



Especialização em Ciência de  
Dados  
(Big Data, Machine Learning e  
Aplicações)

DISCIPLINA: Estatística

## Projeto Prático da Disciplina

---

Neste projeto, você deverá realizar uma análise exploratória dos dados e estudo de regressão linear, a partir da utilização de dados coletados sobre postagens realizadas na rede social Facebook. Os dados encontram-se disponibilizados no Git da disciplina (nome do arquivo: fb.csv).

Para cada conjunto de dados você deve **apresentar o código Python, antecedido pela explicação sobre sua escolha de utilização do código, seguido de sua análise e interpretação dos resultados obtidos após o comando**. Trabalhos que não contenham a implementação e os comentários solicitados serão considerados incompletos.

Ao final você deve **entregar um arquivo em um Jupyter Notebook**, enviando o mesmo para o e-mail da professora Vera Medeiros ([vera.medeiros@maisunifacisa.com](mailto:vera.medeiros@maisunifacisa.com)).

É importante lembrar que este **trabalho é individual**, portanto não será tolerado qualquer tipo de plágio (i.e., comandos e comentários iguais identificados em mais de um trabalho).

O projeto deve ser **entregue até o dia 10 de outubro de 2019**. Perceba que você pode optar por enviar o seu trabalho antes desta data. Caso o trabalho não seja entregue até a data proposta, você está passível a ser reprovado na disciplina.

---

## Algumas considerações sobre a base de dados

A base proposta traz dados sobre postagens realizadas na rede social Facebook. Para facilitar a manipulação da base, as variáveis a ela pertencentes foram renomeadas e devem ser interpretadas da seguinte forma:

<b>Nome</b>	<b>Descrição</b>
Var1	Tipo de postagem (1 = foto; 2 - status; 3 - link; 4 - Vídeo)
Var2	Mês da postagem
Var3	Dia da semana da postagem (sendo 1 referente à segunda-feira e 7 referente ao domingo)
Var4	Hora da postagem
Var5	0, caso a postagem seja classificada como um anúncio pago; 1, caso contrário
Var6	Números total de usuários que receberam e visualizaram a postagem
Var7	Número total de usuários que receberam a postagem, sejam ela via propaganda, story patrocinado ou diretamente na timeline do usuário que a postou
Var8	Número total de usuários que realizaram algum tipo de interação de engajamento (like, compartilhamento ou comentário)
Var9	Númerto total de usuários que clicaram para visualizar a postagem
Var10	Número total de usuários que clicaram e realizaram interações de qualquer tipo na postagem
Var11	Número total de usuários que receberam e deram like na postagem
Var12	Número total de usuário que receberam, visualizaram e deram like na postagem
Var13	Número total de comentários
Var14	Número total de likes
Var15	Número total de compartilhamentos
Var16	Número total de interações

## Atividades

---

1. Realize um estudo inicial sobre as variáveis da amostra proposta. Aponte possíveis relacionamentos entre as variáveis estudadas, fazendo uso das estatísticas calculadas e de gráficos (scatterplot, modelo de regressão, mapa de calor).

2. A partir do estudo realizado na questão anterior, gere um modelo de regressão linear para cada variável alvo a seguir:

- A) Número total de comentários;
- B) Número total de likes;
- C) Número total de comentários;
- D) Número total de interações.

Para cada modelo gerado, aponte quais as variáveis independentes utilizadas e o motivo que o(a) levou escolhe-las. Aponte também qual a equação de regressão gerada e verifique a acurácia do modelo criado realizando previsões para a amostra\*.

3. Realize uma análise detalhada das variáveis escolhidas para cada modelo e aponte possíveis ajustes que podem ser realizados em cada um deles. Caso ache pertinente, gere um novo modelo.

\* Você pode optar por selecionar um subconjunto da amostra fornecida para realizar o treino do teu modelo de regressão, utilizando os demais dados da amostra para realização desta etapa de teste. Lembre-se apenas de verificar o valor real da saída esperada e compará-lo com a saída do teu modelo de regressão.