UNIVERSIDADE
DE ÉVORA

Estatística Aplicada

Trabalho 2

João Lopes 58358

Miguel Grilo 58387

Jorge Couto 58656

Colégio Luís António Verney

Tabela de Conteúdos

Lista de Figuras

Introdução

No âmbito da cadeira de Estatística Aplicada, de semestre ímpar e lecionada principalmente pelo professor Paulo Infante, foi-nos pedida a realização de um segundo, e último, trabalho para a avaliação. Esse trabalho visa a consolidação de assuntos abordados em aula, todavia apenas presentes em trabalhos de casa e frequências no quesito avaliativo, como o modelo logarítmico e a análise de clusters.

O trabalho realizado, com os passos explicados e conclusões retiradas mencionadas neste relatório e o código em R para suporte, encontra-se dividido em duas questões, sendo a primeira relativa às tabelas de contingência, medidas de associação e regressão logística e a segunda relativa à análise de clusters. Para melhor separação das conclusões tiradas, cada uma das alíneas será separada, neste relatório, em uma secção diferente.

Para que um não-estatístico possa, também, compreender as conclusões retiradas, quaisquer conclusões pertinentes serão escritas de modo que um não-estatístico compreenda, em tópico.

Exercício 1

Alínea a)

O primeiro exercício informa-nos de um estudo realizado com a finalidade de conhecer a magnitude da problemática das feridas na área de intervenção da ULSAC. Os dados obtidos do estudo foram armazenados no ficheiro de dados Feridas.csv o qual, no R, apropriadamente chamamos pela variável de mesmo nome.

A primeira alínea, em específico, pede-nos pela associação entre a tipologia das feridas e tanto a sua localização quanto a sua origem. Começando pela associação entre a tipologia das feridas e a sua localização, criamos uma tabela de contingência entre as duas para, com o teste do qui-quadrado, estudar a associação entre as duas. Com um valor p muito inferior a 0.0001, podemos rejeitar a hipótese nula e, assim, admitir que existe associação entre a tipologia das feridas e a sua localização. Vale-se notar que, por falta de avisos dados a realizar o teste do qui-quadrado, não se torna

necessário observar os valores esperados sobre a hipótese de independência, apesar de ainda fazermos isso no código.

Através dos resíduos estandardizados, do V de Cramer e do que já observamos, podemos concluir:

- Existe associação entre a localização da ferida e o tipo de ferida que é;
- Existem mais feridas caracterizadas como cirúrgicas em outro local do que o esperado caso a tipologia das feridas fosse independente da localização, e menos da mesma identificação no abdómen ou em um membro inferior do que o esperado;
- Existem mais feridas caracterizadas como traumáticas em membros superiores do que o esperado caso a tipologia fosse independente da localização, e menos do mesmo tipo no abdómen ou em membros inferiores do que se esperava observar;
- Existem mais feridas caracterizadas como 'Outra' no abdómen ou em membros inferiores do que o esperado caso a tipologia fosse independente da localização, e menos do mesmo tipo em membros superiores ou outros locais do que o esperado;
- Com um V de Cramer de 0.2808, podemos dizer que a associação existente entre a tipologia da ferida e a sua localização é moderada.

Estudando agora a associação entre a tipologia da ferida e a sua origem, seguimos o estudo da associação de modo semelhante àquele já realizado para o estudo da associação entre a tipologia da ferida e a sua localização, obtendo também um valor p no teste qui-quadrado inferior a 0.0001 e, portanto, concluindo que existe associação entre a tipologia e a origem das feridas. Todavia, como desta vez o teste lança um aviso, verificamos se os valores esperados sobre a hipótese de independência são todos superiores a 1. Como, neste caso, de facto o são, podemos prosseguir normalmente. Através dos resíduos estandartizados, do V de Cramer e do que já anteriormente concluímos, podemos dizer:

- A tipologia das feridas e a sua origem estão associadas;
- Existem mais feridas da tipologia cirúrgica com origem no hospital do que o esperado caso não houvesse associação entre

a tipologia e a origem, e menos feridas da mesma tipologia com origem no domicílio;

- Existem mais feridas da tipologia traumática com origem no domicílio do que o esperado caso não houvesse associação entre a tipologia e a origem, e menos feridas da mesma tipologia com origem no hospital;
- Existem mais feridas da tipologia 'outra' com origem no domicílio do que o esperado caso não houvesse associação entre a tipologia e a origem, e menos feridas da mesma tipologia com origem no hospital;
- Existem mais feridas da tipologia 'UP/LH/LPT' com origem em outra localização do que o esperado caso não houvesse associação entre a tipologia e a origem;
- Feridas de origem 'UCSP' não apresentam diferenças significativas entre tipologias;
- Com um V de Cramer de 0.4882, a associação entre a tipologia das feridas e a sua origem é forte.

Alínea b)

A alínea seguinte pede-nos para estudar a associação entre a complexidade das feridas e a ferida ter ou não uma tipologia de úlcera por pressão, lesão por humidade ou lesão dos tecidos profundos (UP/LH/LPT) condicional à origem da ferida. Para realizar tal estudo, criamos um conjunto de tabelas que intercepta a complexidade da ferida com se a ferida é UP/LH/LPT ou não, sendo cada tabela relativa a uma origem da ferida diferente. Uma vez com essas tabelas realizadas, corremos o teste de Mantel-Haenszel que, com um valor p de aproximadamente 0.006 (inferior ao nosso alfa de 0.05), nos permite rejeitar a hipótese nula, concluindo assim que a complexidade está associada à tipologia UP/LH/LPT condicional à origem da ferida. Com o teste de Breslow-Days, podemos ver se essa associação é dependente das categorias da origem ou não. Com um valor p de 0.5114, superior ao nosso alfa de 0.05, não podemos rejeitar a hipótese nula, assim concluindo que essa associação não depende das categorias da Origem.

Visto que a condicional não afeta os resultados obtidos, podemos analisar apenas a associação entre a complexidade da ferida e a tipologia UP/LH/LPT, já obtida no teste de Mantel-Haenszel. Desse modo, concluímos:

- Existe associação entre a complexidade das feridas e se a ferida é do tipo UP/LH/LPT ou não. Essa associação não é afetada pela origem da ferida, ou seja, a ferida ter surgido em diferentes áreas não aumenta ou diminui a chance de ser complexa;
- Com um OR de 2.32, podemos dizer que feridas do tipo UP/LH/LPT têm 2.3x mais chances de serem complexas do que feridas de outro tipo. Com 95% de confiança, dizemos que essa chance está entre as 1.26 e 4.29 vezes.

Alínea c)

Para a alínea c, devemos estudar as relações entre a idade e as variáveis da complexidade das feridas e a tipologia das feridas tornada em uma variável dicotómica (cirúrgica ou não cirúrgica (UP/LH/LPT). Para este exercício, precisamos de pensar no coeficiente correto para cada uma das circunstâncias.

Sabendo que a idade é uma variável contínua, e que a complexidade é uma variável nominal dicotómica, uma vez que a complexidade só têm duas categorias e as suas categorias não são, a nosso ver, ordenáveis (apesar de admitirmos a argumentabilidade da complexidade como ordenável), o coeficiente correto para o estudo da relação seria o Bisserial por Pontos. Com um valor p inferior a 0.001, podemos rejeitar a hipótese nula e, portanto, concluir que existe uma correlação estatisticamente significativa entre a idade e a complexidade das feridas. Com um r de -0.17, podemos dizer também que essa correlação é negativa e fraca.

Para a variável dicotómica de feridas cirúrgicas ou não cirúrgicas, podemos utilizar também o Bisserial por Pontos. Uma vez que a ferida ser, ou não, cirúrgica não é necessariamente ordenável, podemos dizer que é uma variável dicotómica nominal, o que implica também o Bisserial por Pontos. Utilizando o Bisserial por Pontos, obtemos um valor p inferior a 0.001, o que nos permite dizer que eiste correlação estatisticamente significativa entre a idade e se a ferida é, ou não, cirúrgica. Com um r de -0.18, esta correlação é, também, negativa e fraca.

Através das conclusões anteriores, podemos resumir, de um modo compreensível também para um não-estatístico:

- A idade do indivíduo está diretamente correlacionada com a complexidade da ferida e com se a ferida é cirúrgica ou não;
- A correlação da idade do indivíduo com a complexidade da ferida e da idade do indivíduo com se a ferida é cirúrgica ou não é fraca, portanto não influencia demasiado os resultados;
- As duas correlações são negativas, ou seja, a idade é inversamente proporcional com a complexidade da ferida (quanto maior a idade menor a chance da ferida ser complexa) e com se a ferida é cirúrgica ou não (quanto maior a idade menor a chance da ferida ser cirúrgica).

Alínea d)

A quarta e última alínea do primeiro exercício do trabalho pede-nos pela construção de um modelo de regressão logística que tente procurar determinantes para a complexidade das feridas. Para tal, seguimos os passos da modelagem do modelo logístico e, no final, verificamos se o modelo conseguido era, ou não, eficaz na predição. Vale-se ressaltar que, uma vez que não foi pedido, não analisamos probabilidades da ferida ser complexa dados certos parâmetros, e vimos o modelo de um modo mais 'geral'.

Primeiramente, ajustámos o modelo nulo (modelo puramente intercepto) e ajustámos um modelo simples para cada uma das variáveis que temos, comparando-o com o modelo nulo. Se, para alfa igual a 0.20, a comparação do modelo simples com o modelo nulo fosse significativa, então incluiremos a variável desse modelo simples no passo seguinte. Neste passo, apenas a variável Sexo ficou de fora. De seguida, e voltando a trabalhar com um alfa igual a 0.05, ajustámos um modelo com todas as variáveis escolhidas e verificamos quais não eram significativas. Experimentamos remover essas variáveis do modelo em ordem decrescente do valor p, salvo para o caso da variável Tipologia que, apesar de ter uma categoria com um valor p altíssimo, por ter as outras categorias muito significativas foi mantida no modelo sem se testar a sua remoção. O processo de escolha de remoção foi feito de modo semelhante ao passo anterior, comparando o modelo com a variável ao modelo sem, para verificar

se retirar a variável do modelo acaba por afetar significativamente o modelo, ou não. Através desse passo, removemos as variáveis habilitações, idade e origem do modelo.

O passo seguinte, o 4º passo na construção de um modelo de regressão logística, consiste na inclusão das variáveis excluídas no passo dos modelos simples por valor crescente de valor p, verificando se a variável afeta significativamente o novo modelo ou não. No nosso caso, a única variável que excluímos no 4º passo foi a variável sexo, então incluímos ela no modelo e, através de um teste anova, verificamos se a variável afeta significativamente o modelo. Com um valor p de 0.2761, superior ao nosso alfa de 0.05, não podemos rejeitar a hipótese nula, assim concluindo que a variável sexo não afeta significativamente o modelo e, portanto, não deve ser incluída. Com isso, seguimos para o passo seguinte, que dividimos em duas partes: Tentar tornar o modelo mais parcimonioso e validar os pressupostos da linearidade com o logit e da multicolinearidade. Para tornar o modelo mais parcimonioso, verificamos as variáveis categóricas que temos e anotámos as suas estimativas e desvios padrões. Para as mesmas variáveis, tentámos ver que categorias poderíamos e deveríamos unir para simplificar o modelo, vendo se a subtração das estimativas de cada par de categorias era inferior ao melhor desvio padrão (se sim, poderíamos unir). Depois de verificar com todas as categorias da variável tipologia, conseguimos unir as categorias 'Outra' com 'UP/LH/LPT'. Testando a comparação do modelo com as duas categorias unidas ao modelo anterior, obtemos um valor p de 0.2936, superior ao nosso alfa de 0.05, o que não nos permite rejeitar a hipótese nula, assim concluindo que a junção não prejudica significativamente o ajuste, permitindo-nos manter essa junção. Tentamos verificar se havia margem para unir mais categorias, mas infelizmente não podemos fazer mais uniões de categorias. Uma vez terminando de trabalhar na variável tipologia, tentamos verificar a variável Obesidade. No limite da significância, com um valor p de 0.0531, um pouco superior ao nosso alfa de 0.05, queremos saber se a variável é essencial, ou não, no modelo. Para tal, excluímo-la do modelo, testando a comparação do modelo com a variável Obesidade com o modelo sem essa variável. Com um valor p de 0.0285, ou seja, inferior ao nosso alfa de 0.05, rejeitamos a hipótese nula, assim concluindo que a exclusão da variável Obesidade afeta significativamente o ajuste do modelo e, portanto, não deve ser excluída. Finalmente, para a parte seguinte deste passo, verificamos os pressupostos da linearidade com o logit e da

multicolinearidade. Todavia, como temos apenas variávels categóricas no modelo, o pressuposto da linearidade com o logit não se aplica ao nosso modelo, precisando apenas de verificar o pressuposto da multicolinearidade, com o VIF. Caso alguma das variáveis tenha um VIF superior a 10, existem problemas com a multicolinearidade. Se tiverem um VIF superior a 5, devemos calcular os coeficientes de correlação das variáveis. No entanto, no nosso caso, todas as variáveis apresentam um VIF rondando o 1, assim permitindo-nos validar o pressuposto da multicolinearidade, e seguir para o passo seguinte.

O passo seguinte da construção do modelo logístico assenta na procura de interações significativas no modelo. A procura é feita de modo semelhante à procura das variáveis significativas, testando a comparação do modelo atual com o modelo com cada interação. Usando um alfa de 0.05, a única interação que se apresentou como significativa e adicionámos, por isso, ao modelo foi a interação entre Desnutrição e Local_Ferida, com um valor p igual a 0.0134, inferior a 0.05. Vale-se notar que a interação entre Desnutrição e Obesidade apresentou um valor p, quando comparada ao modelo original, de 0. No entanto, pelo valor p não poder ser 0, e pelos graus de liberdade e deviance serem apresentados como 0, considerámos que essa interação não é, portanto, significativa.

Assim, temos o nosso modelo final. Apesar de termos mais por fazer, podemos admitir o nosso modelo logístico final como:

Variável	**Coeficiente**	**Desvio Padrão**	**Valor p**
Constante	-1.6218	0.2844	<0.0001
Desnutrição (Ref: Não)	0.2243	0.4747	0.6366
Local_Ferida (Ref: Abdómen ou Membro Inferior)			
Membro Superior	-3.6115	1.0325	0.0005
Outro Local	-1.62	0.3688	<0.0001
TipologiaNew (Ref: Ferida Cirúrgica)			
Ferida Traumática	0.58	0.4138	0.1610
UP/LH/LPT / Outra	2.513	0.3083	<0.0001
Obesidade (Ref: Não)	1.6104	0.8343	0.0536
Desnutrição (Sim) : Local_Ferida (Membro Superior)	2.496	1.8117	0.1683
Desnutrição (Sim) : Local_Ferida (Outro Local)	2.4928	0.9051	0.0059

Com o modelo final, podemos avançar então para a avaliação da bondade do ajustamento. Através do Teste de Hosmer podemos verificar se

o modelo se ajusta aos dados. Com um valor p igual a 0.7335, superior ao nosso alfa igual a 0.05, não podemos rejeitar a hipótese nula, assim concluindo que o modelo ajusta-se, de facto, aos dados. Com um valor p tão alto, podemos dizer também que parece bem ajustado aos dados. O teste de Cessie-Van Houwelingen não precisa de ser realizado, uma vez que não temos variáveis contínuas no nosso modelo final. Através da curva ROC (Figura 1), podemos verificar a eficácia geral do modelo: Com um valor AUC igual a 0.841, ou seja entre 0.8 e 0.9, podemos dizer que o modelo tem uma capacidade discriminativa muito boa. Para um ponto de corte de 0.325 a sensibilidade (probabilidade do modelo prever um indivíduo positivamente sabendo que é, de facto, positivo) é de 80.7% e a especificidade (probabilidade do modelo prever um indivíduo negativamente sabendo que é, de facto, negativo) é de 78.2%. Através do teste DeLong, podemos dizer que, a 95% de confiança, o valor AUC encontra-se entre 0.8071 e 0.8742.

Para o passo final, da análise de resíduos, devemos verificar se existem padrões influentes ou outliers que afetem significativamente o ajuste do modelo. Através do comando epi.cp, percebemos que o nosso modelo têm 19 padrões e 482 indivíduos. No gráfico da distância de Cook (Figura 2), e usando o comando identify (com uma tolerância de 2, oito vezes superior à referência, porque nenhuma outra tolerância permitia verificar os 3 padrões em destaque...), identificamos três padrões fora do comum: O padrão 6, altamente influente por ter uma distância de cook ao redor dos 10 e os padrões 1 e 5, influentes pela distância de cook ao redor do 4. Com esses três padrões identificados, criamos um subset com uma lista dos 3 padrões, achando isso mais prático do que criar um subset um-a-um. Finalmente, precisamos de ver se esses três padrões devem de facto ser removidos ou podem ser mantidos. Afinal, tendo apenas 19 padrões, retirar 3 deles pode ser bastante prejudicial comparado a remover apenas um deles. Para tal, utilizamos o gráfico da alteração na Deviance (Figura 3) para verificar quais padrões devem ser, ou não, removidos. Por ter uma alteração na Deviance muito grande comparada aos outros padrões, e uma área da circunferência também muito maior que os outros, o padrão 6 parece que deve ser removido. No entanto, o mesmo não se pode dizer sobre os padrões 1 e 5, tendo ambos alterações na Deviance semelhantes aos outros padrões, e uma circunferência não tão grande. Portanto, optamos por remover apenas o padrão 6, mantendo os padrões 1 e 5. Através do cálculo das diferenças dos coeficientes do modelo com e sem o padrão 6, verificamos que a

deviance do modelo reduziu 12.6%, o que melhora significativamente a qualidade do ajuste do modelo.

Finalmente, falta-nos verificar se o modelo que ajustámos é, ou não, significativamente preditivo ou não. Dividindo as nossas amostras em dados de treino e de teste, com 70% das amostras para o treino e 30% para o teste, e usando as amostras de treino para treinar o modelo, podemos verificar a matriz confusão com os dados de teste. Os resultados serão colocados em tópicos, de um modo compreensível por não-estatísticos:

- O modelo final apresenta uma precisão de 77.78%, tendo por isso 77.78% de chances de prever corretamente se, para dadas características, uma ferida é complexa ou não;
- Se o modelo se restringisse a prever todas as feridas como não complexas (por se terem observado mais feridas não complexas do que complexas), teria 59.03% de chances de prever corretamente a complexidade da ferida;
- Com um valor p inferior a 0.0001, podemos dizer que a diferença entre a precisão do modelo e o NIR (No Information Rate, os 59.03% já mencionados) é muito significativa, assim concluindo que o modelo é significativamente preditivo;
- O modelo tem 79.66% de chances de prever uma ferida como complexa sabendo que é, de facto, complexa (true positive);
- O modelo tem 76.47% de chances de prever uma ferida como não complexa sabendo que é, de facto, não complexa (true negative);
- 70.15% das feridas que o modelo preveu como complexas são, de facto, complexas;
- 84.42% das feridas que o modelo preveu como não complexas são, de facto, não complexas;
- 40.97% das feridas são complexas. 32.64% das feridas foram corretamente previstas como complexas;
- O modelo classificou 46.53% das feridas como complexas.

Exercício 2

O segundo, e último, exercício do trabalho pede pela análise de clusters de uma base de dados relacionada à avaliação da qualidade de vida. Apesar de poder ser realizado com recurso ao programa JASP, optamos por utilizar o R pela sua versatilidade.

Lendo o ficheiro de dados, normalizamos e escolhemos apenas as variáveis de cariz sociodemográfico, aquelas que serão usadas para a construção dos clusters. Através do gráfico do cotovelo (Figura 4), percebemos uma grande redução da soma dos quadrados intra-clusters (o eixo das ordenadas) de k = 1 para k = 2, muito superior à redução notada de k = 2 para k = 3. Isso permitiu-nos concluir que o melhor valor de k é 2.

Através de um ciclo for, criamos um gráfico com as médias das silhuetas calculadas para valores de k entre 2 e 6 (Figura 5). Através deste gráfico, percebemos que k = 2 tem a maior silhueta média, pouco superior a 0.30. Apesar de isso representar uma qualidade de clusters fraca a média, é ainda assim superior a todas as outras quantidades de clusters, corroborando a ideia de que o melhor número de clusters é apenas 2. Portanto, correndo o método K-Means com 2 clusters, e criando um gráfico de perfis dos clusters (Figura 6), percebemos que os dois clusters criados variam significativamente em padrão, tornando-se fácil identificar os perfis de cada um dos clusters:

- Através das variáveis do Domínio Físico, Psicológico, Relações Sociais, Meio Ambiente e Geral podemos diferenciar as nossas amostras em dois grupos diferentes;
- O primeiro dos dois grupos é o grupo da qualidade de vida baixa, com menor pontuação em todos os domínios observados, indicando maior vulnerabilidade física, emocional e social;
- O segundo é o grupo da qualidade de vida alta, com maior pontuação em todos os domínios observados, indicando maior bem-estar e uma percepção positiva da vida.

Usando outros gráficos, conseguimos perceber aspetos importantes da análise de clusters acabada de realizar. Por exemplo, através da árvore de decisão (Figura 7), notamos que a variável do domínio psicológico é aquela que é utilizada para dividir a árvore em dois. Para os indivíduos do cluster de qualidade de vida baixa, o Domínio do Meio Ambiente mostra-se decisiva.

Por outro lado, para o cluster de qualidade de vida alta, o Domínio Geral é decisivo na separação.

Através de um gráfico de pizza (Figura 9), com o tamanho dos clusters, percebemos que os clusters estão quase que igualmente distribuídos, tendo os dois aproximadamente 50% das amostras observadas.

O PCA (Figura 10) que realizamos permite-nos confirmar a eficácia do agrupamento realizado pelo K-means, uma vez que os indivíduos apresentam-se bem separados entre os dois clusters, tendo o modelo um poder discriminativo elevado.

Através da análise de um gráfico de silhuetas (Figura 11), notamos que a largura média da silhueta é de 0.31, ou seja, de fraca a média. Ou seja, apesar de haver separação entre clusters, o agrupamento não é extremamente nítido. Ainda por esse mesmo gráfico podemos concluir que a média da silhueta do cluster de qualidade de vida baixa é 0.25, estando, portanto, no limite da qualidade aceitável, enquanto a média da silhueta do cluster de qualidade de vida alta é 0.36, indicando melhor separação em relação ao outro cluster, com mais indivíduos bem agrupados.

Finalmente, fizemos uma tabela final com as médias dos valores das variáveis para cada um dos clusters. Para cada cluster temos:

- O grupo da qualidade de vida baixa têm uma média do domínio físico de 69.7, do domínio psicológico de 68.7, do domínio das relações sociais de 67.1, do domínio do meio ambiente de 67.1 e do domínio geral de 62.1;
- O grupo da qualidade de vida alta têm uma média do domínio físico de 83.9, do domínio psicológico de 86, do domínio das relações sociais de 85.2, do domínio do meio ambiente de 79.4 e do domínio geral de 77.4.

Anexos

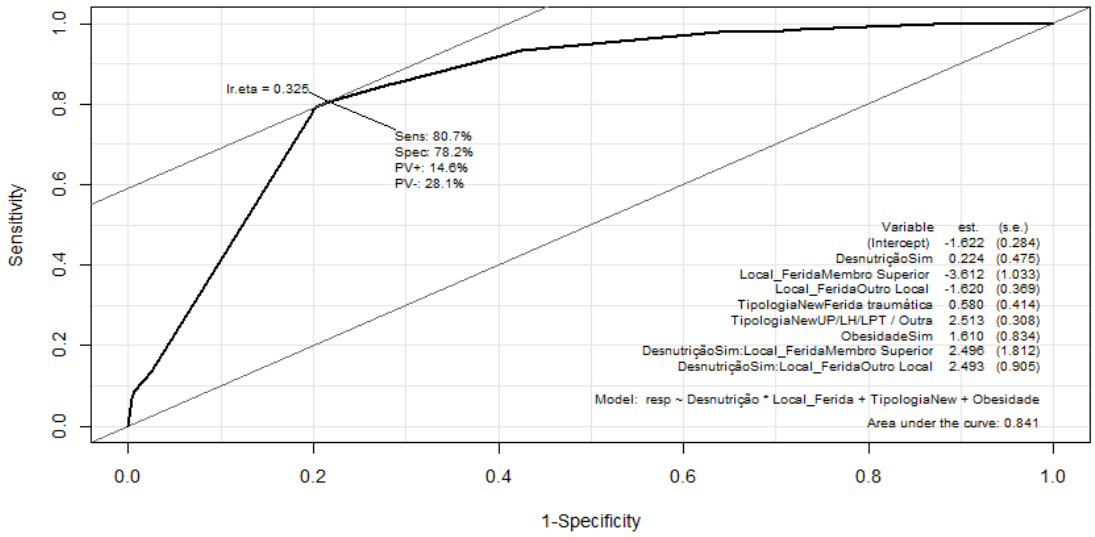

Figura 1 – Gráfico Curva ROC

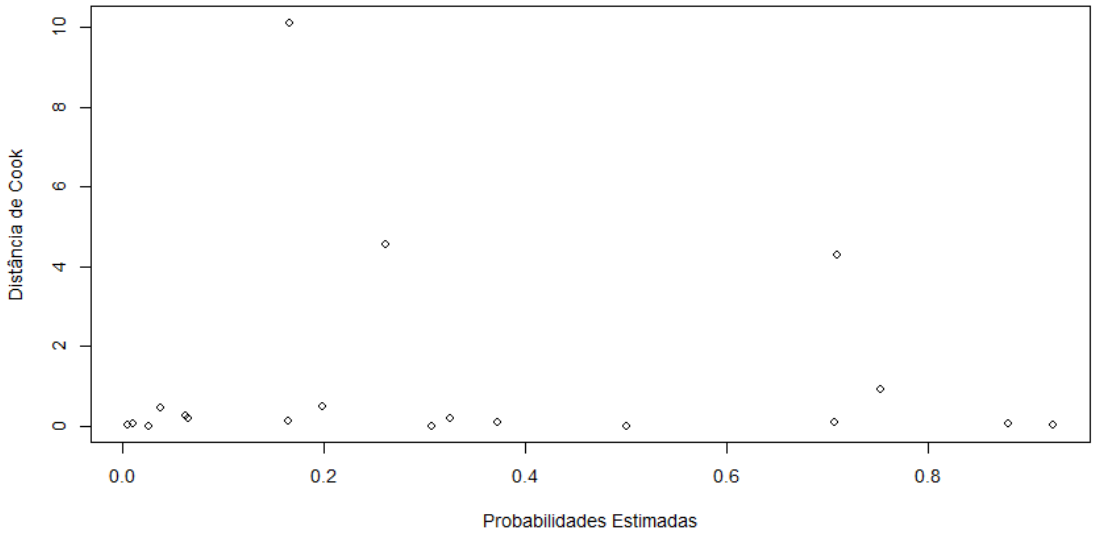

Figura 2 – Gráfico Distância de Cook

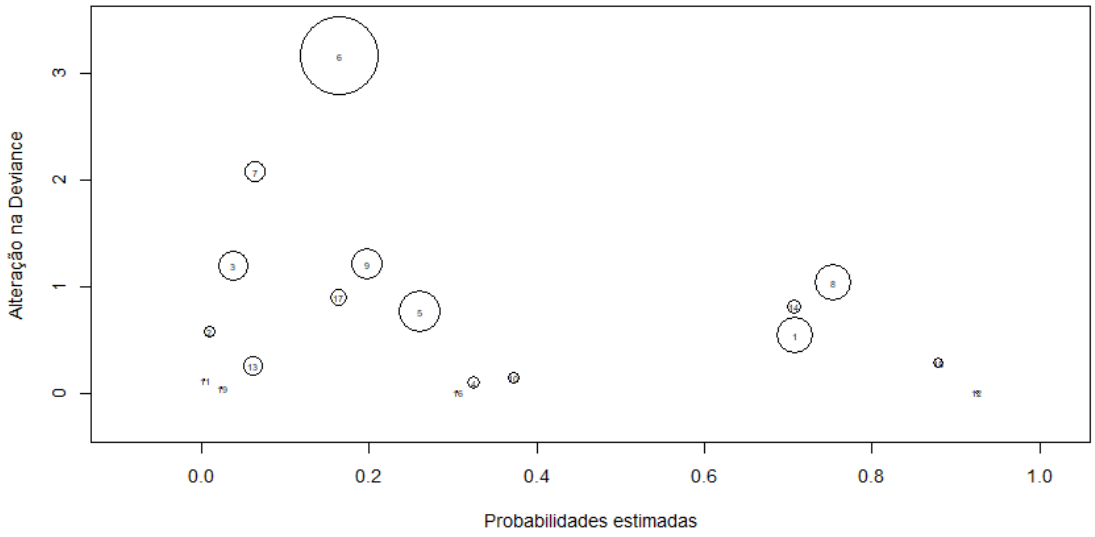

Figura 3 – Gráfico de Padrões Influentes com Deviance e Deltabeta

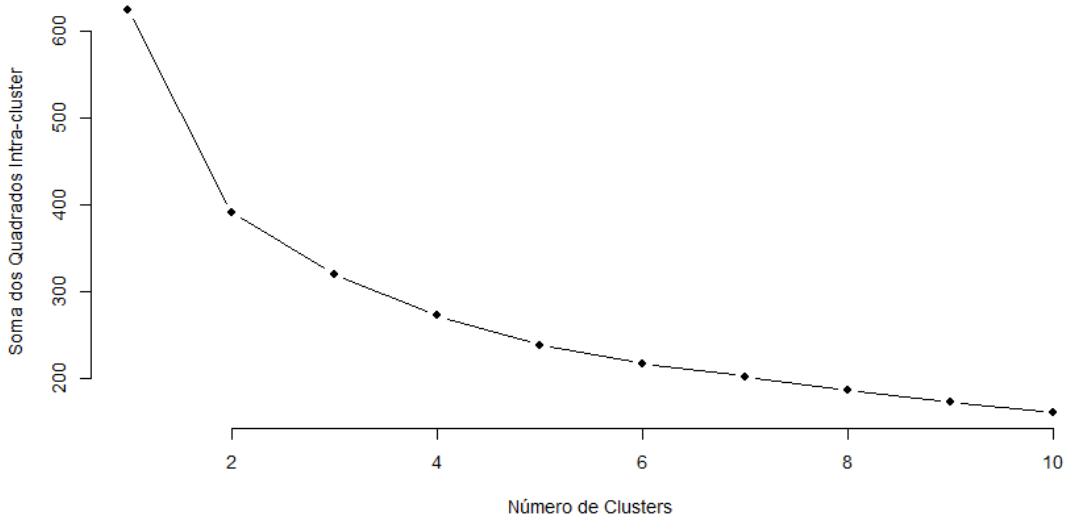

Figura 4 – Gráfico do Método do Cotovelo para Determinação do Número Ótimo de Clusters

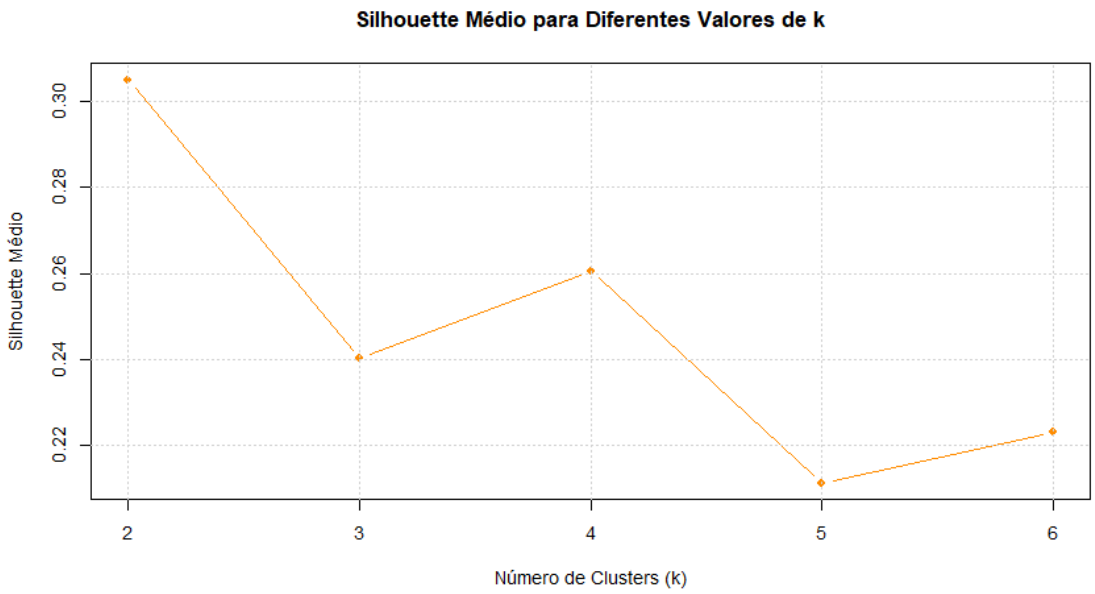

Figura 5 – Gráfico Silhouette Médio para Diferentes Valores de k

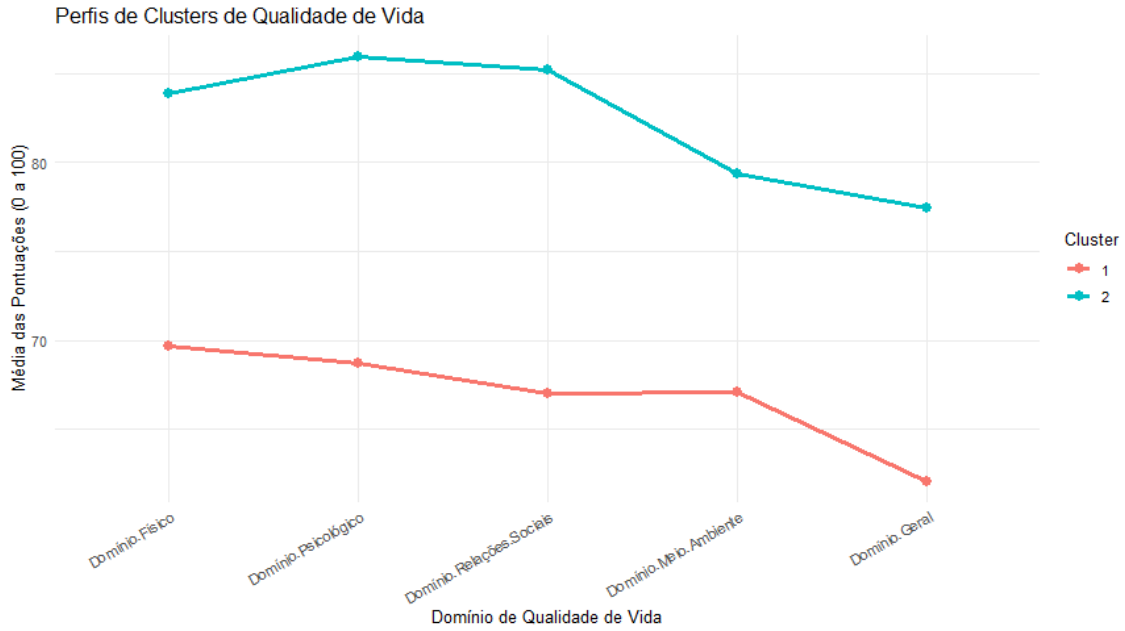

Figura 6 – Gráfico de Perfis dos Clusters

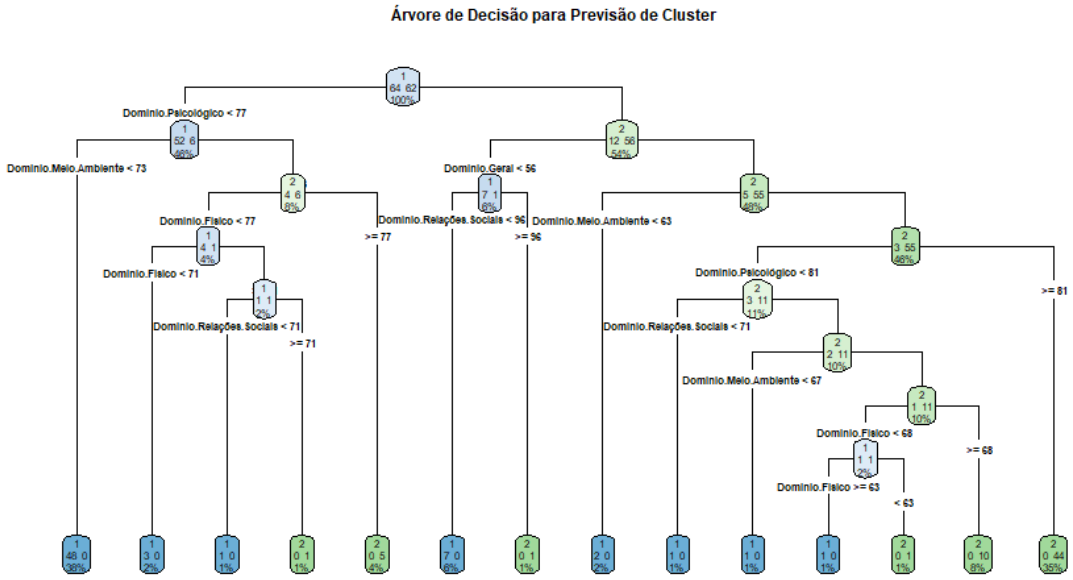

Figura 7 – Árvore de Decisão para Previsão de Cluster

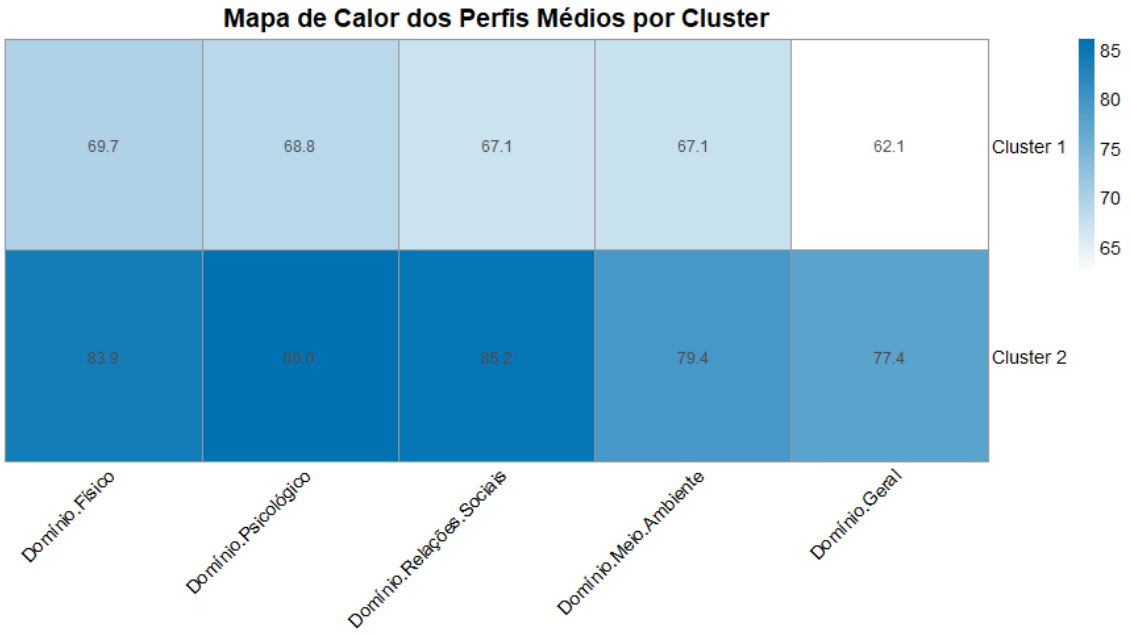

Figura 8 - Mapa de Calor dos Perfis

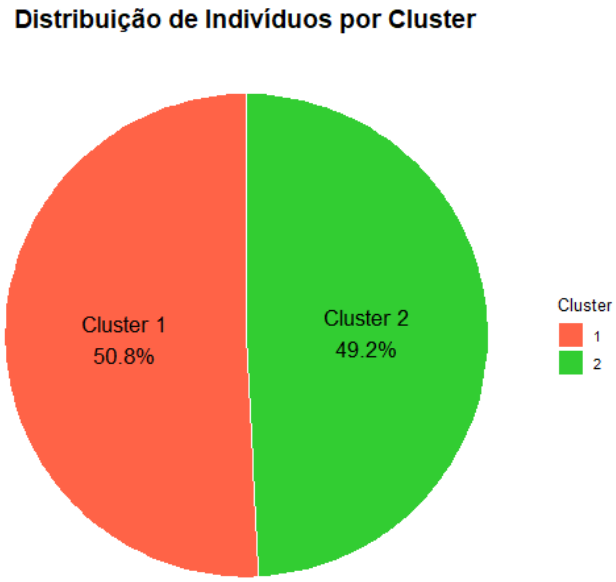

Figura 9 – Distribuição de Indivíduos por Cluster

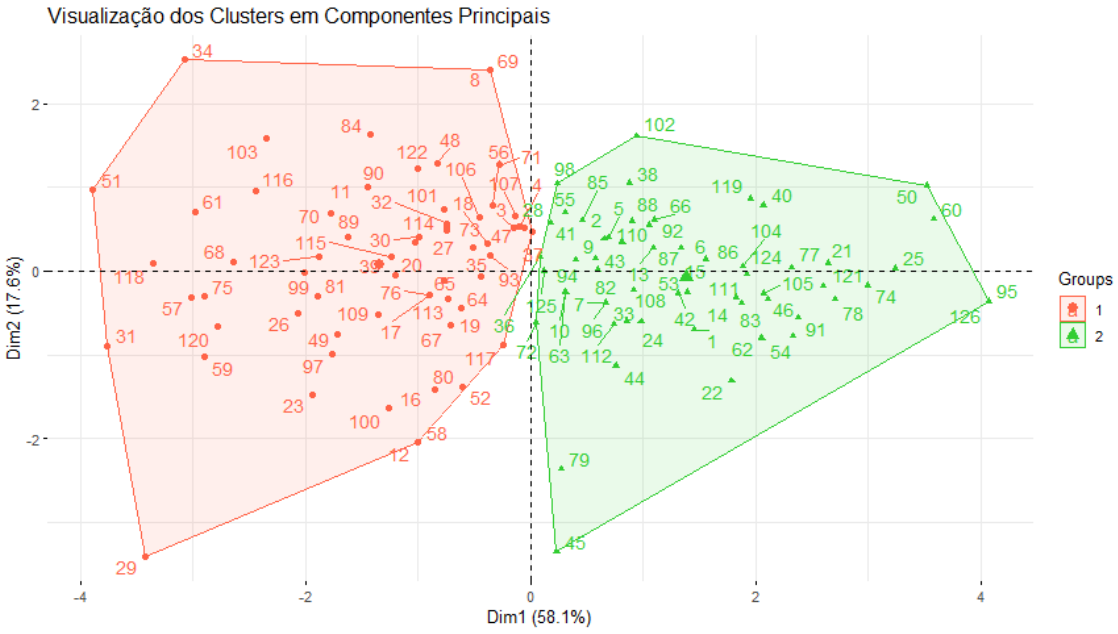

Figura 10 - PCA para Visualizar Clusters

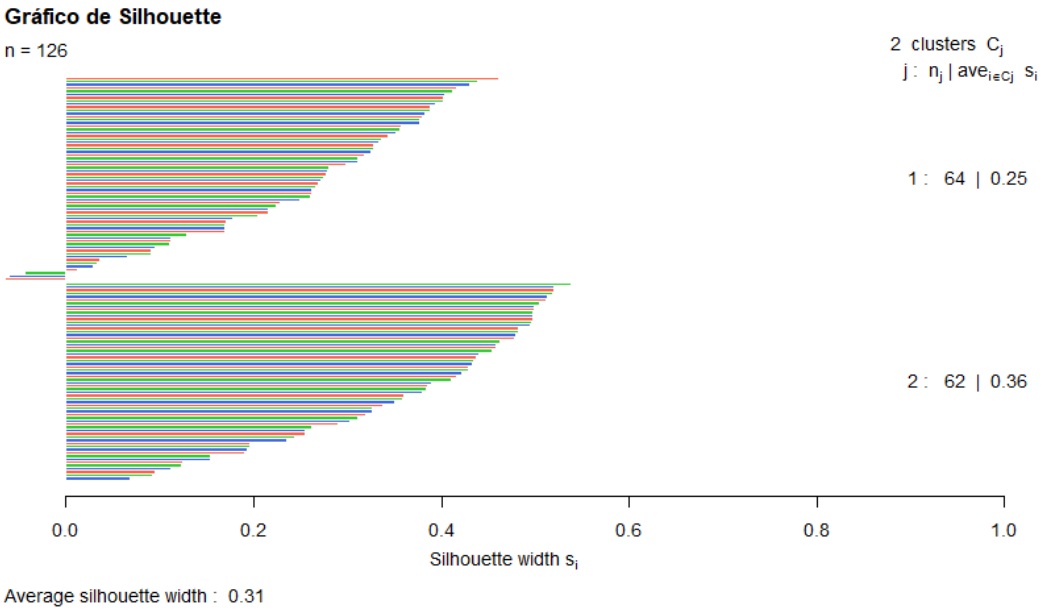

Figura 11 – Gráfico de Silhouette