do perfil da asa de uma aeronave da NASA. Esses testes foram realizados em um túnel de vento anecoico. A intenção deles era a de medir o nível proporcional da pressão sonora, medida em decibéis. A ideia é **prever**, a partir de outros valores medidos, com a frequência e o ângulo de ataque, essa pressão sonora (coluna "*scaled-sound-pressure*").
  - **Importante**: o *dataset* foi ligeiramente alterado em relação à versão disponível na fonte original. Considere para o trabalho somente a versão do *dataset* disponibilizada pela Universidade.

Datasets_AtividadeSomativa_Semana4 1.zip

Aponte a câmera para o código e acesse o link do conteúdo ou clique no código para acessar.