Mineração de Dados Aplicada à Engenharia

Profº - Dr. Thales Levi Azevedo Valente

thales.l.a.valente@gmail.com.br

Governo Federal

Sejam Bem-vindos !

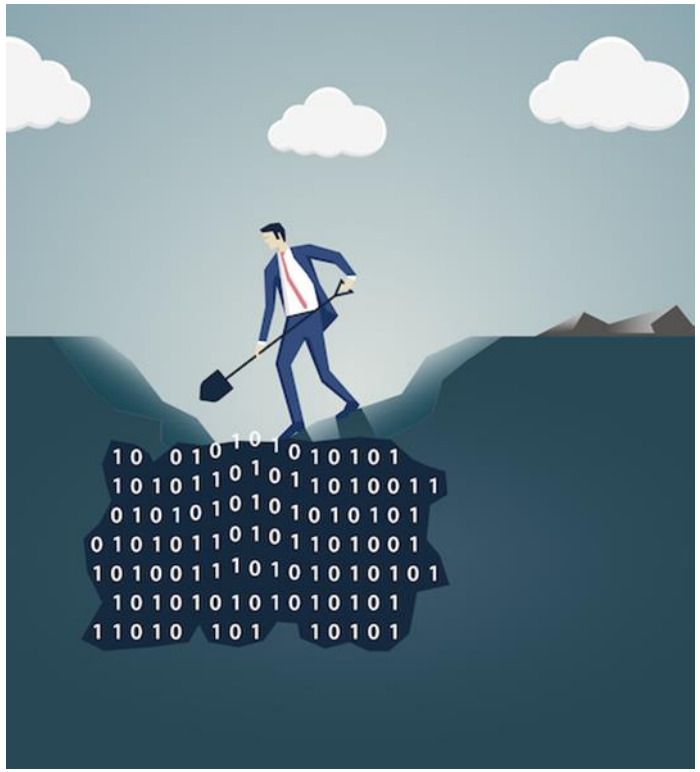
Os celulares devem ficar no silencioso ou desligados

Pode ser utilizado apenas em caso de emergência

Boa tarde/noite, por favor e com licença DEVEM ser usados

Educação é essencial

Na aula de anterior...



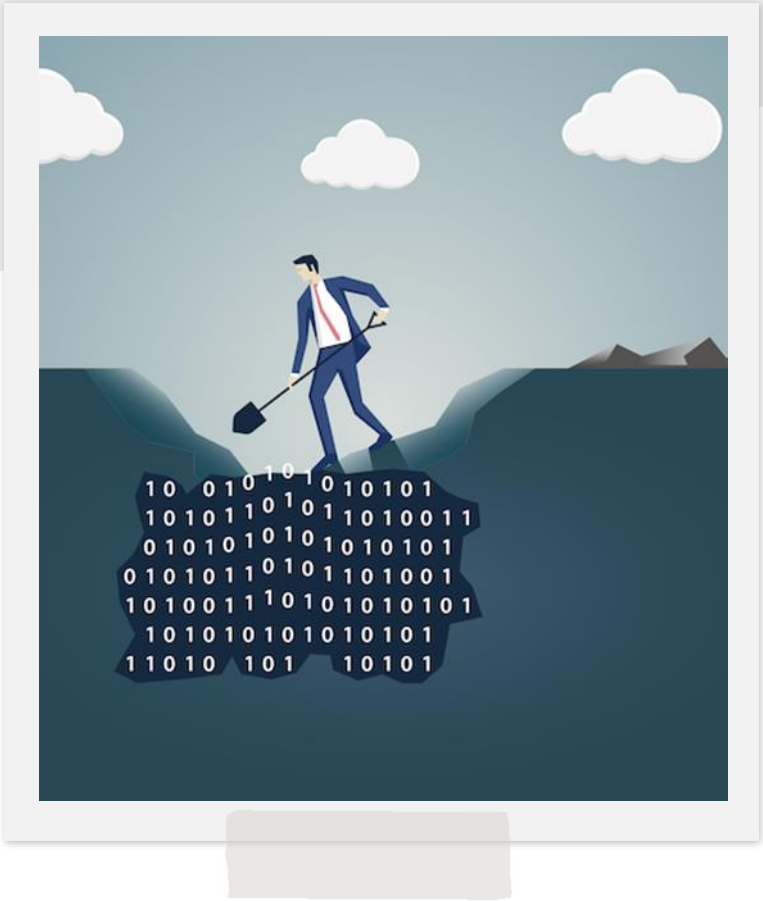
- Dados
- Escalas
- Cardinalidade
- Atividade
- Integração e Transformação
- Limpeza

Objetivos de hoje

Mostrar conceitose métodos redução e discretização de características

Ao final da aula, os alunos terão uma idéia básica de algumas técnicas para seleção e discretização de características relacionadas aos dados

Roteiro Preprocessamento – p2



Redução de Características

Discretização

Redução

- Selecionar um mínimo de atributos que mantenha as distribuições de probabilidade semelhantes as originais
- Em Data Mining a supressão de uma coluna (atributo) é muito mais delicada do que a supressão de uma linha (observação)
- Retirar atributos relevantes ou permanecer com atributos irrelevantes
- Pode implicar na descoberta de padrões de baixa qualidade
  - Daí a necessidade de um estágio de seleção de atributos
- Uma abordagem para a seleção é a manual, baseada em conhecimento especialista *****

Redução: Estratégias

- Bancos de dados muito grandes podem tornar o processo de mineração lento
- Redução de dados
  - Obtém uma representação reduzida do conjunto de dados com menor volume e resultados similares
- Estratégias de redução
  - Redução de dimensionalidade
  - Compressão de dados
  - Redução de casos
  - Discretização

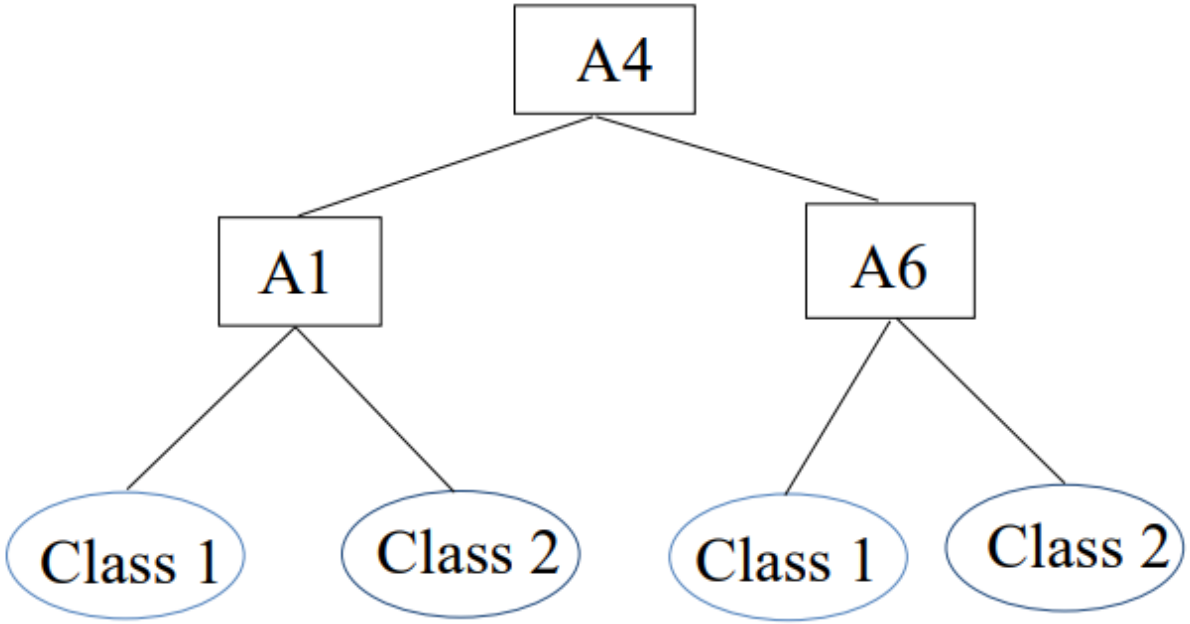
Redução: de Dimensionalidade

- Métodos Heurísticos
  - Indução por Árvore de Decisão
  - Melhores Atributos Individuais
  - Seleção forward ou Eliminação backward

Redução: de Dimensionalidade – Indução por Árvore de Decisão

- Conjunto Inicial de Atributos

{A1, A2, A3, A4, A5, A6}



-------> Reduced attribute set: {A1, A4, A6}

Redução: de Dimensionalidade – Seleção por Heurísticas

- **Seleção forward**

  – a busca é iniciada sem atributos e os mesmos são adicionados um a um

  – cada atributo é adicionado isoladamente e o conjunto resultante é avaliado segundo um critério

- **Eliminação backward**

  – a busca é iniciada com o conjunto completo de atributos e os mesmos são suprimidos um de cada vez

  – cada atributo é suprimido isoladamente e o conjunto resultante é avaliado segundo um critério

Redução: Compressão de Dados Extração de Variáveis

- **Objetivo**

  – obter novas variáveis à partir dos atributos iniciais. Em geral as novas variáveis são combinações lineares das variáveis iniciais

  – limitações: modelo linear (não adequado especialmente para alguns métodos de Data Mining)

- As técnicas de redução de dimensões se propõem a reduzir o número de variáveis com a menor perda possível de informações
- Essas técnicas são úteis também para tratar a redundância de informações (correlação entre variáveis) e ruído

Redução: Compressão de Dados Extração de Variáveis

- Famílias de Métodos
  - Métodos não supervisionados
  - Métodos supervisionados

- Métodos não supervisionados
  - Análise de Componentes Principais (variáveis quantitativas)
  - Análise de Correspondências (variáveis qualitativas)

- Métodos supervisionados
  - Análise Fatorial Discriminante

Análise de Componentes Principais (PCA)

- Componentes Principais (PCs) são tipos específicos de combinações lineares que são escolhidas de tal modo que zp (PCs) tenham as seguintes características

Dado um conjunto D com *n* instâncias e *p* atributos (x_1, x_2,..., x_p), uma transformação linear para um novo conjunto de atributos z_1, z_2,..., z_p pode ser calculada como:

$$z_1 = a_{11} x_1 + a_{21} x_2 + \ldots + a_{p1} x_p$$
$$z_2 = a_{12} x_1 + a_{22} x_2 + \ldots + a_{p2} x_p$$
$$\ldots$$
$$z_p = a_{1p} x_1 + a_{2p} x_2 + \ldots + a_{pp} x_p$$

Análise de Componentes Principais (PCA)

- As p componentes principais (PC) são não-correlacionadas (independentes)
- As PCs são ordenadas de acordo com quantidade da variância dos dados originais que elas contêm (ordem decrescente)
  - A primeira PC “explica” (contém) a maior porcentagem da variabilidade do conjunto de dados original
  - A segunda PC define a próxima maior parte, e assim por diante – Em geral, apenas algumas das primeiras PCs são responsáveis pela maior parte da variabilidade do conjunto de dados
  - O restante das PCs tem uma contribuição insignificante
- PCA é usada em Aprendizado de Máquina principalmente para a redução de dimensionalidade

PCA: Cálculo

- PCA pode reduzida ao problema de encontrar os auto-valores e auto-vetores da matriz de covariância (ou correlação) do conjunto de dados
- A proporção da variância do conjunto de dados originais explicada pela i-ésima PC é igual ao i-ésimo auto-valor divido pela soma de todos os p auto-valores
- Ou seja, as PCs são ordenadas - decrescente - de acordo com os valores dos auto-valores
- Quando os valores dos diferentes atributos estão em diferentes escalas, é preferível usar a matriz de correlação em lugar da matriz de covariância

Análise de Componentes Principais

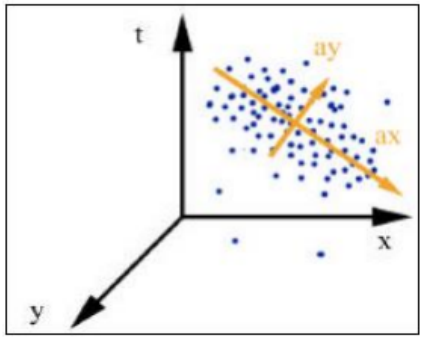
- Principais Limitações
  - – Assume apenas relações lineares entre os atributos
  - – A interpretação dos resultados (e.g., classificador gerado) em termos dos atributos originais pode ficar mais difícil

Análise de Componentes Principais

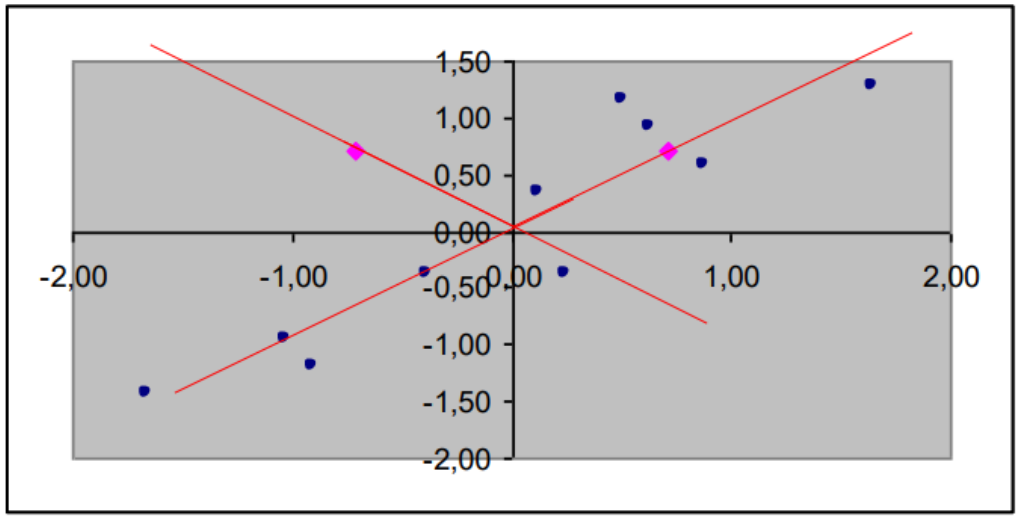
- Análise de Componentes Principais (variáveis quantitativas)
- Ela permite transformar um conjunto de variáveis originais, inter-correlacionadas, num novo conjunto de variáveis não correlacionadas, as componentes principais
- O objetivo mais imediato é verificar se existe um pequeno número das primeiras componentes principais que seja responsável por explicar uma proporção elevada da variação total associada ao conjunto original

Análise de Componentes Principais

PCA



Auto-Vetores e os Dados

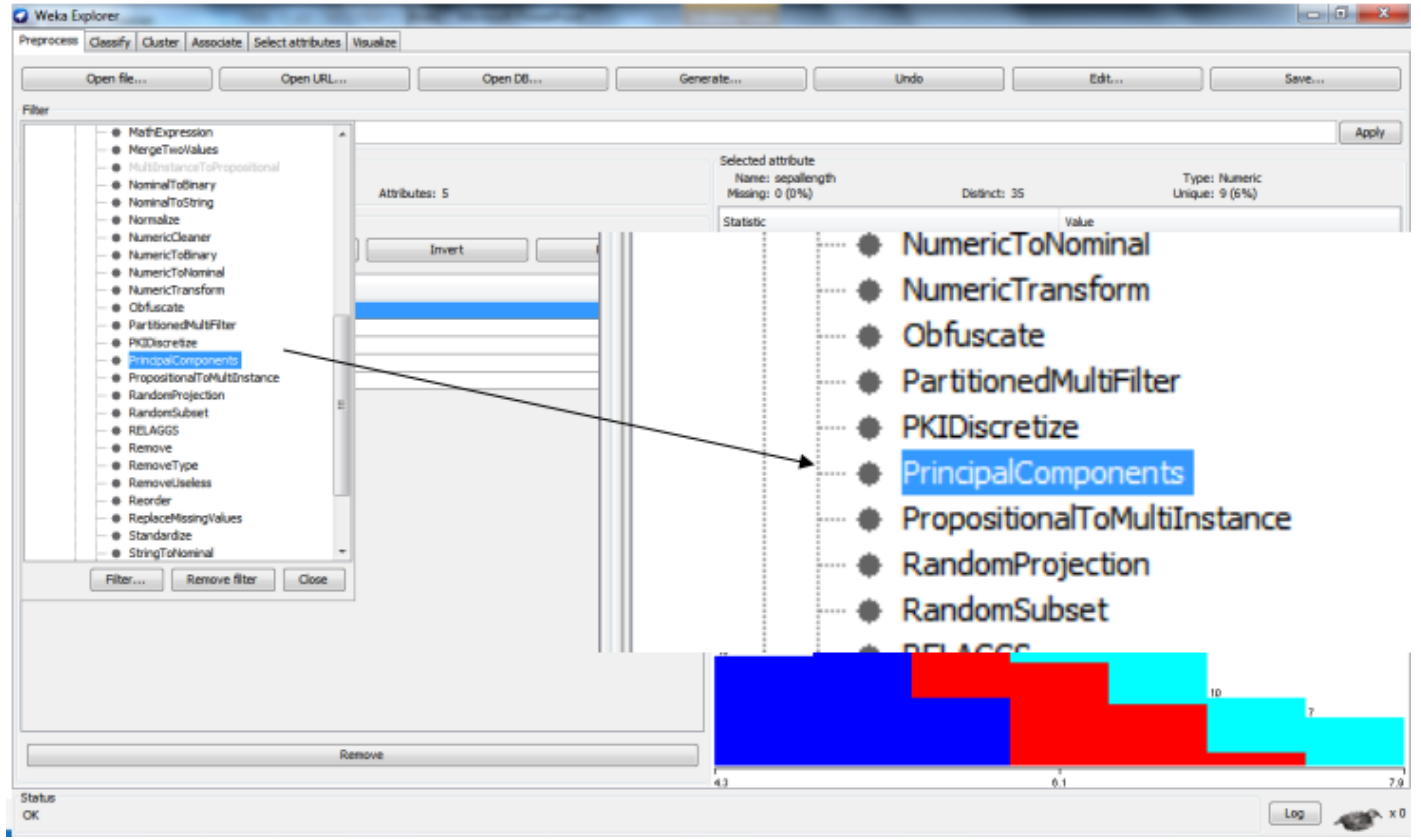


Análise de Componentes Principais (variáveis quantitativas)

- As vantagens são que ao se descorrelacionar os dados, estamos eliminando parte da informação redundante em cada dimensão.
  - os dados podem ser descritos de uma forma mais concisa;
  - certas características escondidas dos dados podem vir à luz depois de transformadas;
  - a distribuição dos dados pode ser representada (aproximadamente) pelas densidades individuais de cada dimensão.

Análise de Componentes Principais (variáveis quantitativas)

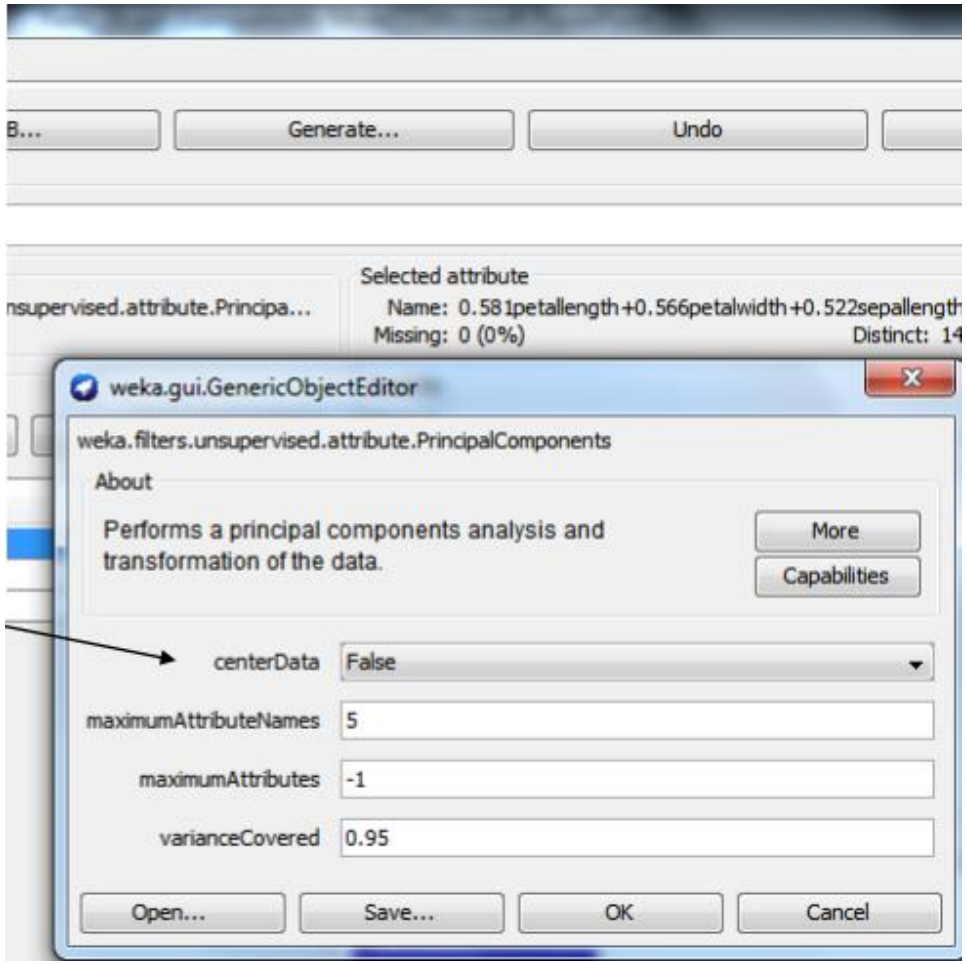
- Filters>unsupervised>attribute>principalcomponents



Análise de Componentes Principais (variáveis quantitativas)

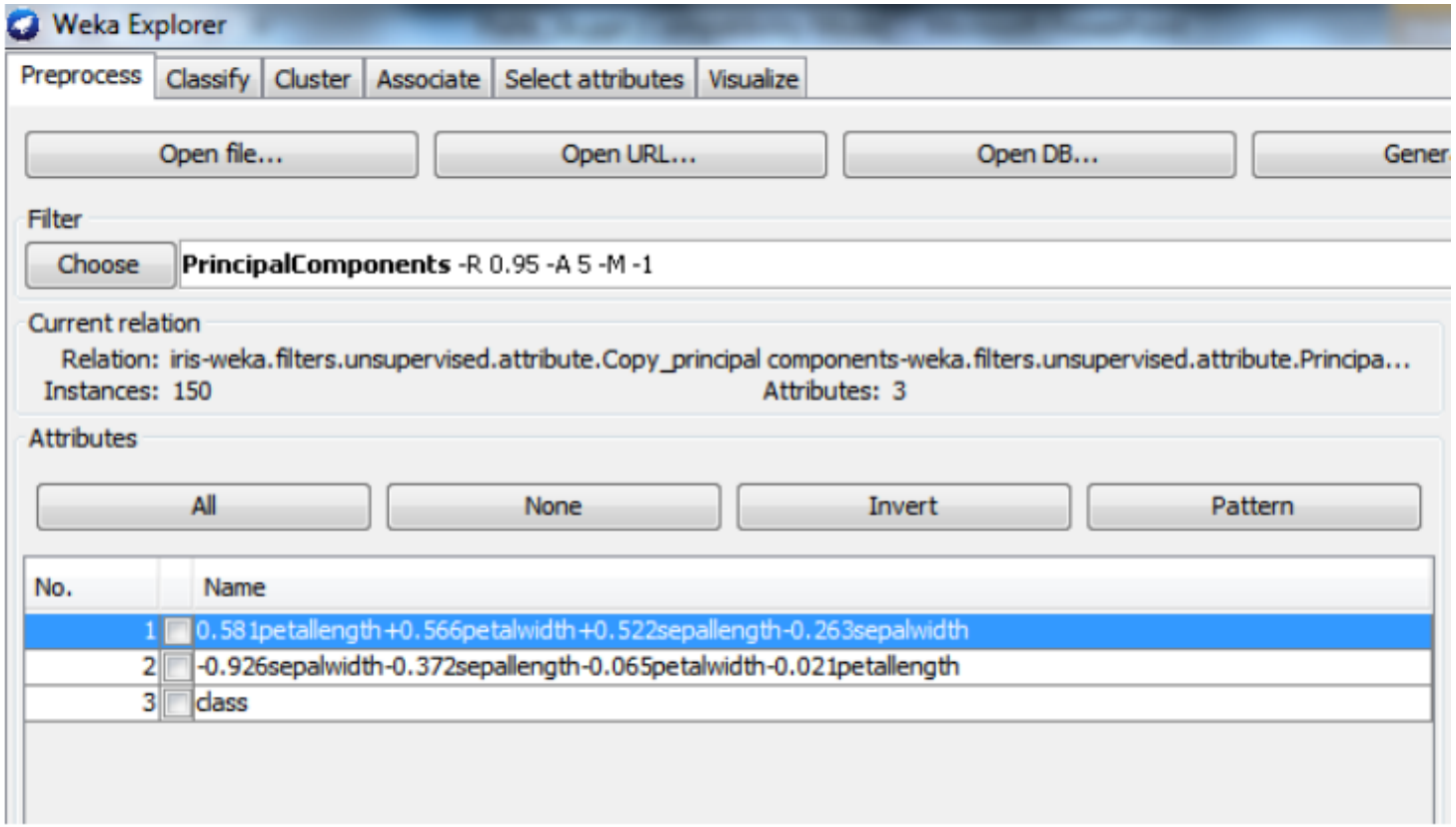
- Se centerData=TRUE e não normalizar ou padronizar seus dados antes, alguns atributos com valores grandes terá maior influência no PCA (matriz de covariância).

- Se centerData=FALSE, Weka padroniza os dados (matriz de correlação)



Análise de Componentes Principais (variáveis quantitativas)

- Filters>unsupervised>attribute>principalcomponents



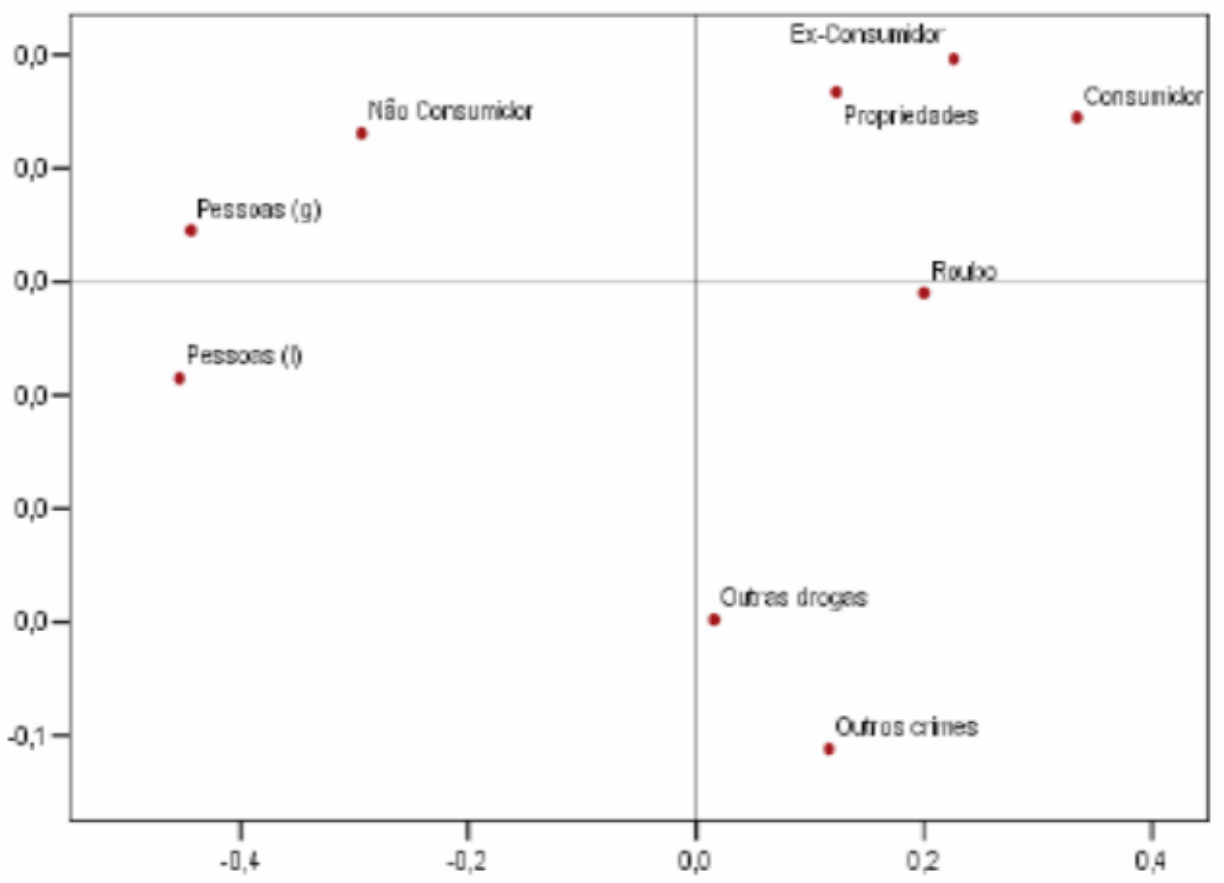
Análise de Correspondências (variáveis qualitativas)

- É uma técnica exploratória que serve para estudar a relação entre duas variáveis categoriais (e.g. nominais).
- Hierarquizar a informação disponível por ordem decrescente e de acordo com o grau de explicação do fenômeno em estudo;
- Representação gráfica da natureza das relações existentes entre as variáveis, ao colocar as categorias semelhantes próximas umas das outras

Análise de Correspondências (variáveis qualitativas)

- Crimes nos EUA nos anos 60

	Tipo de Crime				
Consumo de Heroína	**Pessoas (grave)**	**Roubo**	**Pessoas (leve)**	**Propriedades**	**Outros Crimes**
Consumidor	30	94	14	237	86
Ex-Consumidor	14	20	5	75	27
Outras Drogas	93	94	46	253	124
Não Consumidor	163	79	77	256	93



Análise de Correspondências (variáveis qualitativas)

- load carsmall
- x = [Acceleration Model_Year Horsepower Cylinders Weight];
- x = x(all(~isnan(x),2),:);
- [coefs,score] = princomp(zscore(x));
- vbls = {'Accel','Model_Y','HP','Cylinders','Wgt'};
- biplot(coefs(:,1:2),'scores',score(:,1:2), 'varlabels',vbls);

- Vamos entender cada linha do código

Análise de Correspondências (variáveis qualitativas)

- load carsmall:
  - Carrega o conjunto de dados carsmall, que é um conjunto de dados de exemplo fornecido pelo MATLAB. Ele contém dados sobre diferentes modelos de carros, como aceleração, ano do modelo, potência, número de cilindros e peso.

- x = [Acceleration Model_Year Horsepower Cylinders Weight];
  - Cria uma matriz x que contém as colunas de interesse do conjunto de dados carsmall.

Análise de Correspondências (variáveis qualitativas)

- x = x(all(~isnan(x),2),:);
  - Remove as linhas da matriz x que contêm qualquer valor NaN (Not a Number). Em outras palavras, isso limpa os dados, mantendo apenas as linhas completas sem valores faltantes.

- [coefs,score] = princomp(zscore(x));:
  - zscore(x) padroniza a matriz x para que cada coluna tenha média 0 e desvio padrão 1.
  - princomp realiza a Análise de Componentes Principais (PCA) nos dados padronizados e retorna:
  - coefs: os coeficientes (ou "cargas") dos componentes principais.
  - score: a representação dos dados no novo espaço de componentes principais

Análise de Correspondências (variáveis qualitativas)

- biplot(coefs(:,1:2),'scores',score(:,1:2), 'varlabels',vbls);
  - Cria um biplot usando os dois primeiros componentes principais.
  - Os argumentos passados especificam que ele deve plotar os dois primeiros componentes principais para os dados (score(:,1:2)) e para as variáveis (coefs(:,1:2)).
  - O argumento varlabels é usado para anotar o gráfico com os rótulos das variáveis.

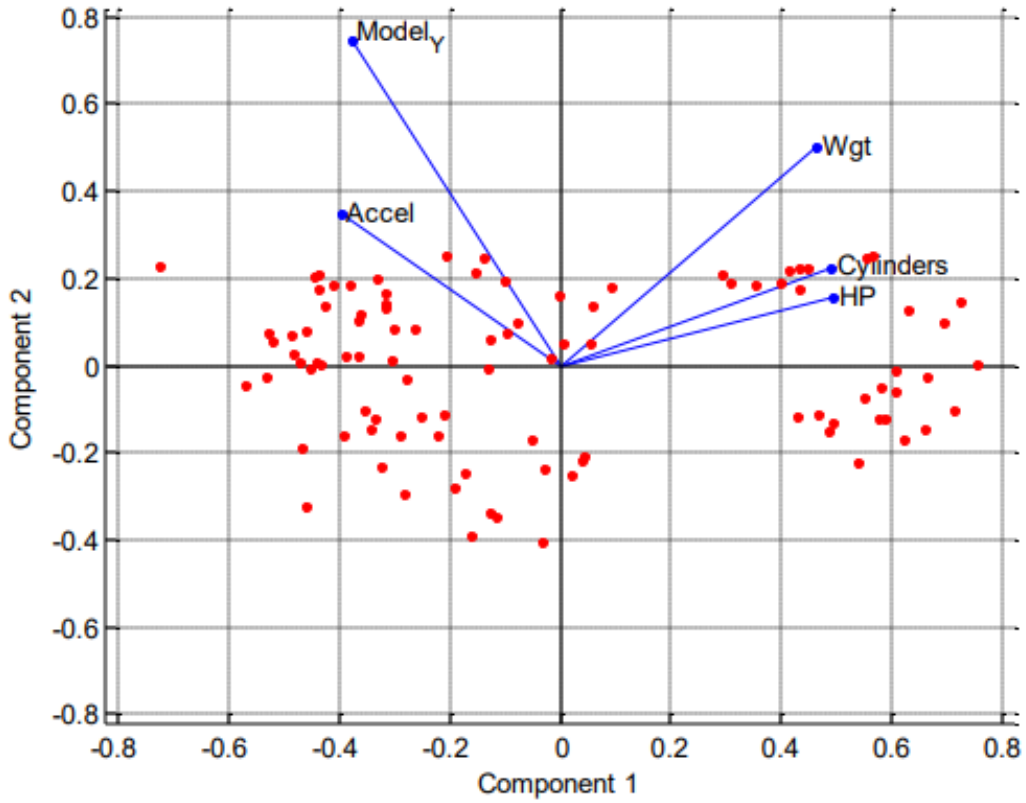
Análise de Correspondências (variáveis qualitativas)

- biplot(coefs(:,1:2),'scores',score(:,1:2), 'varlabels',vbls);
  - Cria um biplot usando os dois primeiros componentes principais.
  - Os argumentos passados especificam que ele deve plotar os dois primeiros componentes principais para os dados (score(:,1:2)) e para as variáveis (coefs(:,1:2)).
  - O argumento varlabels é usado para anotar o gráfico com os rótulos das variáveis.

Análise de Correspondências (variáveis qualitativas)

- Resumo:
  - O código realiza a Análise de Componentes Principais (PCA) no conjunto de dados carsmall (após a limpeza de dados)
  - em seguida, cria um biplot para visualizar os dois primeiros componentes principais e suas relações com as variáveis originais.
  - O biplot pode ajudar a entender as relações entre as variáveis e como elas contribuem para os componentes principais.

Análise de Correspondências (variáveis qualitativas)



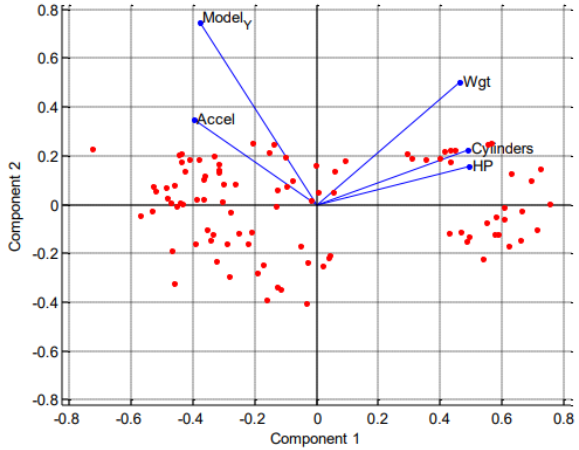
- Modely e wgt foram maiores. O que isso quer dizer?

Análise de Correspondências (variáveis qualitativas)

- No contexto da Análise de Componentes Principais (PCA) e do biplot, se as variáveis Model_Y (ano do modelo) e Wgt (peso) são as mais longas no gráfico, isso tem implicações importantes:
  - Se Model_Y e Wgt têm as projeções mais longas, isso sugere que essas duas variáveis têm a maior variação (ou influência) nos primeiros dois componentes principais.

- Direção das Projeções
  - A direção das projeções no biplot indica a relação entre as variáveis e o componente principal
  - Se projeções de variáveis apontam na mesma direção, isso sugere que essas variáveis estão correlacionadas positivamente
  - Se eles apontam em direções opostas, estão correlacionadas negativamente.

Análise de Correspondências (variáveis qualitativas)



- Interpretação
  - Model_Y tem uma grande projeção positiva no primeiro componente principal, isso pode indicar, por exemplo, que carros mais recentes (com anos de modelo mais altos) têm características distintas que contribuem significativamente para a variação no conjunto de dados.
  - Se Wgt (peso) também tem uma grande projeção, isso pode indicar que o peso do carro é outra característica significativa que varia no conjunto de dados.
  - Se Model_Y e Wgt apontam aproximadamente na mesma direção, isso pode sugerir que carros mais recentes tendem a ser mais pesados (ou vice-versa, dependendo da direção).

Redução: Redução de Casos

- Métodos Paramétricos

  – Assume que os dados podem ser representados por um modelo, armazena os parâmetros do modelo e descarta os dados

  -- Exemplo: regressão (simples e múltipla)

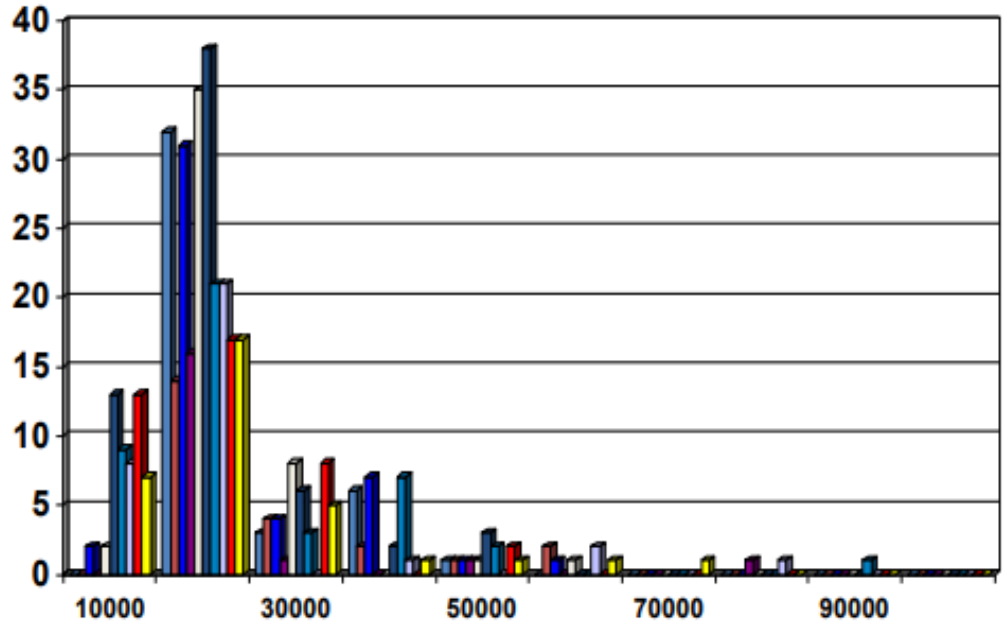
- Métodos Não-Paramétricos

  – Não usa modelos

  – Histogramas, clusterização, amostragem

Redução: Redução de Casos Histogramas

- Divide os dados em blocos e guarda a média de cada bloco
- Representação gráfica da distribuição de cada variável em intervalos de freqüência



Clusterização

- Particiona dados em clusters, e armazena apenas a representação do cluster
- Eficiente se os dados podem ser classificados e não estão muito “espalhados”
- Existem diversos algoritmos de clusterização

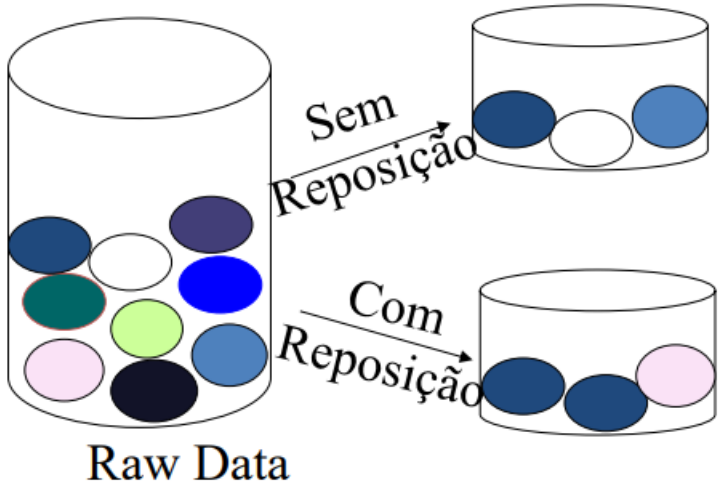
Redução: Redução de Casos Amostragem

- Escolhe um subconjunto de dados representativo
  - Amostragem aleatória simples sem reposição
  - Amostragem aleatória simples com reposição
  - Amostragem de clusters
  - Amostragem estratificada

- Exemplo: amostragem por faixa etária

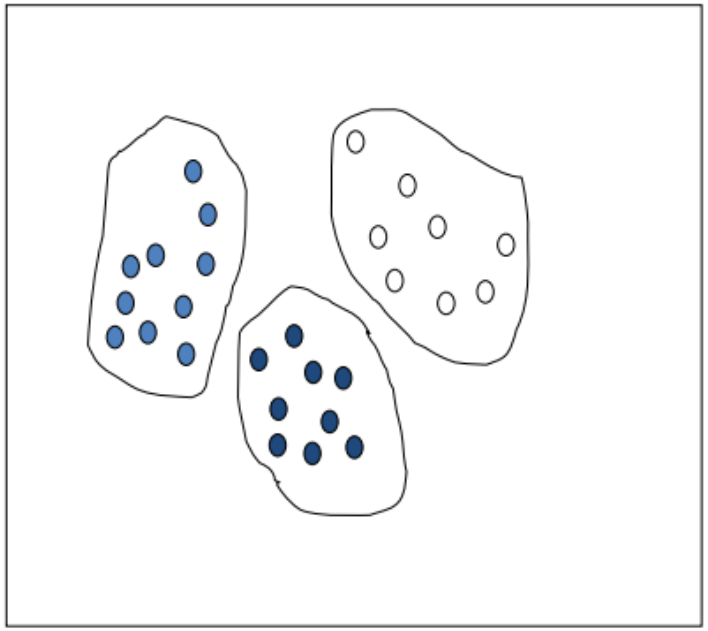
Redução: Redução de Casos
Amostragem – Aleatória

- Na retirada aleatória de elementos, como não há uma lógica para retirada dos pontos, pode-se também excluir dados importantes para o aprendizado do algoritmo.

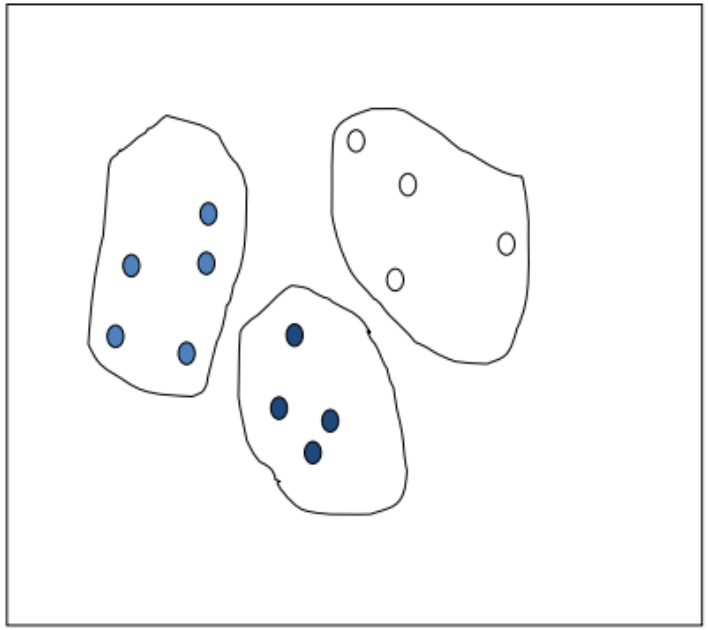


Redução: Redução de Casos
Amostragem - Estratificada

Dados Brutos

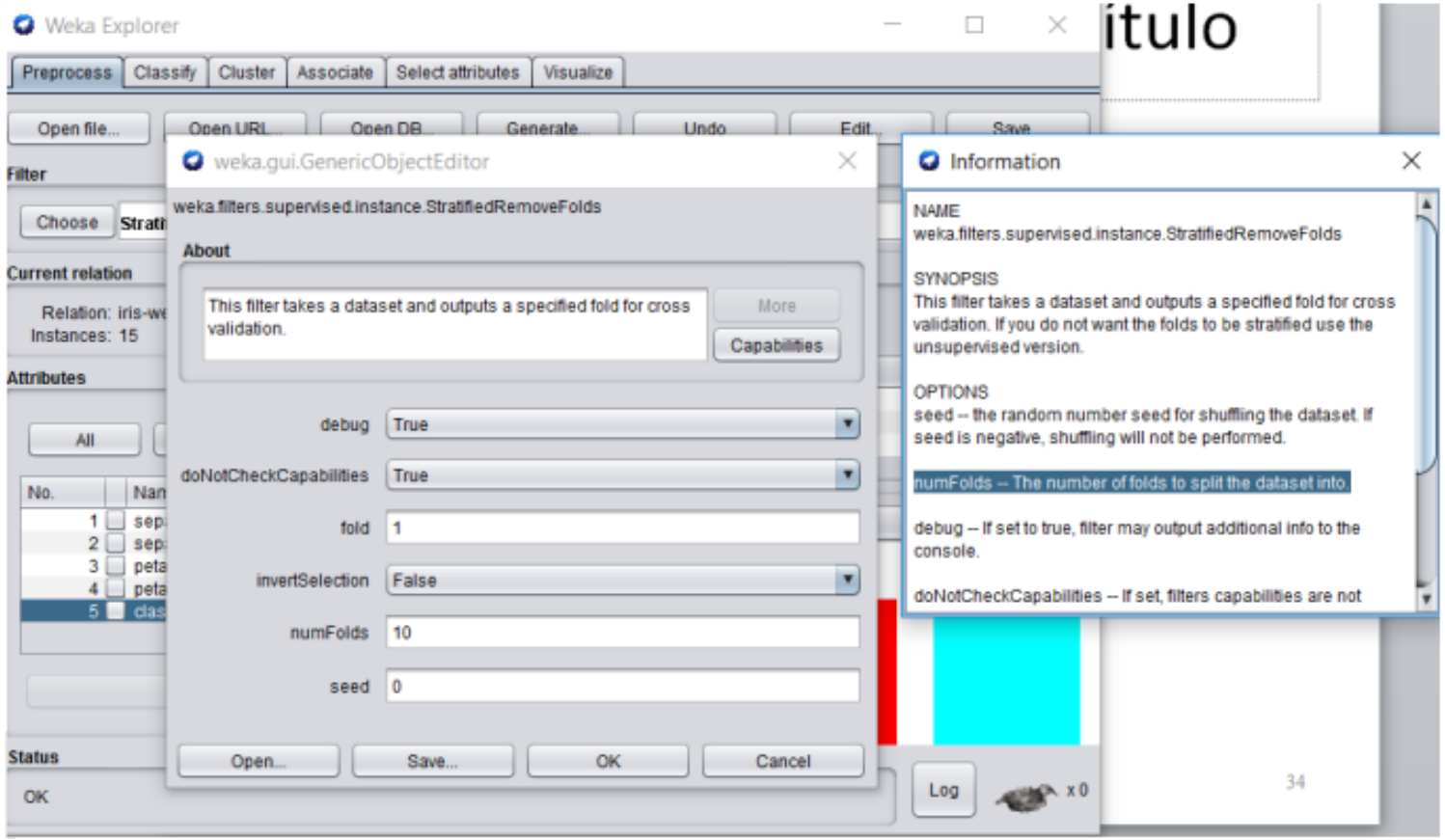


Amostra Estratificada



Redução: Redução de Casos Amostragem - Estratificada

- filters.supervised.instance.StratifiedRemoveFolds
  - fold -- número de folds que deseja manter
  - numFolds – divisão da base em numFolds partes



Redução: Redução de Casos Amostragem - SpreadSubSample

- É usado principalmente para lidar com conjuntos de dados desequilibrados.
  - Exemplo: problemas de classificação, especialmente na área médica ou detecção de fraudes, por exemplo, uma classe pode ser muito mais frequente do que a outra

- **Objetivo**
  - Este filtro seleciona uma amostra de um conjunto de dados usando a técnica de subamostragem. Ele foi projetado para criar uma amostra que é mais equilibrada entre as classes.

Redução: Redução de Casos Amostragem - SpreadSubSample

- Como funciona
  - Ele tenta espalhar a subamostra uniformemente nas regiões de entrada. Isso significa que ele tentará manter a diversidade dos exemplos ao invés de escolher muitos exemplos semelhantes

- Parâmetros importantes:
  - distributionSpread: Este é um valor entre 0 e 1 que determina o grau de espalhamento ou equilíbrio da subamostra entre as classes. Um valor de 1 tentará criar uma subamostra completamente equilibrada, enquanto um valor de 0,5 tentará criar uma subamostra que tem metade da proporção da classe minoritária em relação à classe majoritária.
  - randomSeed: Para a reprodutibilidade, você pode definir uma semente para o gerador de números aleatórios.
  - Other options: Existem outras opções para determinar o tamanho da subamostra e se os exemplos devem ser escolhidos com reposição.

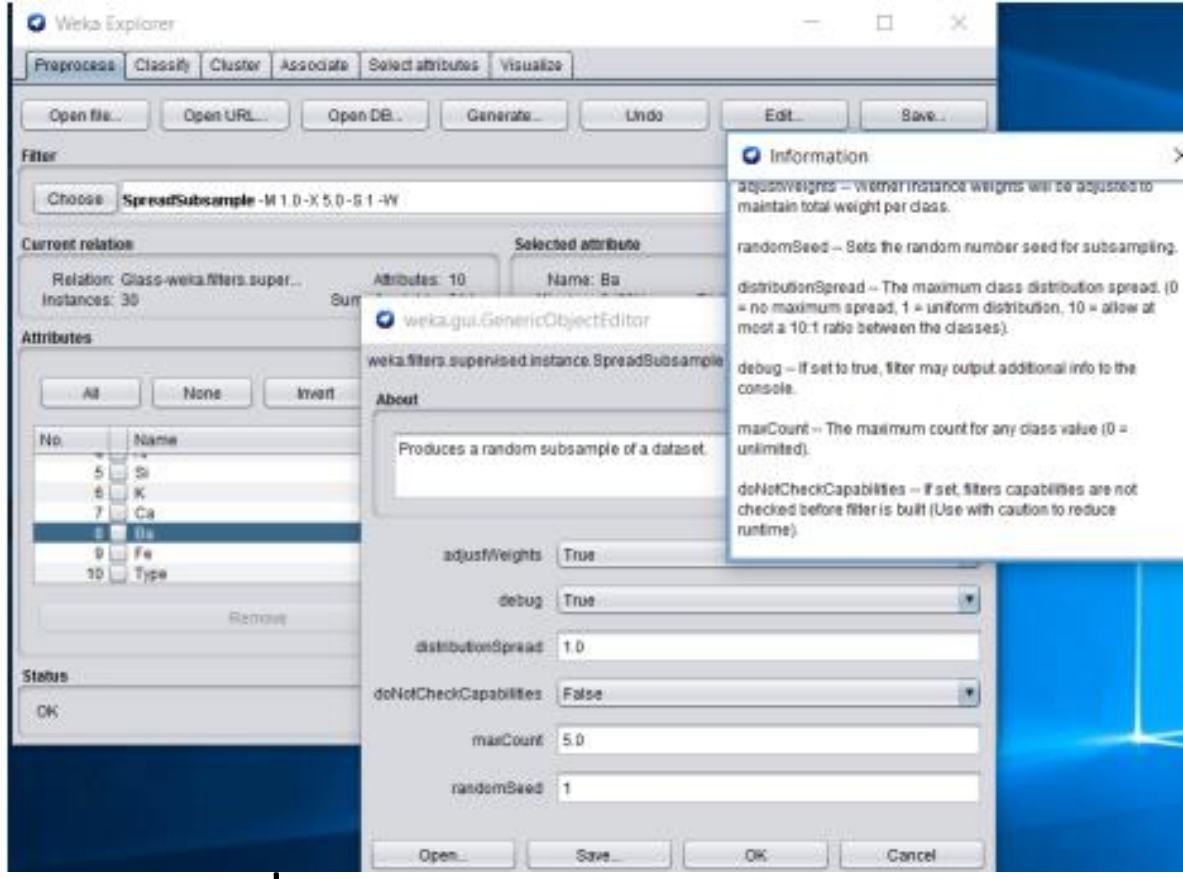
Redução: Redução de Casos Amostragem - SpreadSubSample

- Para entender melhor, vamos considerar um exemplo:

Imagine que você tenha um conjunto de dados com 1000 instâncias, onde 900 são da classe A (classe majoritária) e 100 são da classe B (classe minoritária). A proporção de B para A aqui é 0,10 (ou 10%)

- Se distributionSpread for definido como 1, o filtro tentará criar uma subamostra onde a classe A e a classe B são igualmente representadas. Portanto, se a subamostra contém, digamos, 200 instâncias, tentará ter 100 da classe A e 100 da classe B.
- Se distributionSpread for definido como 0,5, a subamostra terá uma proporção de classe B para classe A que é metade da proporção original no conjunto de dados. Usando o nosso exemplo, a proporção original é 10%, então a subamostra tentará ter uma proporção de 5,0%. Se a subamostra contém 200 instâncias, isso significaria aproximadamente 190 instâncias da classe A e 10 da classe B.

Redução: Redução de Casos Amostragem



- filters.supervised.instance.SpreadSubsa

– *distributionSpread* –

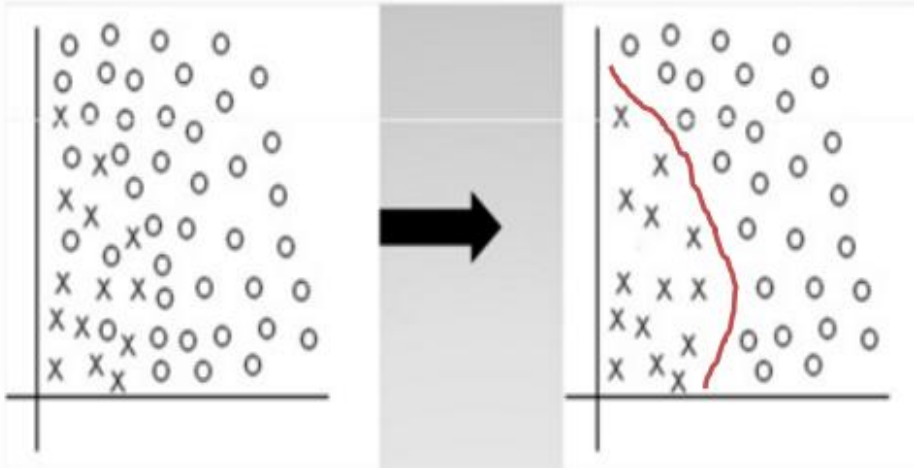
O spread de distribuição de classe máximo. (0 = sem spread máximo, 1 = distribuição uniforme, 10 = permitem no máximo uma razão de 10: 1 entre as classes).

– *maxCount* –

número máximo de registros para qualquer classe (0 = ilimitado).

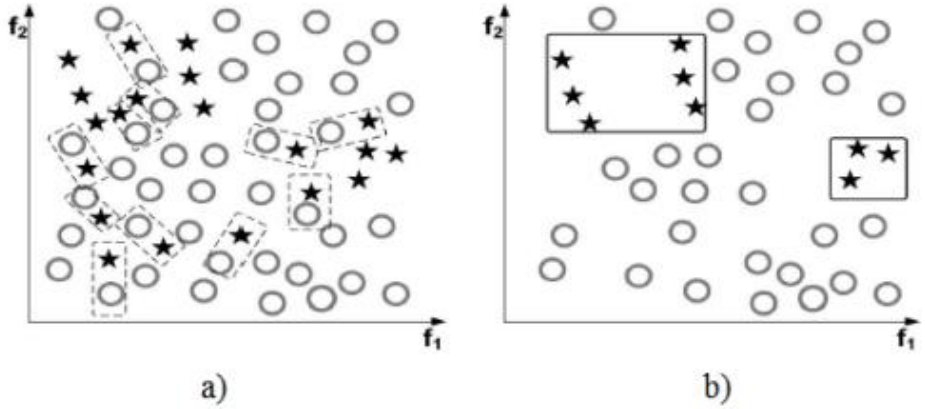
Redução a partir de Clusterização

- As técnicas de redução consistem em reduzir as classes majoritárias, bem como auxiliam na remoção de eventos que estejam na “fronteira” entre as classes (ruídos).
- O objetivo é retirar possíveis dados de borda, que possam confundir o classificador durante o aprendizado.



Redução a partir de Clusterização

- Tomek Links: efetua a remoção dos exemplos da classe majoritária que estão próximos às demais classes.
  - o **círculo** representa a classe **majoritária** e a **estrela** a classe **menor.** Em (a), são efetuados **links** entre os **elementos** das classes **minoritárias** mais **próximo** dos elementos da classe **majoritária**. Em seguida, os elementos da classe **majoritária** são **excluídos**, mantendo somente os pontos das classes menores, de acordo com a figura (b)



Redução a partir de Clusterização

- Edited Nearest Neighbor (ENN): este algoritmo opera da seguinte maneira: se um **elemento** pertence à classe **majoritária** e a classificação dada pelos seus três vizinhos mais próximos contradiz a classe original do elemento, este elemento (**classe majoritária**) é removido.

- Com isso se retira elementos de borda, bem como ruídos presentes na classe majoritária.

- **Podem chamar de algoritmo de remoção da ovelha negra**
  - Rita Lee - "Ovelha Negra" (Ao Vivo) - Multishow Ao Vivo – YouTube **0:30**

Redução a partir de Clusterização

- Neighborhood Cleaning Rule (NCL): o algoritmo inicialmente aplica o algoritmo anterior (ENN – Edited Nearest Neighbour). Se um **elemento** pertencer à classe **minoritária** e seus 3 vizinhos mais próximos o classificarem de forma errada, então **todos estes vizinhos** que pertencem à **classe majoritária** são removidos.
- Ou seja, é semelhante ao algoritmo anterior, porém mais drástico na redução dos exemplos de fronteira.

- O NCL costuma ter melhor desempenho

ligeiramente superior ao ENN.

Redução: Redução de Casos Amostragem

- Amostragem incremental
  - – O treinamento é realizado em amostras aleatórias cada vez maiores de casos, observar a tendência e parar quando não há mais progresso
  - – Um padrão típico de tamanhos de amostras pode ser 10%, 20%, 33%, 50%, 67% e 100%
- Critérios para passar para uma outra amostra
  - – O erro diminuiu? A complexidade do tratamento aumentou mais do que a queda da taxa de erro?
  - – A complexidade da solução atual é aceitável para a interpretação?

Redução: Redução de Casos Amostragem seguida de voto

- Interesse: quando o método de mineração suporta apenas N casos
- O mesmo método de mineração é aplicado para diferentes amostras de mesmo tamanho resultando em uma solução para cada amostra
- Quando um novo caso aparece, cada solução fornece uma resposta
- A reposta final é obtida por votação (classificação) ou pela média (regressão)

Discretização

- Três tipo de atributos
  - Nominais – não ordenável
  - Ordinais – ordenável
  - Contínuos – números reais

- Discretização
  - Divide a faixa de atributos contínuos em intervalos
  - Alguns algoritmos de classificação só aceitam atributos categóricos
  - Reduz o tamanho dos dados por discretização
  - Exemplo: Arredondamento de números

- Hierarquia de Valores
- Hierarquia de Atributos

Discretização de Variáveis

- Transforma atributos contínuos em atributos categóricos.
- Absolutamente essencial se o método inteligente só manuseia atributos categóricos
- Em alguns casos, mesmo métodos que aceitam atributos contínuos podem ter melhor desempenho com atributos categóricos.

Discretização de Variáveis

- Discretização Supervisionada
  - Considera a variável de saída (classe) na discretização
  - Ex: 1R, apesar de simples apresenta resultados similares a arvores de decisão.
- Métodos Não Supervisionados consideram somente o atributo a ser discretizado
  - São a única opção no caso de problemas de agrupamento (clustering), onde não se conhecem as classes de saída

Discretização: 1R

Base de Dados Meteorológicos

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Sol	85	85	Não	Não
Sol	80	90	Sim	Não
Nublado	83	86	Não	Sim
Chuva	70	96	Não	Sim
Chuva	68	80	Não	Sim
Chuva	65	70	Sim	Não
Nublado	64	65	Sim	Sim
Sol	72	95	Não	Não
Sol	69	70	Não	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Nublado	72	90	Sim	Sim
Nublado	81	75	Não	Sim
Chuva	71	91	Sim	Não

Discretização: 1R

Primeiro passo: ordenar pela coluna Temperatura

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Discretização: 1R

Segundo passo: discretizar pela Classe de saída

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Discretização: 1R

Segundo passo: discretizar pela Classe de saída

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Discretização: 1R

Terceiro passo: ajustar divisões

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Discretização: 1R

Terceiro passo: ajustar divisões

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64 (1)	65	Sim	Sim
Chuva	65 (2)	70	Sim	Não
Chuva	68	80	Não	Sim
Sol	69 (3)	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72 (4)	MUITAS DIVISÕES!		Não
Nublado	72			Sim
Chuva	75 (5)	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80 (6)	90	Sim	Não
Nublado	81	75	Não	Sim
Nublado	83 (7)	86	Não	Sim
Sol	85 (8)	85	Não	Não

Discretização: 1R

Quarto passo: mínimo de valores da maior classe (ex: 3)

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Discretização: 1R

Quarto passo: mínimo de valores da maior classe (ex: 3)

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Discretização: 1R

Quarto passo: mínimo de valores (ex: 3)

Tempo	Temperatura		Umidade	Vento	Jogar? (CLASSE)
Nublado	64		65	Sim	Sim
Chuva	65		70	Sim	Não
Chuva	68	1	80	Não	Sim
Sol	69		70	Não	Sim
Chuva	70		96	Não	Sim
Chuva	71		91	Sim	Não
Sol	72		95	Não	Não
Nublado	72	2	90	Sim	Sim
Chuva	75		80	Não	Sim
Sol	75		70	Sim	Sim
Sol	80		90	Sim	Não
Nublado	81	3	75	Não	Sim
Nublado	83		86	Não	Sim
Sol	85		85	Não	Não

Discretização: 1R

Quarto passo: mínimo de valores (ex: 3)

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68 (1)	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72 (2)	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81	75	Não	Sim
Nublado	83 (3)	86	Não	Sim
Sol	85	85	Não	Não

Discretização: 1R

Quarto passo: mínimo de valores (ex: 3)

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68 (1)	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81 (2)	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Discretização: 1R

Quarto passo: mínimo de valores (ex: 3)

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68 (1)	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71			
Sol	72			
Nublado	72			
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81 (2)	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Categoria 1: Temperatura ≤ 77.5

Categoria 2: Temperatura > 77.5

Discretização: 1R

Base de Dados Meteorológicos - Umidade

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Sol	69	70	Não	Sim
Sol	75	70	Sim	Sim
Nublado	81	75	Não	Sim
Chuva	68	80	Não	Sim
Chuva	75	80	Não	Sim
Sol	85	85	Não	Não
Nublado	83	86	Não	Sim
Sol	80	90	Sim	Não
Nublado	72	90	Sim	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Chuva	70	96	Não	Sim

Discretização de Variáveis Contínuas

Base de Dados Meteorológicos - Umidade

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Sol	69	70	Não	Sim
Sol	75	70	Sim	Sim
Nublado	81	75	Não	Sim
Chuva	68	80	Não	Sim
Chuva	75	80	Não	Sim
Sol	85	85	Não	Não
Nublado	83	86	Não	Sim
Sol	80	90	Sim	Não
Nublado	72	90	Sim	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Chuva	70	96	Não	Sim

Discretização de Variáveis

- Discretização Não-Supervisionada
  - O método 1R é supervisionado. Considera a variável de saída (classe) na discretização

- **Métodos Não Supervisionados consideram somente o atributo a ser discretizado**
  - **São a única opção no caso de problemas de agrupamento (clustering), onde não se conhecem as classes de saída**

Métodos de Discretização – Não Supervisionada

- Três abordagens básicas:
  - Número pré-determinado de intervalos
    - uniformes (equal-interval binning)
  - Número uniforme de amostras por intervalo
    - (equal-frequency binning)
  - Agrupamento (clustering): intervalos arbitrários

Métodos de Discretização – Não Supervisionada

- Número pré-determinado de intervalos
  - uniformes (equal-interval binning)

- No exemplo (temperatura):

  64 65 68 69 70 71 72 72 75 75 80 81 83 85

- Bins com largura 6:

  $x \leq 60$

  $60 < x \leq 66$

  $66 < x \leq 72$

  $72 < x \leq 78$

  $78 < x \leq 84$

  $84 < x \leq 90$

Métodos de Discretização – Não Supervisionada

- Número pré-determinado de intervalos
  – uniformes (equal-interval binning)
- No exemplo (temperatura):

  64 65 68 69 70 71 72 72 75 75 80 81 83 85
- Bins com largura 6

  $x \leq 60$: n.a.

  $60 < x \leq 66$: 64, 65

  $66 < x \leq 72$: 68, 69, 70, 71, 72, 72

  $72 < x \leq 78$: 75, 75

  $78 < x \leq 84$: 80, 81, 83

  $84 < x \leq 90$: 85

Métodos de Discretização – Não Supervisionada

- Equal-interval binning: Problemas
  - Como qualquer método não supervisionado, arrisca destruir distinções úteis, devido a divisões muito grandes ou fronteiras inadequadas
- Distribuição das amostras muito irregular, com bins com muitas amostras e outras com poucas amostras

Métodos de Discretização – Não Supervisionada

- Número uniforme de amostras por intervalor
  - (equal-frequency binning)
- Também chamado de equalização do histograma
- Cada bin tem o mesmo número de amostras
- Histograma é plano
- Heurística para o número de bins: raiz de N
  - Onde N = número de amostras

Métodos de Discretização – Não Supervisionada

- Número uniforme de amostras por intervalo
  - (equal-frequency binning)
- No exemplo (temperatura):

  64 65 68 69 |70 71 72 72 |75 75 80| 81 83 85
- 14 amostras: 4 Bins
  - $x \leq 69{,}5$: 64, 65, 68, 69
  - $69{,}5 < x \leq 73{,}5$: 70, 71, 72, 72
  - $73{,}5 < x \leq 80{,}5$: 75, 75, 80
  - $x > 80{,}5$: 81, 83, 85

Hierarquia de Atributos

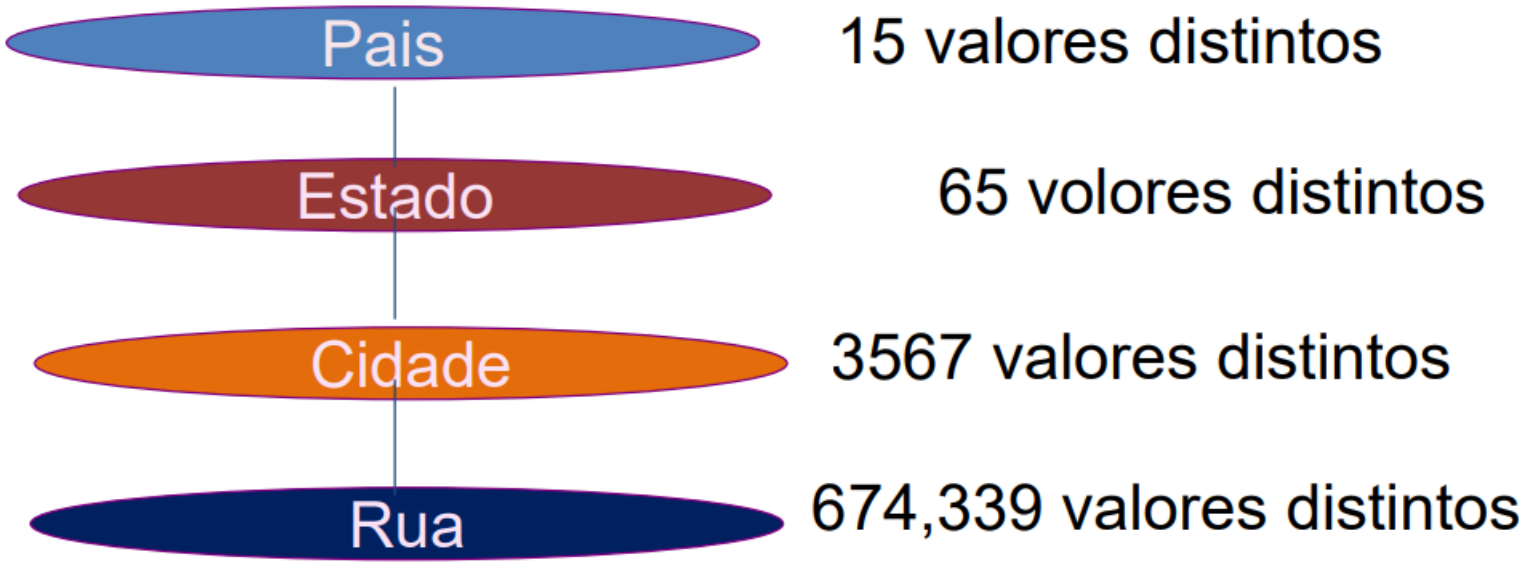
- Especialista do domínio apresenta hierarquia
    - Exemplo
        - Logradouro < Bairro < Cidade < Estado
        - Especialista estabelece nível de corte

Hierarquia de Valores

- Também necessita do especialista
    - Exemplo
        - {tênis, sapato, sandália} = sapato
        - {bermuda, calça, camisa, paletó} = roupa

Hierarquias de conceitos para dados categóricos

- Hierarquia conceitual pode ser gerada automaticamente com base no número de valores distintos por atributo.
- O atributo com o maior número de valores distintos é colocado no nível mais baixo da hierarquia



Métodos de Discretização – Não Supervisionada

- Agrupamento (Clustering)
- No caso unidimensional: para cada grupo (cluster), atribui-se um valor discreto

Principais Bibliotecas para pré-processamento de Dados

- **Pandas**
  - Descrição: Uma das bibliotecas mais populares para manipulação e análise de dados. Fornece estruturas de dados como DataFrame para manipulação eficiente de tabelas de dados.
  - Funcionalidades Principais: Limpeza de dados, filtragem, agregação, fusão e junção de tabelas, manipulação de séries temporais, entre outras.

- **NumPy**
  - Descrição: Biblioteca de álgebra linear e operações matemáticas para arrays multidimensionais.
  - Funcionalidades Principais: Operações matemáticas, geração de números aleatórios, operações em arrays, entre outras.

Principais Bibliotecas para pré-processamento de Dados

- **Scikit-learn (sklearn)**
  - Descrição: Uma das bibliotecas mais populares para aprendizado de máquina em Python.
  - Funcionalidades Principais: Contém uma variedade de ferramentas de pré-processamento, como normalização, padronização, codificação de variáveis categóricas, imputação de dados faltantes, redução de dimensionalidade, entre outras.
- **Statsmodels**
  - Descrição: Biblioteca para estimar e testar modelos estatísticos.
  - Funcionalidades Principais: Análise de regressão, testes estatísticos, exploração de dados, entre outras.

Principais Bibliotecas para pré-processamento de Dados

- **Beautiful Soup:**
  - Descrição: Biblioteca para extração de dados de documentos HTML e XML.
  - Funcionalidades Principais: Parseamento de páginas da web, extração de informações de tags específicas, navegação na árvore do DOM.

- **Scipy:**
  - Descrição: Biblioteca para matemática, ciência e engenharia.
  - Funcionalidades Principais: Contém módulos para otimização, álgebra linear, integração, interpolação, entre outras tarefas.

Principais Bibliotecas para pré-processamento de Dados

- **Missingno**
  - Descrição: Biblioteca especializada na visualização de dados faltantes.
  - Funcionalidades Principais: Visualização gráfica de dados faltantes, matriz de nulidade, entre outras.

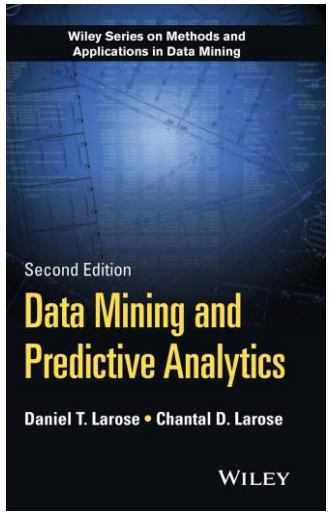
- **Feature-engine**
  - Descrição: Biblioteca focada em engenharia de recursos.
  - Funcionalidades Principais: Transformações de variáveis, codificação, imputação, discretização, entre outras.

Principais Bibliotecas para pré-processamento de Dados

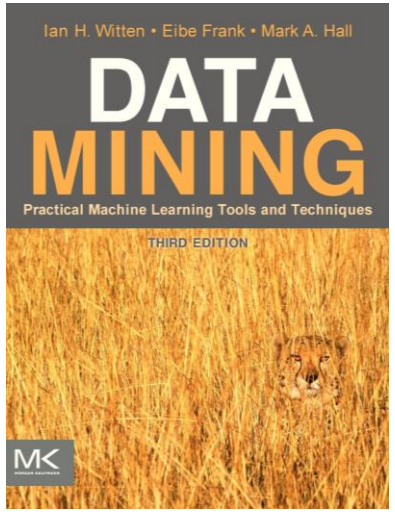
- **Category Encoders:**
  - Descrição: Uma extensão do scikit-learn para codificação de variáveis categóricas.
  - Funcionalidades Principais: Oferece várias técnicas de codificação, como codificação one-hot, codificação ordinal, codificação binária, entre outras.

- **Imbalanced-learn (imblearn)**
  - Descrição: Biblioteca focada em técnicas de reamostragem para conjuntos de dados desequilibrados.
  - Funcionalidades Principais: Subamostragem, superamostragem, geração de novas amostras sintéticas, entre outras.

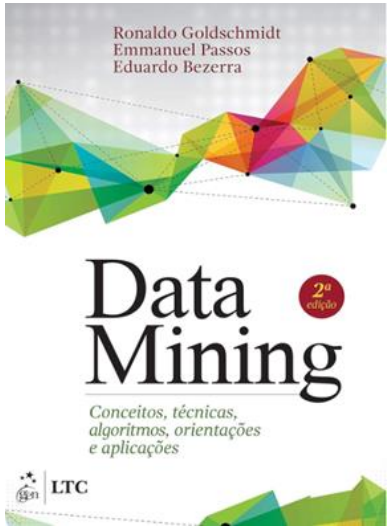
Bibliografia



.



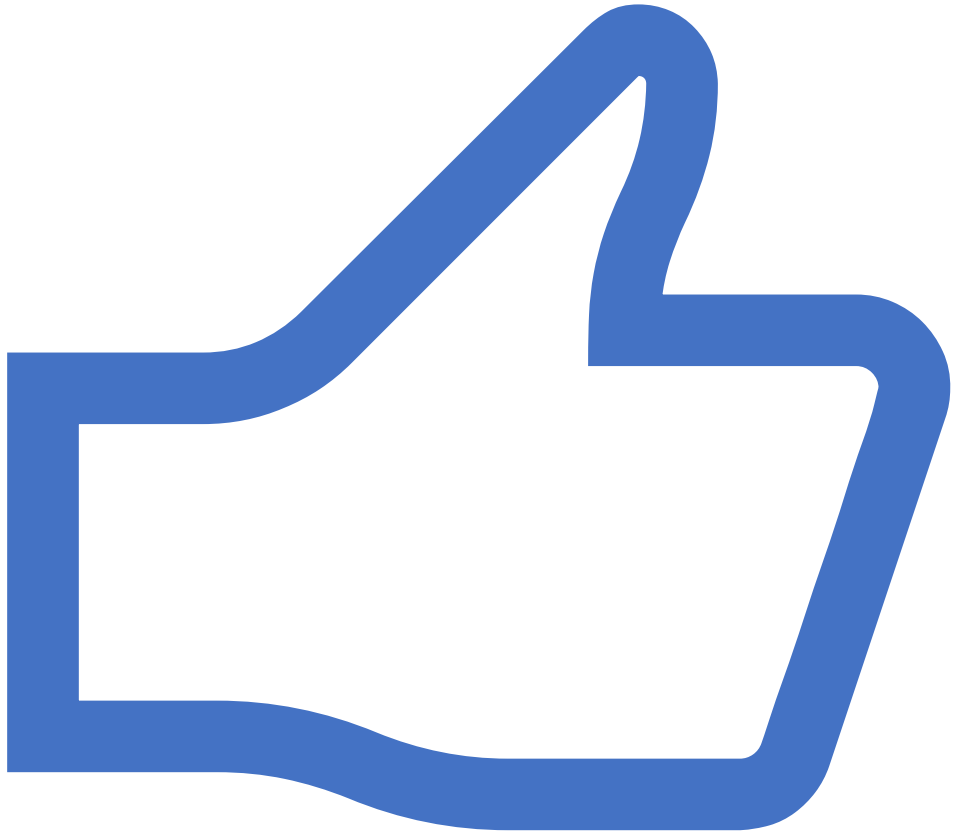
.



.

Dúvidas?

Obrigado !



Apresentador

Thales Levi Azevedo Valente

E-mail:
thales.l.a.valente@gmail.com