

PROJETO 3

MACHINE LEARNING

MODELOS PREDITIVOS DE VARIÁVEIS QUANTITATIVAS E SUAS PERFORMANCES

1. PROBLEMA

O principal objetivo do Projeto 3 é conduzir uma análise de dados (com descritiva e inferência) com grau elevado de autonomia e liberdade de escolha de tema (base de dados) desde que para o mesmo seja possível criar modelos preditivos que, baseados em diversos fatores (variáveis explicativas) de cada unidade amostral, faça uma previsão acerca de uma variável quantitativa (variável resposta).

Para que este fim possa ser alcançado, os estudantes deverão se aprofundar nas técnicas aqui exigidas para realização do projeto.

É importante que o trabalho produza uma conclusão analítica que utilize técnicas inferenciais e vá MUITO além de análise exploratória apenas (isso foi Projeto 1).

Ainda, é extremamente importante ter ciência e real compreensão sobre a definição de plágio tanto quando se envolve texto com código. Leia o [Manual do Aluno disponível no Portal do Aluno do Insper](#) e as regras disponíveis [aqui](#).

2. DESENVOLVIMENTO

Nesse projeto, deverá realizar diversos modelos preditivos para estimar a variável-alvo (obrigatoriamente quantitativa) considerando algumas variáveis explicativas (que podem ser qualitativas ou quantitativas) de acordo com tema escolhido.

A proposta do Projeto 3 foi inspirada em um trabalho que constrói alguns modelos preditivos de notas de redação do ENEM 2015 baseados em diversos fatores acerca de um candidato. Acesse-o [aqui](#).

Assim, seu projeto deverá conter o seguinte desenvolvimento:

A. INTRODUÇÃO

- Detalhar objetivo escolhido para trabalhar neste projeto juntamente com descrição da base de dados. Pesquise trabalhos na literatura que discutem o tema escolhido. Para trabalhos acadêmicos, acesse <https://scholar.google.com.br/>. Guarde as referências estudadas para citá-las no projeto.

B. MINERANDO DADOS e CARACTERÍSTICAS DO DATASET

- Se necessário, faça filtro na base de dados tanto de linhas como de colunas em prol do objetivo traçado anteriormente.
- Descreva as variáveis finais que serão utilizadas a partir deste ponto.
- Faça análise descritiva detalhada das variáveis com olhar ao objetivo do problema.

C. PROCESSO E ESTATÍSTICAS DE VALIDAÇÃO

- Para os modelos preditivos que serão desenvolvidos no **próximo item**, é necessário calcular medidas que informam a *performance* de cada modelo ajustado. Assim, para cada modelo preditivo, faça:
 - Divida a base de dados na parte treinamento e na parte teste. Use a parte treinamento para estimar cada modelo preditivo.

- Estude as medidas coeficiente de determinação (R^2) e raiz do erro médio quadrático (RMS) descritas [aqui](#). Calcule essas medidas tanto para a parte dos dados de treinamento como para a parte dos dados teste.
- Discuta se essas duas medidas se comportam de forma semelhante para as duas partes de dados. Leia o texto disponível [aqui](#) para compreender *overfitting* e *underfitting* e, com isso, refinar senso crítico para discutir sobre as medidas calculadas.
- Extra: Faça o processo de Validação Cruzada utilizando também 10 ciclos e calcule a *performance* média e desvio padrão das duas medidas R^2 e RMS tanto para a parte treinamento como para a parte teste. Discuta com riqueza de detalhes.

D. MODELOS DE PREDIÇÃO

- MODELO DE PREDIÇÃO PELA MÉDIA
- MODELO DE REGRESSÃO LINEAR (Multiple Linear Regression): veja, por exemplo, http://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html

E. CONCLUSÃO

- Faça conclusão final com detalhes levando em consideração todas as interpretações realizadas no decorrer do projeto.

F. REFERÊNCIAS BIBLIOGRÁFICAS

- Todas as pesquisas feitas e estudadas que foram relevantes para o desenvolvimento devem ser citadas no projeto.

Obs.: Neste projeto, pode utilizar bibliotecas prontas disponíveis no Python que façam as modelagens de predição aqui exigidas. Entretanto, é necessário explicar o que cada modelo de predição faz e também explicar como funciona a biblioteca escolhida.

3. BASE DE DADOS

Cada grupo pode trabalhar com uma base de dados de tema livre desde que a mesma ofereça uma variável quantitativa que se deseja prever levando em consideração um conjunto de variáveis. Ainda, é importante ter uma quantidade razoável de tamanho amostral já que a mesma ainda será dividida em base de dados treinamento e base de dados teste.

Seguem alguns links que pode encontrar base de dados:

<http://portal.inep.gov.br/microdados>

<ftp://ftp.ibge.gov.br/>

<https://www.gapminder.org/data/>

<https://www.kaggle.com/>

Ainda, pode utilizar outra base de dados qualquer desde que satisfaça as condições impostas no tópico **PROBLEMA**.

REGRAS:

1. O Projeto 3 é estritamente até TRIO;
2. Use o **arquivo layout** disponibilizado na pasta Projeto3 do GitHub ou Blackboard;
3. O projeto deve ser claro e organizado respeitando o que foi pedido com interpretações e/ou conclusões.

A estrutura do documento deve ser clara e de fácil compreensão da linha de raciocínio. Nesse caso, o notebook não deve haver excesso de impressões não discutidas ou de códigos não utilizados.

Após finalização do projeto, aconselhamos que sua dupla faça uma análise geral e salve com outro nome, limpe seu IPython Notebook apenas com os resultados relevantes e melhore seu texto.

4. Seu projeto deve ser adicionado no seu GitHub dentro de uma pasta chamada Projeto3.
5. O arquivo DEVE ser com extensão .ipynb

CRONOGRAMA:

DATA	Finalização:
22/05	No Github de um aluno do grupo: Publicar itens A e B descritos anteriormente. <i>É importante que as trocas de arquivos desenvolvidos pelos alunos do mesmo grupo sejam feitas pelo Github.</i>
24/05	No Github do mesmo aluno do grupo: Publicar pelo menos um Modelo de predição (item D) e medidas de performance (item C) para tal modelo preditivo. <i>É importante que as trocas de arquivos desenvolvidos pelos alunos do mesmo grupo sejam feitas pelo Github.</i>
01/06 APS 8	No Github do mesmo aluno do grupo: Publicar OS DOIS MODELOS DE PREDIÇÃO (item D) e medidas de performance (item C) para tais modelos preditivos. <i>É importante que as trocas de arquivos desenvolvidos pelos alunos do mesmo grupo sejam feitas pelo Github.</i>
05/06	PROJETO 3 FINALIZADO Fazer git push em seu Github até 23:59 do Projeto 3 finalizado.

RUBRICS DE AVALIAÇÃO DO OBJETIVO DE APRENDIZADO

Objetivo de aprendizado	Insatisfatório (I)	Em desenvolvimento (D)	Essencial (C)	Proficiente (B)	Avançado (A)
<p>Aplicar teoria de probabilidade adequadamente.</p> <p>(Especificar as distribuições de probabilidades adequadas para as variáveis quantitativas discretas e/ou contínuas)</p>	<p>Não consegue trabalhar com bases de dados de forma proficiente.</p> <p>Apresenta problemas com arquivos, formatos de arquivos ou não tem habilidades básicas de separação dos dados.</p>	<p>Descreve com riqueza de detalhes o tópico INTRODUÇÃO e MINERANDO DADOS e CARACTERÍSTICAS DO DATASET descritos no Projeto 3, porém que são já avaliados no Projeto 1.</p> <p>Fora isso, o notebook é praticamente uma coleção de comandos para demais tópicos sem fazer nenhum modelo de predição com qualidade.</p>	<p>Executa os tópicos INTRODUÇÃO e MINERANDO DADOS e CARACTERÍSTICAS DO DATASET de maneira excepcional</p> <p>Descreve com riqueza de detalhes o tópico MODELOS DE PREDIÇÃO descrito no Projeto 3 para todos .</p>	<p>Executa os comportamentos da rubrica C de maneira excepcional.</p> <p>Descreve com riqueza de detalhes o tópico PROCESSO E ESTATÍSTICAS DE VALIDAÇÃO descrito no Projeto 3 (a menos do item Extra).</p> <p>Comunica os resultados com bastante clareza.</p>	<p>Executa os comportamentos da rubrica B de maneira excepcional.</p> <p>Descreve com riqueza de detalhes tanto o item Extra do tópico PROCESSO E ESTATÍSTICAS DE VALIDAÇÃO como os tópicos CONCLUSÃO e REFERÊNCIAS BIBLIOGRÁFICAS descritos no Projeto 3.</p> <p>Expõe os resultados obtidos de maneira excepcional.</p>