Análise de Desempenho de Algoritmos de Aprendizagem Automática

Análise de Dados em Informática – 2º Trabalho Prático

Hugo Daniel Gonçalves Fernandes
Departamento de Engenharia Informática
ISEP
Porto, Portugal
1161155@isep.ipp.pt

Norberto João Gomes Lopes de Sousa
Departamento de Engenharia Informática
ISEP
Porto, Portugal
1120608@isep.ipp.pt

***Abstract*— O presente documento descreve a implementação de um sistema de deteção de Diabetes e subsequente análise dos dados resultantes.**

Keywords*— *algoritmos, aprendizagem automática, CART, LDA, SVM, KNN

I. INTRODUÇÃO

Este documento, realizado no âmbito da cadeira "Análise de Dados em Informática" da Licenciatura em Engenharia Informática do ISEP, tem como objetivo a aplicação de técnicas estatísticas de análise de dados no contexto de análise de desempenho de algoritmos. Estes algoritmos fazem parte de um sistema de deteção de Diabetes em mulheres com herança indígena Pima.

Os Pima são um grupo de nativos americanos que vivem no Arizona. Uma predisposição genética permitiu que este grupo sobrevivesse normalmente com uma dieta pobre em carboidratos durante anos. Recentemente, devido a uma mudança acentuada de dieta e nível de atividade física, existe um aumento de ocorrências de diabetes do tipo 2 e por esta razão foram objeto de vários estudos [1].

II. ESTADO DA ARTE

Este sistema de deteção de Diabetes utiliza os seguintes algoritmos de aprendizagem automática:

- Classification and Regression Tree (CART).
- Linear Discriminant Analysis (LDA).
- Support Vector Machines with Radial Basis Function Kernel (SVM).
- K-Nearest Neighbors (KNN).

Estes algoritmos têm em comum o facto de serem algoritmos de aprendizagem automática supervisionada [2]. Este nome deve-se à semelhança existente entre o seu processo de treino e a relação professor–aluno. Numa primeira fase o algoritmo é treinado usando dados de entrada e saída válidos, o objetivo é aprender o processo de chegar a respostas corretas. Após essa aprendizagem supervisionada terminar, o algoritmo é capaz de aplicar o processo que aprendeu a novos dados.

Os algoritmos de aprendizagem automática supervisionada podem ainda ser separados em dois grupos diferentes, de acordo com os problemas que tratam, classificação e regressão cujos objetivos são respetivamente classificar valores de entrada conforme categorias e identificar valores concretos.

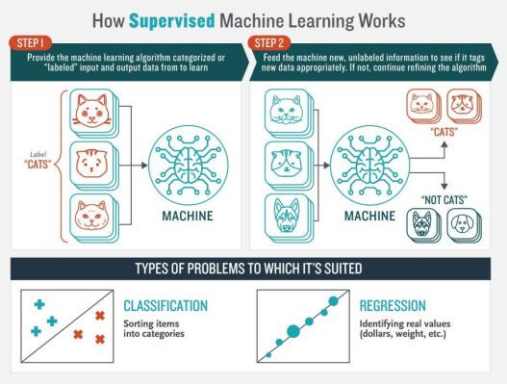

Figura 1 - Funcionamento de algoritmos de aprendizagem automática supervisionada [3].

A. Classification and Regression Tree (CART)

O modelo CART utiliza uma árvore binária para classificar ou prever valores dos dados de entrada. Este modelo gera a árvore binária que vai utilizar criando pontos de divisão nos dados de entrada. Após isso a classificação é um processo simples e direto [4].

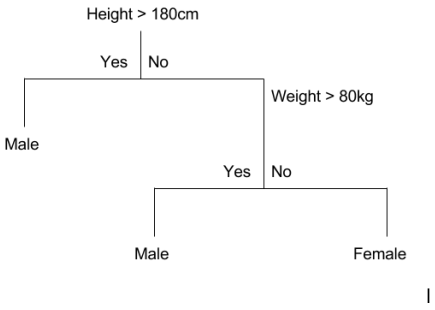

Figura 2 - Árvore binária [4].

B. *Linear Discriminant Analysis (LDA)*

LDA é uma técnica de redução de dimensionalidade usada como uma etapa de pré-processamento em aprendizagem automática e aplicações de classificação de padrões. O objetivo principal de remoção de dimensionalidade é atingido ao remover dados redundantes [5].

Esta abordagem apresenta alguns limites:

- Instável com classes bem separadas - A regressão logística pode tornar-se instável quando as classes estão bem separadas.
- Instável com poucos exemplos – A regressão logística pode tornar-se instável quando há poucos exemplos para estimar os parâmetros.

C. *Support Vector Machines with Radial Basis Function Kernel (SVM)*

O SVM é um algoritmo de aprendizagem automática supervisionada que pode ser usado para problemas de classificação ou regressão, apesar de ser usado principalmente em problemas de classificação. Neste algoritmo, representamos cada dado como um ponto num espaço n-dimensional (onde n é o número de classificações existentes). Depois executa-se a categorização encontrando o hyper-plane que diferencia melhor as duas classes [6].

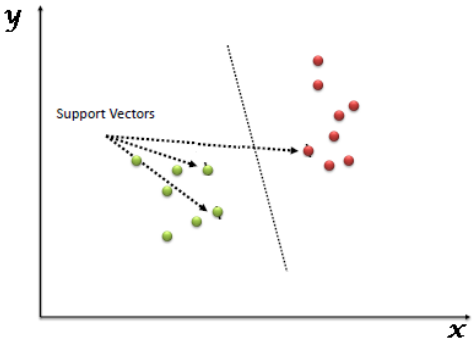

Figura 3 - Vetores de suporte e Hyper-Plane [6].

D. *K-Nearest Neighbors (KNN)*

O algoritmo KNN assume que coisas semelhantes existem em proximidade. Ou seja, coisas semelhantes estão próximas umas das outras. O algoritmo depende deste pressuposto ser verdadeiro a ponto de o algoritmo ser útil. O KNN captura a ideia de similaridade calculando a distância entre os pontos num gráfico [7].

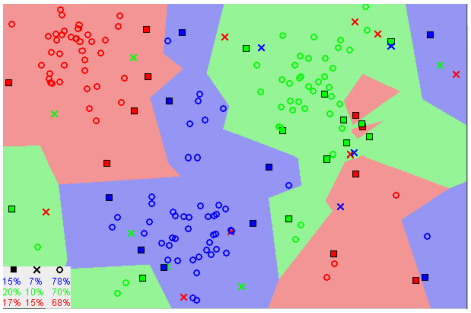

Figura 4 - Pontos de dados semelhantes próximos uns dos outros [7].

III. Análise de desempenho De Algoritmos De Aprendizagem automática

A. *Contextualização do Trabalho*

No âmbito do desenvolvimento de um sistema que possibilita a deteção da Diabetes em mulheres descendentes dos indígenas Pima, pretende realizar-se a análise de desempenho de múltiplos algoritmos, nomeadamente: Classification And Regression Tree (CART), Linear Discriminant Analysis (LDA), Support Vector Machines with Radial Basis Function Kernel (SVM) e k-Nearest Neighbors (KNN).

Para realizar a análise de desempenho, testou-se cada algoritmo com um conjunto de dados provenientes de 11 grupos de mulheres, onde cada grupo é constituído por 100 mulheres, existindo 8 variáveis independentes e uma variável dependente (resposta) estas variáveis são:

- Pregnant – Número de gravidezes.
- Glucose – Concentração de Glicose após 2 horas (teste oral de tolerância à Glicose).
- Pressure – Pressão arterial diastólica (mmHg).
- Tríceps – Espessura da prega cutânea no tríceps (mm).
- Insulin – Insulina sérica após 2 horas (mU/ml).
- Mass – IMC, Índice de Massa Corporal (kg/m^2).
- Pedigree – Função de pedigree de diabetes, representa a probabilidade de obter a doença extrapolando a história dos antepassados.
- Age – Idade (anos).
- Diabetes – (Variável dependente) apresenta “pos” em caso de a inquirida ter diabetes tipo 2, e apresenta “neg” no caso de não ter a doença.

Cada algoritmo para cada grupo que analisa são determinados alguns indicadores que permitem calcular o seu desempenho através da comparação da previsão com o diagnóstico, nomeadamente:

- Verdadeiros Negativos (TN) – a mulher não possui indícios da doença e o algoritmo indicou que não teria indícios.
- Falsos Negativos (FN) – a mulher tem indícios da doença, contudo o algoritmo indicou que ela não teria.
- Falsos Positivos (FP) – a mulher não tem indícios da doença, mas o algoritmo indicou que teria.
- Verdadeiros Positivos (TP) – a mulher tem indícios da doença e o algoritmo indicou que teria.

B. Alínea a)

Com base nos dados disponibilizados, e usando como medida de desempenho a proporção (ponderada) de resultados dada pela fórmula:

$$\frac{0.7\ TN + 0.3\ TP}{TN + FN + FP + TP}$$

Para cada algoritmo foi realizada uma Análise Exploratória de Dados (AED), sendo que aos valores relevantes para cada caso encontram se nas tabelas de 1 a 4. Para cada algoritmo foi também feita uma análise gráfica utilizando histogramas para representar a distribuição dos valores obtidos demostrados nas figuras de 1 a 4.

TABELA 1 - VALORES CARACTERIZANTES DO DESEMPENHO DO ALGORITMO CART

Desvio Padrão	**0.02453309**
Média	0.4755455
Mediana	0.471
Mínimo	0.432
Máximo	0.514
Amplitude Interquartil	0.031

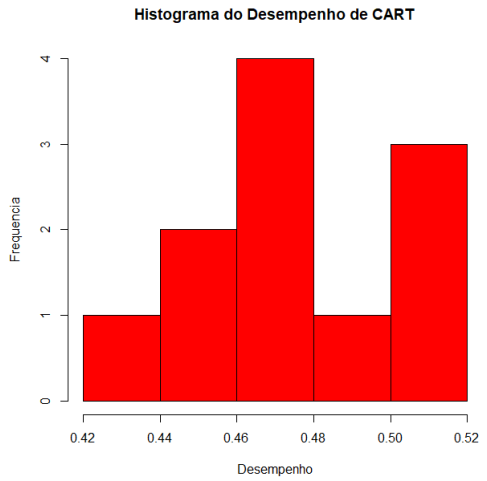

Figura 5 - Histograma dos valores do Desempenho do algoritmo CART

TABELA 2 - VALORES CARACTERIZANTES DO DESEMPENHO DO ALGORITMO LDA

Desvio Padrão	**0.02112861**
Média	0.4572727
Mediana	0.455
Mínimo	0.422
Máximo	0.49
Amplitude Interquartil	0.026

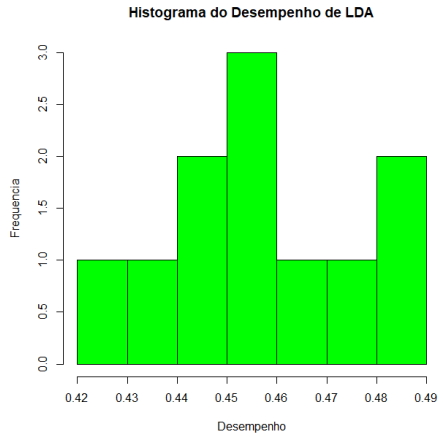

Figura 6 - Histograma dos valores do Desempenho do algoritmo LDA

TABELA 3 - VALORES CARACTERIZANTES DO DESEMPENHO DO ALGORITMO SVM

Desvio Padrão	**0.02107778**
Média	0.4794545
Mediana	0.475
Mínimo	0.446
Máximo	0.52
Amplitude Interquartil	0.0285

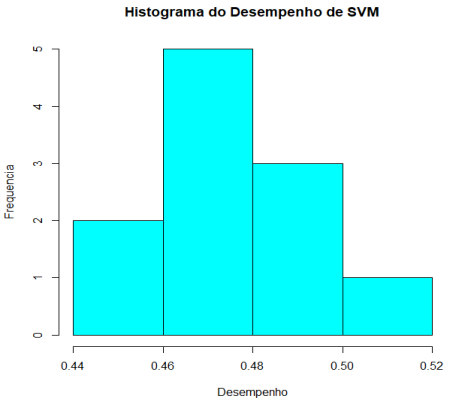

Figura 7 - Histograma dos valores do Desempenho do algoritmo SVM

TABELA 4 - VALORES CARACTERIZANTES DO DESEMPENHO DO ALGORITMO KNN

Desvio Padrão	**0.03013666**
Média	0.4522727
Mediana	0.457
Mínimo	0.399
Máximo	0.495
Amplitude Interquartil	0.0325

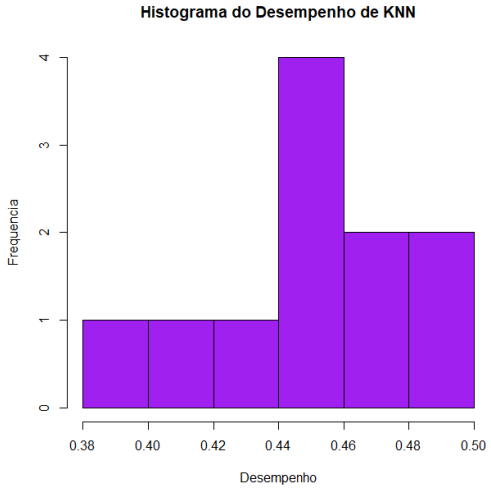

Figura 8 - Histograma dos valores do Desempenho do algoritmo KNN

Para comparar visualmente os desempenhos de cada algoritmo gerou-se um diagrama de extremos e quartis (boxplot) com todos os algoritmos presentes representado na figura 5.

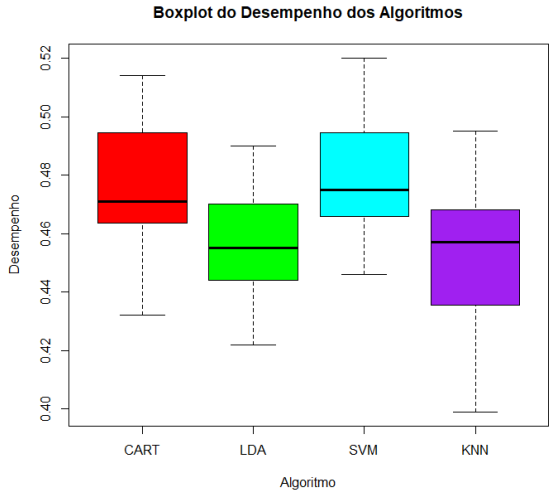

Figura 9 - Diagrama de Extremos e Quartis do Desempenho de todos os Algoritmos

Através da análise da figura 5 e das respetivas tabelas de 1 a 4 podemos concluir que os algoritmos CART e SVM possuem a maior mediana entre os 4.

Entre eles CART e SVM apresentam valores muito próximos de mediana, contudo a amplitude interquartil do SVM é consideravelmente menor (0.0285) que a do CART (0.031), deduzindo assim que o algoritmo SVM parece ser com melhor desempenho.

Como os dados provém de amostras emparelhadas, isto devido ao facto que os 4 algoritmos utilizam os mesmos grupos de mulheres, não podemos recorrer a um teste ANOVA. Assim sendo, recorremos a um teste de Friedemann utilizando as hipóteses:

H_0: Todos os algoritmos apresentam médias, dos resultados obtidos, iguais.

H_1: Pelo menos um dos algoritmos apresenta média dos resultados diferente de algum outro.

Tendo obtido um p-value = 0.006505 com um nível de significância de 0,05 rejeitamos a hipótese nula, e concluímos que os resultados do desempenho de pelo menos um dos algoritmos é diferente dos outros.

Agora poderíamos utilizar uma de duas formas para determinar se existe uma diferença considerável entre os algoritmos ou utilizamos um t-test entre os dois melhores algoritmos (CART e SVM) ou um post hoc de um teste de Friedemann que irá comparar todos os pares de algoritmos possíveis e mostrar os pares em que existe uma diferencia considerável.

Escolhemos utilizar o post hoc de Friedemann proveniente de [8]. As hipóteses nulas e alternativa deste teste são:

H_0: Os dois algoritmos têm médias de desempenho iguais.

H_1: Os dois algoritmos têm médias de desempenho diferentes.

Na fig.6 representamos o plot de cada um dos algoritmos com a sua mediana assinalada e o um gráfico com todos os boxplots das combinações dos algoritmos, a verde estão assinaladas as combinações cujo p-value permite a negação da hipótese nula, enquanto que a cinzento estão assinaladas as que o valor de p-value comprova a hipótese nula.

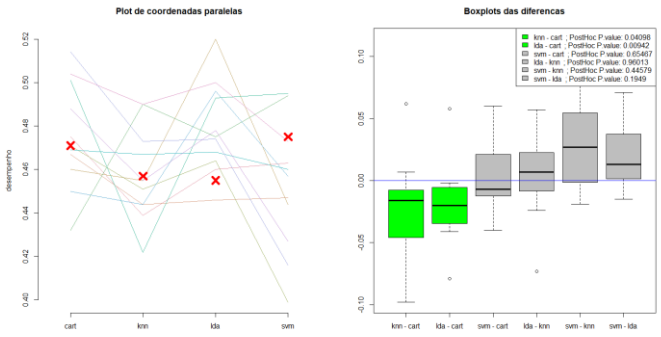

Figura 10 - Plot de coordenadas paralelas, Boxplots das diferenças

Para além da fig.6 representamos na tabela 5 os valores apresentados na fig.6 dos p-values de cada uma das combinações. Seguindo também a separação por cores indicativas de negar ou não a hipótese nula.

TABELA 5 - P-VALUES DO POST HOC PARA CADA PAR DE ALGORITMOS

Par de Algoritmos	P-value do Post Hoc
KNN-CART	0.04098
LDA-CART	0.00942
SVM-CART	0.65467
LDA-KNN	0.96013
SVM-KNN	0.44579
SVM-LDA	0.1949

Como podemos ver pela Tabela 5 apenas duas combinações de algoritmos negam a hipótese nula, o que significa que as outras não apresentam diferenças suficientes entre os algoritmos para dizer que não tem o mesmo desempenho. Estas outras combinações inclui a combinação mais importante para esta análise, CART-SVM o que nos indica que os dois melhores algoritmos têm um desempenho muito próximo quase idêntico até.

Não conseguimos, portanto, tirar conclusões definitivas sobre qual é o melhor algoritmo em termos de desempenho, apesar do que nos é indicado pela fig.5 sugerir que seria o SVM a tabela 5 e a fig.6 indicam que este tem desempenhos equivalentes aos do algoritmo CART.

C. Alínea b

Utilizando os dados obtidos de 768 mulheres de descendência indígena Pima, para construir um modelo para a identificação da Diabetes começamos por identificar o par de variáveis preditoras com maior correlação.

Primeiramente foi criada uma matriz de scatterplots (fig.7), que cruza os plots de cada uma das variáveis com cada uma das outras.

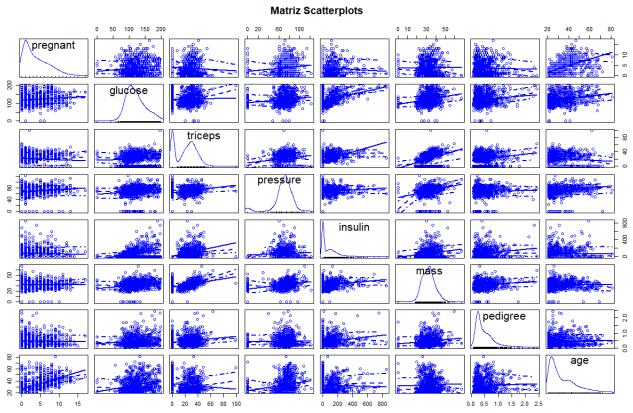

Figura 11 - Matriz de Scatterplots

Podemos ver pelos gráficos de algumas variáveis que demostram uma aparente correlação, para determinarmos se estes exemplos se verificam e saber maior entre estas corelações utilizamos a uma library chamada GGally para dispor todas as corelações num gráfico diferenciando por gradiente de cor a intensidade da correlação demostrado na figura 8.

Figura 12 - Representação gráfica das corelações

Podemos ver claramente na fig.8 que apenas uma relação obteve um valor arredondado de 0.5, sendo que duas obtiveram 0.4 arredondado, os valores exatos destas são dispostos na tabela 6.

TABELA 6 - VALORES EXATOS DAS CORELAÇÕES MAIS ELEVADAS

Par de variáveis	Valor da correlação
Pregnant – Age	0.5443412
Tríceps – Insulin	0.4367826
Tríceps - Mass	0.3925732

Tendo concluído a partir da fig.8 e da tabela 6 que a maior correlação entre variáveis preditoras é entre Age e Pregnant, apesar de mesmo este valor ser bastante afastado de 1, com o intuito de realizar um estudo da regressão entre estas duas variáveis decidimos utilizar a variável Pregnant como variável dependente e Age como independente.

Fazendo a regressão foi utilizado um comando de summary para verificar as seguintes hipóteses:

H_0: A variável independente não é significativa porque compreende valor nulo.

H_1: A variável independente é significativa, não compreende valor nulo.

Com um p-value menor que 2.2e-16 negamos a hipótese nula sabendo assim que a variável independente é significante determinamos os coeficientes da regressão para desenhar a reta

de regressão. A equação da reta resultante foi y= -1.3394071 + 0.1559663x.

Utilizamos um teste de correlação de Pearson para verificar a existência de uma correlação entre as duas variáveis no valor do que nos foi indicado presente na tabela 6, o teste tinha as seguintes hipóteses:

H_0: Não existe uma correlação linear entre as duas variáveis.

H_1: Existe uma correlação linear entre as variáveis.

O p-value obtido é menor que 2.2e-16 indica-nos que podemos negar a hipótese nula e dizer que de facto existe uma correlação linear entre Age e Pregnant com o valor previamente apresentado.

Representamos graficamente a reta de regressão no plot gerado pelas duas variáveis na figura 9.

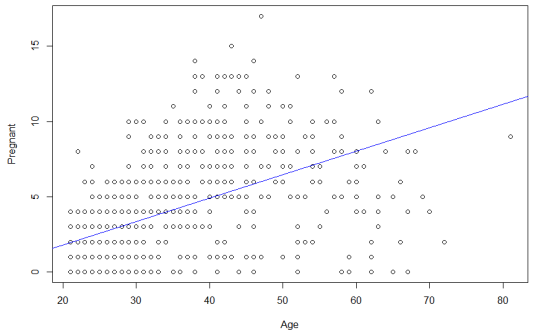

Figura 13 - Reta da regressão representada no plot

Podemos ver pela figura que a reta (y = -1.3394071 + 0.1559663x) não se ajusta bem aos valores apresentados dando a entender que o modelo não possui uma qualidade muito boa.

De forma a avaliarmos ainda mais a qualidade do modelo determinamos os coeficientes de determinação e de determinação ajustado, ambos deram valores bastante afastados de 1 indicando que a qualidade do modelo não é boa, os valores foram, respetivamente: 0.2963074 e 0.2953887.

Em seguida, realizamos uma análise dos resíduos, tentando comprovar os 3 pressupostos: normalidade, homocedasticidade e autocorrelação nula.

Começando pelo pressuposto da normalidade. Utilizando os comandos qqnorm e qqline foi gerada a figura 10.

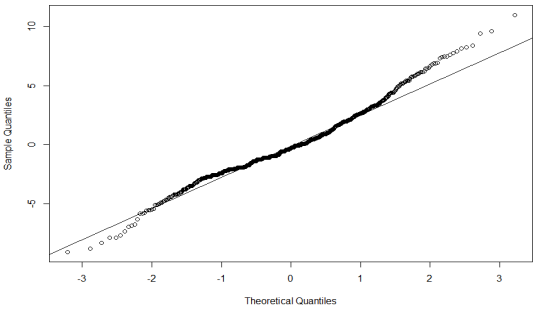

Figura 14 - Q-Q Plot

Observando a fig.10 podemos ver que enquanto que no centro a distribuição dos resíduos encontra se muito próxima da reta os extremos começam a afastar si indicando assim que a normalidade dos dados pode não se verificar.

Para verificar, utilizamos um teste de Shapiro com as hipóteses:

H_0: Os dados seguem uma distribuição normal.

H_1: Os dados não seguem uma distribuição normal

Este teste deu um p-value de 2.803e-09 como tal negamos a hipótese nula e concluímos que a normalidade dos resíduos não se verifica, como tal mesmo sem verificar se a media dos valores é 0 podemos dizer que o primeiro pressuposto não se verifica.

O segundo pressuposto, homocedasticidade para o avaliarmos primeiro representamos os plots entre a variável independente e os resíduos e o plot entre a os resíduos e os valores ajustados (figura 11).

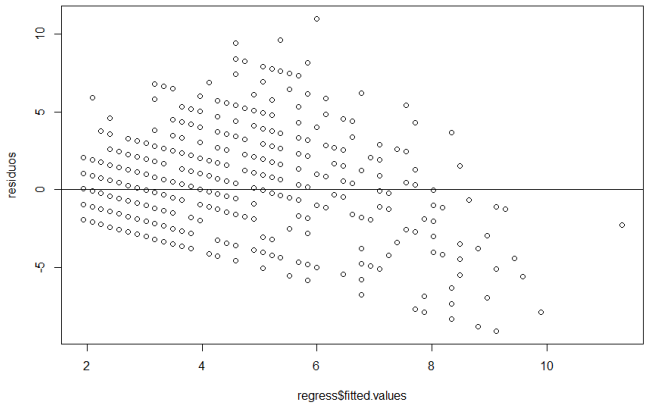

Figura 15 - Plot entre resíduos e os valores ajustados

Podemos ver uma tendência decrescente no gráfico polo que a variância não aparenta ser constante, para confirmar dividimos os resíduos em dois grupos separados pela sua mediana e realizamos um var test para com as seguintes hipóteses:

H_0: A variância dos resíduos é constante.

H_1: A variância dos resíduos não é constante.

Tendo obtido um p-value menor que 2.2e-16 rejeitamos a hipótese nula e confirmamos que de facto a variância não é constante assim como nos indicou a fig.11. Não se verificando assim o segundo pressuposto.

O terceiro, e último pressuposto dos resíduos é a autocorrelação nula, para a testarmos utilizamos um teste de DurbinWatson com as hipóteses:

H_0: Os resíduos são independentes.

H_1: Os resíduos não são independentes.

Tendo obtido um p-value de 0.54, não rejeitamos a hipótese nula o que significa que os resíduos são independentes. E verifica-se o terceiro pressuposto.

Tendo em consideração que apenas 1 dos 3 pressupostos pode ser comprovado e os valores bastante afastados de 1 obtidos a partir dos coeficientes podemos dizer que o modelo encontrado não é de alta qualidade.

Considerando que a variável "pressure" é a variável dependente (e as restantes 7 são as variáveis independentes). Para identificar quais as variáveis independentes que mais influenciam a variável "pressure" e encontrar o modelo de regressão multivariável que depende dessas variáveis e que tenha menor índice de informação de Akaike.

Utilizando o comando summary podemos verificar quais são as variáveis que apresentam um valor de prova abaixo do nível de significância de 0.05 para comprovar ou negar as hipóteses:

H_0: O coeficiente não afeta a variável resposta porque é nulo.

H_1: O coeficiente não é nulo e como tal afeta a variável resposta.

Os valores de prova obtidos para cada variável estão representados na tabela 7.

TABELA 7 - VALORES DE PROVA DAS VARIÁVEIS

Variável	Pr(>|t|)
Pregnant	0.654
Glucose	0.220
Tríceps	4.17e-05
Insulin	0.545
Mass	4.47e-08
Pedigree	0.426
Age	3.86e-08

Podemos ver pela tabela 7 que apenas as variáveis Age Mass e Tríceps negam a hipótese nula e afetam a variável resposta, podemos assim concluir que estas são as variáveis que afetam a variável pressure, sendo a variável que mais afeta a Mass.

Para determinar o novo modelo, melhorado considerando a influencia das variáveis foi utilizado o comando step, este comando utiliza o índice de informação de Akaike (AIC) para editar o modelo e determinar um modelo melhor com aquele conjunto de variáveis, procurando sempre diminuir ao máximo o AIC.

O resultado do comando step encontra-se representado na figura 12, este modelo final demostrou uma diminuição de 5.63 no valor do AIC quando comparado com o valor do modelo inicial.

```
Step:  AIC=4432.2
PimaIndiansDiabetes$pressure ~ triceps + mass + age

           Df Sum of Sq    RSS    AIC
<none>                  243888 4432.2
+ glucose   1     314.1 243574 4433.2
+ pedigree  1     188.1 243700 4433.6
+ pregnant  1      83.8 243804 4433.9
+ insulin   1      29.8 243858 4434.1
- triceps   1    5515.7 249404 4447.4
- mass      1   10946.9 254835 4463.9
- age       1   17466.9 261355 4483.3
```

Figura 16 - Modelo resultante do comando step

Determinamos também a equação correspondente a este modelo que nos deu Y = 34.9501423 + 0.1845835*tríceps + 0.5230561*mass + 0.4100589*age.

Podemos ver que o modelo que foi criado apenas contem as variáveis que nos formam indicadas previamente pelo summary como variáveis que influenciam a pressure na tabela 7.

IV. Conclusão

A partir de toda a análise efetuada podemos sumarizar os seguintes resultados.

Em relação ao desempenho dos algoritmos enquanto que se notam algumas diferenças entre uns deles podemos ver que em múltiplas comparações eles têm desempenhos bastante semelhantes, incluindo nos dois melhores algoritmos CART e SVM cujos resultados são tão semelhantes que com um nível de significância de 5% não podemos dizer que existe diferença. Assim sendo não conseguimos determinar o melhor entre esses dois algoritmos.

Analisando todas as corelações entre as diferentes variáveis podemos ver que a relação entre as variáveis Age e Pregnant é a mais acentuada de todas elas. Mesmo assim o modelo desenhado com base nelas não tem uma qualidade elevada apresentando coeficientes muito afastados de 1 e verificando apenas o pressuposto da autocorrelação nula, não confirmando a normalidade nem a homocesticidade dos resíduos.

Por fim em relação ao modelo que tem a variável pressure como variável dependente utilizando o comando step foi alcançado um modelo melhor que o original, contudo a diferença em valores de AIC foi baixa, enquanto que o modelo é melhor não se confirma ele ter elevada qualidade.

V. References

[1] A. Grandi, "Machine Learning: Pima Indians Diabetes," 14 Abril 2018. [Online]. Available: https://www.andreagrandi.it/2018/04/14/machine-learning-pima-indians-diabetes/. [Acedido em 07 Junho 2019].J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[2] A. Mehta, "Beginner's Guide to Classification and Regression Trees," 23 Janeiro 2019. [Online]. Available: https://www.digitalvidya.com/blog/classification-and-regression-trees/. [Acedido em 07 Junho 2019].

[3] Booz, Allan e Hamilton, "How Machines Learn," [Online]. Available: https://www.boozallen.com/content/dam/boozallen_site/sig/pdf/publications/machine-intelligence-quick-guide-to-how-machines-learn.pdf.

[Acedido em 07 Junho 2019].R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[4] NewTechDojo, “List of Machine Learning Algorithms,” 6 Março 2018. [Online]. Available:https://www.newtechdojo.com/list-machine-learning-algorithms/#Supervised%20Learning. [Acedido em 07 Junho 2019].M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989

[5] J. Brownlee, “Linear Discriminant Analysis for Machine Learning,” 06 Abril 2016. [Online]. Available: https://machinelearningmastery.com/linear-discriminant-analysis-for-machine-learning/. [Acedido em 07 Junho 2019].

[6] S. Ray, "Understanding Support Vector Machine algorithm from examples," 13 Setembro 2017. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/understaing-

[7] O. Harrison, “Machine Learning Basics with the K-Nearest Neighbors Algorithm,” 10 Setembro 2018. [Online]. Available: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761.

[8] T. Galili, “Post hoc analysis for Friedman's Test (R code),” R-statistics blog, 22 fevereiro 2010. [Online]. Available: https://www.r-statistics.com/2010/02/post-hoc-analysis-for-friedmans-test-r-code/. [Acedido em 8 junho 2019].