UNIVERSIDADE FEDERAL DO CEARÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DA COMPUTAÇÃO - PPGEEC
DISCIPLINA DE RECONHECIMENTO DE PADRÕES

Aluna: Kamila Amélia Sousa Gomes
Matrícula: 516916

RELATÓRIO TRABALHO 1

Questão 1

Na questão abordada é utilizada a base de dados aerogerador. Esta base trata-se de um conjunto de dados que possui como variável de entrada: velocidade do vento(m/s) e como variável de saída: potência gerada(kWatts). Os parâmetros são estimados pelo método dos mínimos quadrados.

0.0.1 Divida o conjunto de dados em três regiões e aplique um modelo de regressão linear em cada uma delas. Procure a divisão que proporcione a maior média de R2.

A Regressão Linear é uma equação para se estimar o valor esperado de uma variável y, dados os valores de algumas outras variáveis x. Além disso, ele é capaz de quantificar a relação entre uma ou mais variáveis preditoras e uma variável de resultado.

Nesse primeiro caso, divide-se o modelo em algumas sub-regiões em que o modelo de regressão linear seja mais adequado. Para isso, foram testados alguns intervalos a fins de comparação. Além disso, para avaliar a eficiência do algoritmo, é utilizado o coeficiente de determinação R^2.

O coeficiente de determinação é uma medida de ajuste de um modelo estatístico aos valores observados de uma variável aleatória. Ele varia entre 0 e 1, onde, quanto mais próximo a 1, mais próximo aos dados reais.

A seguir são exibidos os intervalos aplicados no modelo de regressão linear.

$$[R1 : x\epsilon[0-5,5], R2 : x\epsilon[5,5-12],, R3 : x\epsilon[12-14.5] \tag{1}$$

O valor de R2 para Subdivisão 1= 0.840042

O valor de R2 para Subdivisão 2 = 0.956572

O valor de R2 para Subdivisão 3= 0.350633

$$[R1 : x\epsilon[0-4], R2 : x\epsilon[4-7], , R3 : x\epsilon[7-14.5] \tag{2}$$

O valor de R2 para Subdivisão 1= 0.939835

O valor de R2 para Subdivisão 2 = 0.941733

O valor de R2 para Subdivisão 3= 0.948497

$$[R1 : x\epsilon[0-5], R2 : x\epsilon[5-10], , R3 : x\epsilon[10-14.5] \tag{3}$$

O valor de R2 para Subdivisão 1= 0.883880

O valor de R2 para Subdivisão 2 = 0.907028

O valor de R2 para Subdivisão 3= 0.787714

$$[R1 : x\epsilon[0-3.5], R2 : x\epsilon[3.5-8], , R3 : x\epsilon[8-14.5] \tag{4}$$

O valor de R2 para Subdivisão 1= 0.979837

O valor de R2 para Subdivisão 2 = 0.894931

O valor de R2 para Subdivisão 3= 0.935676

$$[R1 : x\epsilon[0-7], R2 : x\epsilon[7-12.5], , R3 : x\epsilon[12.5-14.5] \tag{5}$$

O valor de R2 para Subdivisão 1= 0.856114

O valor de R2 para Subdivisão 2 = 0.957922

O valor de R2 para Subdivisão 3= 0.127505

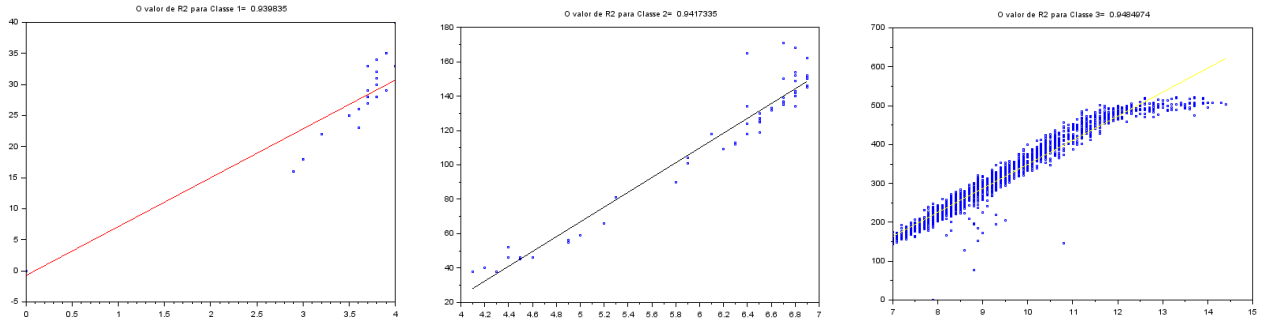
Pode-se notar, que os resultados mais eficientes devem-se aos intervalos que possuem menores quantidades de dados para a subdivisão 1 e maiores para a subdivisão 3. Logo, uma boa divisão seria o aplicado na Equação 2. A Equação 2 traz para a primeira subdivisão, valores para x menores que 4, a segunda subdivisão valores para x maiores que 4 e menores ou iguais a 7 e, por fim, x menor que 7.

Nas Figuras 1a,1b e 1c são apresentadas os gráficos para cada classe com seus respectivos coeficientes de determinação.

A Figura 1a apresenta R2 = 0.939835 para a subdivisão 1, a Figura 1b apresenta R2 = 0.941733 para a subdivisão 2 e a Figura 1c apresenta R2 = 0.948497 para a subdivisão 3. Obtendo uma média de R2 = 0,943355 ou 94%.

Figura 1 – Aplicação da Regressão Linear em Sub-Regiões.

(a) Subdivisão 1 (b) Subdivisão 2 (c) Subdivisão 3



Elaborado pela Autora (2021).

Por fim, sua representação geral é dada na Figura 2. Onde a cor vermelha refere-se a subdivisão 1, a cor preta para subdivisão 2 e cor amarela para a subdivisão 3.

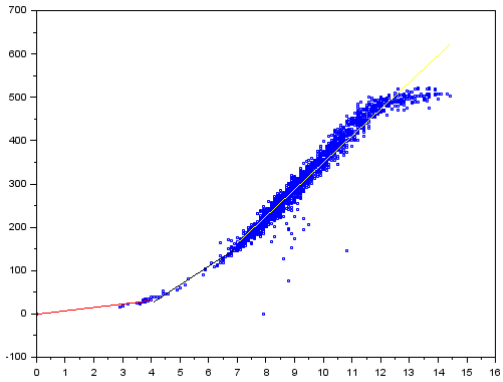


Figura 2 – Aplicação da Regressão Linear.

0.0.2 **Determine modelos de regressão polinomial (graus 2 a 7) e calcule o R2 e R2aj.**

A regressão polinomial é recomendável para tentar representar os dados que não são linearmente comportados. Nesta questão, são aplicados diferentes graus de polinômios para o conjunto de dados do aerogerador. Como mostrado na Tabela 1.

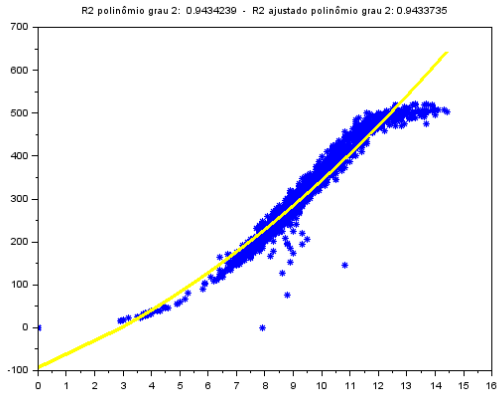
Grau do Polinômio	R2	R2 ajustado
2	0.943424	0.943374
3	0.969023	0.968982
4	0.973724	0.973677
5	0.973726	0.973667
6	0.973761	0.973691
7	0.973857	0.973776

Tabela 1 – Comparação coeficientes de determinação na regressão polinomial

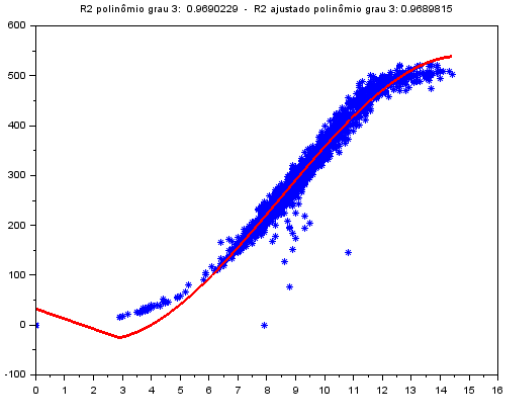
De acordo com a Tabela 1, nota-se que, ao aumentar o grau do polinômio, o resultado tende a se aproximar mais do comportamento real dos dados. Para avaliar a eficiência do algoritmo para cada grau de polinômio, é utilizado o coeficiente de determinação R2 e R2 ajustado. Na Figura 3 serão exibidos os gráficos para cada grau do polinômio na regressão polinomial.

Figura 3 – Aplicação da Regressão Polinomial.

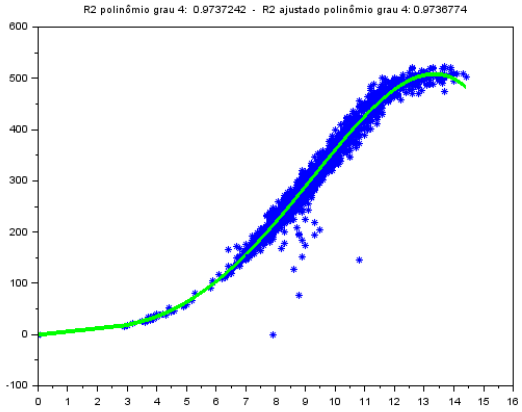
(a) Polinômio grau 2



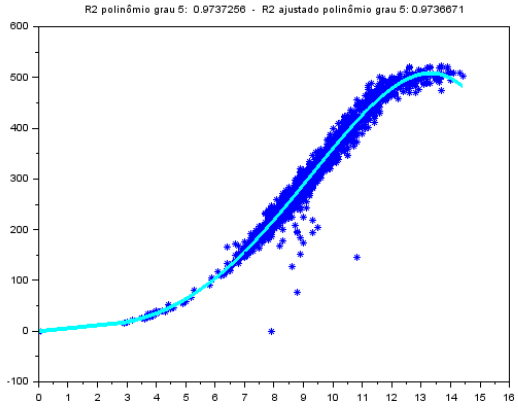
(b) Polinômio grau 3



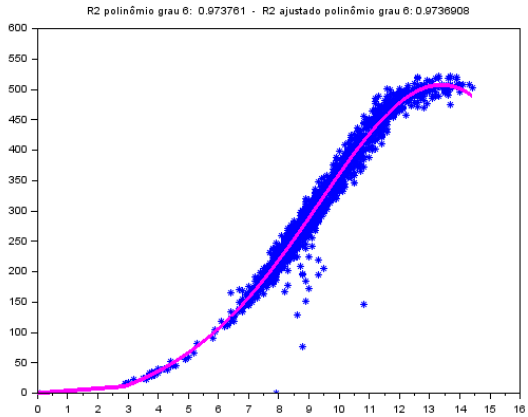
(c) Polinômio grau 4



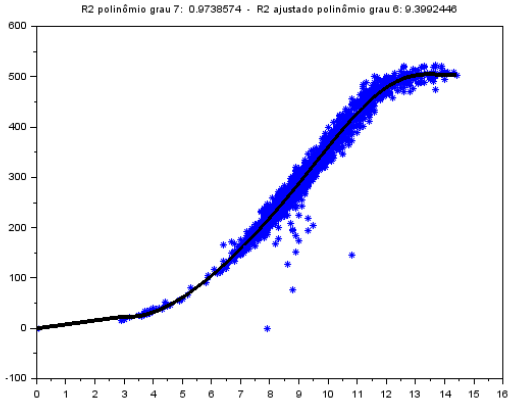
(d) Polinômio grau 5



(e) Polinômio grau 6



(f) Polinômio grau 7



Elaborado pela Autora (2021).

Observa-se, na Figura 3a, que o polinômio de grau 2 não se encaixa ao problema por apresen-

tar uma parábola. Já na Figura 3b, para o polinômio de grau 3, a função tem maior flexibilidade sobre o plano. E, para os valores a partir do grau 4, o coeficiente de determinação obteve pouca alteração. Em suma, nota-se que, quanto maior o termo polinomial, mais a reta se adequa a base de dados.

Por fim, a Figura 4 mostra a Regressão Polinomial em seu comportamento de maneira geral.

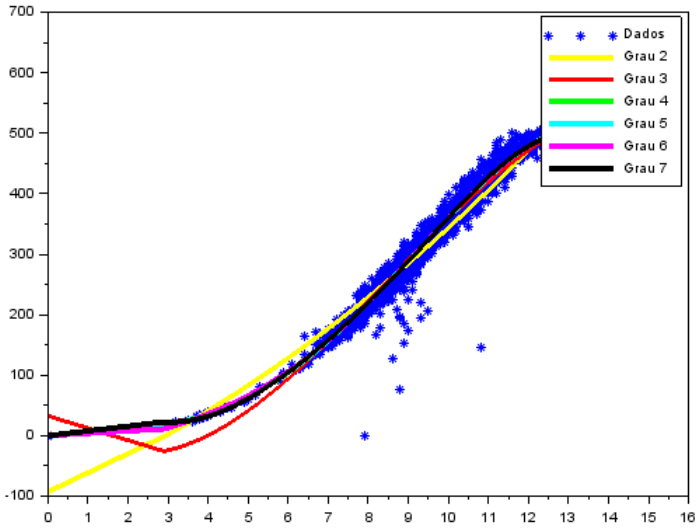


Figura 4 – Regressão Polinomial - Grau 2 a 7

Questão 2

0.0.3 Implemente os métodos LDA e sua versão naive para classificar a base de dados Dermatology. Adotar hold-out com subsampling (20 execuções) e leave-one-out.

O LDA (da sigla, Análise Discriminante Linear), é um método estatístico utilizado para encontrar um combinação linear de recursos que caracterizam duas ou mais classes de objetos. A combinação resultante pode ser usada como um classificador linear. O Naive Bayes é um classificador usado para dados que contenham valores de recursos contínuos. Além disso, o Naive Bayes assume variáveis como independentes, enquanto LDA assume modelos gaussianos de densidade condicional de classe. Nesta questão, eles serão implementados na versão Leave-one-out e Hold-out com subsampling.

Para a execução desta questão, foi utilizado banco de dados dermatologia, disponível em: https://archive.ics.uci.edu/ml/datasets/Dermatology. Ele é constituído pelo diagnóstico de doença de pele com base em informações clínicas (coletadas pelo médico no consultório) e informações hispatológicas (resultantes de uma biópsia – análise do tecido em um laboratório de patologia). O conjunto de dados possui 6 classes, 34 atributos e 366 amostras. Como mostrado na Tabela 2.

Classes	Quantidade de amostras	Patologias
1	111	Psoríase
2	60	Dermatite Seborréica
3	71	Líquen plano
4	48	Pitiríase rósea
5	48	Dermatite Crônica
6	20	Rubrar Pilar

Tabela 2 – Informações sobre o conjunto de dados dermatologia.

Na Tabela 2, não são quantificados 366 amostras, pois existem amostras com valores do tipo *Nan* (nulos), o que deixa o conjunto de dados com 358 amostras. Para isso, após a leitura da base de dados, houve uma etapa de pré-processamento, que excluiu os valores nulos da mesma. Em seguida, a base foi permutada aleatoriamente para não haver enviesamento.

A versão *Leave-one-out* é uma implementação de validação cruzada com o nome bem sugestivo: deixar um de fora. Ou seja, o aproximador de função é treinado em todos os dados, exceto para um ponto e uma previsão é feita para esse ponto. Neste caso, têm-se:

Acurácia LDA Versão Naive utilizando Leave-one-out: 95.810056%

O algoritmo pode ser encontrado em: lda-leaveoneout.sce.

Acurácia LDA utilizando Leave-one-out: 96.368715 %

O algoritmo pode ser encontrado em: ldanaive-leaveoneout.sce.

O Método *Hold-Out* é um método simples de validação cruzada, que consiste na ideia básica de dividir o conjunto de dados de treinamento em duas partes, ou seja, treinar e testar o modelo. A subamostragem aleatória realiza 'k' iterações de todo o conjunto de dados. O modelo é ajustado ao conjunto de treinamento de cada iteração e uma estimativa do erro de predição é obtida de cada conjunto de teste. Neste caso, serão efetuados 20 execuções.

Para a efetuação do hold-out com 20 execuções, foi utilizado o método *grand*, método de geração de números aleatórios, que tem como argumento o tamanho da matriz desejada, a distribuição *'prm'* (permutações aleatórias) e a matriz/vetor. Inicialmente, este método também foi utilizado para permutar a base aleatoriamente.

Durante as execuções, a **Média de Acurácia LDA Versão Naive com Hold-Out** variou entre 90% e 100%. E a **Média de Acurácia LDA com Hold-Out** variou entre 80% e 100%. Os algoritmos estão disponíveis em *lda-holdout.sce* e *ldanaive-holdout.sce*.

OBS: Questões implementadas na linguagem de programação Scilab.