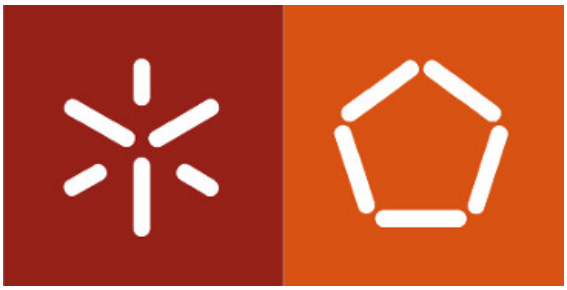


Universidade do Minho
Mestrado Integrado em Engenharia Informática
Ciências de Dados

Personal Medical Cost

Aprendizagem Automática I

Carolina Cunha, A80142
Bruno Veloso, A78352

27 de dezembro de 2020

Conteúdo

Lista de Figuras

Resumo

O presente projeto consistiu na análise de um conjunto de dados real, de forma a prever os custos médicos individuais cobrados pelo seguro de saúde. Para este efeito, foi realizado, numa primeira instância, um tratamento de dados que permitisse inferir de que forma os diferentes preditores afetam a variável de interesse. Subsequentemente, foram realizados quatro modelos supervisionados de regressão linear, aos quais se removeu, de forma gradual, as variáveis menos significativas para o estudo. Sobre estes modelos foram realizados testes de avaliação de qualidade através do *Validation Set Approach* e, posteriormente, de validação cruzada, recorrendo aos métodos *K-fold Cross Validation* e *Leave-One-Out Cross Validation*, para comparar os diferentes modelos ajustados. Dada a quantidade de erros associados à previsão da variável de interesse resultante dos modelos ajustados, seguiu-se uma manipulação da variável de interesse, tomando uma abordagem logarítmica sobre a mesma. Verificou-se uma distribuição Gaussiana dos valores da nova variável de interesse, contrariamente à inclinação para a direita dos valores da abordagem anterior. Para esta interpretação, foram realizados dois modelos supervisionados de regressão linear, aos quais se removeu, de igual forma, os preditores com menor significância. Posteriormente, procedeu-se à realização dos mesmos testes de avaliação de qualidade. Foi observada uma melhoria da qualidade dos novos modelos relativamente à abordagem anterior. Apesar disso, a regressão linear mostrou-se insuficiente para o sucesso da previsão da variável de interesse. Por este motivo, optou-se por realizar um estudo de regressão do *K-Nearest-Neighbors* sobre as duas abordagens para a previsão da variável de interesse. Com base na análise dos resultados, observou-se que o modelo capaz de responder às questões de interesse com maior exatidão consistia naquele que recorria à variável de interesse original, através da regressão do *K-Nearest-Neighbors*.

Capítulo 1

1.1 Introdução

Para a realização deste projeto, pretende-se, a um conjunto de dados reais, experimentar vários métodos de aprendizagem estatística abordados na unidade curricular de Aprendizagem Automática I. Com o objetivo de responder a questões concretas, foram desenvolvidos modelos de aprendizagem com a capacidade de prever os resultados de conjuntos de dados, recorrendo à linguagem de programação R.

1.2 Descrição do Problema

O conjunto de dados utilizados pertence ao *dataset* **Medical Cost Personal** [1]. O objetivo deste estudo é inferir de que forma as diferentes variáveis afetam os custos médicos individuais cobrados pelo seguro de saúde, assumindo o valor da variável de interesse Y.

1.3 Descrição do Conjunto de Dados

O conjunto de dados real escolhido é constituído por 1338 registos (linhas) e 7 covariáveis (colunas). De seguida serão apresentados, detalhadamente, os preditores existentes neste conjunto de dados, bem como a variável de decisão.

1.3.1 Preditores

- *age* – Idade do beneficiário principal (variável numérica);
- *sex* – Género do contratante de seguro (variável categórica: *male, female*);
- *bmi* – Índice de massa corporal, em kg/m2, relacionando a altura com o peso (variável numérica);
- *children* – Número de filhos cobertos pelo seguro / Número de dependentes (variável numérica);
- *smoker* – Indicador de tabagismo (variável categórica: *yes, no*);
- *region* – Área residencial do beneficiário nos EUA (variável categórica: *northeast, northwest, southeast, southwest)*);

1.3.2 Variável de Interesse

Dada pela variável *charges*, variável numérica que representa os custos médicos individuais cobrados pelo seguro de saúde.

1.4 Questões de Interesse

1. Que variáveis influenciam o valor dos custos médicos individuais cobrados pelo seguro de saúde?
2. O seguro de saúde é mais vantajoso para famílias numerosas?
3. De que forma é que os níveis de índice de massa corporal influenciam os custos médicos individuais cobrados pelo seguro?
4. Indivíduos fumadores estão sujeitos a custos mais elevados?

Capítulo 2

2.1 Análise Exploratória dos Dados

De forma a estudar do conjunto de dados escolhido é necessário realizar uma análise exploratória dos dados, de modo a compreender o que significam, bem como as relações que apresentam entre si. Os gráficos 2.1 apresentam as variáveis discretas do modelo.

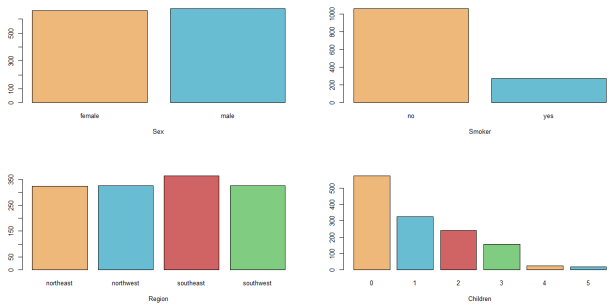


Figura 2.1: Valores das variáveis discretas.

Através da análise dos gráficos previamente apresentados, observa-se uma distribuição simétrica dos dados pelos sexos feminino e masculino e pelas diferentes regiões de residência. Por outro lado, evidencia-se uma grande discrepância nos dados relativos à informação de fumador. Verifica-se que o número de dados relativos a famílias é progressivamente menor à medida que aumenta o número de dependentes.

No que diz respeito às variáveis contínuas, é imprescindível analisar a existência de *outliers* - observações que têm resíduos de valor elevado quando comparados com outras observações - nos preditores. Pela análise da Figura 2.2, verifica-se a existência de uma elevada quantidade de possíveis *outliers* na variável de interesse. Quanto à variável *Bmi*, verificam-se alguns possíveis valores *outlier*.

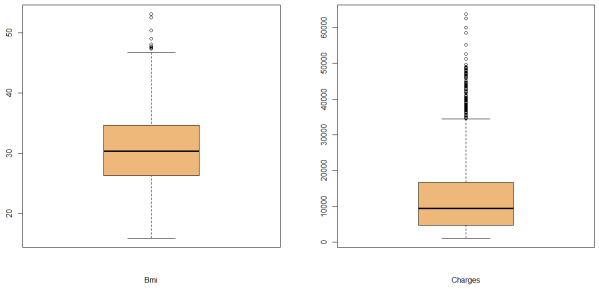


Figura 2.2: Valores das variáveis contínuas.

Relativamente à variável de interesse, existe uma evidente inclinação para a direita dos seus valores (Figura 2.3).

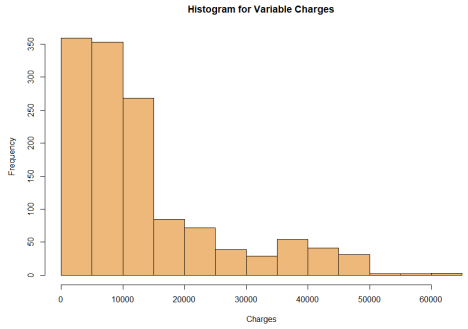


Figura 2.3: Histograma da Variável *Charges*.

2.2 Correlação entre os Preditores e a Variável de Interesse

A observação dos gráficos 2.4 permite aferir custos médicos mais elevados para fumadores em relação a não fumadores e para famílias com menor número de dependentes, em relação a maior número de dependentes.

Não se observa relação entre custos médicos e as variáveis sexo e região de residência.

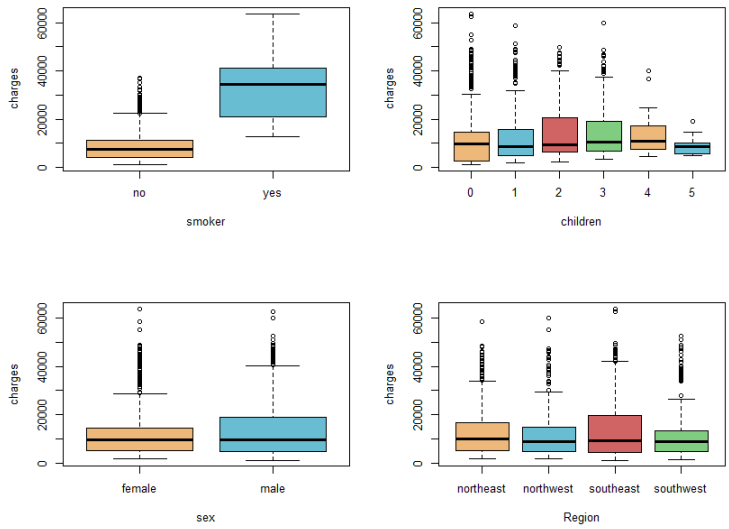


Figura 2.4: Correlação entre as variáveis discretas e a variável de interesse

A análise da Figura 2.5 permite verificar o aumento do custo médico individual com a idade e índice de massa corporal.

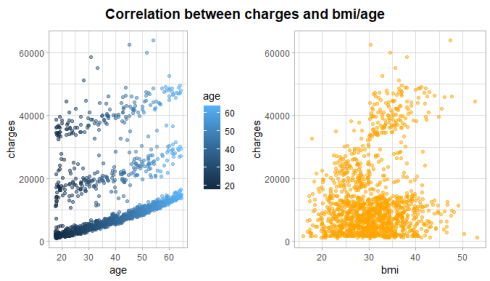


Figura 2.5: Correlação entre bmi e charges e age e charges

2.3 Regressão Linear

O estudo da amostra tratou-se segundo um problema de **regressão linear múltipla supervisionado**, uma vez que a variável de interesse - custo médico cobrado pelo seguro de saúde - é uma variável aleatória contínua e admite erros gaussianos. Propomos prever de que forma os preditores influenciam esta variável.

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon^1 \tag{2.1}$$

Pela Equação 2.1, pode deduzir-se que β_j é o aumento médio em Y quando X_j é aumentado por uma unidade, e todos os restantes X são mantidos constantes.

[1] p corresponde ao número de covariáveis existentes no modelo

2.4 Ajuste do Modelo de Regressão

2.4.1 Primeiro Ajuste

De forma a ajustar o modelo de regressão linear, foi utilizada a função *lm* sobre a amostra, a fim de prever a variável *charges*. No primeiro ajuste, foram utilizados todos os preditores do modelo, permitindo excluir variáveis não significativas para a variável de interesse.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
regionsouthwest   -960.0      477.9  -2.009 0.044765 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Figura 2.6: Modelo de Regressão - Totalidade de preditores

A significância de cada preditor determina-se pelo teste de hipóteses. Desta forma, considere-se

$$H_0 : \beta_j = 0 \text{ vs } H_1 : \beta_j \neq 0$$

As variáveis cujo *p_ value* - probabilidade da hipótese nula - seja menor do que 5%, serão consideradas significativas para o resultado final. Com base na Figura 2.6, verifica-se que o preditor *sexmale* apresenta um *p_ value* superior a 0,05, pelo que a variável sexo foi excluída. A variável *regionnorthwest* apresenta um *p_ value* elevado, pelo que será agrupado com a variável padrão (*regionnortheast*), numa tentativa de melhorar o ajuste do modelo, sem eliminar potenciais variáveis significativas.

O **coeficiente de determinação** R^2 foi utilizado para avaliar o peso da variabilidade de X que explica Y. Assim, este modelo explica a variabilidade dos dados em 74.94%. A qualidade dos ajustes do modelo de regressão foi avaliada pelo AIC *(Akaike Information Criterion)* que, para este modelo, é 23316.43.

2.4.2 Segundo Ajuste

Para o segundo ajuste, foi removido o preditor *sex*, obtendo-se os seguintes resultados.

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11990.27     978.76 -12.250  < 2e-16 ***
age                256.97      11.89  21.610  < 2e-16 ***
bmi                338.66      28.56  11.858  < 2e-16 ***
children           474.57     137.74   3.445 0.000588 ***
smokeryes        23836.30     411.86  57.875  < 2e-16 ***
regionnorthwest   -352.18     476.12  -0.740 0.459618
regionsoutheast  -1034.36     478.54  -2.162 0.030834 *
regionsouthwest   -959.37     477.78  -2.008 0.044846 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6060 on 1330 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7496
F-statistic: 572.7 on 7 and 1330 DF,  p-value: < 2.2e-16
```

Figura 2.7: Modelo de Regressão - Remoção do preditor sex

Com base nos resultados, verifica-se uma melhoria na qualidade do ajuste após remoção deste preditor (R^2 de 74.96%). O AIC deste modelo é de 23314.58. De acordo com a equação da **probabilidade relativa de um modelo**, este modelo tem 0,3965 vezes mais probabilidade de minimizar a perda de informação.

$$exp(\frac{AIC_{min} - AIC_i}{2})$$

2.4.3 Terceiro Ajuste

Como terceiro ajuste, optou-se por agrupar os dados relativos às regiões norte, tendo sido divididos os valores de *region* por *north, southeast, southwest*. Os resultados obtidos foram os seguintes:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -12165.38     949.54 -12.812  < 2e-16 ***
age               257.01      11.89  21.617  < 2e-16 ***
bmi               338.64      28.55  11.860  < 2e-16 ***
children          471.54     137.66   3.426 0.000632 ***
smokeryes       23843.87     411.66  57.921  < 2e-16 ***
regionsoutheast   -858.47     415.21  -2.068 0.038873 *
regionsouthwest   -782.75     413.76  -1.892 0.058734 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6059 on 1331 degrees of freedom
Multiple R-squared:  0.7508,    Adjusted R-squared:  0.7497
F-statistic: 668.3 on 6 and 1331 DF,  p-value: < 2.2e-16
```

Figura 2.8: Modelo de Regressão - Junção das regiões norte

Registou-se um incremento no R^2 para 74.97%, e no valor de AIC para 23313.13.

2.4.4 Quarto Ajuste

Numa tentativa de aprimorar a qualidade do ajuste do modelo, removeu-se o preditor *region*. No entanto, verificou-se paradoxalmente redução da qualidade (Figura 2.9). Além disso, o valor de AIC aumenta para 23314.96, inviabilizando a qualidade deste ajuste. Assim sendo, consideramos o modelo optimizado após o terceiro ajuste.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12102.77     941.98 -12.848  < 2e-16 ***
age            257.85      11.90  21.675  < 2e-16 ***
bmi            321.85      27.38  11.756  < 2e-16 ***
children       473.50     137.79   3.436 0.000608 ***
smokeryes    23811.40     411.22  57.904  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6068 on 1333 degrees of freedom
Multiple R-squared:  0.7497,    Adjusted R-squared:  0.7489
F-statistic: 998.1 on 4 and 1333 DF,  p-value: < 2.2e-16
```

Figura 2.9: Modelo de Regressão - Remoção do preditor region

2.5 Avaliação da Qualidade

A qualidade das predições geradas a partir dos modelos ajustados foi avaliada pelos **métodos de reamostragem**.

2.5.1 Validation Set Approach

Nesta primeira abordagem, foi utilizada a função *predict*, que gera previsões a partir de um modelo de regressão linear. Os valores reais do conjunto de dados de teste (20% da amostra) foram comparados com os resultados obtidos (dados de treino, 80% da amostra) , permitindo avaliar a exatidão do modelo.

		Modelo 1	Modelo 2	Modelo 3	Modelo 4
Com *Outliers*	R^2	0.7479	0.7481	0.7483	0.748
	AIC	18611.12	18609.13	18607	18606.51
	RMSE	6492.8559	6493.5776	6496.2903	6521.0413
	Accuracy (Yes/No)	60 / 208	60 / 208	61 / 207	64 / 204
Sem *Outliers*	R^2	0.7608	0.761	0.7609	0.7608
	AIC	18472.43	18470.55	18469.91	18468.51
	RMSE	5937.7281	5932.0520	5925.5114	5952.6318
	Accuracy (Yes/No)	49 / 217	49 / 217	51 / 215	50 / 216

Figura 2.10: Resultados das Previsões com e sem *outliers*

Considerando os valores superiores a 50000$ como *outliers*, verifica-se, pela análise da Figura 2.10, que o resultado das previsões é melhor quando são removidos os *outliers*, à custa de uma menor exatidão, para um erro de 1000$. Na figura seguinte, encontram-se representados os resultados das previsões do melhor modelo ajustado (modelo 3).

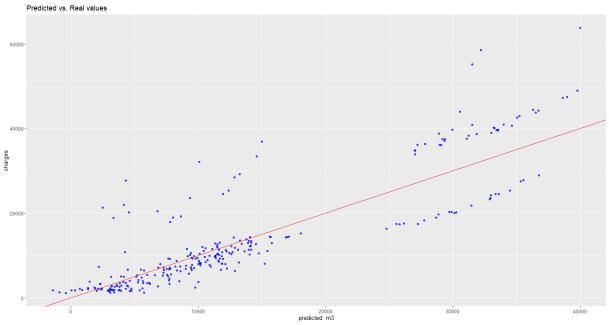


Figura 2.11: Resultado Gráfico das Previsões

2.5.2 Cross Validation

K-fold Cross Validation

O algoritmo deste método divide aleatoriamente o conjunto de dados em K subconjuntos, ajustando o modelo com K-1 subconjuntos. Para cada iteração, é escolhido um subconjunto que será utilizado para testar o modelo. Para medir a qualidade do modelo durante a validação cruzada, foram consideradas as seguintes variáveis [2]:

1. **R-squared** (R^2) - Representa a correlação quadrada entre os valores de resultados observados e os valores previstos pelo modelo. Quanto maior o valor de R^2 ajustado, melhor é o modelo;
2. **Root Mean Squared Error** (RMSE) - Mede o erro médio de predição feito pelo modelo ao predizer o resultado de uma observação. Quanto menor o seu valor, melhor é o modelo;
3. **Mean Absolute Error** (MAE) - Alternativa ao RMSE, menos senvível a *outliers*. Corresponde à diferença absoluta média entre os resultados observados e previstos. Quanto menor o seu valor, melhor é o modelo.

		Modelo 1	Modelo 2	Modelo 3	Modelo 4
K = 5	RMSE	6076.406	6073.937	6071.139	6066.238
	R^2	0.7484877	0.7476499	0.7478529	0.7504719
	MAE	4196.064	4201.173	4201.044	4189.172
K = 10	RMSE	6070.575	6060.319	6056.156	6070.119
	R^2	0.7499458	0.7464654	0.7468764	0.7504596
	MAE	4200.105	4203.399	4202.151	4193.364

Figura 2.12: Resultados do K-fold Cross Validation

Leave-One-Out Cross Validation

Neste modelo, os dados de tamanho N são divididos em N-1 para o conjunto de dados de treino, pelo que o conjunto de dados de teste é composto por apenas um dado. O modelo é ajustado usando os dados de treino, cuja validação é realizada através do dado de teste. Este processo é repetido N vezes.

	Modelo 1	Modelo 2	Modelo 3	Modelo 4
RMSE	6087.388	6083.14	6079.641	6083.813
R^2	0.7471345	0.7474863	0.7477757	0.7474283
MAE	4202.09	4199.76	4201.052	4197.328

Figura 2.13: Resultados do Leave-One-Out Cross Validation

Utilizando este método, obtêm-se valores muito elevados de erros.

2.6 Manipulação da Variável de Interesse

A avaliação da qualidade do modelo revelou uma previsão inadequada da variável de interesse tendo por base os preditores do modelo. Por este motivo, foi aplicada a função *logaritmo* sobre a variável

de interesse, com o intuito de optimizar a previsão do modelo. Os valores obtidos seguem uma distribuição Gaussiana, visível na Figura 2.14.

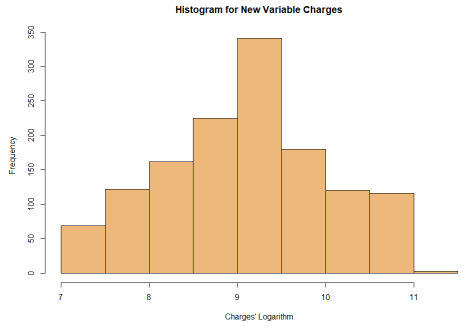


Figura 2.14: Histograma da Variável de Interesse - Logaritmo de Charges

A nova variável de interesse (Figura 2.15) permitiu eliminar todos os *outliers* existentes na variável de interesse anterior.

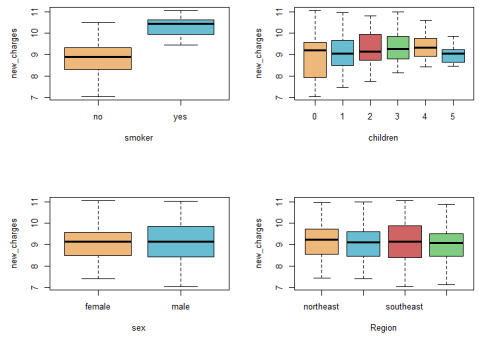


Figura 2.15: Correlação entre as variáveis discretas e a nova variável de interesse

Através da Figura 2.16, verifica-se uma maior dispersão dos valores da variável new_charges para os diferentes valores de índice de massa corporal.

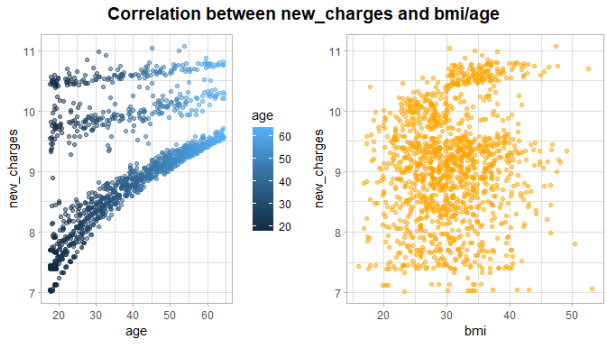


Figura 2.16: Correlação entre as variáveis contínuas e a nova variável de interesse

2.7 Ajuste do Novo Modelo de Regressão

2.7.1 Primeiro Ajuste

Para o novo modelo de regressão linear, foi utilizada a função *lm* sobre a amostra, a fim de prever a variável *new_charges*. No primeiro ajuste, foram utilizados todos os preditores do modelo.

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.0305581  0.0723960  97.112  < 2e-16 ***
age              0.0345816  0.0008721  39.655  < 2e-16 ***
sexmale         -0.0754164  0.0244012  -3.091 0.002038 **
bmi              0.0133748  0.0020960   6.381 2.42e-10 ***
children         0.1018568  0.0100995  10.085  < 2e-16 ***
smokeryes        1.5543228  0.0302795  51.333  < 2e-16 ***
regionnorthwest -0.0637876  0.0349057  -1.827 0.067860 .
regionsoutheast -0.1571967  0.0350828  -4.481 8.08e-06 ***
regionsouthwest -0.1289522  0.0350271  -3.681 0.000241 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4443 on 1329 degrees of freedom
Multiple R-squared:  0.7679,    Adjusted R-squared:  0.7666
F-statistic: 549.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Figura 2.17: Modelo de Regressão - Totalidade de preditores

Este modelo explica a variabilidade dos dados em 77.66%, com AIC de -2162.046. Este modelo revelou superioridade face aos modelos anteriormente estudados.

2.7.2 Segundo Ajuste

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.9987690  0.0703368  99.504  < 2e-16 ***
age              0.0345877  0.0008728  39.628  < 2e-16 ***
sexmale         -0.0752309  0.0244224  -3.080  0.00211 **
bmi              0.0133699  0.0020979   6.373 2.55e-10 ***
children         0.1013081  0.0101039  10.027  < 2e-16 ***
smokeryes        1.5556774  0.0302970  51.348  < 2e-16 ***
regionsoutheast -0.1253389  0.0304715  -4.113 4.14e-05 ***
regionsouthwest -0.0969605  0.0303652  -3.193  0.00144 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4447 on 1330 degrees of freedom
Multiple R-squared:  0.7674,    Adjusted R-squared:  0.7661
F-statistic: 626.7 on 7 and 1330 DF,  p-value: < 2.2e-16
```

Figura 2.18: Modelo de Regressão - Junção dos valores *northwest e northeast*

Realizou-se um segundo ajuste agrupando as regiões norte, visto que não revelavam diferenças significativas entre si (explicado anteriormente). Este ajuste não trouxe qualquer benefício para a qualidade (R2 76.61% e AIC -2160.688).

2.8 Avaliação da Qualidade do Novo Modelo

2.8.1 Abordagem da Validação do Conjunto

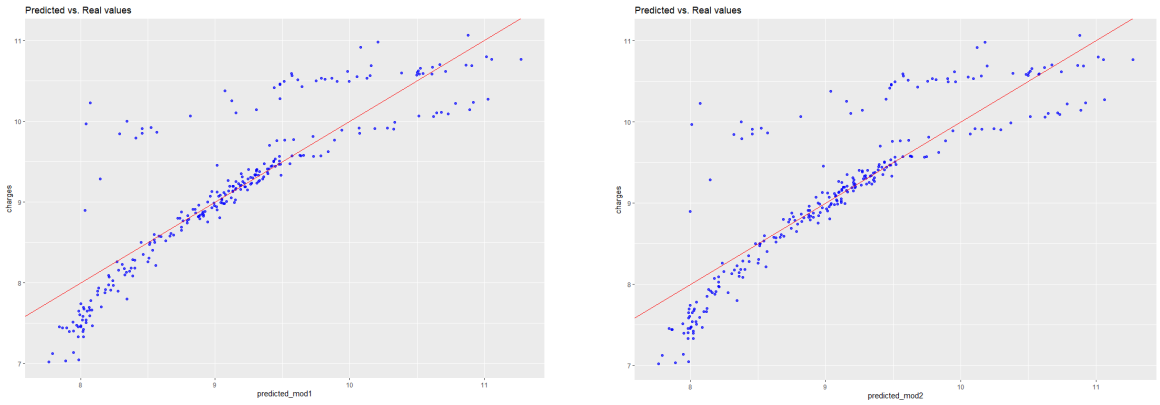
Foi utilizada a função *predict*, que utiliza um modelo de regressão linear. Os valores reais do conjunto de dados de teste foram comparados com os resultados obtidos, permitindo avaliar a exatidão do modelo. Para a avaliação da qualidade dos modelos, o conjunto de dados foi dividido em 80% para dados de treino e 20% para dados de teste. Os resultados dos vários modelos ajustados foram analisados.

	Modelo 1	Modelo 2
R^2	0.7686	0.7682
AIC	-1774.151	-1773.051
RMSE	0.482632	0.482817
Accuracy (Yes/No) - 0.5	206 / 62	206 / 62
Accuracy (Yes/No) - 0.3	175 / 93	178 / 90

Figura 2.19: Resultados das Previsões do Novo Modelo

Como se pode verificar pela Figura 2.19, o resultado das previsões é melhor quando são não são agrupadas as regiões *northeast* e *northwest*. Em contrapartida, o segundo modelo ajustado apresenta uma exatidão mais elevada quando o erro é de 0.3.

Na figura que se segue, encontram-se representados os resultados das previsões do melhor modelo ajustado (modelo 3).



(a) Modelo 1 (b) Modelo 2

Figura 2.20: Resultado Gráfico das Novas Previsões

2.8.2 Cross Validation

K-fold Cross Validation

		Modelo 1	Modelo 2
K = 5	RMSE	0.4463375	0.4450865
	R^2	0.7671421	0.7662693
	MAE	0.2800966	0.2803187
K = 10	RMSE	0.4435475	0.44435
	R^2	0.7671734	0.7637346
	MAE	0.2792226	0.2799781

Figura 2.21: Resultados do K-fold Cross Validation para o Novo Modelo

Leave-One-Out Cross Validation

	Modelo 1	Modelo 2
RMSE	0.4458979	0.4461195
R^2	0.76468	0.764445
MAE	0.2796067	0.2800271

Figura 2.22: Resultados do Leave-One-Out Cross Validation para o Novo Modelo

Analisando os resultados obtidos, confirma-se uma ligeira melhoria no ajuste do modelo 1 relativamente ao segundo modelo. No entanto, os erros em ambos os modelos revelam-se inaceitavelmente elevados para valores altos de Y.

A regressão linear mostrou-se insuficiente para a previsão adequada da variável de interesse, pelo que se prosseguiu para o estudo da regressão dos **K-Nearest-Neighbors**.

2.9 K-Nearest-Neighbors

A regressão KNN (*K-Nearest-Neighbors*) é um método não paramétrico que aproxima a associação entre as variáveis independentes e o resultado contínuo pela média das observações na mesma vizinhança [3].

2.9.1 Variável de Interesse Original

Para utilizar o modelo numa regressão do *K-Nearest-Neighbors*, converteram-se os dados categóricos em valores booleanos ou caracterizações não binárias, de forma a trabalhar apenas com atributos numéricos. Seguidamente, dividiu-se o conjunto de dados em dados de treino e dados de teste, na proporção 80/20. Utilizou-se a função *knn.reg*, para ajuste do modelo. Considerando um erro de 100$, foi calculada a exatidão de cada ajuste para os diferentes valores de K. Os resultados são visíveis na Figura 2.23.

	K = 1	K = 2	K = 3
Accuracy (Yes/No)	252 / 16	257 / 11	249 / 19
RMSE	199.1646	301.889	332.2719

Figura 2.23: Resultados da Regressão KNN para diferentes valores de K

Analisando os resultados obtidos, conclui-se que o valor de K para o qual o modelo de regressão é melhor ajustado é 1, uma vez que apresenta erros de previsão menores. Na Figura 2.24, verifica-se que o modelo é menos eficaz em prever corretamente valores mais elevados dos custos cobrados. Mesmo assim, os resultados obtidos pela regressão KNN são notoriamente mais favoráveis do que aqueles obtidos pela regressão linear.

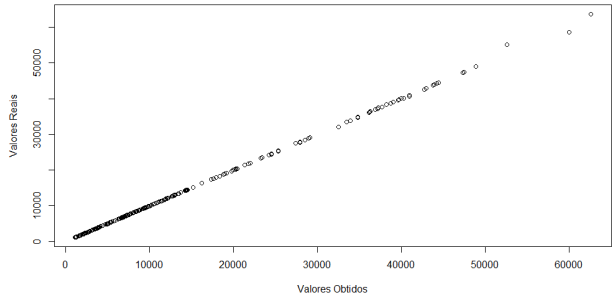


Figura 2.24: Regressão KNN - Resultados obtidos

2.9.2 Variável de Interesse Manipulada

Experimentou-se, ainda, ajustar o segundo modelo estudado numa regressão do *K-Nearest-Neighbors*. Numa primeira abordagem, recorreu-se a um erro estipulado de 0.2, após normalização da amostra. Este não aparentou ser muito significativo para valores baixos do logaritmo dos custos. Contudo, para valores superiores a 9, este valor é significativo, tornando as previsões imprecisas. Assim sendo, foi utilizada a razão entre os valores estimados e os valores reais da variável de interesse (Equação 2.2) para a margem de erro calculada, que deve ser menor do que 5% para que seja válida. No entanto, esta abordagem não obteve resultados mais favoráveis que a anterior.

$$\varepsilon = \frac{\hat{Y} - Y}{Y} \tag{2.2}$$

	K = 1	K = 6	K = 7
Accuracy (Yes / No)	258 / 10	262 / 6	262 / 6
RMSE	0.08740456	0.07536328	0.07265539

Figura 2.25: Resultados da Regressão KNN para diferentes valores de K

Analisando a Figura 2.25, conclui-se que os valores de K para o qual o modelo de regressão é melhor ajustado são 6 e 7, uma vez que apresenta menores erros de previsão. Na Figura 2.26, verifica-se uma ligeira disparidade entre os valores reais e os valores previstos. Apesar disso, é notório o contraste dos resultados obtidos entre os dois modelos de regressão de KNN.

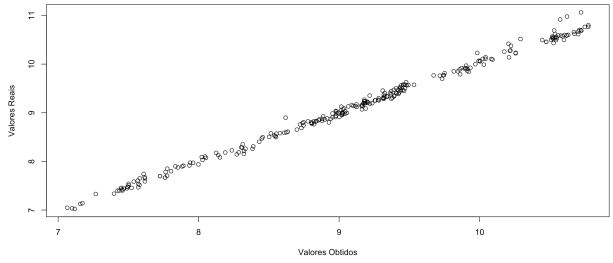


Figura 2.26: Regressão KNN - Resultados obtidos

Capítulo 3

3.1 Resposta às Questões de Interesse

1. Que variáveis influenciam o valor dos custos médicos individuais cobrados pelo seguro de saúde?

Após análise dos dados, observa-se que ser fumador e ter filhos são as variáveis que mais influenciam positivamente os custos individuais médicos. Por outro lado, indivíduos residentes das regiões sul do país evidenciam custos aproximadamente 1000 dólares inferiores aos restantes. O sexo não está relacionado com os custos.

2. O seguro de saúde é mais vantajoso para famílias numerosas?

Não, uma vez que é esperado um custo de 475 dólares por dependente do agregado familiar, independentemente do número de dependentes. Se excluirmos o fator de confundimento *tabagismo*, verificamos que existe uma relação positiva entre o número de dependentes e os custos médicos associados ($p = 0.01$).

3. De que forma é que os níveis de índice de massa corporal influenciam os custos médicos individuais cobrados pelo seguro?

No geral, verifica-se um aumento dos custos estatisticamente significativo com o aumento de índice de massa corporal. Analisando o gráfico (Figura 3.1), verifica-se, no entanto, que há dois sub-grupos claramente demarcados. A análise, tendo em conta as outras variáveis, revela que o tabagismo influencia de forma linear a maneira como o peso irá aumentar os custos ($p = 2 \times 10^{-}16$ vs $p = 0.439$ para não fumadores).

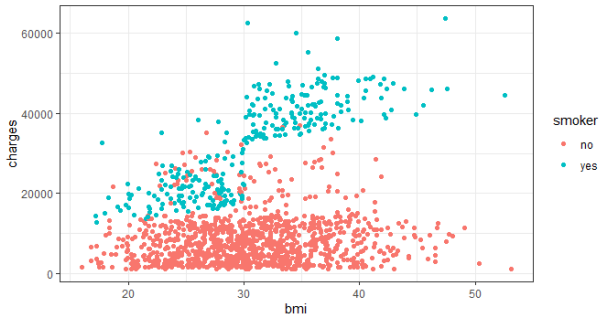


Figura 3.1: Correlação entre BMI e Tabagismo

4. Indivíduos fumadores estão sujeitos a custos mais elevados?

Sim. Prevê-se que cada indivíduo fumador pague cerca de 24000 dólares adicionais.

3.2 Conclusões

Este projecto permitiu-nos consolidar os vários métodos de aprendizagem estatística abordados na unidade curricular de Aprendizagem Automática I.

A abordagem não logarítmica utilizada inicialmente impedia a previsão dos custos médicos individuais, pelo que se testou o logaritmo da variável de interesse, já capaz de prever resultados individuais. No entanto, estes resultados eram ainda pouco precisos, visto que a margem de erro encontrada se tornava inaceitavelmente elevada quando exposta ao exponencial (erros na ordem dos 4000 a 6000 dólares).

A regressão KNN apresentou-se como a abordagem que melhor previa o custo médico, sobretudo quando a variável de interesse assumia unidades na ordem dos milhares, tendo sido, portanto, abandonada a variável logarítmica.

Em suma, foi possível prever, com sucesso, a variável de interesse, através do KNN, apesar das dificuldades colocadas pelo *dataset* e pela não-linearidade dos dados.

Bibliografia

[1] Kaggle.com. *Medical Cost Personal Datasets*. 2020. URL: https://www.kaggle.com/mirichoi0218/insurance (acedido em 22/11/2020) (ver p. 1).

[2] Kassambara, Janeman, kamenrider Cahya, Julie, Kassambara, Sfd e Visitor. *Cross-Validation Essentials in R*. Mar. de 2018. URL: http://www.sthda.com/english/articles/38-regression-model-validation/157-cross-validation-essentials-in-r/ (acedido em 19/12/2020) (ver p. 6).

[3] Armando Teixeira-Pinto. *Machine Learning for Biostatistics*. Ago. de 2020. URL: https://bookdown.org/tpinto_home/Regression-and-Classification/k-nearest-neighbours-regression.html (acedido em 19/12/2020) (ver p. 9).

Apêndice A

Código R desenvolvido

```
# Library used to plot
library(ggplot2)
# Library used to assist ggplot2 library
library(cowplot)

Health <- read.csv("C:\\Users\\f7car\\Desktop\\UM\\CD\\AA1\\Projeto\\insurance.csv")
Health <- read.csv("/home/dreamerz/Desktop/CD/Aprendizagem Automatica
    I/Trabalho/insurance.csv")

dim(Health)
attach(Health)

plot(Health)
summary(Health)

# ------------------------------------------------------------------------------
# Analise Exploratoria dos Dados - Regressao Linear
# ------------------------------------------------------------------------------

hist(charges,
     main="Histogram for Variable Charges",
     xlab="Charges",
     border="#000000",
     col="#EDB879"
)

par(mfrow=c(2,2))

counts <- table(sex)
barplot(counts,xlab = "Sex", col=c("#EDB879", "#69BDD2"))

smoke <- table(smoker)
barplot(smoke,xlab = "Smoker",col=c("#EDB879", "#69BDD2"))

regiao <- table(region)
barplot(regiao,xlab = "Region",col=c("#EDB879", "#69BDD2", "#D06363", "#80CC80"))

filhos <- table(children)
barplot(filhos,xlab = "Children", col=c("#EDB879", "#69BDD2", "#D06363", "#80CC80"))

par(mfrow=c(1,2))
boxplot(bmi, xlab = "Bmi", col=c("#EDB879"))
boxplot(charges, xlab = "Charges", col=c("#EDB879"))

##### Analise Exploratoria dos Dados com a Variavel de Interesse

x <- ggplot(Health, aes(age, charges)) +
  geom_jitter(aes(color = age), alpha = 0.5) +
  theme_light()

y <- ggplot(Health, aes(bmi, charges)) +
  geom_jitter(color = "orange", alpha = 0.5) +
  theme_light()
```

```
p2 <- plot_grid(x,y)
    title <- ggdraw() + draw_label("Correlation between charges and bmi/age", fontface='bold')
    plot_grid(title, p2, ncol=1, rel_heights=c(0.1, 1))

ggplot(data=Health, mapping = aes(x = bmi, y = charges)) +
    geom_point(aes(color = smoker)) +
    theme_bw()

par(mfrow=c(2,2))
boxplot(charges ~ smoker, col=c("#EDB879","#69BDD2"))
boxplot(charges ~ children, col=c("#EDB879", "#69BDD2", "#D06363", "#80CC80"))
boxplot(charges ~ sex, col=c("#EDB879","#69BDD2"))
boxplot(charges ~ region, col=c("#EDB879", "#69BDD2", "#D06363", "#80CC80"), xlab="Region")

# ------------------------------------------------------------------------
# Ajuste do Modelo de Regressao Linear
# ------------------------------------------------------------------------

# Modelo com todos os preditores
m1 <- lm(charges ~., Health)
summary(m1)    # Adjusted R-squared:  0.7494
extractAIC(m1) # 23316.43
coef(m1)
confint(m1)

# Modelo sem genero
m2 <- lm(charges ~.-sex, Health)
summary(m2)    # Adjusted R-squared:  0.7496
extractAIC(m2) # 23314.58
coef(m2)
confint(m2)
confint(m2,level=0.97)

# Modelo sem genero e regioes norte agrupadas
Health_north <- Health
Health_north$region<-ifelse(Health_north$region == "northwest","north",
            ifelse(Health_north$region == "northeast" ,"north",
            ifelse(Health_north$region == "southwest","southwest",
            ifelse(Health_north$region == "southeast","southeast","erro"
            ))))

detach(Health)
attach(Health_north)

m3 <- lm(charges ~.-sex, Health_north)
summary(m3)    # Adjusted R-squared:  0.7497
extractAIC(m3) # 23313.13
coef(m3)
confint(m3)

# Modelo sem regiao e genero
m4 <- lm(charges ~.-(sex+region), Health)
summary(m4)    # Adjusted R-squared:  0.7489
extractAIC(m4) #23314.96
coef(m4)
confint(m4)
confint(m4, level=0.99)

# ------------------------------------------------------------------------
# Preparar Dataset para Predicao
# ------------------------------------------------------------------------

Health <- read.csv("C:\\Users\\f7car\\Desktop\\UM\\CD\\AA1\\Projeto\\insurance.csv")
```

```r
Health <- read.csv("/home/dreamerz/Desktop/CD/Aprendizagem Automatica
    I/Trabalho/insurance.csv")
Health_outlier <- Health
Health_north <- Health
Health_north$region<-ifelse(Health_north$region == "northwest","north",
                    ifelse(Health_north$region == "northeast" ,"north",
                    ifelse(Health_north$region == "southwest","southwest",
                    ifelse(Health_north$region == "southeast","southeast","erro"
                    ))))

attach(Health)

par(mfrow=c(1,2))
boxplot(bmi, xlab = "Bmi", col=c("#EDB879"))
boxplot(charges, xlab = "Charges", col=c("#EDB879"))

#Possiveis outliers
boxplot.stats(bmi)$out
boxplot.stats(charges)$out

################################## BMI Outliers ###################################

# Linhas dos possiveis outliers bmi
outlier_bmi <- boxplot.stats(bmi)$out
outlier_bmi_line <- which(bmi %in% c(outlier_bmi))
outlier_bmi_line
Health_outlier[outlier_bmi_line, ]

# Remover possiveis outliers de bmi
Health_outlier <- Health_outlier[-outlier_bmi_line, ]

# Remover entradas baseadas num valor de bmi (considerado como outlier)
Health_outlier<- Health_outlier[Health_outlier$bmi > 30, ]

############################### Charges Outliers ##################################

# Linhas dos possiveis outliers charges
outlier_charges <- boxplot.stats(charges)$out
outlier_charges_line <- which(charges %in% c(outlier_charges))
outlier_charges_line
Health_outlier[outlier_charges_line, ]

# Remover possiveis outliers de charges
Health_outlier <- Health_outlier[-outlier_charges_line, ]

# Remover entradas baseadas num valor de charges (considerado como outlier)
Health_outlier <- Health_outlier[Health_outlier$charges < 50000, ]
Health <- Health[Health$charges < 50000, ]
Health_north <- Health_north[Health_north$charges < 50000, ]

# ------------------------------------------------------------------------------
# Treino e Teste do modelo
# Dividir os dados em treino (80%) + teste (20%)
# ------------------------------------------------------------------------------

set.seed(1)

# Obter linhas que serao usadas para treino
train <- sample(1:nrow(Health), round(0.8 * nrow(Health)))

#Obter dados de treino
Health_train <- Health[train,]
Health_train_north <- Health_north[train,]
dim(Health_train)
fix(Health_train)
```

```r
fix(Health_train_north)

#Obter dados de teste
Health_test <- Health[-train,]
Health_test_north <- Health_north[-train,]
dim(Health_test)
fix(Health_test)
fix(Health_test_north)

# Guarda os valores reais dos charges dos dados de teste
# para comparar no fim do treino
valores_reais <- Health_test$charges
valores_reais_north <- Health_test_north$charges


#################################### M1 ######################################

# Treinar para o modelo m1 com todos os preditores
model_train_m1 <- lm(charges ~., data = Health_train)

summary(model_train_m1)    # Adjusted R-squared:  0.7479
extractAIC(model_train_m1) # 18611.12

test_result_m1 <- predict(model_train_m1,Health_test)
residuals_m1 <- Health_test$charges - test_result_m1
rmse_m1 <- sqrt(mean(residuals_m1^2))

Health_test$predicted_m1 <- predict(model_train_m1, Health_test)
ggplot(Health_test, aes(x = predicted_m1, y = charges)) +
  geom_point(color = "blue", alpha = 0.7) +
  geom_abline(color = "red") +
  ggtitle("Predicted vs. Real values")

# Testar exatidao com base no erro predefinido
Health_test$Accuracy_m1 <- ifelse(abs(Health_test$predicted_m1 - Health_test$charges) <
    1000,1,0)
Health_test$Accuracy_m1 <- as.factor(Health_test$Accuracy_m1)
levels(Health_test$Accuracy_m1) <- c("no","yes")

summary(Health_test$Accuracy_m1) # no - 208; yes - 60

#################################### M2 ######################################

# Treinar para o modelo m2 com todos os preditores exceto genero
model_train_m2 <- lm(charges ~.-sex, data = Health_train)

summary(model_train_m2)    # Adjusted R-squared:  0.7481
extractAIC(model_train_m2) # 18609.13

test_result_m2 <- predict(model_train_m2,Health_test)
residuals_m2 <- Health_test$charges - test_result_m2
rmse_m2 <- sqrt(mean(residuals_m2^2))

Health_test$predicted_m2 <- predict(model_train_m2, Health_test)
ggplot(Health_test, aes(x = predicted_m2, y = charges)) +
  geom_point(color = "blue", alpha = 0.7) +
  geom_abline(color = "red") +
  ggtitle("Predicted vs. Real values")

# Testar exatidao com base no erro predefinido
Health_test$Accuracy_m2 <- ifelse(abs(Health_test$predicted_m2 - Health_test$charges) <
    1000,1,0)
Health_test$Accuracy_m2 <- as.factor(Health_test$Accuracy_m2)
levels(Health_test$Accuracy_m2) <- c("no","yes")
```

```r
summary(Health_test$Accuracy_m2) # no - 208; yes - 60

##################################### M3 #####################################

# Treinar para o modelo m3 com nortes agrupados e removendo o preditor genero
model_train_m3 <- lm(charges ~.-sex, data = Health_train_north)

summary(model_train_m3)    # Adjusted R-squared:  0.7483
extractAIC(model_train_m3) # 18607.19

test_result_m3 <- predict(model_train_m3,Health_test_north)
residuals_m3 <- Health_test_north$charges - test_result_m3
rmse_m3 <- sqrt(mean(residuals_m3^2))

Health_test_north$predicted_m3 <- predict(model_train_m3, Health_test_north)
ggplot(Health_test_north, aes(x = predicted_m3, y = charges)) +
  geom_point(color = "blue", alpha = 0.7) +
  geom_abline(color = "red") +
  ggtitle("Predicted vs. Real values")

# Testar exatidao com base em erro predefinido
Health_test_north$Accuracy_m3 <- ifelse(abs(Health_test_north$predicted_m3 -
    Health_test_north$charges) < 1000,1,0)
Health_test_north$Accuracy_m3 <- as.factor(Health_test_north$Accuracy_m3)
levels(Health_test_north$Accuracy_m3) <- c("no","yes")

summary(Health_test_north$Accuracy_m3) # no - 207; yes - 61

##################################### M4 #####################################

# Treinar para o modelo m4 com todos os preditores exceto genero e regiao
model_train_m4 <- lm(charges ~.-(sex+region), data = Health_train)

summary(model_train_m4)    # Adjusted R-squared:  0.748
extractAIC(model_train_m4) # 18606.51

test_result_m4 <- predict(model_train_m4,Health_test)
residuals_m4 <- Health_test$charges - test_result_m4
rmse_m4 <- sqrt(mean(residuals_m4^2))

Health_test$predicted_m4 <- predict(model_train_m4, Health_test)
ggplot(Health_test, aes(x = predicted_m4, y = charges)) +
  geom_point(color = "blue", alpha = 0.7) +
  geom_abline(color = "red") +
  ggtitle("Predicted vs. Real values")

# Testar exatidao com base no erro predefinido
Health_test$Accuracy_m4 <- ifelse(abs(Health_test$predicted_m4 - Health_test$charges) <
    1000,1,0)
Health_test$Accuracy_m4 <- as.factor(Health_test$Accuracy_m4)
levels(Health_test$Accuracy_m4) <- c("no","yes")

summary(Health_test$Accuracy_m4) # no - 204; yes - 64

# ------------------------------------------------------------------------------
# K-Fold e LOOCV
# ------------------------------------------------------------------------------

library(tidyverse)
library(caret)

Health <- read.csv("C:\\Users\\f7car\\Desktop\\UM\\CD\\AA1\\Projeto\\insurance.csv")
Health <- read.csv("/home/dreamerz/Desktop/CD/Aprendizagem Automatica
    I/Trabalho/insurance.csv")
attach(Health)
```

```r
set.seed(1)

############################ K-Fold e LOOCV para M1 ############################

train_control <- trainControl(method = "cv", number = 5)
model_kf_m1 <- train(charges ~., data = Health, method = "lm",trControl = train_control)
print(model_kf_m1)

train_control <- trainControl(method = "LOOCV")
model_loocv_m1 <- train(charges ~., data = Health, method = "lm",trControl = train_control)
print(model_loocv_m1)

############################ K-Fold e LOOCV para M2 ############################

train_control <- trainControl(method = "cv", number = 5)
model_kf_m2 <- train(charges ~.-sex, data = Health, method = "lm",trControl = train_control)
print(model_kf_m2)

train_control <- trainControl(method = "LOOCV")
model_loocv_m2 <- train(charges ~.-sex, data = Health, method = "lm",trControl =
    train_control)
print(model_loocv_m2)

############################ K-Fold e LOOCV para M3 ############################

Health_aux <- Health
Health_aux$region<-ifelse(Health_aux$region == "northwest","north",
                  ifelse(Health_aux$region == "northeast" ,"north",
                  ifelse(Health_aux$region == "southwest","southwest",
                  ifelse(Health_aux$region == "southeast","southeast","erro"
                  ))))

attach(Health_aux)

train_control <- trainControl(method = "cv", number = 5)
model_kf_m3 <- train(charges ~.-sex, data = Health_aux, method = "lm",trControl =
    train_control)
print(model_kf_m3)

train_control <- trainControl(method = "LOOCV")
model_loocv_m3 <- train(charges ~.-sex, data = Health_aux, method = "lm",trControl =
    train_control)
print(model_loocv_m3)

# ------------------------------------------------------------------------------
# KNN
# ------------------------------------------------------------------------------

library(FNN)
library(class)


Health_knn <- read.csv("C:\\Users\\f7car\\Desktop\\UM\\CD\\AA1\\Projeto\\insurance.csv",
    stringsAsFactors = TRUE)
Health_knn <- read.csv("/home/dreamerz/Desktop/CD/Aprendizagem Automatica
    I/Trabalho/insurance.csv", stringsAsFactors = TRUE)

set.seed(1)

Health_knn$smoker <- ifelse(Health_knn$smoker == "yes", 1, 0)
Health_knn$sex <- ifelse(Health_knn$sex == "male", 1, 0)
Health_knn$region<-ifelse(Health_knn$region == "northwest",1,
                  ifelse(Health_knn$region == "northeast" ,2,
                  ifelse(Health_knn$region == "southwest",3,
```

```
                    ifelse(Health_knn$region == "southeast",4,0
                    ))))

train <- sample(1:nrow(Health_knn), round(0.8 * nrow(Health_knn)))

Health_knn_train <- Health_knn[train,]
Health_knn_teste <- Health_knn[-train,]

# Fit a KNN regression with k = 3
# using the knn.reg() function from the FNN package
knn_charges <- knn.reg(train=Health_knn_train[,-Health_knn_train$charges],
                     y=Health_knn_train$charges,
                     test= Health_knn_teste[,-Health_knn_teste$charges],
                     k=1)

summary(knn_charges)
knn_charges$pred

Health_knn_teste$predicted <- knn_charges$pred

Health_knn_teste$Accuracy <- ifelse(abs(Health_knn_teste$predicted -
    Health_knn_teste$charges) < 100,1,0)
Health_knn_teste$Accuracy <- as.factor(Health_knn_teste$Accuracy)
levels(Health_knn_teste$Accuracy) <- c("no","yes")

summary(Health_knn_teste$Accuracy)

plot(Health_knn_teste$charges)
lines(Health_knn_teste$predicted)
plot(Health_knn_teste$charges ~ Health_knn_teste$predicted)

knn_rmse  <- sqrt(mean((Health_knn_teste$predicted - Health_knn_teste$charges)^2) )
knn_rmse

# ------------------------------------------------------------------------------
# Alteracao da variavel de interesse
# ------------------------------------------------------------------------------

Health <- read.csv("C:\\Users\\f7car\\Desktop\\UM\\CD\\AA1\\Projeto\\insurance.csv")
Health <- read.csv("/home/dreamerz/Desktop/CD/Aprendizagem Automatica
    I/Trabalho/insurance.csv")

newdata <- Health
new_charges <- log(Health$charges)
newdata$charges <- log(newdata$charges)
attach(newdata)

hist(charges,
     main="Histogram for New Variable Charges",
     xlab="Charges' Logarithm",
     border="#000000",
     col="#EDB879"
    )

x <- ggplot(newdata, aes(age, charges)) +
  geom_jitter(aes(color = age), alpha = 0.5) +
  theme_light()

y <- ggplot(newdata, aes(bmi, charges)) +
  geom_jitter(color = "orange", alpha = 0.5) +
  theme_light()

p2 <- plot_grid(x,y)
title <- ggdraw() + draw_label("Correlation between new_charges and bmi/age", fontface='bold')
plot_grid(title, p2, ncol=1, rel_heights=c(0.1, 1))
```

```r
par(mfrow=c(2,2))
boxplot(charges ~ smoker, col=c("#EDB879","#69BDD2"))
boxplot(charges ~ children, col=c("#EDB879", "#69BDD2", "#D06363", "#80CC80"))
boxplot(charges ~ sex, col=c("#EDB879","#69BDD2"))
boxplot(charges ~ region, col=c("#EDB879", "#69BDD2", "#D06363", "#80CC80"), xlab="Region")

# ------------------------------------------------------------------------------
# Ajuste do modelo
# ------------------------------------------------------------------------------

newdata_north <- newdata
summary(Health)
summary(newdata)
summary(newdata_north)

# Modelo com todos os preditores

mod1 <- lm(charges ~., newdata)
summary(mod1)    # Adjusted R-squared:  0.7666
extractAIC(mod1) # -2162.046
coef(mod1)
confint(mod1)

# Modelo com regioes norte agrupadas

newdata_north$region<-ifelse(newdata_north$region == "northwest","north",
            ifelse(newdata_north$region == "northeast" ,"north",
            ifelse(newdata_north$region == "southwest","southwest",
            ifelse(newdata_north$region == "southeast","southeast","erro"
            ))))

mod2 <- lm(charges ~., newdata_north)
summary(mod2)    # Adjusted R-squared:  0.7661
extractAIC(mod2) # -2160.688
coef(mod2)
confint(mod2)

# ------------------------------------------------------------------------------
# Treino e Teste do modelo
# Dividir os dados em treino (80%) + teste (20%)
# ------------------------------------------------------------------------------

set.seed(1)

# Obter linhas que serao usadas para treino
train <- sample(1:nrow(newdata), round(0.8 * nrow(newdata)))

#Obter dados de treino
newdata_train <- newdata[train,]
newdata_train_north <- newdata_north[train,]
dim(newdata_train)
dim(newdata_train_north)

#Obter dados de teste
newdata_test <- newdata[-train,]
newdata_test_north <- newdata_north[-train,]
dim(newdata_test)
dim(newdata_test_north)

################################### Mod1 ######################################

# Treinar para o modelo m1 com todos os preditores
model_train_mod1 <- lm(charges ~., data = newdata_train)
```

```r
summary(model_train_mod1)
extractAIC(model_train_mod1)

test_result_mod1 <- predict(model_train_mod1,newdata_test)
residuals_mod1 <- newdata_test$charges - test_result_mod1
rmse_mod1 <- sqrt(mean(residuals_mod1^2))

newdata_test$predicted_mod1 <- predict(model_train_mod1, newdata_test)
ggplot(newdata_test, aes(x = predicted_mod1, y = charges)) +
  geom_point(color = "blue", alpha = 0.7) +
  geom_abline(color = "red") +
  ggtitle("Predicted vs. Real values")

# Testar exatidao com base num erro pre-definido

# Erro como 0.30 ou 0.5
newdata_test$Accuracy_mod1 <- ifelse((abs(newdata_test$charges - newdata_test$predicted_mod1)
    < 0.30),1,0)
newdata_test$Accuracy_mod1 <- as.factor(newdata_test$Accuracy_mod1)
levels(newdata_test$Accuracy_mod1) <- c("no","yes")

# Erro atraves da razao
newdata_test$Accuracy_mod1 <- ifelse((((abs(newdata_test$charges -
    newdata_test$predicted_mod1)) / (newdata_test$charges)) < 0.30),"true","false")
newdata_test$Accuracy_mod1 <- as.factor(newdata_test$Accuracy_mod1)

summary(newdata_test$Accuracy_mod1)

################################### Mod2 ####################################

# Treinar para o modelo m2 com todos os preditores mas north juntos
model_train_mod2 <- lm(charges ~., data = newdata_train_north)

summary(model_train_mod2)
extractAIC(model_train_mod2)

test_result_mod2 <- predict(model_train_mod2,newdata_test_north)
residuals_mod2 <- newdata_test_north$charges - test_result_mod2
rmse_mod2 <- sqrt(mean(residuals_mod2^2))

newdata_test_north$predicted_mod2 <- predict(model_train_mod2, newdata_test_north)
ggplot(newdata_test_north, aes(x = predicted_mod2, y = charges)) +
  geom_point(color = "blue", alpha = 0.7) +
  geom_abline(color = "red") +
  ggtitle("Predicted vs. Real values")

# Testar exatidao com base em erro pre-definido

# Erro como 0.3 ou 0.5
newdata_test_north$Accuracy_mod2 <- ifelse((abs(newdata_test_north$charges -
    newdata_test_north$predicted_mod2) < 0.30),1,0)
newdata_test_north$Accuracy_mod2 <- as.factor(newdata_test_north$Accuracy_mod2)
levels(newdata_test_north$Accuracy_mod2) <- c("no","yes")

summary(newdata_test_north$Accuracy_mod2)

# Erro atraves da razao
newdata_test_north$Accuracy_mod2 <- ifelse((((abs(newdata_test_north$charges -
    newdata_test_north$predicted_mod2)) / (newdata_test_north$charges)) <
    0.30),"true","false")
newdata_test_north$Accuracy_mod2 <- as.factor(newdata_test_north$Accuracy_mod2)

summary(newdata_test_north$Accuracy_mod2)
```

```r
# ------------------------------------------------------------------------------
# K-Fold e LOOCV
# ------------------------------------------------------------------------------

library(tidyverse)
library(caret)

Health <- read.csv("C:\\Users\\f7car\\Desktop\\UM\\CD\\AA1\\Projeto\\insurance.csv")
Health <- read.csv("/home/dreamerz/Desktop/CD/Aprendizagem Automatica
    I/Trabalho/insurance.csv")
newdata <- Health
new_charges <- log(Health$charges)
newdata$charges <- log(newdata$charges)
newdata_north <- newdata
newdata_north$region<-ifelse(newdata_north$region == "northwest","north",
                      ifelse(newdata_north$region == "northeast" ,"north",
                      ifelse(newdata_north$region == "southwest","southwest",
                      ifelse(newdata_north$region == "southeast","southeast","erro"
                      ))))

set.seed(1)

############################ K-Fold e LOOCV para Mod1 ############################

train_control <- trainControl(method = "cv", number = 5)
model_kf_mod1 <- train(charges ~., data = newdata, method = "lm",trControl = train_control)
print(model_kf_mod1)

train_control <- trainControl(method = "LOOCV")
model_loocv_mod1 <- train(charges ~., data = newdata, method = "lm",trControl = train_control)
print(model_loocv_mod1)

############################ K-Fold e LOOCV para Mod2 ############################

train_control <- trainControl(method = "cv", number = 5)
model_kf_mod2 <- train(charges ~., data = newdata_north, method = "lm",trControl =
    train_control)
print(model_kf_mod2)

train_control <- trainControl(method = "LOOCV")
model_loocv_mod2 <- train(charges ~., data = newdata_north, method = "lm",trControl =
    train_control)
print(model_loocv_mod2)

# ------------------------------------------------------------------------------
# KNN
# ------------------------------------------------------------------------------

library(FNN)
library(class)

Health_knn <- read.csv("C:\\Users\\f7car\\Desktop\\UM\\CD\\AA1\\Projeto\\insurance.csv",
    stringsAsFactors = TRUE)
#Health_knn <- read.csv("/home/dreamerz/Desktop/CD/Aprendizagem Automatica
    I/Trabalho/insurance.csv", stringsAsFactors = TRUE)

set.seed(1)
Health_knn$charges <- log(Health_knn$charges)
Health_knn$smoker <- ifelse(Health_knn$smoker == "yes", 1, 0)
Health_knn$sex <- ifelse(Health_knn$sex == "male", 1, 0)
Health_knn$region<-ifelse(Health_knn$region == "northwest",1,
                  ifelse(Health_knn$region == "northeast",2,
                  ifelse(Health_knn$region == "southwest",3,
```

```
                  ifelse(Health_knn$region == "southeast",4,0
                  ))))

normalize <- function(x) { (x-min(x)) / (max(x)-min(x))}
test <- Health_knn$charges
Health_knn <- as.data.frame(lapply(Health_knn[,c(1:6)],normalize))
Health_knn$charges <- test

train <- sample(1:nrow(Health_knn), round(0.8 * nrow(Health_knn)))

Health_knn_train <- Health_knn[train,]
Health_knn_teste <- Health_knn[-train,]

# Fit a KNN regression with k = 1
# using the knn.reg() function from the FNN package
knn_charges <- knn.reg(train=Health_knn_train[,-Health_knn_train$charges],
                       y=Health_knn_train$charges,
                       test= Health_knn_teste[,-Health_knn_teste$charges],
                       k=1)

summary(knn_charges)
knn_charges$pred

Health_knn_teste$predicted <- knn_charges$pred

# Erro pre-definido
Health_knn_teste$Accuracy <- ifelse(abs(Health_knn_teste$predicted -
    Health_knn_teste$charges) < 0.2,1,0)
Health_knn_teste$Accuracy <- as.factor(Health_knn_teste$Accuracy)
levels(Health_knn_teste$Accuracy) <- c("no","yes")

summary(Health_knn_teste$Accuracy)

Health_knn_teste$graph_maior <- Health_knn_teste$predicted + 0.3
Health_knn_teste$graph_menor <- Health_knn_teste$predicted - 0.3
plot(Health_knn_teste$charges)
plot(Health_knn_teste$predicted,col="green")
lines(Health_knn_teste$graph_maior,col="red")
lines(Health_knn_teste$graph_menor,col="red")
plot(Health_knn_teste$charges ~ Health_knn_teste$predicted)

knn_rmse  <- sqrt(mean((Health_knn_teste$predicted - Health_knn_teste$charges)^2) )
knn_rmse

# Razao
Health_knn_teste$Accuracy <- ifelse(abs(Health_knn_teste$predicted -
    Health_knn_teste$charges) / Health_knn_teste$charges < 0.05,"true","false")
Health_knn_teste$Accuracy <- as.factor(Health_knn_teste$Accuracy)
levels(Health_knn_teste$Accuracy) <- c("no","yes")

summary(Health_knn_teste$Accuracy)

plot(Health_knn_teste$charges)
lines(Health_knn_teste$predicted)
plot(Health_knn_teste$charges ~ Health_knn_teste$predicted)

knn_rmse  <- sqrt(mean((Health_knn_teste$predicted - Health_knn_teste$charges)^2) )
knn_rmse

################################################################################
# Justificacao resposta 2
################################################################################

Health_child <- Health[Health$children > 3, ]
dim(Health_child)
```

```r
attach(Health_child)

Health_aux2 <- Health_child[Health_child$smoker == "no",  ]
Health_aux2 <- Health_aux2[,-5]
dim(Health_aux2)
attach(Health_aux2)

m <- lm(charges ~., Health_aux2)
summary(m) # p_value children: 0.011

#############################################################################
# Influencia do tabagismo na significancia do bmi para variavel charges
#############################################################################

Health_aux3 <- Health[Health$smoker == "no",  ]
attach(Health_aux3)
Health_aux3 <- Health_aux3[,-5]
mnosmoker <- lm(charges ~., Health_aux3)
summary(mnosmoker)    # p_value bmi:  0.439265

Health_aux5 <- Health[Health$smoker == "yes",  ]
attach(Health_aux5)
Health_aux5 <- Health_aux5[,-5]
msmoker <- lm(charges ~., Health_aux5)

summary(msmoker)    # p_value bmi:  <2e-16
```