**FA 084 – Introdução à Mineração de Dados**

**AP 522 –Mineração de Dados**

**ATIVIDADE PRÁTICA – AULA 09**

Gustavo Boaventura Cruz, 169227

**2. PERGUNTAS PARA SEREM RESPONDIDAS.**

**2.a) Quais as dimensões do novo conjunto de dados (número de linhas e de colunas)?**

Após toda a manipulação, o novo conjunto de dados apresenta com 8565 linhas e 13 colunas.

**2.b) Qual a Média da Produtividade considerando TODOS OS ESTÁGIOS e TODAS AS VARIEDADES?**

Para a Media da Produtividade, foi utilizado o comando df.tch.mean(), onde “tch” se tratava da produtividade, resultando em 62.522490367775944 (62.52).

**2.c) Qual a Média da Produtividade DE CADA VARIEDADE?**

Para as três variedades, foram constatadas as respectivas médias:

RB855453 78.541131

RB867515 61.895681

SP813250 57.511456

**2.d) Qual a Média da Produtividade EM CADA ESTÁGIO DE CORTE?**

Para os seis estágios de corte, foram constatadas as respectivas médias:

1 73.986447

2 69.696946

3 64.364203

4 55.887034

5 53.098862

6 52.010959

**3. ATIVIDADES PARA SEREM REALIZADAS (CONSTRUÇÃO DE MODELOS).**

Para as seguintes atividades, o conjunto foi divididos em dois subconjuntos, de treino e teste, respectivamente na proporção 70/30 com random\_state = 2020.

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X,y, test\_size = 0.3, random\_state = 2020)

Foi solicitado a estimativas de modelos, sendo estes listados abaixo. Para cada modelo, foi adicionado (.append) ao conjunto para poder analisar a diferença de estimativa para cada modelo.

**3.a) Mod1: Qual a estimativa desse "modelo" para a produtividade dos talhões?**

Mod1 = 62.52249

**3.b) Mod2: Qual a estimativa desse "modelo" para a produtividade dos talhões de cada variedade?**

Variedad/ Mod2

RB855453 78.541131

RB867515 61.895681

SP813250 57.511456

**3.c) Mod3: Qual a estimativa desse "modelo" para a produtividade dos talhões em cada corte?**

Estágio/ Mod3

1 73.986447

2 69.696946

3 64.364203

4 55.887034

5 53.098862

6 52.010959

**3.d) Mod4: Qual foi o melhor K?**

Para a Técnica de KNN, analisando K entre 1 e 9 (kvizinhos = list(range(1,10,2))), K = 1 apresentou a melhor acurácia.

**3.e) Mod5: Quais foram os valores dos coeficientes do modelo para cada atributo?**

Os valores de, b1b2 para coeficientes angulares e b0 para linear, são mostrados abaixo.

b1b2 (coeficientes angulares) = ([-1.33803465e+00, 1.23369714e+00, -2.58919531e+00, -6.19616819e-01, 1.65257875e-01, -1.73354329e+00, -7.25167808e-03, 9.37743920e-03, 1.79098153e-02, -4.48489236e+00, -2.12417741e+00, 3.44460534e+00, -1.32042793e+00, 8.42891681e+00, -2.03926614e+00, -6.38965066e+00])

b0 (coeficiente linear) = 1534.5368426361156

**4. ATIVIDADES PARA SEREM REALIZADAS (AVALIAÇÃO DOS MODELOS).**

**4.a) Qual o valor de RMSE para os cinco modelos?**

RMSE Mod1: 20.525623803564347

RMSE Mod2: 19.604823007301757

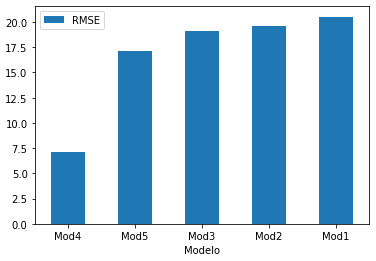
RMSE Mod3: 19.097721183557514

RMSE Mod4: 7.1396551080073865

RMSE Mod5: 17.09703551054584

**4.a.2) Gráfico de barras**

É demonstrado o gráfico de barras demonstrando a relação de RMSE para cada Modelo feito.



**4.a.3) Qual foi o pior modelo? Faça uma análise crítica dos resultados, em especial com relação ao pior modelo.**

O pior modelo se deu pelo 1 (Mod1), no qual apresentou o maior RMSE, logo como o maior erro. Como já se era esperado, pois ao calcular a média de todos os estágios com todas as variedades ao mesmo tempo, sua precisão não tende a ser das melhores em relação aos outros modelos. Os outros modelos em relação ao primeiro, foram melhores, dados com menos erros. Em especial aos que foram usados com técnicas de machine learning, como o Modelo 4, que foi usado a técnica de KNN para regressão, onde apresentou um erro muito menor em relação aos outros.

**4.b) Considere os QUATRO MELHORES modelos de acordo com os resultados obtidos no item 4.a. Renomeie-s para Modelos A, B, C e D, sendo A para o melhor, B para o segundo melhor, etc.**

**4.b.1) Apresente a relação do número do modelo para a letra (Por exemplo, Mod1=ModA, Mod2=ModB, etc.)**

ModA = y\_pred4 (Mod4)

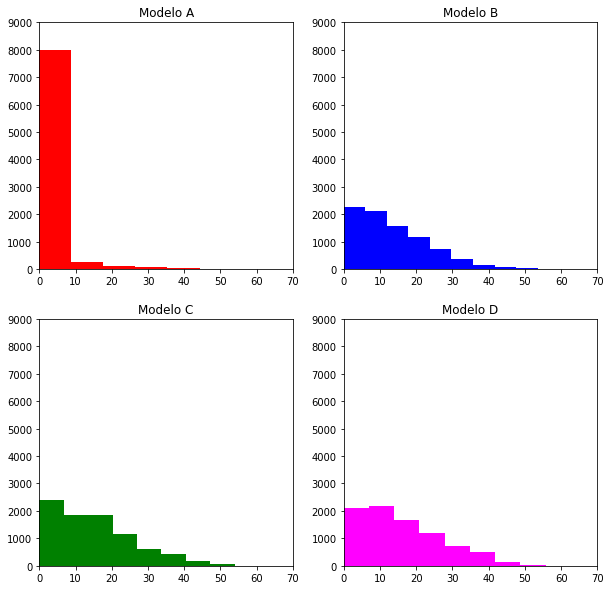
ModB = y\_pred5 (Mod5)

ModC = df.Mod3 (Mod3)

ModD = df.Mod2 (Mod2)

**4.b.2) Histograma.**

Foi feito um Histograma para os 4 melhores modelos, estes são demonstrados abaixo.



**4.b.3) Faça uma discussão crítica com relação aos gráficos apresentados no item 4.b.2.**

Foram feitos quatro histogramas, tendo analisar os erros de cada modelo. O primeiro gráfico, mostra que a ocorrência de erros de valor de 0 a 10 são muito altas, contendo a grande maioria dos dados neste intervalo, mostrando sua grande eficiência. Ainda podemos analisar os seguintes gráficos, que apresentam uma distribuição similar, isto também já era esperado, já que o erro (RMSE) de seus respectivos modelos, apresentam um valor bem próximo um dos outros.

**4.b.4) Histograma.**

Para cada modelo, foi construído um gráfico, onde traçando uma reta na diagonal (45º), podemos comparar o conjunto de dados originais com os modelos sintetizados. Vale a ressalva, que quanto mais os pontos estiverem próximos a reta diagonal, será melhor o modelo, evidenciando o modelo A como o melhor destes.

