Relatório de Pré-Processamento

Lia Sucupira Furtado

Dezembro 2018

1 Introdução

Devida a grande quantidade de dados em um dataset,é importante explorar e analisar os dados para compreender melhor o que eles representam. Muitas vezes, podem ocorrer incoerências, como o aparecimento de outliers ou dados redundantes que atrapalham o processamento e podem originar em um modelo preditivo sem qualidade.Por isso aqui discutimos várias técnicas de pré-processamento dos dados, para que eles fiquem aptos para serem usados nos mais diversos modelos preditivos.

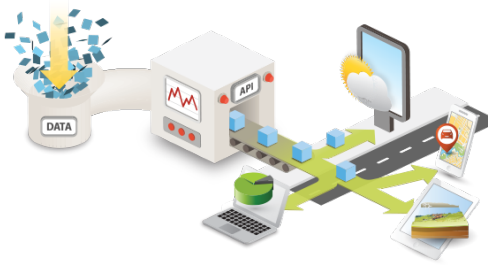


Figura 1: Ilustração de pré-processamento

2 Metodologia

Análise exploratória dos dados

Inicialmente, o primeiro passo para se familiarizar com os dados é realizar uma análise inicial com as técnicas de estatísticas. Ou seja, deve-se calcular a média, desvio padrão, análise inter-quartis, obliquidade. Além de ser importante plotar o histograma para ver a distribuição dos dados.

A análise exploratória de dados emprega grande variedade de técnicas gráficas e quantitativas, visando maximizar a obtenção de informações ocultas na sua

estrutura, descobrir variáveis importantes em suas tendências, detectar comportamentos anômalos do fenômeno, testar se são válidas as hipóteses assumidas, escolher modelos e determinar o número ótimo de variáveis.

Redimensionamento dos dados

Este passo é muito importante quando se lida com parâmetros de diferentes unidades e escalas.Principalmente quando deve-se comparar valores e para isso eles precisam ter a mesma escala para acarretar bons resultados.

Essas manipulações são geralmente usadas para melhorar a estabilidade numérica do modelo.

Normalização

A normalização redimensiona os valores dos atributos de seus intervalos originais em um intervalo de [0,1]. Isso pode ser útil em alguns casos em que todos os parâmetros precisam ter a mesma escala positiva. No entanto, os outliers do conjunto de dados são perdidos.

$$\boxed{x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}} \tag{1}$$

Padronização

A padronização redimensiona os dados para ter uma média (μ) de 0 e desvio padrão (σ) de 1 (variação unitária). Para padronizar os dados faz-se a centralização, em que o valor médio do preditor é subtraído de todos os valores. Como resultado dessa centralização, o preditor tem uma média zero.

E também faz-se o escalonamento em que cada valor da variável preditora é dividido por seu desvio padrão. Escalar os dados é forçar os valores a terem um desvio padrão comum de um.

Assim, nossa equação é:

$$\boxed{x_{new} = \frac{x - \mu}{\sigma}} \tag{2}$$

Para a maioria das aplicações, a padronização é recomendada.Pois os dados não estão limitados, ao contrário da normalização.

É especialmente valiosa para os métodos que calculam distâncias entre atributos. Por exemplo, um método como o k-vizinhos mais próximos tende a dar mais importância para os atributos que possuem um intervalo maior de valores.

A única desvantagem real dessas transformações é a perda de interpretabilidade dos valores individuais, uma vez que os dados não estão mais nas unidades originais. É importante padronizar os dados de teste da mesma forma que se padroniza os dados de treino.

Outliers

Outliers são dados cujos valores são muito diferentes em relação aos demais ou que estão fora dos intervalos aceitáveis do conjunto de dados. As anomalias dos outliers podem enviesar o resultado e consequentemente fazer com que ele apresente distorções.

Detectando outliers

Por meio de gráficos

Um dos métodos mais simples para detectar outliers é o uso de gráficos de caixa. Um box plot é uma exibição gráfica para descrever a distribuição dos dados. Box plots usam a mediana e os quartis inferior e superior.

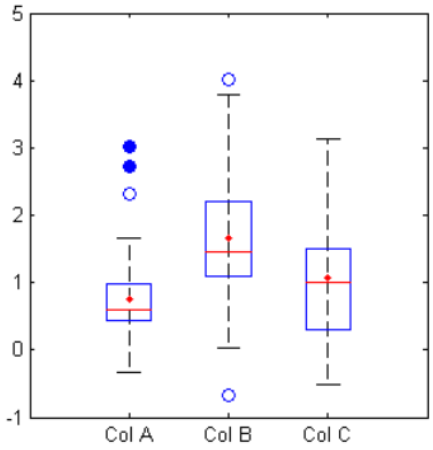


Figura 2: Gráfico Box Plot

O método de Tukey define um outlier como aqueles valores do conjunto de dados que estão distantes do ponto central, que é mediana.

A distância máxima até o centro dos dados que serão permitidos é chamada de parâmetro de limpeza,caso um dado esteja fora desse alcance ele será classificado como um outlier. Se o parâmetro de limpeza for muito grande, o teste se tornará menos sensível aos outliers. Pelo contrário, se for muito pequeno, muitos valores serão detectados como outliers.

Esse método funciona bem quando tem-se valores de anomalias extremas e podem ser facilmente detectadas.

Por meio de funções matemáticas

A intuição por trás do Z-score é descrever qualquer ponto de dados encontrando sua relação com o Desvio Padrão e a Média do grupo de pontos de dados.

O valor do Z-score representa o numero de desvios padrões que estão acima ou abaixo da média da população.

$$\boxed{Z = \frac{x - \mu}{\sigma}} \tag{3}$$

O Z-score utiliza a distribuição normal como o 'ideal' então devemos ter para os dados uma média é 0 e um desvio padrão 1.

Na maioria dos casos, um limite de 3 ou -3 é usado, ou seja, se o valor da pontuação Z for maior ou menor que 3 ou -3, respectivamente, esse ponto de dados será identificado como outliers.

Removendo outliers

Pode-se remover os outliers definindo um ponto de corte dos outliers, tanto usando o escore Z como o método de Tukey. Para uma faixa escolhida será feito a retirada dos dados que foram detectados como outliers.

Mas, deve-se tomar muito cuidado para não remover os valores apressadamente, especialmente se o tamanho da amostra for pequeno. Pois essa anomalia pode ter uma relação direta com o conjunto de dados, especialmente se existir uma alta correlação, e pode-se perder informações importantes.

Transformações para resolver outliers

Ao invés de simplesmente remover os valores discrepantes dos dados, considera-se o conjunto de valores discrepantes e altera seus valores para algo mais representativo no conjunto de dados. É uma pequena distinção, mas importante: quando você recorta os dados, os valores extremos são descartados. Quando você altera os valores, os valores extremos são substituídos por certos percentis (o mínimo e o máximo aparados).

Métodos comuns para fazer a **imputação** dos valores são o uso da média de uma variável ou o uso de um modelo de regressão para prever o valor ausente.

Outra forma de resolver os outliers é **reduzindo seu impacto no modelo**. O erro de Minkowski é um índice de perda que é mais insensível a valores discrepantes do que o erro de soma quadrática padrão. Isso reduz a contribuição de outliers para o erro total.

Então é válido calcular o erro do modelo com essa medida.

Se um modelo é considerado sensível a outliers, uma transformação de dados que pode minimizar o problema é o **spatial sign** (Serneels et al. 2006). Este procedimento projeta os valores do preditor em uma esfera multidimensional. Isto tem o efeito de fazer todas as amostras a mesma distância do centro da esfera. Matematicamente, cada amostra é dividida por sua norma quadrada, ou seja as amostras são normalizadas. Dessa forma o conjunto de dados não terá mais valores discrepantes.

$$x_{ij}^{*} = \frac{x_{ij}}{\sum_{j=1}^{P} x_{ij}^{2}} \tag{4}$$

Como o denominador destina-se a medir a distância ao quadrado da distribuição do preditor, é importante centralizar e dimensionar os dados do preditor antes de usar essa transformação. Observe que, ao contrário de centralização ou dimensionamento, essa manipulação dos preditores os transforma em um grupo.

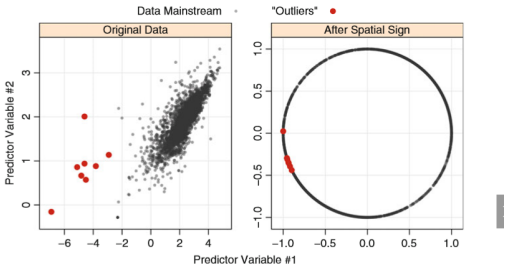


Figura 3: Formula do spatial sign

Lidando com valores em falta

Em muitos casos, alguns preditores têm valores em falta para uma determinada amostra.Podem ter vários motivos para a ausência desses dados, como porque o valor falha na extração do dado ou não foi determinado no momento de construção do modelo.

Descartar os dados

Para grandes conjuntos de dados, a remoção de amostras com base em valores omissos não é um problema, supondo-se que a falta de dados não seja informativa.

Em alguns casos, a porcentagem de dados ausentes é substancial o suficiente para remover esse preditor das atividades de modelagem subsequentes.

Em conjuntos de dados menores, há um preço alto na remoção de amostras; algumas das abordagens alternativas descritas abaixo podem ser mais apropriadas.

Preencher com novos valores

Dessa forma,dados perdidos podem ser imputados. Nesse caso, podemos usar as informações dos preditores do conjunto de treinamento para, em essência, estimar os valores de outros preditores. Isso equivale a um modelo preditivo dentro de um modelo preditivo.

Muitas técnicas têm sido aplicadas, sendo algumas delas bastante simples, como a substituição dos valores desconhecidos pela média ou moda do atributo.

De uma forma geral, deve-se evitar utilizar alguns métodos mais simples, como a imputação pela média ou moda, por serem métodos que podem distorcer os dados

Outras técnicas mais elaboradas podem ser implementadas e avaliadas experimentalmente. Por exemplo, pode-se substituir os valores desconhecidos por valores preditos utilizando um algoritmo de aprendizado.

Segundo Batista (2), o método de imputação com base no algoritmo K-vizinhos mais próximos obteve resultados que foram, na maioria das vezes, superiores aos demais métodos analisados.

Um limitação é que normalmente os métodos predizem valores mais bem comportados do que os valores reais (não conhecidos) seriam. Dessa forma, os classificadores induzidos tendem a se tornar mais simples quanto maior for a quantidade de valores desconhecidos tratados. Esse fato pode levar ao risco de simplificar excessivamente o problema que esta sendo estudado.

Outra alternativa para lidar com os valores faltante é verificar se não existe um outro atributo com informações similares, isto é, alta correlação, no conjunto de dados.

Resolvendo a obliquidade

Uma conjunto de dados que tem uma distribuição irregular é oblíqua.Isso significa que a probabilidade de um lado da distribuição fica muito maior do que o outro. O objetivo é ter uma distribuição simétrica com uma obliquidade próxima de zero. A obliquidade é calculada com as equações abaixo:

$$obliquidade = \frac{\sum(x_i - \bar{x})^3}{(n-1)v^{3/2}} \tag{5}$$

onde

$$v = \frac{\sum(x_i - \bar{x})^2}{(n-1)} \tag{6}$$

Substituir os dados pelo log, raiz quadrada ou inversa pode ajudar a remover o desvio.

Box e Cox (1964) propõem uma família de transformações

$$x^* = \begin{cases} \frac{x^\lambda - 1}{\lambda} & se\ \lambda \neq 0 \\ log(x) & se\ \lambda = 0 \end{cases} \tag{7}$$

Além da transformação logarítmica, esta família pode identificar a transformação quadrada ($\lambda = 2$), raiz quadrada (λ= 0 .5), inversa ($\lambda = -1$) e outras entre si. Box e Cox (1964) mostram como usar a estimação por máxima verossimilhança para determinar o parâmetro de transformação (λ). Esse procedimento seria aplicado independentemente a cada dado do preditor que contenha valores maiores que zero.

Correlação

Primeiro, menos preditores significam menor tempo computacional e complexidade. Se dois preditores são altamente correlacionados, isso implica que eles estão medindo a mesma informação subjacente. A remoção de um não deve comprometer o desempenho do modelo e pode levar a um modelo mais parcimonioso e interpretável.

Para visualizar a correlação dos dados é utilizada a matriz de correlação. Cada correlação pareada é calculada a partir dos dados de treinamento e colorida de acordo com sua magnitude. Esta visualização é simétrica: as diagonais superior e inferior mostram informações idênticas. As cores azul-escuras indicam fortes correlações positivas, vermelho-escuro é usado para fortes correlações negativas e branco não implica relação empírica entre os preditores.

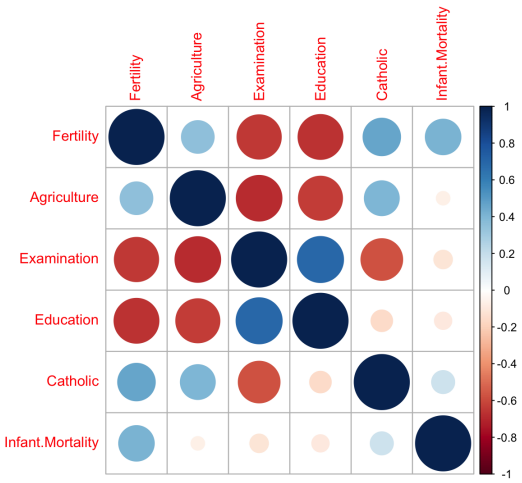


Figura 4: Matriz de correlação

Resolvendo problemas de alta correlação

A multicolinearidade forte é problemática porque pode aumentar a variância dos coeficientes de regressão, tornando-os instáveis.

Removendo os preditores

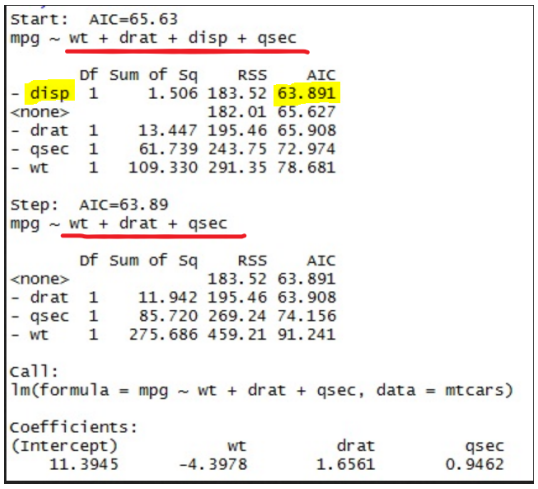
Se dois preditores são altamente correlacionados, isso implica que eles estão fornecendo informações redundantes. A remoção de um não deve comprometer o desempenho do modelo, normalmente não reduz drasticamente o R^2, e pode levar a um modelo mais interpretável.

Para escolher quais variáveis remover pode-se pensar um métodos como a regressão stepwise.

Os métodos stepwise avaliam vários modelos usando procedimentos que adicionam e / ou removem preditores para identificar um subconjunto útil de preditores e encontrar a combinação ideal que maximiza o desempenho do modelo. Os preditores são avaliados (um de cada vez) no atual modelo de regressão linear. Dessa forma, este método só funciona com modelos que a saída já é conhecida, pois com base nela o modelo de regressão linear em cada preditor será ajustado. Na forward selection, o modelo inicia sem nenhum preditor e vai sendo adicionada confirme o valor de t-value aumenta e o valor de AIC diminui.Além disso, para verificar se cada um dos preditores recém-adicionados é estatisticamente significativo se utiliza o p-value. Ou seja, se um preditor tiver um valor-p abaixo do limiar pré-definido esse preditor poderá ser adicionado ao modelo pois ele terá um papel significativo.

Na backward selection, o modelo inicial contém todos os preditores de P, que são então removidos iterativamente para determinar quais não estão contribuindo significativamente para o modelo.

Na foto abaixo podemos ver que todos os preditores são inicialmente adicionadas e depois o preditor de menor AIC é removido do modelo.



```
Start:  AIC=65.63
mpg ~ wt + drat + disp + qsec

       Df Sum of Sq    RSS    AIC
- disp  1     1.506 183.52 63.891
<none>              182.01 65.627
- drat  1    13.447 195.46 65.908
- qsec  1    61.739 243.75 72.974
- wt    1   109.330 291.35 78.681

Step:  AIC=63.89
mpg ~ wt + drat + qsec

       Df Sum of Sq    RSS    AIC
<none>              183.52 63.891
- drat  1    11.942 195.46 63.908
- qsec  1    85.720 269.24 74.156
- wt    1   275.686 459.21 91.241

Call:
lm(formula = mpg ~ wt + drat + qsec, data = mtcars)

Coefficients:
(Intercept)           wt         drat         qsec
    11.3945      -4.3978       1.6561       0.9462
```

Figura 5: Backward Selection

Reduzindo o número de preditores

Quando os preditores são muito correlacionados, isso pode acarretar em redundâncias no conjunto de dados e prejudicar a predição do modelo. Ao perceber essa correlação pode-se tomar medidas para diminui-la.Dessa forma, é aplicado a Análise de componentes principais que é uma técnica não supervisionada, pois não considera a variável de resposta ao resumir a variabilidade.

Este método procura encontrar combinações lineares dos preditores, conhecidos como componentes principais (PCs), que capturam a maior variância

possível. O primeiro componente principal é o que representa melhor os dados pois é responsável pela maior quantidade de informação, e uma maior variância. A forma que ele é calculado é demonstrado abaixo:

$$\boxed{Y_1 = a_{11}X_1 + a_{12}X_2 + a_{1q}X_q} \tag{8}$$

Para escolher a quantidade de PCs ideal que irá representar o conjunto de dados é feito a cross-validation.

As principal vantagem do PCA é que além dele reduzir a dimensionalidade do conjunto de dados, ele cria componentes não correlacionados.Além disso, a PCA caracteriza os preditores associados a cada componente, o que é possível já que cada componente é combinação linear dos preditores e o coeficiente de cada preditor é chamado de peso.

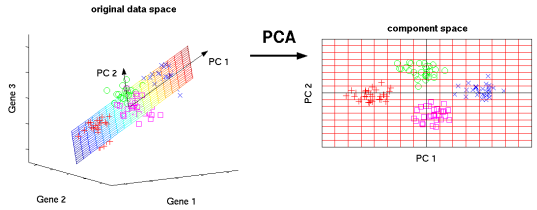


Figura 6: Visualização dos pesos e componentes principais de um conjunto de dados

Antes de se aplicar a PCA deve-se transformar primeiro os preditores com alta obliquidade e depois centralizar e dimensionar. A centralização e o dimensionamento permitem que a PCA encontre os relacionamentos subjacentes nos dados sem ser influenciada pelas escalas de medidas originais.

Alternativamente, temos outro modelos que reduz a dimensionalidade, o PLS. O modelo de mínimos quadrados parciais (PLS) é essencialmente uma versão supervisionadas da análise de componentes principais (PCA). O PLS encontra combinações lineares dos preditores. O que ele faz de diferente da PCA é que a PLS é escolhida para resumir ao máximo a covariância com a resposta. Isso significa que o PLS encontra componentes que resumem ao máximo a variação dos preditores e, ao mesmo tempo, exigem que esses componentes tenham correlação máxima com a resposta. O PLS, portanto, estabelece um compromisso entre os objetivos da redução da dimensão do espaço do preditor e um relacionamento preditivo com a resposta. No gráfico abaixo, vemos que é utilizado a saída Y e com as observações de X é calculada as componentes t1 e t2.

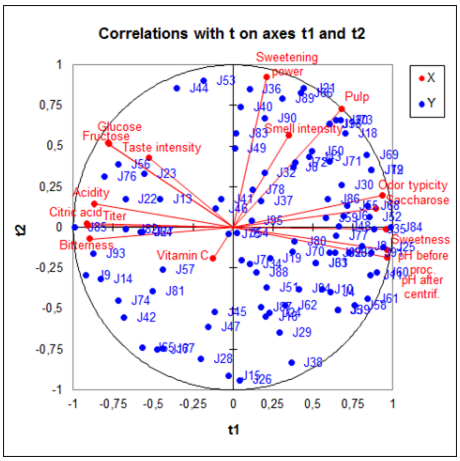


Figura 7: Visualização dos pesos e componentes principais de um conjunto de dados

A validação cruzada é usada para determinar o número ideal de componentes PLS a reter que minimizam o RMSE.

Além disso, a diferença principal entre esses modelos é que a redução da dimensão supervisionada encontra um RMSE mínimo com componentes significativamente menores do que a redução de dimensão não supervisionada.

Ou seja embora a capacidade preditiva desses modelos seja próxima, o PLS encontra um modelo mais simples que utiliza muito menos componentes do que o PCA.

ANOVA

Análise da Variância (ANOVA) é um método para testar a igualdade de três ou mais médias populacionais, baseado na análise das variâncias amostrais.

Para realizar a análise existe os seguintes pressupostos básicos:

- As amostras são aleatórias e independentes
- As populações têm distribuição normal
- As variâncias populacionais são iguais

Dessa forma, nomeamos estes pressupostos de hipótese inicial e ao longo da análise de variância pode-se mostrar que eles são aceitos ou rejeitados.

Caso a hipótese seja rejeitada isso significa que existe pelo menos uma diferença entre os grupos.

A ANOVA é baseada em estimativas de variância.A variação total é separada em duas fontes diferentes de variação:a variação intra-grupo, em que se analisa a variação somente daquele grupo e a variação inter grupos, que calcula a variação entre todos os grupos.

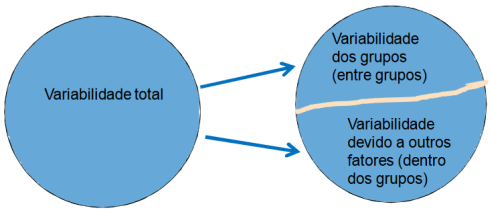


Figura 8: Fontes de variação

Para calcular essas variações é utilizado as equações abaixo.
Estimativa do desvio-padrão Entre-Grupos (s_E):

$$s_E^2 = \frac{\sum_{j=1}^{k} n_j(\bar{X}_j - \bar{\bar{X}})^2}{k-1} \tag{9}$$

Estimativa do desvio-padrão Intra-Grupo (s_E):

$$s_D^2 = \frac{\sum_{j=1}^{k}(n_i - 1)s_i^2}{n-k} \tag{10}$$

A estatística F mede se a variância inter-grupos é maior do que a variância intra-grupos, caso ela for podemos dizer que existe uma diferença entre as médias globais. Ou seja,quanto maior o F for maior a probabilidade da Hipótese inicial ser rejeitada. Podemos ver abaixo como é calculado esse valor de F.

$$F = \frac{s_E^2}{s_D^2}\begin{pmatrix} Variância\ entre\ amostras \\ Variância\ dentro\ das\ amostras \end{pmatrix} \tag{11}$$

Utilizamos um F crítico para comparar o valor de F obtido no nosso conjunto de dados, e ver se ele está dentro da faixa esperada. Para calcular o F crítico usamos os graus de liberdade das variâncias.

O df da variância entre grupos é:

- $df_E = n - k$

para qual n é a soma do número de elementos de todas as amostras e k quantidade de amostras de cada grupo.

O df da variância intra-grupos é:

- $df_D = k - 1$

Assim,utilizamos esses valores na tabela de distribuição F para descobrir qual é o valor correspondente de F crítico.

Se o valor de F for menor do que o F crítico, não conseguimos rejeitar a hipótese nula, ou seja, podemos afirmar que não existe uma diferença significativa entre os grupos já que os valores de suas médias são muito similares.

Como forma de resumir todas as informações que são calculadas na ANOVA é construída uma tabela com os dados.Ela tem a estrutura a seguir:

Fonte de variação	Soma quadrada	Grau de liberdade	Média quadrática	F
Entre-Grupos	SQ_E	k - 1	$S_E^2 = SQ_E$ / k - 1	S_E^2/S_D^2
Intra-Grupo	SQ_D	n - k	$S_D^2 = SQ_D$ / n - k	S_E^2/S_D^2
Total	SQ_{total}	n - 1		

Tabela 1: Tabela da Analise de Variância (ANOVA).

Outra forma de determinar se alguma das diferenças entre as médias é estatisticamente significativa é ao invés de calcular a F critica, calcular-se o P-value. Esse valor também é obtido da tabela de distribuição F, pois se refere a área a direita do valor F. Ao comparar o valor de p com o seu nível de significância pode-se avaliar a hipótese inicial.

Ao ter esse valor p, podemos analisa-lo da seguinte forma:

- Se o valor p for menor ou igual ao nível de significância (α),é rejeitado a hipótese inicial e concluído que nem todas as médias populacionais são iguais, pois existe uma diferença entre os conjuntos que estão sendo testados. As diferenças entre algumas das médias são estatisticamente significantes.

- Se o valor de p for maior do que o nível de significância Se o valor p for maior que o nível de significância (α),não tem-se evidência suficiente para rejeitar a hipótese nula, então a diferença entre as médias não são estatisticamente diferentes e podem ser consideradas iguais.

Normalmente, um nível de significância (denotado como α) de 0,05 funciona bem. Um nível de significância de 0,05 indica um risco de 5 % de concluir que existe uma diferença quando não há diferença real.

Referências

[1] M. Kuhn and K. Johnson. *Applied Predictive Modeling* , 2014.

[2] Pré-processamento de dados em aprendizado de máquina supervisionado. Batista, 2003