Lisbon School of Economics & Management
Universidade de Lisboa

Mestrado em Métodos Quantitativos para a Decisão Económica e Empresarial

Caso de Estudo

Modelação e Análise Multivariada

Acácio dos Santos Carriço Rebocho, nº48976

André Almeida Catarino, nº56788

2 de junho de 2022

1. *Introdução*

Com o presente trabalho pretende-se, a partir de um conjunto de dados previamente recolhido que nos é disponibilizado com indicadores financeiros de várias empresas, criar um indicador financeiro que permita construir um ranking das empresas. Para o efeito, recorremos à Análise de Componentes Principais (ACP) de forma a reduzir a dimensionalidade e a captar os padrões da variância dos dados com a construção de componentes principais, procurando minimizar a quantidade de informação perdida. Foi nos também proposto desenvolver modelo(s) preditivo(s) para prever se determinada empresa entrará em falência, ou não, no próximo período, tendo sido analisado dois métodos de previsão distintos, Análise Discriminante (AD) e Regressão Logística (RL), com a respetiva seleção do melhor modelo. Por fim iremos realizar uma Análise de Clusters aos dados, de forma a agrupá-los em clusters em que as observações (empresas) em cada grupo sejam o mais homogéneas possível, e que sejam simultaneamente o mais diferente das observações noutros grupos. A análise e cálculos referenciados neste relatório foram realizados com recurso do software RStudio.
A base de dados disponibilizada conta com 1000 observações, neste caso empresas, onde cada uma delas é caracterizada por 52 variáveis observáveis, sendo 3 delas categóricas:
X0 - variável binária com o valor 1 no caso de a empresa estar em falência;
X43 - Passivo-ativos: 1 se o total de passivos exceder o total de ativos, 0 caso contrário;
X50 - Lucro Líquido: 1 se o Lucro Líquido for Negativo nos últimos dois anos, 0 caso contrário.

2. *Análise de Componentes Principais (ACP) para a criação de um Ranking Financeiro*

A ACP é uma técnica que permite a construção de novas variáveis explicativas, resultantes de combinações lineares das variáveis originais, sendo que as novas variáveis não apresentam correlação, e ao contrário das originais. Essa construção será feita de modo que um menor número de variáveis explique um valor satisfatório da variabilidade dos dados. Onde a primeira nova variável explique o máximo possível da variabilidade dos dados, a segunda variável irá explicar o máximo possível da variabilidade dos dados não explicada pela primeira variável, e assim progressivamente até chegarmos à última variável, que irá apresentar pouco poder explicativo.
E assim sendo, se pretendêssemos explicar uma variabilidade total dos dados suficientemente grande, mas menor que a total, poderíamos omitir as últimas novas variáveis de forma a reduzir a dimensionalidade dos dados.

Primeiramente, e na filtração inicial dos dados, removemos as variáveis categóricas X0, X43 e X50 pois não têm influência nesta análise. Sendo apenas analisadas as restantes variáveis na ACP que exporemos de seguida.

2.1. Estandardização dos Dados

Uma vez que a ACP é sempre realizada com dados corrigidos pela média, bastava então averiguar se deveríamos estandardizar, ou não, os dados. Para tomar esta decisão teremos que ter em conta a variância relativa das variáveis originais, relativamente à variância total. Sendo estandardizamos os dados se existir uma grande disparidade entre as variáveis, ver Figura 1 – Anexos.

Neste caso, primeiro definimos a matriz de variâncias e covariâncias, e através do traço desta mesma matriz (cálculo da variância total), procedemos ao cálculo das variâncias relativas para cada variável original, representadas na figura 1.

Através da interpretação da anterior figura, podemos rapidamente concluir que a estandardização dos dados é necessária.

2.2. Obtenção das Componentes Principais

Para obter as componentes principais iremos resolver o problema de Valores e Vetores próprios associado à matriz de correlações, no caso em análise, uma vez que os dados foram estandardizados, caso os dados não o fossem iremos utilizar em alternativa a matriz de variâncias e covariâncias dos dados.

Na figura 2 - Anexos, estão representadas as Variâncias associadas a cada Componente Principal, novas Variáveis, indo decrescendo o seu valor com o aumento do índice da Componente Principal, tal como é pretendido.

2.3. Definição do número de componentes principais a reter

Existem várias técnicas utilizadas na prática para decidir o número de Componentes Principais a reter, mas a que apresenta maior consenso entre os diversos autores é a Análise Paralela de Horn, que decidimos então implementar, ver Figura 3 – Anexos.

Segundo esta técnica iremos reter apenas os Componentes Principais cujos Valores Próprios definidos a partir da base de dados superam a média dos Valores Próprios dos dados simulados. A representação gráfica na Figura 3 simplifica esse estudo, sendo a média dos Valores Próprios simulados, representada pela linha vermelha.

Facilmente concluímos que deveríamos reter no mínimo 10 Componentes Principais, sendo que a 11ª e 12ª Componentes teriam que ser comparadas pelo seu valor, visto estarem bastante próximas da média referente aos dados simulados.

Depois de proceder a uma análise mais detalhada, conseguimos apurar que iremos manter as 11 primeiras Componentes Principais uma vez que a 11ª Componente Principal têm um Valor Próprio de 1,2197 e a média nesse ponto é 1,2020, e a 12ª Componente Principal têm um Valor Próprio de 1,1797 sendo a média nesse ponto de 1,1840, ver Tabela 1 – Anexos.

Para além disso, podemos ainda verificar através da tabela anterior que ao utilizarmos apenas 11 Componentes Principais, iremos conseguir explicar, aproximadamente, 65% da Variância Total dos

dados, onde a maior parte advém da 1ª Componente Principal, sendo responsável por cerca de 20 pontos percentuais dos 65% explicados.

Perde-se assim 35% da informação referente à variabilidade.

2.4. Interpretação dos Componentes Principais

De forma a podermos interpretar os Componentes Principais iremos recorrer aos PC-Scores, que são as coordenadas dos dados na nova base, de vetores próprios, representando basicamente as novas variáveis. E também à interpretação dos Loadings, que correspondem ao valor da correlação de determinada Variável j com um determinado Componente Principal i, podendo como é obvio ser positivos, no casso dessa variável influenciar positivamente essa Componente Principal, ou negativos, no caso contrário. Sendo que quanto maior for o valor do Loading, em modulo, maior também irá ser a sua influência nessa Componente Principal.

Ao valor de corte atribuímos um valor de 0,5, para definir um determinado Loading como relevante ou não, estando realçados a verde as correlações positivas e a vermelho as correlações negativas, ver Figura 7 – Anexos.

Em relação aos Scores, no nosso caso devido ao elevado número de Componentes Principais não nos irá ser possível realizar uma interpretação relevante para esta base de dados.

Mas através da interpretação dos Loadings poderemos atribuir um significado prático a cada Componente principal, sendo que:

- A 1ª C.P. é afetada positivamente pelas variáveis X0, X2, X3, X10, X13, X14, X15, X17, X18, X31, X36, X42, X44, X47. Sendo esta afetada positivamente, e maioritariamente, por Variáveis relacionadas com o lucro ou rácios que utilizam o lucro da empresa, poderemos atribuir-lhe o valor de índice de lucro generalizado por empresa em estudo;
- A 2ª C.P. é afetada positivamente pelas variáveis X16, X26, X33, X35 e negativamente pelas variáveis X2, X3, X27, X47, X51. Sendo esta afetada maioritariamente por Variáveis relacionadas com as vendas ou rácios que utilizam as vendas, poderemos atribuir-lhe o valor de índice de vendas, geral.
- A 3ª C.P. é afetada positivamente pelas variáveis X4, X5, X6, X7. Sendo esta afetada maioritariamente, por Variáveis relacionadas com fatores após impostos, como Lucros Líquidos, iremos atribuir-lhe o impacto dos impostos nas empresas em estudo;
- A 4ª C.P. é afetada negativamente pelas variáveis X29, X30, X41. Sendo esta afetada negativamente, e maioritariamente, por Variáveis relacionadas com a contabilização de longo prazo ou ao longo do tempo iremos atribuir-lhe o impacto das Taxas de juro nas empresas em estudo;
- A 6ª C.P. é afetada positivamente pelas variáveis X28 e X37. Não conseguimos chegar a uma interpretação plausível desta Componente Principal;
- As 7ª e 8ª C.P. são ambas afetadas positivamente pelas variáveis X20 e X21. Sendo elas afetadas positivamente apenas por Variáveis que representam rácios do crescimento do Lucro ao longo do tempo, poderá representar a Componente relacionada com o crescimento da operação das empresas em estudo;

- As restantes CP's , não apresentam nenhum Loading relevante, segundo o valor de corte previamente estabelecido.

2.5. Conclusões

Tendo em conta os significados práticos que conseguimos atribuir a alguns dos Componentes Principais bem como os pesos das variáveis originais j em relação a determinada Componente principal i, Figura 8 - Anexos, Wij, a finalidade principal desta Análise de Componentes Principais parece ter sido concluída com algum sucesso. Uma vez que a criação de um Índice financeiro para criar um ranking das empresas representadas na nossa base de dados poderá ser simplificado, visto que ao passar de 51 variáveis originais para classificar as 1000 empresas, passaremos para apenas 11 variáveis para a realização do ranking.

Sendo o lugar de cada empresa definido pela substituição do seu PC-Score na correspondente variável de cada Componente principal, sendo depois aplicada uma média ponderada face ao valor da variância relativa de cada Componente Principal. A empresa com Valor mais alto estaria em 1º lugar do ranking, a empresa com 2º valor mais alto estaria em 2º lugar do ranking, etc...

Não consideramos que esta metodologia de classificação ou ranking financeiro das empresas possa ser classificada com um sucesso total, uma vez que ao passarmos para as 11 Componentes principais se irá perder uma significativa quantidade de variabilidade dos dados originais, cerca de 35%.

3. Previsão de possível falência no próximo período

De modo a podermos realizar esta previsão, recorremos a dois métodos distintos, a Análise Discriminante (AD) e a Regressão Logística (RL).

No caso da Análise Discriminante (AD) é uma técnica que possibilita definir as características a partir dos quais iremos distinguir os grupos, no nosso caso serão dois grupos utilizando a variável X0 para os definir, para que ao conhecer as características de um novo objeto, e tendo em conta o eixo de separação dos grupos definido por este método, se possa prever a que grupo pertencerá essa mesma observação. Tempo nos tido em consideração que a Análise Discriminante só deverá ser usada quando os dados seguem uma distribuição normal multivariada, hipótese esta que é violada sempre que existem variáveis categóricas, como é o nosso caso. Sendo que ao ser violada esta hipótese tanto os métodos de classificação como os testes de significância das variáveis seriam afetados.

Como alternativa surge a Regressão Logística (RL) técnica esta que é recomenda sempre que a hipótese de normalidade dos dados, que falamos anteriormente, não é respeitada. Este método de previsão não irá fazer assim quaisquer assunções sobre a distribuição dos dados, sendo à partida uma melhor solução ao nosso problema de previsão. Onde através das variáveis observáveis iremos realizar uma regressão linear onde a função dependente não é uma função linear, para definir em que grupo determinada observação irá entrar.

3.1. Análise Discriminante

De forma a podermos aplicar a Análise Discriminante de forma coerente, decidimos, de forma a cumprir a hipótese de normalidade multivariada dos dados que este método assume, remover da base de dados inicial as variáveis X43 e X50, ambas categóricas.

Deixando apenas uma variável categórica binaria, X0 - variável com o valor 1 no caso da empresa estar em falência ou 0 no caso contrário, que irá definir então os nossos dois grupos de estudo sobre os quais iremos classificar as nossas observações.

3.1.1. Analisar Poder Discriminante das Variáveis

Primeiramente, e de forma a verificar se cada variável, de forma isolada, apresenta poder discriminante, realizamos uma análise univariada para cada variável em estudo, através do Teste-t, onde se testa a diferença de médias entre os 2 grupos, para cada variável, assumindo-se uma variância igual e uma hipótese nula de igualdade de médias entre os 2 grupos, que ao se rejeitar revela que existe poder discriminante nessa variável.

Com base na Figura 4 - Anexos podemos facilmente chegar à conclusão de que as variáveis X8, X9, X11, X16, X19, X24, X25, X28, X34, X40, X45, X48, X49 não têm poder discriminante, onde eventualmente poderia haver dúvida seria nas Variáveis X28 e X40, mas por terem Valor-P igual a 0,05071 e 0,05002 também se enquadram nas anteriores.

Já para o teste da significância conjunta da análise multivariada realizamos o teste de Wills, semelhante ao Teste-t anterior tendo apenas a diferença de se testar todas as variáveis em conjunto.

Ao realizarmos o teste para todas as variáveis, incluindo as não discriminantes, obtivemos um Valor-P = 1, muito superior a 0,05, e assim não poderíamos rejeitar a Hipótese nula e logo as variáveis não teriam significância em conjunto. Para ultrapassar este problema realizamos então o mesmo teste, mas apenas para as variáveis que apresentavam poder discriminante, tendo então agora obtido um Valor-P < 2.2*10-16 que é muito inferior a 0,05 e assim já podemos afirmar que existe no conjunto existe poder discriminante. Indo utilizar a partir daqui apenas as variáveis discriminantes.

3.1.2. Identificar Eixo de Separação

Chegando ao momento de definir o eixo de separação dos 2 grupos, definido pela função discriminante temos que ter em conta que para obter o angulo que maximiza a separação dos grupos teremos que por sua vez maximizar $\gamma = \frac{SS_b}{SS_w}$, onde SSb é Soma dos Quadrados Corrigida pela média entre grupos e SSw a Soma dos Quadrados Corrigida pela média dentro dos grupos.

Sendo este eixo de separação, Z, nada mais que uma combinação linear das variáveis discriminantes que maximiza o rácio γ. Ao calcularmos a função discriminante no RStudio também nos foi indicado que 900 observações correspondem a empresas que não estão falidas e 100 observações correspondem a empresas que estão falidas.

3.1.3. Classificação de Observações Futuras

Definindo o valor de corte, cutoff, representado pela linha azul a tracejado na Figura 5, que será basicamente o Score a partir do qual determinada observação futura ira ser englobada num dos dois grupos. Sendo que o Score de cada observação futura irá ser calculado através da função discriminante, definida anteriormente.

No nosso caso, o valor de corte foi definido em aproximadamente 0, 9.118×10^{-14} , sendo calculado usando a seguinte formula: $cutoff = \frac{n_1 \times \overline{Z_1} + n_2 \times \overline{Z_2}}{n_1 + n_2}$, onde $\overline{Z_g}$=Valor Médio dos Scores do grupo g e n_g = nº observações do grupo g, g = [1;2].

Realizamos então uma previsão dentro da amostra, In-Sample, e através da análise do gráfico da Figura 5 representante da distribuição dos scores, é possível ver que os mesmos se encontram bem repartidos de acordo com cada um dos 2 grupos, a preto as observações do grupo "Não Falida" e a vermelho as do Grupo "Falida" separados pelo valor de corte.

Apesar de existirem de alguns pontos fora do seu grupo, sendo mais difíceis de prever, algo normal visto que a previsão nunca terá uma precisão de 100%, ver Figura 5 – Anexos e Tabela 2 – Anexos.

Em suma, para a classificação In-Sample verificou-se uma precisão de 93,4% sendo apenas mal classificadas 66 empresas, das quais 21 Não Falidas foram englobadas nas Falidas e 45 Falidas foram englobadas nas Não Falidas.

Já para a classificação Out-Sample, obtida também através da função discriminante, mas desta vez com validação cruzada, verificou-se uma precisão muito inferior, como era esperado devido a existir um sobre ajustamento da função aos dados de estimação, de 49,4% sendo mal classificadas 506 empresas, das quais 449 Não Falidas foram englobadas nas Falidas e 57 Falidas foram englobadas nas Não Falidas.

Ou seja, a precisão da previsão das observações diminui bastante de dentro para fora da amostra, descendo em cerca de 44 pontos percentuais, sendo que em ambos os casos uma pior previsão ocorre na classificação das empresas falidas, com uma precisão de 55% In-Sample e 43% Out-Sample.

3.2. Regressão Logística

Ao contrário da análise discriminante, a regressão logística não assume hipóteses distribucionais sobre os dados, é possível incluir todas as variáveis categóricas, uma vez que a hipótese da normalidade multivariada dos dados pode ser violada.

A estimação do modelo de regressão logística é feita pelo método de máxima verosimilhança, sendo os estimadores dos coeficientes consistentes, eficientes e seguem assimptoticamente uma distribuição normal. Tal como na Análise Discriminante anterior será a variável categórica diagnóstico, a variável dependente, X0 que irá definir então os nossos dois grupos de estudo sobre os quais iremos classificar as nossas observações.

3.2.1. Teste à significância individual dos coeficientes

Procedemos então à estimação no RStudio pelo método da máxima verosimilhança, utilizando o comando GLM (generalized linear model) para a família de distribuições binomiais. Obtendo o seguinte Output na consola, Figura 9 - Anexos, referente aos coeficientes estimados,β_j , e seus respetivos testes de significância individual, teste estes onde se pretende rejeitar $H_0: \beta_j = 0$, ou seja, que o Valor-p, Pr(>|z|), seja inferior a 0.05 se a significância for de 5% ou 0.01 se for 1%. Verificamos que apenas os repressores das variáveis X8, X12, X20, X23, X26, X28, X33, X35, X39 apresentaram significância individual a 10%, 5% ou 1%. Ponderamos retirar as variáveis cujos repressores não tinham significância, mas ao fazê-lo verificamos que tanto a precisão do modelo como o valor do pseudo R^2 de McFadden diminuíam, respetivamente para 93,3% e 0,5082767, tornando o modelo mais fraco.

3.2.2. Significância global da Regressão, Adequação do Modelo e Intervalos de Confiança para os Parâmetros

Tendo em conta a significância global da regressão procedeu-se à comparação do modelo sem regressões com o modelo com repressores. Iremos então através da Estatística de teste Likelihood ratio, com H_0: $\beta_1 = \beta_2 = \cdots = \beta_j = 0$, tendo o Valor-P para este teste sido de aproximadamente 0, ou seja, rejeitamos a Hipótese nula, a 5% neste caso, existindo assim significância global dos regressores.

Relativamente à adequação global do modelo, aplicamos o método do pseudo R^2 de McFadden, obtendo-se um valor de 0,6337606 que parece indicar um ajuste razoável. Mas sendo só por si de difícil interpretação, podemos apenas notar que o seu valor é superior ao do modelo apenas com os regressores com significância individual.

Os intervalos de confiança para os parâmetros, a 95%, estão presentes na Figura 10 - Anexos.

3.2.3. Classificação de Observações

Conhecendo os valores das variáveis independentes, recorrendo à regressão logística, podemos estimar a probabilidade de a empresa falir. Ao definirmos o valor de corte entre grupos a 0,5 podemos observar como ficam classificadas as empresas da nossa amostra. Podemos então verificar graficamente falsos positivos e falsos negativos, erros de previsão, mas na sua generalidade a parte superior do gráfico corresponde as empresas que tem probabilidade de estar falidas e a inferior ao oposto (Ver Figura 6 – Anexos). Ou seja, é uma classificação In-Sample, usando-se o modelo para classificar as observações da amostra. Sendo expressos na Tabela 3 – Anexos.

De forma a podermos comparar a precisão da Regressão Logística com a da Análise Discriminante teremos que realizar uma classificação Out-Sample também por sua vez para a Relação Logística, optando por utilizar os dados, recorrendo à técnica da validação cruzada, tal como foi aplicada na Análise Discriminante para obter os dados Out-Sample. Obtendo os resultados da Tabela 4 – Anexos.

Em suma, tanto para a classificação In-Sample como Out-Sample verificou-se uma precisão média de 94,8% sendo mal classificada 52 empresas, na análise In-Sample , 17 empresas não falidas foram

englobadas nas falidas e 35 falidas foram englobadas nas não falidas, já na análise Out-of-Sample, 35 empresas não falidas foram englobadas nas falidas e 17 falidas foram englobadas nas não falidas.

3.3. Conclusão: Análise Discriminante vs Regressão Logística

Ambos os modelos são aplicados com o principal objetivo de classificar futuras observações nos grupos, empresa falida ou não falida.

Optámos por recorrer a ambas as técnicas, uma vez que ambas têm especificidades diferentes (como por exemplo a possibilidade de na regressão logística ser possível obter o poder preditivo de cada uma das variáveis, Figura 11 - Anexo) e pretendemos analisar qual dos modelos teria uma maior capacidade preditiva, dado o nosso dataset, comparando a precisão da previsão in-sample e out-of-sample.

Tendo então em conta que a Análise Discriminante apresentou uma precisão In-Sample de 93,4% e apresentado uma queda elevada na precisão Out-Sample passando para 49,4%, já a Regressão Logística apresentou uma precisão tanto In-Sample como Out-Sample de 94,8%.

4. Analise de Clusters

Técnica que tem como objetivo o agrupamento de observações em grupos de forma que as observações em cada grupo sejam o mais homogéneo possível, em relação a determinadas características, e que as de um determinado grupo sejam o mais diferente possível das dos outros grupos.

Medida de similaridade escolhida foi a distância euclidiana:

$$D_{ij}^2 = (x_{i1} - x_{j1})^2 + \cdots + (x_{ip} - x_{jp})^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$$

Para efetuar uma análise de grupos foi delineada uma análise de clusters hierárquica aglomerativa, em que a cada iteração o número de clusters é diminuído de uma unidade (formação de n-1 clusters, para n observações).

Método de agrupamento selecionado:

```
>  #calculate agglomerative coefficient for each clustering linkage method
>  sapply(m, ac)
  average    single  complete      ward
0.9884788 0.9886981 0.9923560 0.9990436
```

Para selecionar o método de agrupamento foi analisado o coeficiente aglomerativo com maior desempenho. O coeficiente aglomerativo mede a qualidade de agrupamento aglomerativo, apresentando valores entre 0 e 1, em que valores baixos do coeficiente correspondem a estruturas em que nenhum agrupamento foi encontrado, e valores próximos de 1 representam estruturas claras que foram identificadas.

Apesar dos vários métodos de agrupamento testados apresentarem semelhantes coeficientes aglomerativos, o agrupamento por método Ward apresenta o maior coeficiente.

No método de Ward não se calculam distâncias entre clusters, sendo formado o cluster que tem a maior homogeneidade entre as observações que nele estão incluídos. A heterogeneidade de uma solução é medida pelo Error of Sum Squares (ESS).

$$ESS = \sum_{i=1}^{nclus} \sum_{j=1}^{n} \sum_{k=1}^{p} (X_{ijk} - \bar{X}_{i \cdot k})^2$$

Neste método de agrupamento, formação de um novo cluster é efetuada de forma a minimizar o ESS.

4.1 Métodos para seleção do número ótimo de clusters

- Gap Statistic
- Elbow Method
- Silhouette Coefficient

(ver Figura 12 – Anexos)

Gap Statistic:

O conceito da abordagem é encontrar forma de comparar a homogeneidade do cluster com uma referência em que os dados não apresentam um agrupamento facilmente identificável. O número ótimo de clusters por esta estimativa é atingido quando existe a maior homogeneidade possível dentro dos clusters formados relativamente aos dados inicialmente.

Método Elbow:

É o método mais comum para determinar o número ótimo de clusters. Este método é baseado no WSS (Within-Cluster-Sum of Squared Errors) que é calculado para um número diferente de clusters (k) de modo a selecionar k para qual a variação de WSS passe a ser mínima.

O conceito subjacente ao método “Elbow” é que a variância intra-clusters é altamente influenciada para um número pequeno de clusters, sendo a influência minimizada em função de um maior número de clusters, o que leva à formação de um cotovelo na curvatura. O número ótimo de clusters está localizado no referido cotovelo.

Coeficiente da Silhueta:

Refere-se a um método de interpretação e validação de consistência dentro de agrupamentos de dados. A técnica, fornece uma representação gráfica sucinta de quão bem cada observação foi classificada, comparando com os clusters vizinhos, recorrendo a um coeficiente que varia entre 0 e

1 (mais próxima de 1 indica que a observação se encontra no cluster em que a sua distância média às outras observações dentro do mesmo cluster é menor do que nos restantes clusters).

Na figura 12, encontram-se os resultados da implementação de todos os métodos com um número ótimo de clusters(k) de 3, 7 e 10, respetivamente, em que foi definido um k máximo de 10.

Selecionámos 3 clusters para a formação do dendrograma (k=3), selecionando o resultado do método “Elbow” como critério de decisão, uma vez que era de mais fácil identificação dos clusters no dendrograma gerado (Figura 13), quando comparado com o número elevado de clusters indicado pelos restantes métodos

ANEXOS

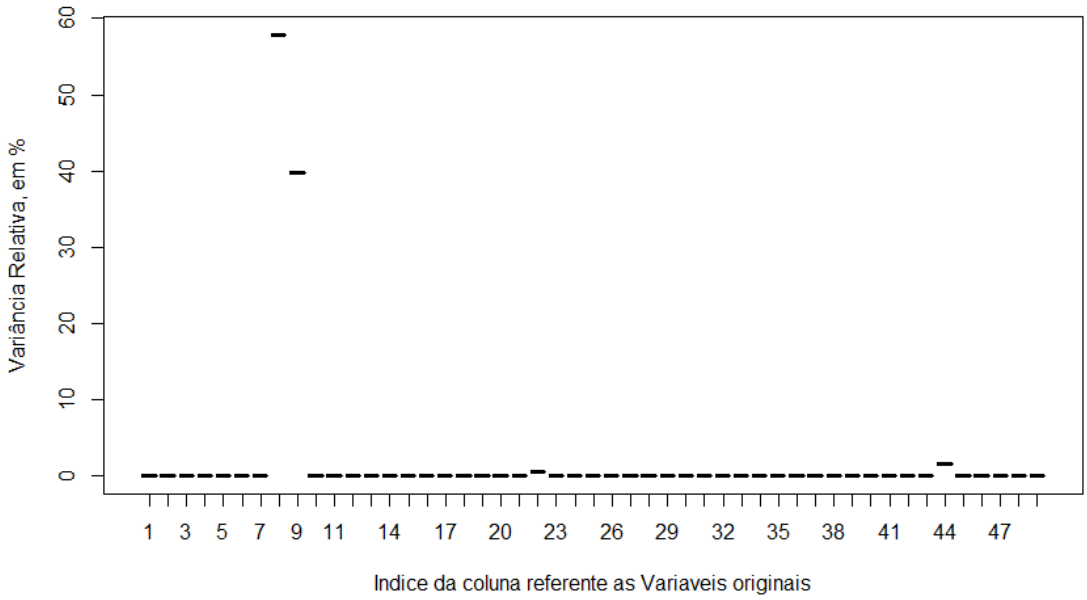

Figura 1 - Variância Relativa das Variáveis Originais

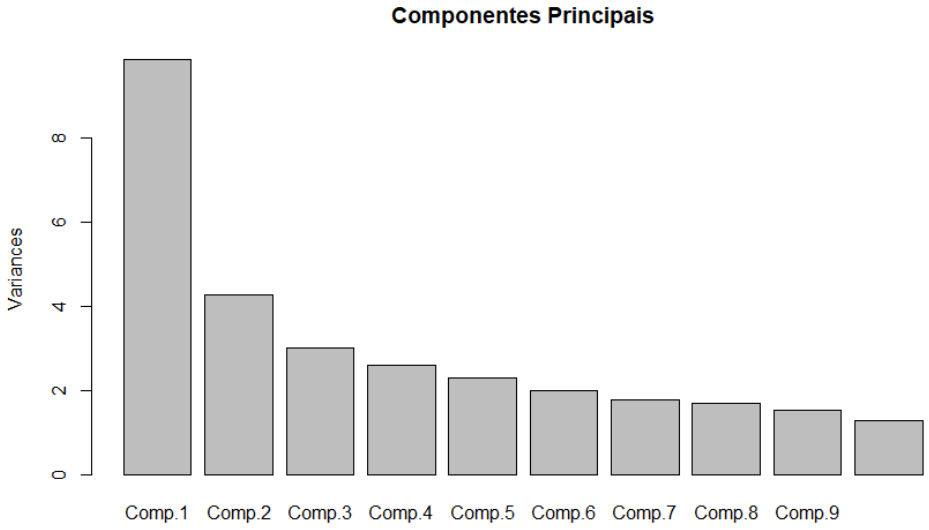

Figura 2 - Valores Próprios, ou variâncias, dos Componentes Principais

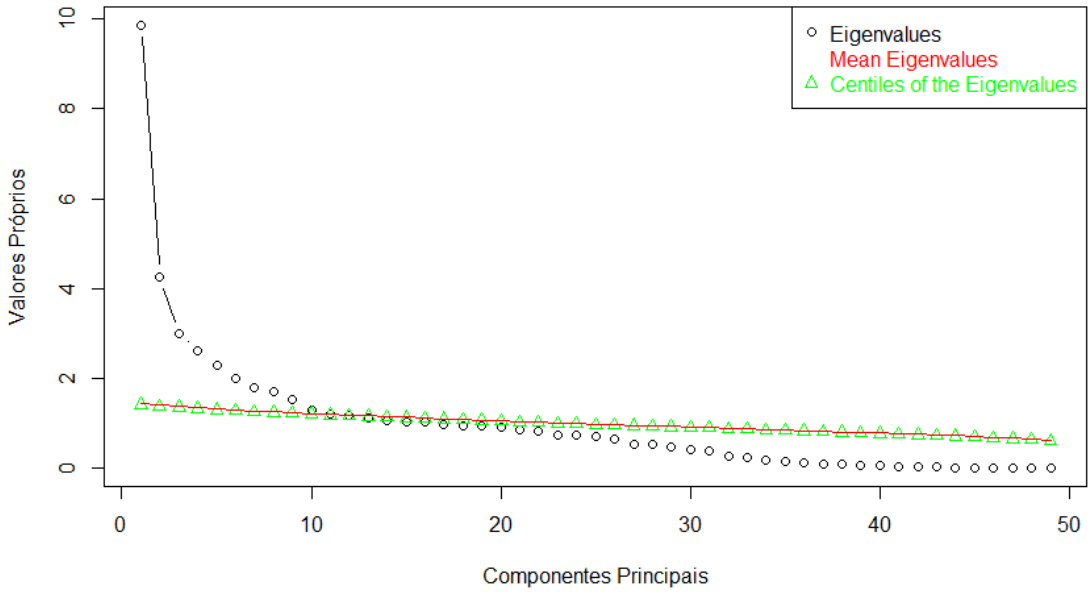

Figura 3 - Análise Paralela de Horn

```
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8     Comp.9    Comp.10    Comp.11
Standard deviation     3.139838 2.06527039 1.73277293 1.61745545 1.51391417 1.41660491 1.33665758 1.30530437 1.24142450 1.13871169 1.1038559
Proportion of Variance 0.201397 0.08713493 0.06133689 0.05344451 0.04682103 0.04099547 0.03649882 0.03480663 0.03148321 0.02648903 0.0248922
Cumulative Proportion  0.201397 0.28853195 0.34986884 0.40331335 0.45013438 0.49112985 0.52762867 0.56243530 0.59391851 0.62040754 0.6452997
```

Tabela 1 - Importância dos Componentes Principais

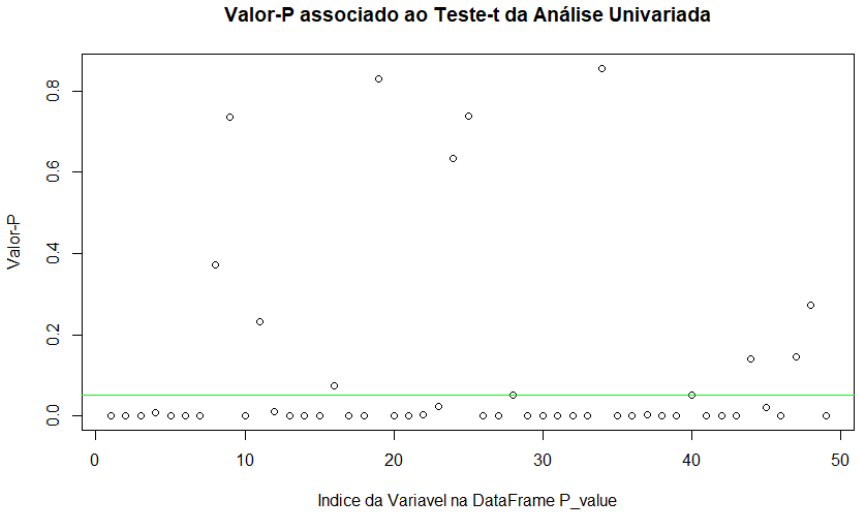

Figura 4 - Valores-P associados ao teste de cada variável em estudo; Linha verde é o Valor-P associado a uma confiança de 5%

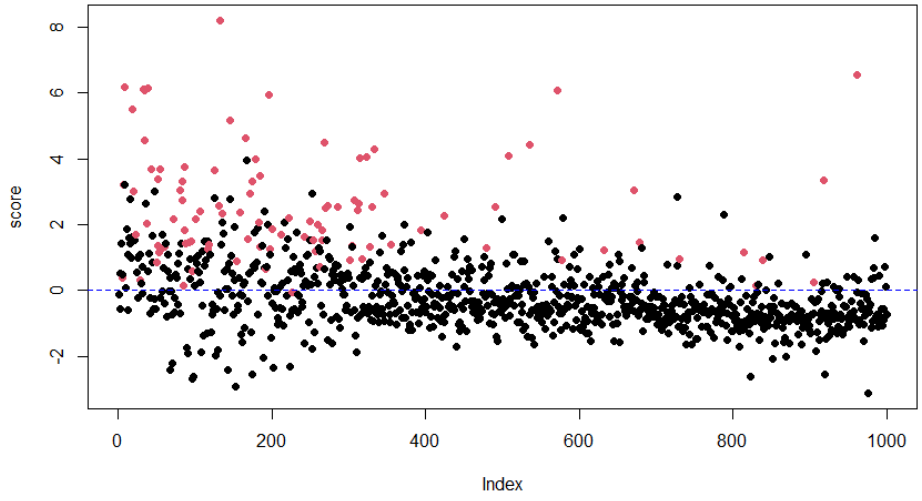

Figura 5 - Gráfico de distribuição dos scores dentro da amostra, In-Sample

MATRIZ DE CLASSIFICAÇÃO IN-SAMPLE			
GRUPO A QUE PERTENCE	Número de Observações	Previsão	
		Empresa Não Falida (0)	Empresa Falida (1)
EMPRESA NÃO FALIDA (0)	900	879	21
EMPRESA FALIDA (1)	100	45	55
MATRIZ DE CLASSIFICAÇÃO OUT-SAMPLE			
GRUPO A QUE PERTENCE	Número de Observações	Previsão	
		Empresa Não Falida (0)	Empresa Falida (1)
EMPRESA NÃO FALIDA (0)	900	451	449
EMPRESA FALIDA (1)	100	57	43

Tabela 2 - Matriz de Classificação In-Sample e Out-Sample

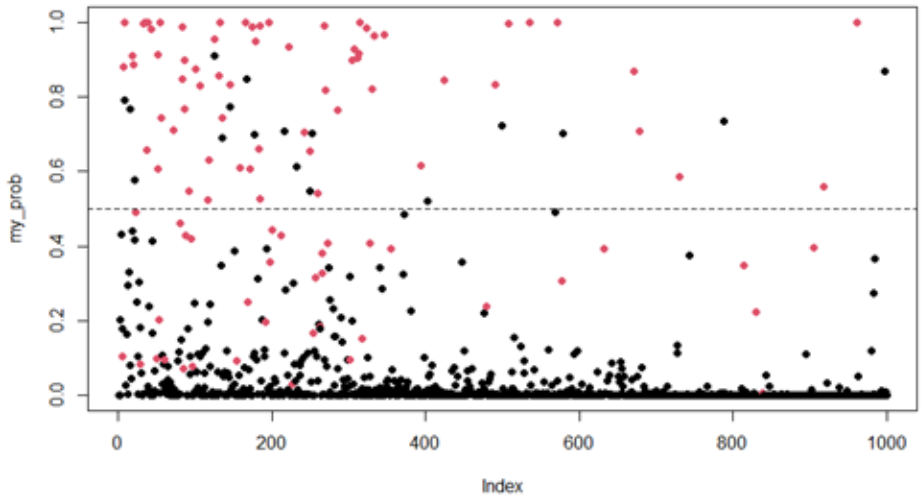

Figura 6 - Probabilidade de determinada empresa falir

Grupo a que pertence	Número de Observações	Previsão	
		Empresa não falida (0)	Empresa falida (1)
Empresa não falida (0)	900	883	17
Empresa falida (1)	100	35	65

Tabela 3 - Matriz de Classificação In-Sample

Grupo a que pertence	Número de Observações	Previsão	
		Empresa não falida (0)	Empresa Falida (1)
Empresa não falida (0)	918	883	35
Empresa Falida (1)	82	17	65

Tabela 4 - Matriz de Classificação Out-Sample

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11
X1	0,85	0,12	0,07	0,05	0,15	0,04	0,07	0,01	0	0,01	0,01
X2	0,61	-0,51	-0,08	-0,39	-0,18	-0,08	0,16	-0,17	0,16	-0,08	-0,04
X3	0,61	-0,51	-0,08	-0,39	-0,18	-0,08	0,16	-0,18	0,16	-0,08	-0,04
X4	0,48	0,07	0,55	0	-0,11	-0,01	0,13	-0,28	0,04	-0,11	0,25
X5	0,44	0,11	0,77	0,05	-0,35	0	-0,17	0	0,08	0,09	-0,05
X6	0,19	0,07	0,55	0,07	-0,37	0,01	-0,34	0,23	0,07	0,2	-0,29
X7	0,43	0,11	0,76	0,07	-0,34	0	-0,18	-0,08	0,08	0,08	-0,04
X8	0,05	0,4	0,12	0,21	0,15	0	0,15	0,13	0,12	-0,34	0,02
X9	0,06	0,2	0,09	0,2	0,22	0,04	0,13	0,15	0,25	-0,17	0,06
X10	0,54	-0,21	0,19	0,07	0,36	0,03	0,03	-0,17	-0,16	0,02	0,01
X11	-0,01	0	0	0	-0,08	-0,03	-0,06	0,01	-0,09	0,02	-0,11
X12	0,18	0,11	0,01	0,04	-0,02	-0,06	0,04	0,04	-0,05	-0,46	-0,25
X13	0,73	0,2	-0,22	0	0,08	0,08	-0,2	0,1	0,09	0,05	-0,01
X14	0,87	0,22	-0,27	-0,01	0,04	0,06	-0,13	0,08	0,12	0,06	0,01
X15	0,58	0,02	-0,04	0,06	0,37	0	0	-0,2	-0,03	0,23	-0,06
X16	0,24	0,67	-0,21	-0,09	-0,31	-0,18	-0,03	-0,14	-0,34	-0,08	0
X17	0,85	0,29	-0,3	-0,02	0	0,03	-0,07	0	0,09	0,04	0,02
X18	0,86	0,23	-0,23	-0,01	0,01	0,06	-0,13	0,16	0,11	0	0
X19	0	0	-0,02	-0,04	-0,1	-0,01	0,06	0,1	0,01	0,07	0,09
X20	0,17	-0,01	0,09	-0,21	-0,19	-0,1	0,53	0,58	-0,15	0,15	0,02
X21	0,23	0	0,08	-0,15	-0,11	-0,09	0,57	0,54	-0,2	0,21	-0,02
X22	-0,08	0	-0,02	0,1	-0,1	-0,02	0,07	-0,17	0,11	0,09	0,07
X23	0,34	0	0,15	0,25	0,18	-0,13	0,3	-0,3	-0,4	0,1	-0,16
X24	0,02	0,02	0,01	0,03	0,03	-0,12	0,03	-0,03	-0,01	-0,16	-0,29
X25	0	-0,13	0,01	-0,01	0	0,01	-0,18	0,13	-0,28	-0,35	0,39
X26	-0,49	0,65	0,06	-0,03	0	0	0,24	-0,22	0,24	0	0,05
X27	0,49	-0,65	-0,06	0,03	0	0	-0,24	0,22	-0,24	0	-0,05
X28	0	0	-0,05	-0,13	-0,23	0,9	0,13	-0,1	-0,17	-0,05	-0,09
X29	-0,24	0,27	0,2	-0,77	0,3	0,04	-0,15	0,09	0,06	0,02	0,07
X30	-0,12	0,2	0,23	-0,66	0,37	-0,04	-0,15	-0,02	-0,22	-0,01	-0,2
X31	0,85	0,29	-0,3	-0,02	0,01	0,03	-0,07	-0,01	0,09	0,05	0,02
X32	-0,13	0,27	0	-0,5	-0,02	0,09	-0,11	0,04	0,22	0,04	0,29
X33	0,15	0,56	-0,15	-0,11	-0,45	-0,22	0	-0,11	-0,37	-0,05	0,02
X34	-0,13	0,11	-0,03	0,23	0,13	0,23	-0,33	0,15	-0,17	0,16	0
X35	0,19	0,59	-0,19	-0,05	-0,27	-0,13	-0,03	-0,14	-0,29	-0,1	0,02
X36	0,77	0,21	-0,24	-0,03	0,02	0,03	-0,08	0,01	0,05	0,03	-0,01
X37	-0,03	0,01	-0,02	-0,11	-0,19	0,92	0,17	-0,1	-0,12	-0,06	-0,09
X38	0,42	0,04	0,02	0,01	0,04	0,05	-0,11	0,21	0,2	-0,12	0
X39	-0,31	-0,27	-0,32	-0,25	-0,37	-0,17	-0,14	-0,29	0,03	0,2	-0,13
X40	0,25	-0,05	0,37	-0,07	-0,1	-0,03	0,07	-0,11	-0,03	-0,23	0,42
X41	-0,2	0,27	0,27	-0,62	0,44	-0,02	-0,08	0	-0,15	0	-0,17
X42	0,65	-0,06	0,22	0,12	0,41	0	0,15	-0,28	-0,16	0,12	0
X44	0,86	0,1	0,11	0,01	0,06	0,05	0,08	0,16	-0,04	-0,02	0,03
X45	-0,13	0,04	-0,15	-0,11	-0,16	-0,05	-0,09	0,11	0,08	0	-0,06
X46	0,05	-0,03	0,05	-0,02	-0,12	-0,1	-0,2	0,11	-0,05	-0,12	-0,11
X47	0,61	-0,51	-0,08	-0,39	-0,18	-0,08	0,16	-0,17	0,16	-0,08	-0,04
X48	-0,02	0,04	0,05	0,02	0,04	0	0,09	0,05	0,21	-0,37	-0,3
X49	0,05	0	0	0,04	0,02	-0,03	0,01	0,03	0,01	-0,33	-0,37
X51	0,23	-0,63	0	-0,06	0	0,01	-0,32	0,13	-0,35	-0,24	0,18

Figura 7 - Loadings

	Vetores Próprios (Wij)										
	i										
j	1	2	3	4	5	6	7	8	9	10	11
1	-0,27	-0,05	-0,04	0,03	0,1	-0,03	-0,05	0	0	0,01	-0,01
2	-0,19	0,24	0,05	-0,24	-0,12	0,06	-0,12	-0,13	0,13	-0,07	0,04
3	-0,19	0,24	0,05	-0,24	-0,12	0,06	-0,12	-0,13	0,13	-0,07	0,04
4	-0,15	-0,03	-0,31	0	-0,07	0	-0,09	-0,21	0,03	-0,09	-0,23
5	-0,14	-0,05	-0,44	0,03	-0,23	0	0,13	0	0,06	0,08	0,05
6	-0,06	-0,03	-0,32	0,04	-0,25	-0,01	0,25	0,17	0,06	0,18	0,26
7	-0,13	-0,05	-0,44	0,04	-0,22	0	0,14	-0,06	0,06	0,07	0,04
8	-0,01	-0,19	-0,07	0,13	0,1	0	-0,11	0,1	0,09	-0,3	-0,02
9	-0,02	-0,1	-0,05	0,12	0,14	-0,02	-0,1	0,11	0,2	-0,15	-0,06
10	-0,17	0,1	-0,11	0,04	0,24	-0,02	-0,02	-0,13	-0,13	0,02	-0,01
11	0	0	0	0	-0,05	0,02	0,04	0,01	-0,07	0,02	0,1
12	-0,05	-0,05	-0,01	0,02	-0,01	0,04	-0,03	0,03	-0,04	-0,4	0,22
13	-0,23	-0,09	0,13	0	0,05	-0,05	0,14	0,07	0,07	0,04	0,01
14	-0,27	-0,11	0,15	0	0,02	-0,04	0,1	0,06	0,09	0,05	-0,01
15	-0,18	-0,01	0,02	0,04	0,24	0	0	-0,16	-0,02	0,2	0,06
16	-0,07	-0,32	0,12	-0,05	-0,2	0,13	0,02	-0,1	-0,27	-0,07	0
17	-0,27	-0,14	0,17	-0,01	0	-0,02	0,05	0	0,07	0,04	-0,02
18	-0,27	-0,11	0,13	0	0	-0,04	0,09	0,12	0,09	0	0
19	0	0	0,01	-0,02	-0,07	0,01	-0,04	0,08	0,01	0,06	-0,08
20	-0,05	0	-0,05	-0,13	-0,12	0,07	-0,4	0,44	-0,12	0,13	-0,01
21	-0,07	0	-0,04	-0,09	-0,07	0,06	-0,43	0,41	-0,16	0,18	0,02
22	0,02	0	0,01	0,06	-0,06	0,01	-0,05	-0,13	0,09	0,08	-0,06
23	-0,1	0	-0,09	0,15	0,12	0,09	-0,22	-0,23	-0,32	0,09	0,15
24	0	-0,01	0	0,02	0,02	0,08	-0,02	-0,02	0	-0,14	0,26
25	0	0,06	0	0	0	0	0,13	0,1	-0,23	-0,31	-0,36
26	0,15	-0,31	-0,03	-0,02	0	0	-0,18	-0,16	0,19	0	-0,04
27	-0,15	0,31	0,03	0,02	0	0	0,18	0,16	-0,19	0	0,04
28	0	0	0,03	-0,08	-0,15	-0,63	-0,1	-0,08	-0,14	-0,04	0,08
29	0,07	-0,13	-0,11	-0,47	0,2	-0,03	0,11	0,07	0,05	0,02	-0,07
30	0,03	-0,09	-0,13	-0,4	0,24	0,03	0,11	-0,01	-0,18	-0,01	0,18
31	-0,27	-0,14	0,17	-0,01	0	-0,02	0,05	-0,01	0,07	0,04	-0,02
32	0,04	-0,13	0	-0,31	-0,01	-0,06	0,08	0,03	0,18	0,04	-0,26
33	-0,04	-0,27	0,09	-0,07	-0,3	0,16	0	-0,08	-0,3	-0,04	-0,02
34	0,04	-0,05	0,02	0,14	0,09	-0,16	0,24	0,11	-0,14	0,14	0
35	-0,06	-0,28	0,11	-0,03	-0,18	0,09	0,02	-0,11	-0,24	-0,09	-0,01
36	-0,24	-0,1	0,13	-0,02	0,01	-0,02	0,06	0,01	0,04	0,02	0,01
37	0,01	0	0,01	-0,07	-0,12	-0,65	-0,12	-0,08	-0,1	-0,05	0,08
38	-0,13	-0,02	-0,01	0	0,03	-0,04	0,08	0,16	0,16	-0,1	0
39	0,1	0,13	0,18	-0,15	-0,24	0,12	0,1	-0,22	0,02	0,17	0,12
40	-0,08	0,02	-0,21	-0,04	-0,06	0,02	-0,05	-0,09	-0,03	-0,2	-0,38
41	0,06	-0,13	-0,16	-0,38	0,29	0,02	0,06	0	-0,12	0	0,15
42	-0,2	0,03	-0,12	0,07	0,27	0	-0,11	-0,21	-0,13	0,11	0
43	-0,27	-0,05	-0,06	0	0,04	-0,03	-0,06	0,12	-0,03	-0,01	-0,02
44	0,04	-0,02	0,09	-0,06	-0,1	0,03	0,06	0,08	0,06	0	0,06
45	-0,01	0,01	-0,02	-0,01	-0,08	0,07	0,15	0,08	-0,04	-0,11	0,1
46	-0,19	0,24	0,05	-0,24	-0,12	0,06	-0,12	-0,13	0,13	-0,07	0,04
47	0	-0,02	-0,03	0,01	0,02	0	-0,07	0,04	0,17	-0,32	0,27
48	-0,01	0	0	0,02	0,01	0,02	-0,01	0,02	0,01	-0,29	0,34
49	-0,07	0,3	0	-0,04	0	0	0,24	0,1	-0,28	-0,21	-0,16

Figura 8 - Pesos ou Vetores Próprios

```
Call:
glm(formula = X0 ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 +
    X10 + X11 + X12 + X13 + X14 + X15 + X16 + X17 + X18 + X19 +
    X20 + X21 + X22 + X23 + X24 + X25 + X26 + X27 + X28 + X29 +
    X30 + X31 + X32 + X33 + X34 + X35 + X36 + X37 + X38 + X39 +
    X40 + X41 + X42 + X43 + X44 + X45 + X46 + X47 + X48 + X49 +
    X50 + X51, family = binomial(link = logit), data = dados)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.19289  -0.15813  -0.03766  -0.00602   3.15946

Coefficients: (2 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.756e+02  1.496e+03  -0.585  0.55830
X1          -1.027e+01  8.007e+00  -1.282  0.19968
X2           7.795e+04  8.721e+04   0.894  0.37142
X3           7.592e+00  5.999e+02   0.013  0.98990
X4           4.413e+03  2.820e+03   1.565  0.11756
X5          -3.597e+03  2.738e+03  -1.314  0.18888
X6           1.830e+03  1.330e+03   1.376  0.16872
X7           2.899e+01  1.072e+03   0.027  0.97843
X8           1.021e-10  5.966e-11   1.711  0.08716 .
X9           7.556e-11  6.948e-11   1.087  0.27684
X10         -6.657e+01  4.345e+01  -1.532  0.12547
X11         -6.541e-10  3.930e-09  -0.166  0.86782
X12          3.228e+00  1.541e+00   2.094  0.03622 *
X13         -1.351e+01  1.369e+01  -0.986  0.32393
X14         -4.363e+01  3.955e+01  -1.103  0.27003
X15         -4.343e+01  3.849e+01  -1.128  0.25915
X16         -2.386e+00  1.349e+01  -0.177  0.85955
X17         -1.953e+02  2.703e+02  -0.723  0.46998
X18          4.913e+00  3.580e+01   0.137  0.89085
X19          2.824e+01  3.055e+01   0.924  0.35533
X20         -2.355e+02  1.320e+02  -1.784  0.07446 .
X21         -6.455e+00  2.300e+01  -0.281  0.77898
X22          1.502e-09  7.361e-07   0.002  0.99837
X23          3.059e+01  1.697e+01   1.803  0.07134 .
X24         -8.152e+00  1.768e+01  -0.461  0.64478
X25          7.899e-09  3.585e-06   0.002  0.99824
X26          2.595e+01  1.114e+01   2.329  0.01987 *
X27                 NA         NA      NA       NA
X28         -1.082e+02  6.404e+01  -1.689  0.09115 .
X29         -1.139e+01  4.321e+01  -0.263  0.79217
X30          4.914e+02  3.261e+02   1.507  0.13188
X31          2.192e+02  2.742e+02   0.799  0.42411
X32         -1.206e+01  2.966e+01  -0.406  0.68446
X33         -1.970e+01  6.068e+00  -3.246  0.00117 **
X34         -2.129e-08  5.035e-06  -0.004  0.99663
X35          3.529e+01  7.933e+00   4.449 8.63e-06 ***
X36          1.326e+01  1.793e+01   0.740  0.45938
X37          3.449e-08  1.887e-06   0.018  0.98542
X38          3.137e+02  9.941e+02   0.316  0.75232
X39          2.249e+01  8.397e+00   2.678  0.00740 **
X40         -1.417e+03  2.234e+03  -0.634  0.52604
X41         -2.897e+00  1.620e+01  -0.179  0.85806
X42         -3.302e+00  1.220e+01  -0.271  0.78664
X43         -1.273e+00  2.149e+03  -0.001  0.99953
X44          1.829e+01  1.170e+01   1.564  0.11788
X45          3.578e-10  4.948e-10   0.723  0.46958
X46         -9.556e+00  3.250e+01  -0.294  0.76875
X47         -7.796e+04  8.719e+04  -0.894  0.37127
X48         -2.270e+01  4.652e+01  -0.488  0.62555
X49         -1.075e+01  2.827e+01  -0.380  0.70369
X50                 NA         NA      NA       NA
X51         -3.229e+01  2.644e+01  -1.221  0.22201
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 650.17  on 999  degrees of freedom
Residual deviance: 238.12  on 950  degrees of freedom
AIC: 338.12

Number of Fisher Scoring iterations: 17
```

Figura 9 - Teste de Significância Individual

	2.5 %	97.5 %
(Intercept)	-4.029388e+03	5.405406e+01
X1	-1.528332e+01	4.751787e+00
X2	1.115225e+05	-4.319739e+02
X3	2.419938e+02	-2.415966e+02
X4	5.927472e+03	2.018573e+03
X5	-9.009588e+03	-1.870669e+03
X6	8.766669e+02	4.453200e+03
X7	3.879186e+02	-8.557843e+02
X8	1.526619e-10	5.257357e-11
X9	2.960372e-11	2.122153e-10
X10	-1.634210e+02	-5.242239e+01
X11	-2.912805e-09	4.155179e-09
X12	1.336359e-01	4.197885e+00
X13	-5.033648e-01	4.884856e+00
X14	-1.158854e+02	-3.182945e+01
X15	-1.137725e+02	-3.155640e+01
X16	2.519257e+00	-7.292311e+00
X17	-2.534244e+02	1.756256e+02
X18	-3.215483e+00	6.615870e+01
X19	7.253958e+00	9.244693e+01
X20	-3.170466e+02	-1.138550e+01
X21	-1.422264e+01	4.255531e+01
X22	-9.325430e-09	1.258038e-08
X23	-6.190131e+00	3.615242e+01
X24	-2.039413e+00	-1.444371e+01
X25	-1.066770e-05	-1.472483e-06
X26	3.909980e+00	2.851376e+01
X27	NA	NA
X28	-1.443975e+02	1.599553e+00
X29	-2.243088e+01	6.557493e+01
X30	4.325573e+02	1.121550e+03
X31	-1.517793e+02	2.808460e+02
X32	-6.593884e+01	7.431459e+00
X33	-2.063433e+01	-8.652114e+00
X34	-1.342414e-05	-4.281031e-07
X35	2.042039e+01	3.966677e+01
X36	-2.212944e+01	2.542435e+01
X37	2.564180e-07	6.033582e-07
X38	6.498815e+02	-5.051853e+02
X39	2.637505e+01	1.534876e+01
X40	-2.868631e+03	2.810230e+03
X41	-3.193174e+01	9.725142e-01
X42	-1.088560e+01	2.036710e+01
X43	1.440078e+03	4.292458e+03
X44	-5.677662e+00	2.519374e+01
X45	-4.793623e-10	4.312438e-10
X46	1.330478e+01	-3.418312e+01
X47	-5.514643e+03	-1.173869e+05
X48	9.032724e-01	-6.384233e+01
X49	3.343090e+00	-3.302512e+01
X50	NA	NA
X51	-9.309786e+00	-5.427726e+01

Figura 10 – Intervalos de Confiança a 95%

```
> varImp(fitControl)
glm variable importance

  only 20 most important variables shown (out of 49)

    Overall
X35  100.00
X33   72.97
X39   60.20
X26   52.34
X12   47.07
X23   40.53
X20   40.09
X8    38.44
X28   37.96
X4    35.17
X44   35.14
X10   34.43
X30   33.86
X6    30.93
X5    29.52
X1    28.82
X51   27.44
X15   25.35
X14   24.78
X9    24.43
```

Figura 11 – Importância das Variáveis na regressão

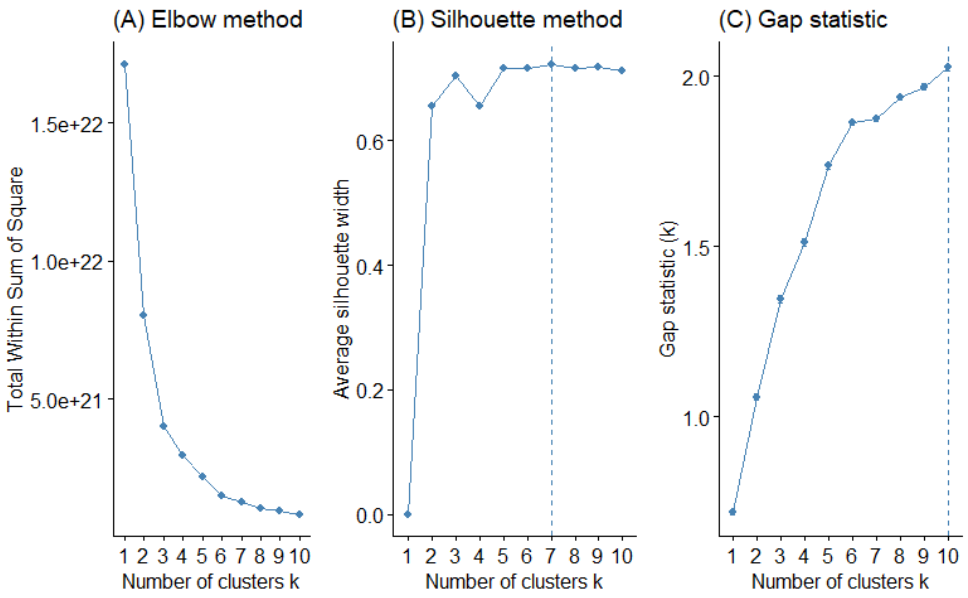

Figura 12 – Métodos estatísticos de escolha do número de clusters

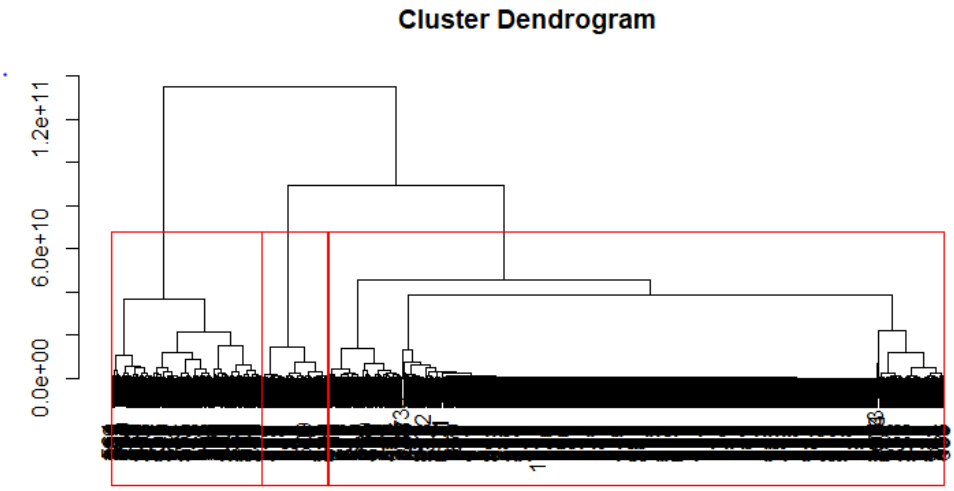

Figura 13 – Dendrograma dos 3 Clusters