Aprendizagem Automática: Projeto 1

Grupo 29: Alexandre Monforte (54491)(15H), João Braz (60419)(13H), Joel Oliveira (59442)(15.5H)

A partir de Parkinsons Telemonitoring Data Set realizou-se um ciclo completo de testes e validação de seleção e de avaliação de modelos usando modelos de árvores de decisões e de modelos lineares.

Objetivo 1- Fase de Tratamento dos Dados

O objetivo deste primeiro exercício é produzir o melhor modelo de regressão para tentar prever a variável *motor_ UPDRS*. O Data set disponibilizado, *parkinsons_ updrs.data* é composto por um total de 5875 linhas (cabeçalho não incluído), correspondentes a 42 pacientes, e por 22 colunas (a descrição das colunas estão presentes no documento *parkinsons_ updrs.names*).

Antes de começar a testar e decidir quais os modelos a serem analisados, começando pela fase de tratamento dos dados, foi necessária a criação de um Data set de **Treino** e de um Data set **Independente** a partir do Data set inicialmente fornecido. O Data set de Treino será usado para treinar, testar e avaliar os nossos diferentes modelos, o Data set Independente, apenas será usado no processo de validação final do nosso modelo mais promissor. A nossa escolha foi retirar de forma aleatória um total de 361 linhas do Data set inicial de forma a compor o Data set Independente. Esta escolha das 361 linhas não foi aleatória e baseou-se na seguinte equação (1), onde TA é o tamanho da amostra, z é é o número de desvios padrão entre determinada proporção e a média (valor usado foi 1,96), p é o grau de confiança, e é a margem de erro e N é o tamanho da população. De facto, conhecendo o número da nossa população, criou-se uma amostra que tivesse um grau de confiança de 95% e uma margem de erro de 5%. O grau de confiança corresponde à probabilidade que a amostra represente com precisão as características da população e a margem de erro corresponde à variação que a população possa ter em relação à amostra.

$$TA = \frac{\frac{z^2 \times p(1-p)}{e^2}}{1 + \frac{z^2 \times p(1-p)}{e^2 N}} \quad (1)$$

Na fase de tratamento dos dados, tivemos de analisar o que correspondiam as colunas, de forma a descartar aquelas que não devem estar presentes aquando na fase de teste. Escolhemos descartar as colunas: *subject#*,*test_ time*, *Total_ UPRDS*. Retirámos *subject#* pela razão desta corresponder aos Id's dos pacientes que não têm relação com a resposta e, portanto, estes nunca devem estar presentes em qualquer uma das fases, ou seja, fase de teste e validação; *test_ time* pois apenas corresponde ao tempo passado, em dias, desde o início do recrutamento até à realização dos testes; por último *Total_ UPRDS* não deve ser usado como variável independente pois estamos a estudar as relações das variáveis com o *motor_ UPRDS*.

Por último, passámos à decisão de quais os modelos que iríamos anlisar tendo sido escolhidos como tipo de modelo de Árvore a **Árvore de Regressão** e como tipo de modelos lineares o **Lasso** e o **Ridge**. Resumidamente, a árvore de regressão irá tentar prever o valor de uma variável de destino (*motor_ UPRDS*) com base em várias variáveis de entrada; Lasso é um método de análise de regressão que realiza uma seleção e uma regularização de variáveis para melhorar a precisão da previsão e a interpretabilidade do modelo estatístico resultante, por fim, Ridge é um método para estimar os coeficientes de modelos de múltipla regressão em situações onde as variáveis independentes estão altamente correlacionadas.

Objetivo 1- Fase de teste e Resultados intermédios

Iniciámos a fase de teste realizando um K-fold do nosso Data set Teste subdividindo-o em 16 partes. Tal como anteriormente para a separação do Data set inicial entre um Data set Treino e Independente escolheu-se que cada bloco de teste resultante do K-fold tivesse aproximadamente 360 linhas retiradas aleatoriamente do Data set Treino. De modo a testar os nossos modelos, usámos a técnica do **Cross-Validation**. Cross-Validation consiste em usar um dos blocos como um Data set Teste e o resto dos dados como Data set Treino e realiza-se a fase de treino do modelo realizando o mesmo processo para cada um dos blocos. Nesta situação, iremos obter um total de 16 modelos distintos. De modo a avaliar os modelos fizeram-se variar os **Hiperparâmetros** que compõem os diferentes modelos analisados, que iremos explicitar posteriormente neste relatório. Assim, realizaram-se vários ciclos em que se fizessem variar os hiperparâmetros. De modo a comparar os diferentes modelos, utilizámos os seguintes critérios [1]: **Pearson Correlation**, **root mean square error** do Data set Teste, **root mean square error** do Data set Treino, **Max error** e **Ratio of the Variance Explained**. Tal como dito anteriormente, para cada ciclo de testes iremos obter 16 modelos diferentes, por esta razão calculámos os critérios citados anteriormente para cada um dos modelos individualmente e realizou-se a média desses resultados. Foi a média dos resultados que foi utilizada na comparação dos diferentes modelos com os variados hiperparâmetros.

Para o modelo Lasso, o hiperparâmetro testado foi o α presente no termo: $\alpha \sum_{i=1}^{n} |\theta_i|$, onde θ corresponde ao coeficiente de índice i. Variou-se α de 0 até 5 por passos de 0,05 e obtiveram-se os seguintes gráficos presentes na Fig.(1). Repare-se que não existe em nenhum ponto do gráfico um *sweet spot*, que corresponde a um ponto no gráfico onde o erro do teste é mínimo. Este resultado mostra que para o melhor modelo seria aquele que tive-se um valor de α igual a 0, fazendo perder o propósito do modelo Lasso, pois assim apenas se realiza uma simples regressão linear. Ainda foi considerada a possibilidade do *sweet spot* encontrar-se entre os valores 0 e 1, portanto realizou-se um

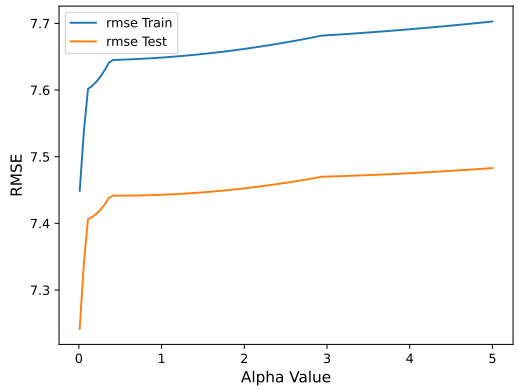

Figura 1. Teste do modelo Lasso fazendo variar o valor de α entre 0 e 5 por passos de 0,05 calculando o respetivo rmse do Data set Treino e Teste.

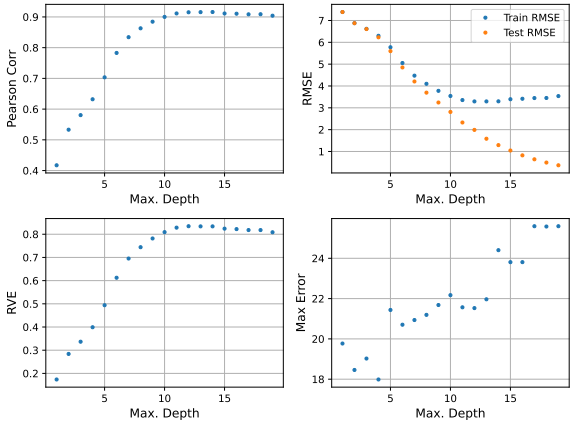

Figura 2. Teste do modelo Lasso fazendo variar o valor de α entre 0 e 5 por passos de 0,05 calculando o respetivo rmse do Data set Treino e Teste.

estudo mais específico fazendo variar os valores de α entre 0 e 1 por passos de 0,001. Novamente, não encontrámos qualquer *sweet spot* nessa região.

Para o modelo Ridge, o hiperparâmetro testado foi novamente o α presente no seguinte termo: $\frac{\alpha}{2}\sum_{i=1}^{n}\theta_i^2$. Seguiu-se exatamente do mesmo modo que no estudo do Lasso, obtendo-se um gráfico semelhante ao da Fig.(1) também sem a presença de um *sweet spot* (não foi apresentado o gráfico por falta de espaço).

Por fim, para o modelo da árvore de regressão estudaram-se os hiperparâmetros presentes na Tab.(I). É de notar que estudou-se cada hiperparâmetro individualmente, ou seja, estudava-se o hiperparâmetro A e retirava-se o melhor valor consoante os critérios usados para comparar presentes em [1], depois fixava-se o melhor valor do hiperparâmetro A e fazia-se variar o hiperparâmetro B e assim sucessivamente. Na Fig.(2) apresentamos a variação dos critérios [1] fazendo variar o hiperparâmetro *max_ depth*. É de notar que existe um *sweet spot* perto do valor 13 do *max_ depth* no gráfico que relaciona o rmse do Data set Treino e Teste com a profundidade máxima dos ramos da árvore. Para cada um dos hiperparâmetros o melhor valor foi determinado por esta forma. Nas situações em que se podiam considerar um certo número de pontos no *sweet spot* usou-se aquele que tivesse um menor Max error.

Tabela I. Apresentação dos hiperparâmetros testados para o modelo da árvore de regressão, juntamente com o correspondente intervalo de teste, os passos das iterações e o melhor valor obtido com base nos critérios apresentados em [1].

Hiperparâmetro	Intervalo	Passos	Melhor Valor
max_depth	[2;20]	1	13
min_samples_split	[2;50]	1	34
min_samples_leaf	[1;10]	1	2
max_features	—	—	None
max_leaf_nodes	[2;50]	1	None
min_impurity_decrease	[0;0,05]	0,001	0,025
ccp_alpha	[0;0,05]	0,001	0

Apresentamos na Tab.(II) os valores obtidos para cada critério para cada um dos melhores modelos de Lasso, Ridge e árvore de regressão. Por análise dessa tabela, conclui-se que o melhor modelo corresponde ao modelo da árvore de regressão tendo sido obtido os melhores valores para cada critério comparativamente com os dois outros modelos.

Tabela II. Apresentação dos resultados obtidos para cada um dos melhores modelos de Lasso, Ridge e Árvore de Regressão (A.R) segundo os critérios Pearson Correlation (P.C), RMSE Teste, RMSE Treino, Max Error e RVE.

Modelo	P.C	RMSE Teste	RMSE Treino	Max Error	RVE
Lasso	0,47	7,38	7,24	22,90	0,21
Ridge	0,47	7,38	7,24	22,90	0,21
A.R	0,93	2,97	2,57	21,00	0,86

Objetivo 1- Fase de Validação e Resultados Finais

A última fase para concluírmos o Objetivo 1 é a fase de Validação. Nesta fase aplicámos o melhor modelo que treinámos, criámos e avaliámos com o Data set Independente. Os resuldados obtidos foram as seguintes: Pearson Correlation de 0,93; RMSE Teste de 2,91; RMSE Treino de 2,54; Max Error de 14,22; RVE de 0,87.

Objetivo 1- Discussão e Conclusão

Da observação da Fig.(1) e da Tab.(II) ficou claro que os modelos de Lasso e Ridge não são bons modelos para tentar prever a variável dependente *motor_ UPRDS* pois o melhor parâmetro que implica um menor erro é aquele que corresponde um $\alpha = 0$ ficando uma regressão linear simples dos dados, daí obtermos exatamente os mesmos valores para cada critério entre o Lasso e o Ridge na Tab.(II), pois sem

a componente α são o mesmo modelo. Este resultado já era espectável pois, por observação de gráficos que relacionassem cada variável com o *motor_ UPRDS*, não parecia existir uma relação de linearidade entre as variáveis independentes com a dependente, parecendo-se como "dados dispersos". Ambos os modelos são mais eficazes quando existe alguma linearidade entre as variáveis em estudo.

Em relação ao modelo de Árvore de Regressão, este obteve melhores previsões e melhores valores nos critérios em estudo que o Lasso e Ridge. De facto, este é um modelo que se consegue adaptar a um vasto tipo de situações relacionais distintas, incluindo relações lineares ou não lineares. Por esta razão, já se esperavam melhores resultados vindos deste modelo que nos outros dois anteriores, que acabou por ser corroborado pelos resultados obtidos.

Em suma, consoante a situação que se pretenda estudar, devemos priorizar o estudo do modelo de árvore de regressão ao invés de modelos lineares como o Lasso e Ridge, no caso de não haver uma relação de linearidade entre as variáveis independentes com a variável dependente, ou ainda no caso de haver uma fraca relação de linearidade entre as mesmas.

Objetivo 2- Fase de Tratamento dos Dados

O objetivo deste segundo exercício é produzir o melhor modelo de classificação binário assumindo como positivo sempre que o valor de *total_ UPRDS>40* e como negativo todos os casos restantes. Tal como para o Objetivo 1, foi necessária uma fase de tratamento dos dados. Iniciámos por retirar do nosso Data set inicial 361 linhas de modo a criar o nosso Data set Independente, pela mesma razão que explicada anteriormente no Objetivo 1. Analizando novamente as colunas, descartámos as colunas *subject#*,*test_ time*, *motor_ UPRDS*, pelas mesmas razões que explicadas anteriormente mas desta vez descartou-se a coluna *motor_ UPRDS* pois estamos a usar como variável dependente *total_ UPRDS*. Seguidamente, para irmos ao encontro do que é pedido neste exercício, alterou-se a coluna correspondente aos dados de *total_ UPRDS* para valores 1 se *total_ UPRDS>40* e 0 caso contrário.

Por fim, decidímos testar os seguintes modelos: Árvore de Decisão e Regressão Logística. A Árvore de Decisão tem o mesmo funcionamento que a Árvore de Regressão mas ao invés da previsão ser um número real, nesta situação será uma situação binária de 0 ou 1. Por sua vez, Regressão Logística é um processo que modela a probabilidade de um resultado discreto, neste caso 0 ou 1, dada uma ou várias variáveis de entrada.

Objetivo 2- Fase de teste e Resultados intermédios

Semelhante ao primeiro objectivo, foi realizado um *Cross-validation 16-fold* ao Data set de Teste em que cada bloco tivesse, tal como anteriormente, aproximadamente 360 linhas retiradas aleatoriamente, obtendo-se assim 16 modelos diferentes.

Os **Hiperparâmetros** para a árvore de decisão utilizados para este problema foram os mesmos do Objectivo 1 (Tab.(I)), enquanto que no modelo de regressão logística não existem hiperparâmetros a serem estudados. É de notar que no processo de *model-fitting* foi necessário que a matriz com as variáveis independentes fosse ajustada para média de 0 com variância de 1, ou seja, standardizar os dados.

De forma a comparar os diferentes modelos, foram utilizados os seguintes parâmetros[2]: **Precision, Recall, F1-Score e o Phi coefficient (MCC).**

Análogo ao procedimento realizado para os modelos da árvore de decisão, foi estudado cada hiperparâmetro individualmente, fixando primeiramente o melhor valor de um hiperparâmetro A, variando-se o hiperparâmetro B, e assim sucessivamente. A Tab.(III) mostra os intervalos utilizados para cada hiperparâmetro testado usando os parâmetros apresentados em [2], enquanto que a Fig.(3) compara vários valores possíveis para o hiperparâmetro Max Depth. No entanto, neste caso, este procedimento foi realizado duas vezes com dois tipos de critérios, o ***Gini*** e a **Entropia**, de forma a averiguar qual o melhor modelo.

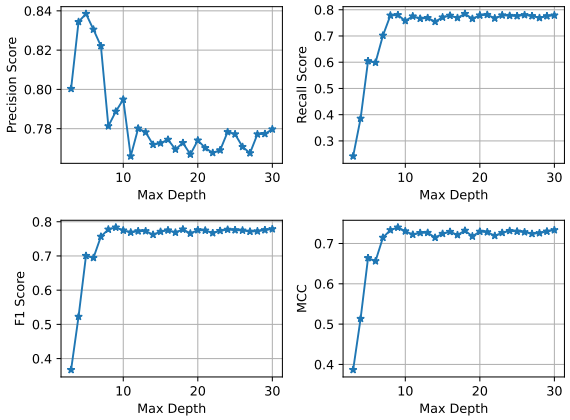

Figura 3. Estudo da variação de Precision Score, F1 Score, Recall Score e MCC usando a árvore de decisão com a variação do hiperparâmetro Max depth no intervalo [2, 30] por passos de 1.

Tabela III. Apresentação dos hiperparâmetros testados para o modelo da árvore de decisão, juntamente com o correspondente intervalo de teste, os passos das iterações e o melhor valor obtido com base nos critérios apresentados em [2].

Hiperparâmetro	Intervalo	Passos	Gini	Entropia
max_depth	[3;30]	1	9	7
min_samples_split	[2;50]	1	4	29
min_samples_leaf	[1;50]	1	20	2
max_leaf_nodes	[2;50]	1	34	28
min_impurity_decrease	[0;0,05]	0,0005	0	0

O modelo de regressão logística obteve, em média para os 16 modelos, os seguintes resultados: *Precision* de 0,32; *Recall* de 0,01; *F1 Score* de 0,01; *MCC* de 0,01.

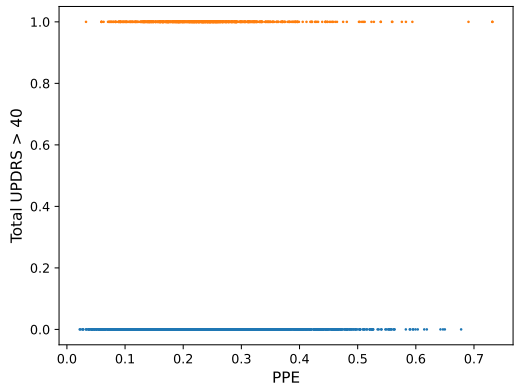

Figura 4. Aplicação do modelo de regressão logística aos dados da coluna *PPE*.

Objetivo 2- Fase de Validação e Resultados Finais

O melhor modelo obtido com os dados de teste foi a Árvore de Decisão. De forma a validar-mos a performance deste, avaliámo-lo com o Data set Independente, tendo obtido os seguintes resultados: Precision de 0,83; Recall de 0,79; F1 Score de 0,81; Matthews Corr. Coef. de 0,77.

Objetivo 2- Discussão e Conclusão

Repare-se que pela Fig.(4) não foram obtidos bons resultados para o modelo da regressão logística. Naturalmente, não possível mostrar o tratamento dos dados pela regressão logística pois esta trabalha com todas as dimensões em simultâneo, daí apenas apresentar com uma das variáveis. Contudo, conclui-se que o problema persiste para todas as variáveis pelos resuldados obtidos para cada um dos parâmetros usados para analisar a qualidade do modelo. Existem várias razões para o insucesso do modelo sendo a mais relevante a existência de multicolinearidade entre as variáveis independentes. Num primeiro contacto com o Data set inicial, observou-se como as diferentes variáveis relacionavam-se entre si e entre algumas delas existia uma forte correlação, por exemplo entre as diferentes colunas do tipo *Jitter* e entre o tipo *Shimmer*. Estas correlações prejudicam o bom desempenho do modelo, daí o seu insucesso neste estudo. Outra verificação foi a existência ou não de linearidade dos dados com a transformação **log odds**, que acabou por não ser verificada, sendo este outro requisito importante para o sucesso do modelo da regressão logística.

Em relação ao modelo de Árvore de Decisão, este obteve melhores previsões e melhores valores nos parâmetros estudados que o modelo de regressão logística. Tal como no caso do modelo da árvore de regressão analisada no Objetivo 1, este é um modelo que se consegue adaptar a um vasto tipo de situações relacionais distintas, incluindo relações lineares ou não lineares. Por esta razão, já se esperavam melhores resultados vindos deste modelo que no outro apresentado anteriormente, que acabou por ser corroborado pelos resultados obtidos apresentados na secção **Objetivo 2- Fase de teste e Resultados intermédios**.

Em conclusão, para o bom sucesso do modelo de regressão logística, este necessita que os dados cumpram um certo número de requisitos tais como: a natureza binária da variável dependente, as observações (variáveis independentes) serem independentes umas das outras, não pode haver ou no máximo haver uma fraca multicolinearidade entre as variáveis independentes, linearidade das variáveis independentes aquando da transformação *log odds* e um número relativamente grande de elementos na amostra (que nesta situação em estudo foi respeitada). No caso do não cumprimento desses requisitos devemos escolher antes um modelo de árvore de decisão, sendo este modelo facilmente adaptável a qualquer tipo de dados.

Conclusão

Este trabalho permitiu identificar algumas limitações que constituem os modelos lineares. Estes têm a vantagem de serem de fácil implementação, contudo exigem a existência de alguma linearidade entre as variáveis independentes com a variável dependente. Nesta situação em estudo, verificou-se que a falta de linearidade tornou os modelos lineares ineficazes para realizarem boas previsões. Uma solução a este problema é o uso de modelos de Árvores de Decisão, que, pela sua natureza de grande adaptabilidade, conseguem encontrar relações entre os dados que possam-se relacionar de diversas formas distintas e ainda assim realizar boas previsões.