ESCOLA

DE ENGENHRARIA

DA UNIVERSIDADE

DO MINHO

M

Mestrado

Engenharia Informática

“Parkinson Telemonitoring”

Trabalho Realizado por:

Bruno Santos PG44414

Guilherme Palumbo PG42832

João Silva PG42834

Docente:

Raquel Menezes

Perfil de Especialização em Ciência de Dados

Aprendizagem Automática 1

4º/1º Ano, 1º Semestre

Ano letivo 2020/2021

Resumo

Cada vez mais, os métodos de aprendizagem estatísticas, tem assumido um papel relevante na extração de informação e conhecimento a partir de base de dados, e assim ajudar na tomada de decisão. A saúde é sempre um tema preocupante e que exige os melhores métodos e práticas, sendo que para isso é necessário estar constantemente informado. É neste sentido que se começa a reconhecer a importância de identificar os fatores que colocam em risco a saúde da sociedade em geral. Para isso, nada melhor do que usar um método que o permita saber com antecedência.

Quanto mais eficaz for o método a identificar os fatores de risco, maior será o retorno para a sociedade. Sendo o Parkinson uma doença que afeta cada vez mais a sociedade em geral, com este projeto pretendeu-se estudar o comportamento de uma série de medições biomédicas da voz de 42 pessoas com doença de Parkinson em estado inicial que foram recrutadas para um teste de seis meses utilizando um dispositivo de telemonitoramento para o monitoramento remoto da progressão dos sintomas.

Posto isto, começamos por fazer uma análise exploratória aos dados e preparar os mesmos de forma a terem mais qualidade no desenvolvimento dos modelos. De seguida construímos alguns modelos para fazer a previsão da pontuação UPDRS total baseados em técnicas de regressão. Os modelos desenvolvidos neste projeto podem ser agrupados em 2 grupos: modelos fixos e modelos mistos. Por último comparamos os resultados obtidos para cada um dos modelos de acordo com as melhores métricas de avaliação.

Palavras-Chave: Parkinson, Regressão, Aprendizagem Automática.

Conteúdo

Lista de Figuras

Capítulo 1

Introdução

1.1 Identificação do Projeto e Objetivos

No âmbito da unidade curricular "Aprendizagem Automática 1" do perfil "Ciência de Dados", foi nos proposto a seleção de um conjunto de dados real, de onde pudemos extrair questões interessantes para responder. Para responder às questões devemos experimentar/considerar os vários métodos de aprendizagem estatística.

Posto isto, o principal objectivo deste projeto é analisar e explorar um *dataset* com o auxílio dos métodos de aprendizagem estatística de forma a obter algum conhecimento a partir dos dados. No final deste trabalho prático, deve ser entregue um relatório onde é descrito todo o trabalho desenvolvido, assim como todos os documentos provenientes do mesmo.

1.2 Estrutura do Relatório

Para além deste capítulo, este relatório está organizado em cinco capítulos, são eles:

- **Capítulo 2: Análise Exploratória** – capítulo que apresenta uma análise exploratória dos dados, através de gráficos e medidas estatísticas e que suporta algumas questões e hipóteses formuladas.
- **Capítulo 3: Análise Preditiva** – capítulo que apresenta o treino dos modelos preditivos e descreve o propósito de cada modelo.
- **Capítulo 4: Análise de Resultados** - capítulo que apresenta os resultados obtidos de cada modelo e retira conclusões sobre os mesmos.
- **Capítulo 5: Conclusão** – Apresentação de sugestões e recomendações após análise dos resultados obtidos e dos modelos desenvolvidos.
- **Capítulo 6: Referências Bibliográficas** – neste capítulo estão referenciadas, de acordo com as normas APA, todas as fontes utilizadas.

Capítulo 2

Análise Exploratória

O *dataset Parkinsons Telemonitoring Data Set*, foi retirado do site UCI Machine Learning Repository e é composto por uma série de medições biomédicas da voz de 42 pessoas com doença de Parkinson em estado inicial que foram recrutadas para um teste de seis meses utilizando um dispositivo de telemonitoramento para o monitoramento remoto da progressão dos sintomas. As gravações foram capturadas automaticamente nas residências do paciente.

Posto isto, temos um total de 5875 registos, caracterizados por 22 atributos, sendo estes:

subject (1) ID do paciente.

age (2) Idade.

sex (3) Sexo masculino "0", Sexo feminino "1".

test_time (4) Tempo desde o recrutamento para o ensaio. A parte inteira é o número de dias desde o recrutamento.

Motor_UPDRS (5) Pontuação UPDRS motora, interpolada linearmente (UPDRS => Unified Parkinson's Disease Rating Scale).

Total_UPDRS (6) Pontuação UPDRS total, interpolada linearmente.

Jitter (7-11) Várias medidas de variação na frequência fundamental (Jitter => Medida utilizada para medir a qualidade de voz).

- **Jitter (7)** percentagem.
- **Jitter (Abs) (8)** Absolute.
- **Jitter:RAP (9)** Relative Average Perturbation.
- **Jitter: PPQ5 (10)** Five-point Period Perturbation Quotient.
- **Jitter:DDP (11)** Difference of Differences of Periods.

Shimmer (12-17) Várias medidas de variação da amplitude da fala.

- **Shimmer (12)** average absolute difference between the amplitudes of consecutive periods.
- **(Shimmer(dB) (13)** average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods).
- **Shimmer: APQ3 (14)** Fhree-point Shimmer Amplitude Perturbation Quotient values.
- **Shimmer: APQ5 (15)** Five-point Shimmer Amplitude Perturbation Quotient values.
- **Shimmer: APQ11 (16)** Eleven-point Shimmer Amplitude Perturbation Quotient values.
- **Shimmer:DDA (17)** average absolute difference between consecutive differences between the amplitudes of consecutive periods.

NHR (18) Medida de proporção de ruído para componentes tonais na voz (Noise-to-Harmonics Ratio).

HNR (19) Medida de proporção de ruído para componentes tonais na voz (Harmonic-to-Noise Ratio).

RPDE (20) Uma medida de complexidade dinâmica não linear (Recurrence Period Density Entropy).

DFA (21) Expoente de escala do fractal de sinal (Detrended Fluctuation Analysis).

PPE (22) Uma medida não linear da variação da frequência fundamental (Pitch period entropy).

Dado a conhecer a constituição do *dataset* passamos para a análise da correlação entre os diversos atributos, para tal e aplicando uma função de correlação e com o suporte de uma ferramenta externa para visualização da mesma *(Knime)* concluímos que todos os atributos, num momento inicial são relevantes para o problema. Contudo, notamos que a variável *test_ time* quase não possui qualquer correlação com as restantes variáveis, mas tomamos a decisão de não a retirar nesta primeira fase tendo em conta o problema e crendo que o tempo possa vir a ser um factor importante no desenvolvimento de qualquer doença (Figura 2.1).

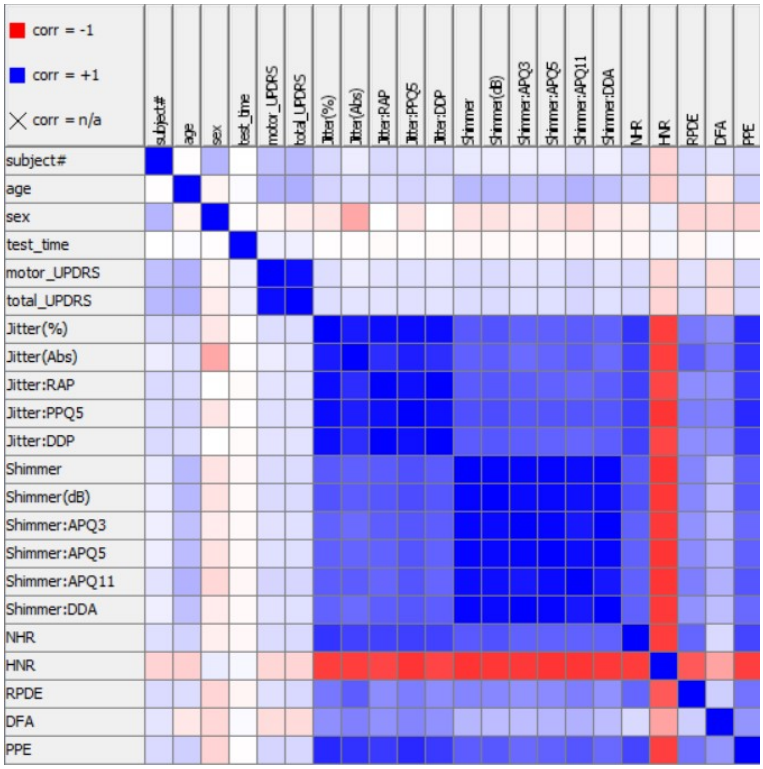
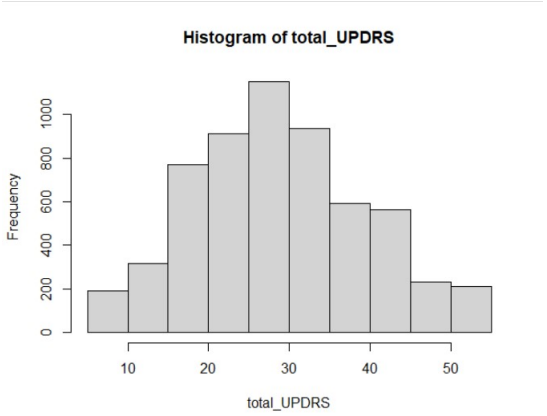
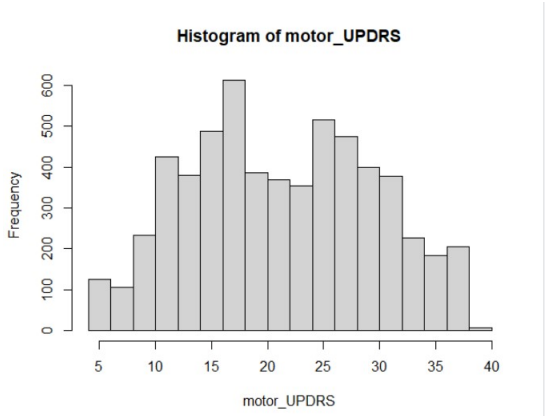


Figura 2.1: Correlação entre Variáveis

Uma vez determinada a correlação entre variáveis passamos para a selecção da variável de interesse que iremos prever. Ora, como possuímos duas variáveis dependentes *Motor_ UPDRS* e *Total_ UPDRS* achamos por bem seleccionar apenas aquela que teria mais valor significativo e com resultados mais confiáveis, dessa forma, através da análise da distribuição de ambas as variáveis (Figura 2.2) concluímos que a melhor variável para se prever seria a variável *Total_ UPDRS* por possuir uma distribuição gaussiana como ilustrado na Figura 2.2(a).



(a) Histograma da variável total_UPDRS



(b) Histograma da variável motor_UPDRS

Figura 2.2: Histogramas das variáveis de interesse

Na continuidade do processo de exploração do dados foram realizadas várias tarefas que iriam enriquecer a informação que possuímos sobre o *dataset* e por conseguinte o tratamento dos dados dependendo dos resultados obtidos. Desse modo, identificamos se o *dataset* possuía qualquer valor em falta, eliminamos os registos com *test_ time* inferior a zero, representamos visualmente todos os *outliers* das variáveis existentes e exploramos se o *dataset* tinha qualquer registo duplicado.

Posteriormente ao tratamento inicial dos dados, respondemos a algumas questões que achamos relevantes para uma melhor compreensão e análise do *dataset* que estamos a utilizar. As questões que proporcionaram uma melhor compreensão do *dataset* foram as seguintes:

- Existe algum resíduo (*outlier*) presente nas observações? Se sim, que medidas devemos tomar para o tratamento desse resíduo.?

  Após a descoberta dos *outliers* existentes no *dataset*, ou seja, as observações que tinham resíduos de valor elevado quando comparados com outras observações da mesma variável, averiguamos que não iríamos eliminá-los do modelo, visto que esses valores não significam que são más observações. Por outro lado, concluímos que no caso da variável *age* por exemplo, notamos que o *outlier* representa um valor muito importante, onde mostra que existem indivíduos muito novos que apresentam a doença de Parkinson em estado inicial (Figura 2.3).

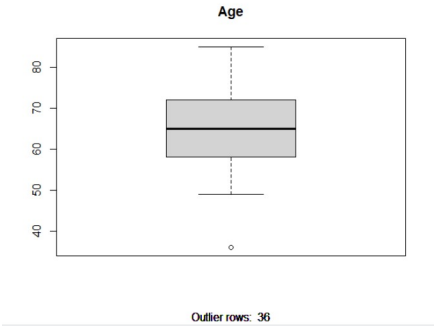
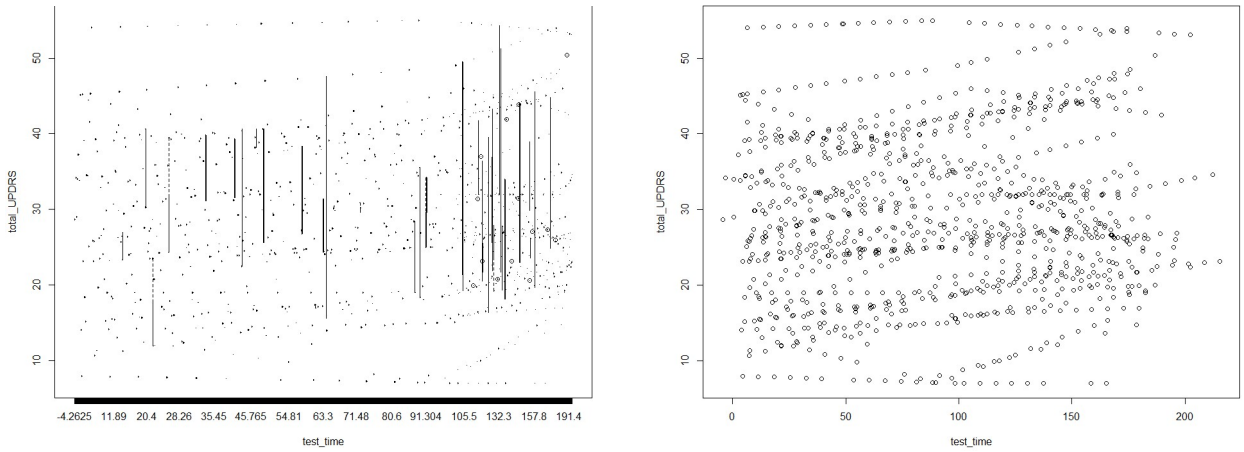


Figura 2.3: BoxPlot da variável *Age*

- De que forma a progressão do *test_ time* influência o *total_ updrs*?

  Através da análise da Figura 2.4 não conseguimos tirar conclusões sobre a relação entre o *test_ time* e o *total_ updrs*. A explicação para este resultado é devido à constituição do *dataset*. Como se trata de um conjunto de dados referente a apenas 42 pacientes o tempo entre testes realizados, apesar de assumirmos que possui uma alta correlação com o *total_ updrs* para cada

paciente individualmente, ao analisarmos a um nível global não podemos afirmar que o mesmo aconteça.



(a) Boxplot entre test_time e total_UPDRS

(b) Plot entre test_time e total_UPDRS

Figura 2.4: Plots entre test_time e total_UPDRS

- Qual o atributo ou atributos que mais influenciam a variável *total_ updrs*?

  Analisando as correlações de forma individual e com suporte à Figura 2.1 podemos concluir que os atributos que mais influenciam a variável *total_ updrs* são os atributos *motor_ updrs* com 0.95 de correlação e a *age* com 0.31 de correlação. Por outro lado, podemos também concluir que o atributo *HNR* possui uma correlação de -0.16, sendo o atributo com maior correlação negativa à variável *total_ updrs*.

- Em que medida a idade de um individuo afeta o *total_ updrs*?

  Como referido anteriormente, podemos concluir que a idade tem uma influência significativa sobre a variável dependente. Com isso podemos concluir que num modo geral, quanto mais velho for o individuo maior a pontuação UPDRS de parkinson.

- Em média os homens têm uma tendência a ter *total_ updrs* maior que as mulheres?
  De acordo com a Figura 2.5 podemos concluir que os homens em média têm uma ligeira tendência a apresentar valores de *total_ updrs* mais elevados, sendo o único sexo a apresentar valores superiores a 50. No entanto, devido à disparidade de muitas instâncias do sexo masculino relativamente ao sexo feminino este resultado pode não ser real.

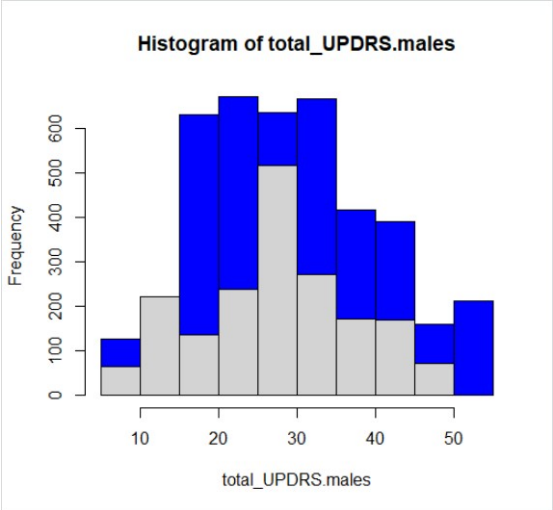


Figura 2.5: Histograma comparativo entre homem e mulher total_updrs

- Quais são as características mais comuns de um individuo que possui um *total_ updrs* superior a 50 UPDRS (Escala Unificada de Avaliação da Doença de Parkinson)?

  Após uma análise dos dados para todos os pacientes com um *total_ updrs* superior a 50, podemos concluir que as características mais frequentes são do sexo masculino e uma idade compreendida entre os 70 e 72 anos. Possuem um *Jitter...* entre [0.00, 0.01], um *Jitter.Abs.* compreendido entre [0e+00, 0.5e-02] e um *Jitter.RAP* entre [0.00, 0.005]. Para *Shimmer* os valores mais comuns estão entre [0.02, -0.05], para *Shimmer.APQ5* [0.00, 0.04] e para *Shimmer.APQ11* [0.02, 0.04]. O *NHR* está compreendido entre [20, 22], o *RPDE* entre [0.45, 0.5], o *DFA* entre [0.72, 0.74] e o *PPE* entre [0.2, 0.3]. Por último, as características mais comuns para *motor_ UPDRS* estão compreendidas no intervalo [36, 38]. (Apenas foram analisados os atributos considerados mais relevantes).

Capítulo 3

Análise Preditiva

3.1 Análise Preditiva utilizando o conjunto de dados total

Concluída a análise exploratória do dados passamos para a análise preditiva dos mesmos. Nesta fase foram desenvolvidos três macro-processos diferentes onde foram aplicadas formas diferentes de gerar o modelo de Regressão Linear e onde os dados analisados terão dimensões diferentes, consequentemente obtendo assim três resultados diferentes. Ambos os resultados serão analisados no capítulo seguinte. Para explanação desta fase, o primeiro processo que foi elaborado foi uma regressão linear com todos os dados do *dataset*, por outras palavras, não possui um *dataset* de treino nem de teste. Foi concebido, numa fase primária, uma Regressão Linear com todas as variáveis e introduzido um termo de interacção entre *age* e *sex* que achamos relevante para o problema actual (Figura 3.1).

```
#Regressão com todos os atributos sem dataset de treino e teste
#adicionamos age*sex pois achamos ser um bom termo de interação
x1 <- lm(total_UPDRS ~ subject. + age + sex + test_time + motor_UPDRS + Jitter...
                  + Jitter.Abs. + Jitter.RAP + Jitter.PPQ5 + Jitter.DDP + Shimmer
                  + Shimmer.dB. + Shimmer.APQ3 + Shimmer.APQ5 + Shimmer.APQ11 + Shimmer.DDA
                  + NHR + HNR + RPDE + DFA + PPE + age*sex, data=parkinsons_updrs)
summary(x1)
```

Figura 3.1: Primeira Regressão Linear

Numa fase seguinte, neste mesmo modelo foi desenvolvida de forma automática a melhor Regressão Linear com as melhores variáveis possíveis através da função *step* com direcção *"backwards"*. O resultado da Regressão Linear mais adequada está representada na Figura 3.2.

```
#Podemos concluir qual o melhor modelo através do step
step(x1 , direction = "backward")
#Resultado:
x7 <- lm(formula = total_UPDRS ~ subject. + age + sex + test_time +
          motor_UPDRS + Jitter... + Jitter.Abs. + Jitter.RAP + Shimmer +
          Shimmer.APQ5 + Shimmer.APQ11 + HNR + RPDE + DFA + PPE + age:sex,
        data = parkinsons_updrs)
summary(x7)
```

Figura 3.2: Melhor Regressão Linear

Comparativamente podemos ver que houve uma evolução entre a primeira Regressão Linear desenvolvida e a melhor Regressão desenvolvida possível (Figuras 3.3(a) e 3.3(b)).

```
> summary(x1)

Call:
lm(formula = total_UPDRS ~ subject. + age + sex + test_time +
    motor_UPDRS + Jitter... + Jitter.Abs. + Jitter.RAP + Jitter.PPQ5 +
    Jitter.DDP + Shimmer + Shimmer.dB. + Shimmer.APQ3 + Shimmer.APQ5 +
    Shimmer.APQ11 + Shimmer.DDA + NHR + HNR + RPDE + DFA + PPE +
    age * sex, data = parkinsons_updrs)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6453 -2.1197 -0.4687  1.3377 11.2680

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -8.577e-01  1.111e+00  -0.772 0.440362
subject.       4.647e-02  3.855e-03  12.054  < 2e-16 ***
age            1.187e-01  6.900e-03  17.199  < 2e-16 ***
sex            4.314e+00  6.479e-01   6.659 3.01e-11 ***
test_time      2.676e-03  7.859e-04   3.405 0.000665 ***
motor_UPDRS    1.208e+00  5.793e-03 208.447  < 2e-16 ***
Jitter...     -3.927e+02  7.008e+01  -5.603 2.20e-08 ***
Jitter.Abs.    1.731e+04  3.268e+03   5.297 1.22e-07 ***
Jitter.RAP     5.256e+03  1.534e+04   0.343 0.731922
Jitter.PPQ5    7.459e+01  6.220e+01   1.199 0.230502
Jitter.DDP    -1.570e+03  5.115e+03  -0.307 0.758907
Shimmer       -4.712e+01  2.132e+01  -2.210 0.027118 *
Shimmer.dB.    5.464e-02  1.591e+00   0.034 0.972611
Shimmer.APQ3  -1.747e+04  1.541e+04  -1.134 0.256945
Shimmer.APQ5   1.051e+02  1.819e+01   5.777 7.98e-09 ***
Shimmer.APQ11 -4.588e+01  8.179e+00  -5.610 2.12e-08 ***
Shimmer.DDA    5.822e+03  5.137e+03   1.133 0.257114
NHR           -3.510e+00  2.051e+00  -1.711 0.087137 .
HNR           -1.064e-01  2.284e-02  -4.657 3.28e-06 ***
RPDE           3.418e+00  6.015e-01   5.683 1.39e-08 ***
DFA           -3.247e+00  7.793e-01  -4.167 3.13e-05 ***
PPE           -4.233e+00  9.643e-01  -4.390 1.15e-05 ***
age:sex       -9.376e-02  9.861e-03  -9.508  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.191 on 5852 degrees of freedom
Multiple R-squared:  0.9114,    Adjusted R-squared:  0.9111
F-statistic:  2736 on 22 and 5852 DF,  p-value: < 2.2e-16
```

(a) Resultado da Primeira Regressão Linear

```
> summary(x7)

Call:
lm(formula = total_UPDRS ~ age + sex + motor_UPDRS + Jitter... +
    Jitter.Abs. + Jitter.RAP + Shimmer + Shimmer.APQ5 + Shimmer.APQ11 +
    HNR + RPDE + DFA + PPE, data = parkinsons_updrs)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0403 -2.0781 -0.4749  1.3835 11.1367

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.203e+00  1.044e+00   2.110  0.03486 *
age            6.836e-02  5.122e-03  13.345  < 2e-16 ***
sex           -1.398e+00  1.025e-01 -13.644  < 2e-16 ***
motor_UPDRS    1.227e+00  5.667e-03 216.573  < 2e-16 ***
Jitter...     -2.697e+02  5.259e+01  -5.127 3.03e-07 ***
Jitter.Abs.    1.403e+04  3.086e+03   4.545 5.61e-06 ***
Jitter.RAP     3.819e+02  8.700e+01   4.390 1.15e-05 ***
Shimmer       -4.653e+01  1.047e+01  -4.446 8.91e-06 ***
Shimmer.APQ5   8.920e+01  1.625e+01   5.490 4.20e-08 ***
Shimmer.APQ11 -3.427e+01  6.833e+00  -5.014 5.48e-07 ***
HNR           -9.958e-02  2.275e-02  -4.377 1.22e-05 ***
RPDE           3.177e+00  6.030e-01   5.269 1.42e-07 ***
DFA           -2.136e+00  7.085e-01  -3.015  0.00258 **
PPE           -3.842e+00  9.541e-01  -4.027 5.72e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.258 on 5861 degrees of freedom
Multiple R-squared:  0.9075,    Adjusted R-squared:  0.9073
F-statistic:  4425 on 13 and 5861 DF,  p-value: < 2.2e-16
```

(b) Resultado da Melhor Regressão Linear

Figura 3.3: Resultados dos modelos ajustados

De forma sucinta notamos que foram eliminados todas as variáveis com *p-value* superior a 0.05, ou seja, todas as variáveis em que a probabilidade de obter resultados seja pelo menos tão extremos quanto os resultados observados de um teste de hipótese estatística. Por último, utilizando a função *confint* pudemos calcular os coeficientes com 95(porcento) de certeza que não será necessário eliminar nenhum preditor em nenhum dos casos pois o zero não está contido em nenhum dos intervalos das variáveis, logo podemos concluir que são todos bons preditores adicionando mais uma camada de análise e consistência ao *dataset* (Figura 3.4)).

```
> confint(x7,level=0.95)
                    2.5 %        97.5 %
(Intercept)      0.15669677  4.249519e+00
age              0.05831463  7.839826e-02
sex             -1.59917927 -1.197369e+00
motor_UPDRS      1.21622250  1.238442e+00
Jitter...     -372.75976572 -1.665591e+02
Jitter.Abs.   7976.28531462  2.007736e+04
Jitter.RAP     211.36014685  5.524667e+02
Shimmer        -67.04788433 -2.601474e+01
Shimmer.APQ5    57.34712782  1.210565e+02
Shimmer.APQ11  -47.66190624 -2.086981e+01
HNR             -0.14417238 -5.498261e-02
RPDE             1.99479228  4.358926e+00
DFA             -3.52492718 -7.471328e-01
PPE             -5.71295354 -1.972037e+00
```

Figura 3.4: Intervalos de confiança para os parâmetros do modelo ajustado

3.2 Análise Preditiva utilizando um conjunto de dados de treino e teste

O segundo processo desenvolvido envolve um conjunto de treino e um conjunto de teste, onde o conjunto de treino é um conjunto de dados real utilizado para treinar o modelo, o modelo vai ver e aprender com esses dados, o *dataset* foi dividido em 75% treino, 25% teste. O conjunto de teste é um conjunto de dados de teste que descreve o *golden standard* utilizado para avaliar o modelo. O conjunto de teste só é utilizado quando um modelo está completamente treinado. Para tal foi desenvolvido o seguinte processo (Figura 3.5).

```
#Criar training e test data
## 75% of the sample size
smp_size <- floor(0.75 * nrow(m1))

## set the seed to make your partition reproducible
set.seed(123)
train_ind <- sample(seq_len(nrow(m1)), size = smp_size)

train <- m1[train_ind, ]
test <- m1[-train_ind, ]

dim(train)
dim(test)
dim(m2)
```

Figura 3.5: Conjunto de treino e teste

Com os *datasets* de treino e teste definidos de forma aleatória (75/25) começamos por gerar uma função de Regressão Linear tal como no processo anterior onde introduzimos também um termo de interacção entre *age* e *sex* por ser relevante (Figura 3.6(a)).

```
> summary(y1)

Call:
lm(formula = total_UPDRS ~ subject. + age + sex + test_time +
    motor_UPDRS + Jitter... + Jitter.Abs. + Jitter.RAP + Jitter.PPQ5 +
    Jitter.DDP + Shimmer + Shimmer.dB. + Shimmer.APQ3 + Shimmer.APQ5 +
    Shimmer.APQ11 + Shimmer.DDA + NHR + HNR + RPDE + DFA + PPE +
    age * sex, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0584 -2.1022 -0.4849  1.3796 11.5893

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.257e+00  1.282e+00  -0.981  0.32681
subject.       4.636e-02  4.458e-03  10.399  < 2e-16 ***
age            1.152e-01  7.991e-03  14.412  < 2e-16 ***
sex            4.050e+00  7.487e-01   5.409 6.67e-08 ***
test_time      2.868e-03  9.085e-04   3.157  0.00161 **
motor_UPDRS    1.214e+00  6.667e-03 182.155  < 2e-16 ***
Jitter...     -2.412e+02  8.111e+01  -2.973  0.00296 **
Jitter.Abs.    1.500e+04  3.647e+03   4.112 3.99e-05 ***
Jitter.RAP     5.340e+03  1.767e+04   0.302  0.76253
Jitter.PPQ5   -4.654e+01  7.365e+01  -0.632  0.52751
Jitter.DDP    -1.644e+03  5.891e+03  -0.279  0.78022
Shimmer       -5.413e+01  2.405e+01  -2.250  0.02447 *
Shimmer.dB.    1.039e+00  1.866e+00   0.557  0.57759
Shimmer.APQ3  -9.531e+03  1.769e+04  -0.539  0.59005
Shimmer.APQ5   1.346e+02  2.192e+01   6.139 9.05e-10 ***
Shimmer.APQ11 -6.481e+01  1.035e+01  -6.264 4.12e-10 ***
Shimmer.DDA    3.174e+03  5.896e+03   0.538  0.59045
NHR           -5.575e+00  2.343e+00  -2.379  0.01738 *
HNR           -1.050e-01  2.632e-02  -3.991 6.70e-05 ***
RPDE           3.364e+00  6.978e-01   4.821 1.47e-06 ***
DFA           -2.740e+00  8.946e-01  -3.063  0.00221 **
PPE           -3.594e+00  1.105e+00  -3.253  0.00115 **
age:sex       -8.851e-02  1.140e-02  -7.765 1.01e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.184 on 4383 degrees of freedom
Multiple R-squared:  0.913,     Adjusted R-squared:  0.9126
F-statistic:  2092 on 22 and 4383 DF,  p-value: < 2.2e-16
```

(a) Resultado da Primeira Regressão Linear

```
> summary(y6)

Call:
lm(formula = total_UPDRS ~ subject. + age + sex + test_time +
    motor_UPDRS + Jitter... + Jitter.Abs. + Jitter.RAP + Shimmer +
    Shimmer.APQ5 + Shimmer.APQ11 + NHR + HNR + RPDE + DFA + PPE +
    age * sex, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8740 -2.1015 -0.4903  1.3930 11.5538

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.295e+00  1.268e+00  -1.022 0.307038
subject.       4.622e-02  4.438e-03  10.416  < 2e-16 ***
age            1.151e-01  7.972e-03  14.434  < 2e-16 ***
sex            4.060e+00  7.474e-01   5.432 5.86e-08 ***
test_time      2.860e-03  9.070e-04   3.153 0.001627 **
motor_UPDRS    1.215e+00  6.660e-03 182.362  < 2e-16 ***
Jitter...     -2.635e+02  6.381e+01  -4.129 3.71e-05 ***
Jitter.Abs.    1.540e+04  3.462e+03   4.450 8.80e-06 ***
Jitter.RAP     4.002e+02  1.048e+02   3.819 0.000136 ***
Shimmer       -4.895e+01  1.166e+01  -4.197 2.76e-05 ***
Shimmer.APQ5   1.263e+02  1.898e+01   6.655 3.18e-11 ***
Shimmer.APQ11 -6.072e+01  8.611e+00  -7.051 2.05e-12 ***
NHR           -5.593e+00  2.035e+00  -2.748 0.006025 **
HNR           -1.025e-01  2.601e-02  -3.942 8.20e-05 ***
RPDE           3.407e+00  6.893e-01   4.942 8.01e-07 ***
DFA           -2.817e+00  8.810e-01  -3.198 0.001394 **
PPE           -3.462e+00  1.088e+00  -3.181 0.001480 **
age:sex       -8.853e-02  1.138e-02  -7.781 8.91e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.182 on 4388 degrees of freedom
Multiple R-squared:  0.913,     Adjusted R-squared:  0.9127
F-statistic:  2710 on 17 and 4388 DF,  p-value: < 2.2e-16
```

(b) Resultado da Melhor Regressão Linear

Figura 3.6: Resultados dos modelos ajustados utilizando dataset treino/teste

Porém em vez de gerarmos automaticamente a melhor função de Regressão Linear através da função *step*, tomamos a iniciativa de desenvolver manualmente essa função. Para tal, retiramos manualmente sempre a variável com *p-value* maior e criamos uma nova função de Regressão Linear. Fizemos esse processo até não haver qualquer variável com *p-value* superior a 0.05 eliminando um total de 6 variáveis (Figura 3.6(b)) e analisamos os coeficientes com 95(porcento) de certeza através da função *confint* (Figura 3.7).

```
> #Calcular os coeficientes com 95% de certeza
> confint(y6,level=0.95)
                    2.5 %        97.5 %
(Intercept)   -3.780718e+00  1.190384e+00
subject.       3.752437e-02  5.492534e-02
age            9.944106e-02  1.307003e-01
sex            2.594804e+00  5.525326e+00
test_time      1.081603e-03  4.637886e-03
motor_UPDRS    1.201475e+00  1.227589e+00
Jitter...     -3.885556e+02 -1.383569e+02
Jitter.Abs.    8.617858e+03  2.219084e+04
Jitter.RAP     1.947788e+02  6.056667e+02
Shimmer       -7.180957e+01 -2.608280e+01
Shimmer.APQ5   8.908541e+01  1.634952e+02
Shimmer.APQ11 -7.760189e+01 -4.383794e+01
NHR           -9.583405e+00 -1.602411e+00
HNR           -1.535079e-01 -5.153843e-02
RPDE           2.055407e+00  4.758242e+00
DFA           -4.544717e+00 -1.090265e+00
PPE           -5.595476e+00 -1.327876e+00
age:sex       -1.108349e-01 -6.622316e-02
```

Figura 3.7: Intervalos de confiança para os parâmetros do modelo ajustado com treino/teste

3.3 Análise Preditiva utilizando um modelo misto

O nosso conjunto de dados apresenta apenas 42 indivíduos diferentes para um total de mais de 5 mil registos, o que nos leva a crer que alguns registos, não são exatamente independentes. Posto isto, decidimos desenvolver um modelo misto, uma vez que este garante uma certa independência entre os dados. O modelo é chamado de misto porque possui efeitos fixos, como na regressão linear, e efeitos aleatórios. O útlimo (efeito aleatório) permite assumir um valor de resposta diferente para cada fator, assumindo diferentes interceptos (aleatórios) para cada resposta. Os modelos lineares tradicionais são desenvolvidos no R da seguinte forma: y x1 + x2*x3. Os modelos mistos, por sua vez, devem ser construídos y x1 + efeito fixo | efeito aleatório.

Posto isto, desenvolvemos o nosso modelo misto onde definimos os atributos *subject.* e *test_ time* como efeito aleatório e os restantes atributos como efeito fixo, sendo que para isso usamos o comando lmer.

Decidimos colocar como efeito aleatório *subject.* e *test_ time*, porque acreditamos serem os únicos atributos não independentes. Para além disso, desenvolvemos dois modelos que usam o efeito aleatório. Num dos modelos selecionamos todos os atributos (model1) e no outro modelo selecionamos apenas os atributos do melhor modelo do subcapítulo anterior (model2). Para a execução destes modelos optamos, ainda, por usar os dados de treino e teste.

```
> anova(model1,model2)
Data: train
Models:
model2: total_UPDRS ~ subject. + age + sex + test_time + motor_UPDRS +
model2:     Jitter... + Jitter.Abs. + Jitter.RAP + Shimmer + Shimmer.APQ5 +
model2:     Shimmer.APQ11 + NHR + HNR + RPDE + DFA + PPE + (1 | subject.) +
model2:     (1 | test_time)
model1: total_UPDRS ~ subject. + age + sex + test_time + motor_UPDRS +
model1:     Jitter... + Jitter.Abs. + Jitter.RAP + Jitter.PPQ5 + Jitter.DDP +
model1:     Shimmer + Shimmer.dB. + Shimmer.APQ3 + Shimmer.APQ5 + Shimmer.APQ11 +
model1:     Shimmer.DDA + NHR + HNR + RPDE + DFA + PPE + (1 | subject.) +
model1:     (1 | test_time)
       npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
model2   20 3408.9 3536.6 -1684.4   3368.9
model1   25 3415.7 3575.5 -1682.9   3365.7 3.1209  5     0.6814
```

Figura 3.8: Análise Preditiva utilizando um modelo misto

Para a avaliação do desempenho destes dois modelos usamos o ANOVA e ainda o RMSE. A partir do ANOVA conseguimos extraiar o AIC.

Capítulo 4

Análise de Resultados

Neste capítulo iremos apresentar o desempenho de todos os modelos executados, sendo que para isso usamos medidas estatísticas como *p-value*, *Akaike information criterion (AIC), Adjusted R-squared* e o *Root Mean Square Error (RMSE)*. Apesar de existirem outras medidas estatísticas acreditamos que estas são as que mais se adequam aos nossos modelos. As três primeiras dão-nos informações relevantes para qualquer modelo, seja ele um modelo de classificação ou regressão. Já a última, *Root Mean Square Error*, é mais adequada para problemas de regressão quando comparada com a *accuracy* que se adequa mais a problemas de classificação.

Ao analisar os modelos em que não dividimos os dados em treino e teste, percebemos que só por si o modelo já tem um *Ajusted R-squared* elevado, ou seja, os nossos preditores são muito bons a explicar a variável Y (*total_UPDRS*). No entanto, devemos privilegiar modelos mais simples pelo que analisando o *p-value* e o *AIC* percebemos que o melhor modelo é o x6, ou seja, o modelo que usa os preditores *"age", "sex", "motor_UPDRS", "Jitter...", "Jitter.Abs.", "Jitter.RAP", "Shimmer", "Shimmer.APQ5", "Shimmer.APQ11", "HNR", "RPDE", "DFA", "PPE", "subject."* e *"test_time"*. Este modelo apresenta um *AIC* de 13618.72. Posto isto, ao usar este modelo para fazer a predição percebemos que apresenta um *RMSE* de 3.185192.

À semelhança do que acontece com os modelos em que não dividimos os dados em treino e teste, os modelos com dados de treino e teste também apresentam um *Adjusted R-squared* elevado. Os preditores para o melhor modelo que usa dados de teste e treino são iguais ao do modelo que não usa dados de treino e teste, sendo estes *"age", "sex", "motor_UPDRS", "Jitter...", "Jitter.Abs.", "Jitter.RAP", "Shimmer", "Shimmer.APQ5", "Shimmer.APQ11", "HNR", "RPDE", "DFA", "PPE", "subject."* e *"test_time"*.Este modelo apresenta um *AIC* de 10252.02. Ao usar este modelo para fazer a predição temos um *RMSE* de 3.165032.

Nos modelos mistos, o desempenho melhora de forma significativa em relação aos modelos anteriores. Posto isto, e analisando o *AIC*, verificamos que o melhor modelo é o "model2", ou seja, o modelo que usa os preditores *"age", "sex", "motor_UPDRS", "Jitter...", "Jitter.Abs.", "Jitter.RAP", "Shimmer", "Shimmer.APQ5", "Shimmer.APQ11", "HNR", "RPDE", "DFA", "PPE", "subject."* e *"test_time"*. Este modelo apresenta uma *AIC* de 3408.9 enquanto que o "model1"(usa todos os preditores) apresenta um *AIC* de 3415.7.

No entanto, se analisarmos o valor do *RMSE* para ambos os modelos preditivos percebemos que o "model1"apresenta um valor ligeiramente melhor que o "model2"(1.011055 e 1.011119, respetivamente).

Em suma os melhores modelos para cada uma das três abordagens usam sempre os mesmos preditores. A principal diferença está nos valores de *AIC* e *RMSE* que variam de modelo para modelo.

Capítulo 5

Conclusão

O principal objetivo deste projeto era analisar e explorar um *dataset* com o auxílio de métodos de aprendizagem estatística de forma a obter algum conhecimento a partir dos dados. Para isso utilizamos o *dataset Parkinsons Telemonitoring Data Set*, onde aplicamos um conjunto de tarefas para extrair conhecimento.

Através dos resultados obtidos, e analisando os modelos desenvolvidos, concluímos que o facto de dividirmos (ou não) o conjunto de dados em treino e teste, privilegiar modelos mais simples e aplicar efeito aleatório foi determinante para o desempenho dos modelos. Se dividirmos os dados em treino e teste temos modelos com *AIC* e *RMSE* melhor quando comparado com os modelos que não utilizam dados de treino e teste, não só obtemos um melhor resultado como o modelo é mais realista e não se corre o risco de *overfitting*. Por outro lado, ao utilizar modelos mais simples (com menos preditores) o desempenho dos modelos melhora, ainda que ligeiramente. Por último, e se assumirmos que há dados não independentes, os modelos que aplicam efeitos aleatórios e efeitos fixos melhoram de forma significativa o valor de *AIC* e *RMSE* em comparação com os modelos que apenas efeitos fixos.

Posto isto, e tendo em conta o valor de *AIC* e *RMSE*, o melhor modelo é o "model2", ou seja, aquele que usa efeitos fixos e aleatórios.

5.1 Limitações e Recomendações para Trabalhos Futuros

O presente trabalho respondeu às questões centrais levantando, no entanto, alguns pontos que podem ser aprofundados. O primeiro ponto incide sobre a parametrização dos algoritmos. Apesar de não serem considerados descabidos os resultados obtidos, não se exclui a hipótese de que, com outro tipo de afinamento, os algoritmos apresentados poderiam obter melhores resultados, principalmente no *lmer*. O segundo ponto recai sobre a divisão do *dataset* em treino e teste. Existe várias formas de dividir os dados em treino e teste, sendo que, algumas delas poderão melhorar os resultados obtidos. Por exemplo, se usássemos o *cross validation* é possível que os resultados obtidos fossem melhores. No entanto, também devemos perceber se a diferença compensa tendo em conta que o *cross validation* é computacionalmente mais caro.

Capítulo 6

Bibliografia

Leite, M., Gaiarsa, M., Medeiros, L. (2018). Introdução aos Modelos Mistos, Grupo de discussão Modelos Mistos da LAGE - IB/USP. Disponível online em: https://rpubs.com/melinatarituba/309285?fbclid=IwAR0jQx0q0xIAkX2QfMvek9X_wlZ2bgKggFlrs8UxnOL8NiSSjwQ1uadHLo0

Winter, B., (2013). Linear models and linear mixed effects models in R with linguistic applications, University of California, Merced, Cognitive and Information Sciences. Disponível online em: https://arxiv.org/ftp/arxiv/papers/1308/1308.5499.pdf?fbclid=IwAR1-gklBgrncdmWikmWSjqDB3I-oysynKgLlHvfvYrWVIzZTakVlRHTSZa4