

Preparação dos dados

Disciplina: Mineração de Dados

Prof. Braian Varjão

Principais técnicas de pré-processamento de dados

1. Identificação de variáveis;
2. Análise de variáveis;
3. Tratamento de dados faltantes;
4. Tratamento de outliers;
5. Transformação de atributos;
6. Derivação de atributos;
7. Reamostragem;
8. Redução de dimensionalidade.



Principais técnicas de pré-processamento de dados

1. **Identificação de variáveis;**
2. Análise de variáveis;
3. Tratamento de dados faltantes;
4. Tratamento de outliers;
5. Transformação de atributos;
6. Derivação de atributos;
7. Reamostragem;
8. Redução de dimensionalidade.



Identificação de variáveis

Preditores vs. Objetivo

Student_ID	Gender	Prev_Exam_Marks	Height (cm)	Weight Caregory (kgs)	Play Cricket
S001	M	65	178	61	1
S002	F	75	174	56	0
S003	M	45	163	62	1
S004	M	57	175	70	0
S005	F	59	162	67	0



Preditor



Objetivo

Identificação de variáveis

Tipos de dados

Student_ID	Gender	Prev_Exam_Marks	Height (cm)	Weight Caregory (kgs)	Play Cricket
S001	M	65	178	61	1
S002	F	75	174	56	0
S003	M	45	163	62	1
S004	M	57	175	70	0
S005	F	59	162	67	0

String

Inteiro

Identificação de variáveis

Atributos categóricos vs. contínuos

Student_ID	Gender	Prev_Exam_Marks	Height (cm)	Weight Caregory (kgs)	Play Cricket
S001	M	65	178	61	1
S002	F	75	174	56	0
S003	M	45	163	62	1
S004	M	57	175	70	0
S005	F	59	162	67	0

-  Categórico
-  Contínuo

Principais técnicas de pré-processamento de dados

1. Identificação de variáveis;
2. **Análise de variáveis;**
3. Tratamento de dados faltantes;
4. Tratamento de outliers;
5. Transformação de atributos;
6. Derivação de atributos;
7. Reamostragem;
8. Redução de dimensionalidade.



Análise univariada



Principais perguntas:

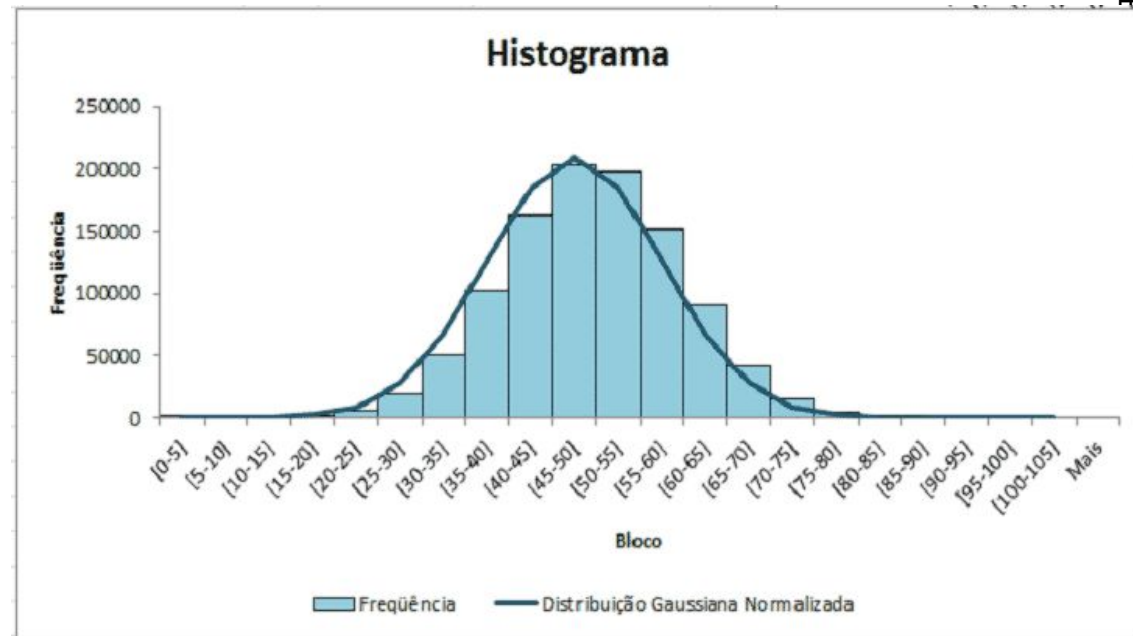
- Os atributos estão nos tipos corretos?
- Existem dados faltantes?
- Existem outliers? De que tipo?
- A amostra dos dados representa o mundo real?
- Existe um desbalanceamento importante?
- Em que valores os dados se concentram?
- É possível simplificar os dados?
- É possível/necessário criar atributos derivados?

Análise univariada

Atributos contínuos

- Média, mediana, máximo, mínimo...

```
>>> s = pd.Series([1, 2, 3])
>>> s.describe()
count      3.0
mean       2.0
std        1.0
min        1.0
25%        1.5
50%        2.0
75%        2.5
max         3.0
dtype: float64
```

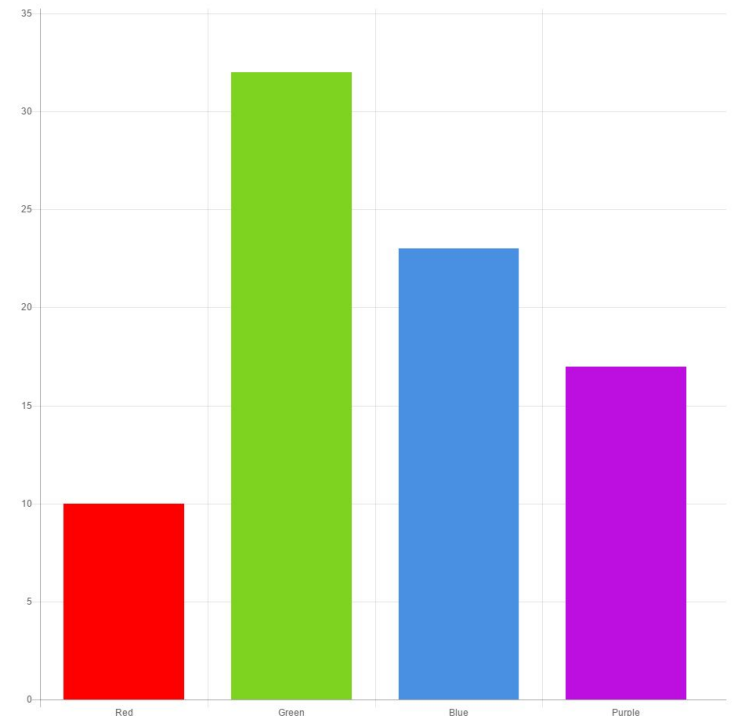
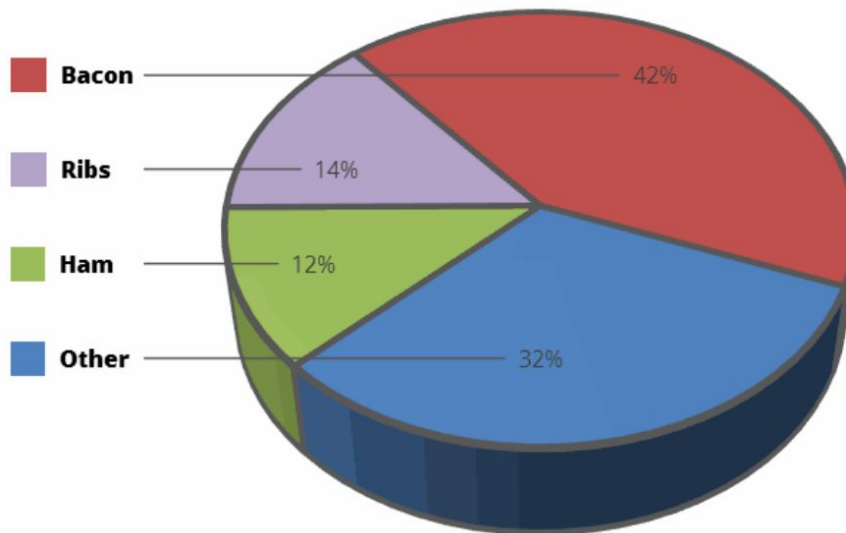


Análise univariada

Atributos categóricos

- Contagem e percentual de ocorrência.

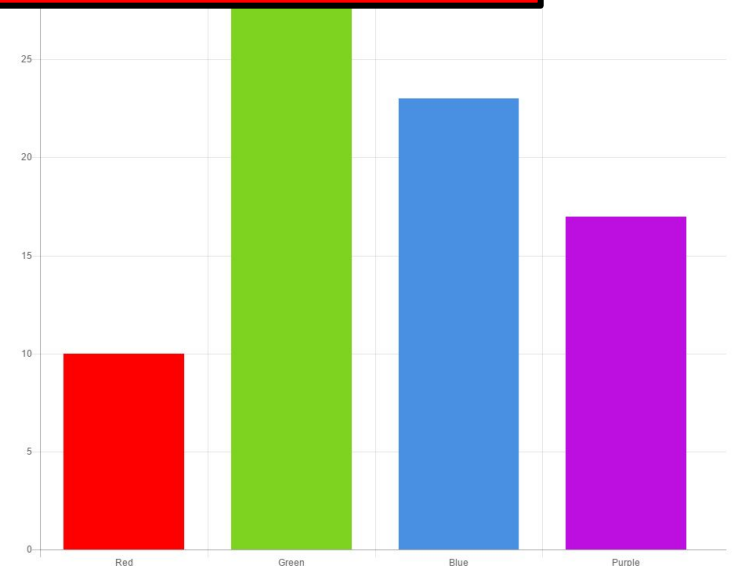
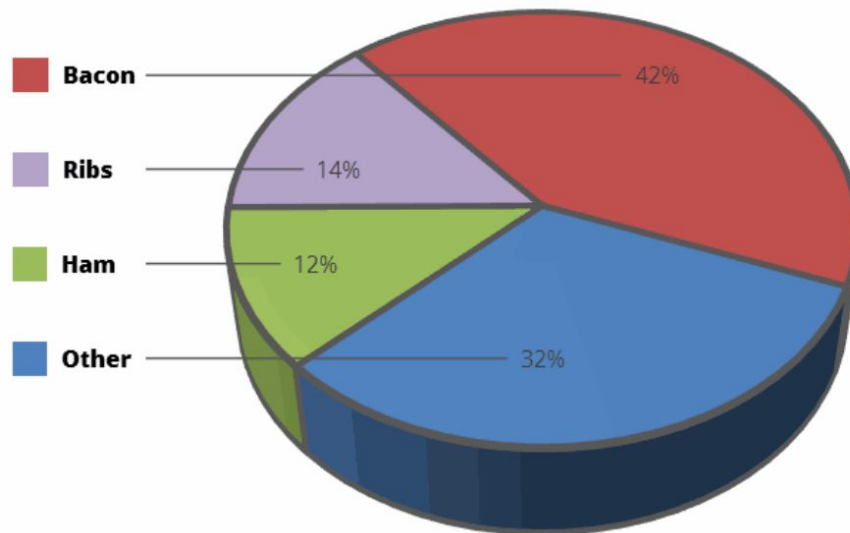
Pig Meat Preferences



Análise univariada

Atributos categóricos

Atenção: evitar gráficos de pizza e não utilizar estilos 3D em gráficos!



Análise bivariada



Principais perguntas:

- Existe correlação / dependência entre os atributos?
- Separando os dados a partir de um atributo x , como a distribuição do atributo k varia em cada grupo?
- Quais outros insights eu consigo obter das relações entre os dados?

Análise bivariada



- ▷ Contínua & Contínua;
- ▷ Categórica & Categórica;
- ▷ Contínua & Categórica.

Análise bivariada: Contínua & Contínua

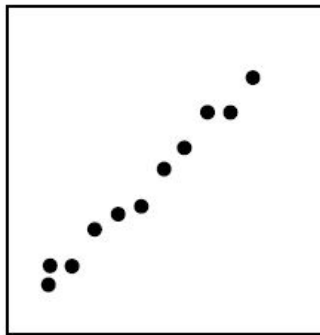
Índice de correlação

- +1: correlação linear positiva perfeita;
- 0: sem correlação;
- -1: correlação linear negativa perfeita.

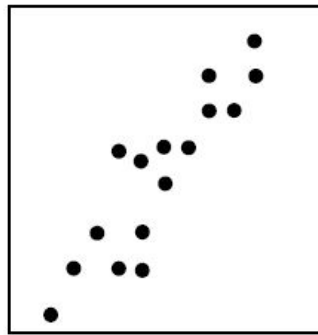
X	65	72	78	65	72	70	65	68
Y	72	69	79	69	84	75	60	73

Metrics	Formula	Value
Co-Variance (X,Y)	=COVAR(E6:L6,E7:L7)	18.77
Variance (X)	=VAR.P(E6:L6)	18.48
Variance (Y)	=VAR.P(E7:L7)	45.23
Correlation	=G10/SQRT(G11*G12)	0.65

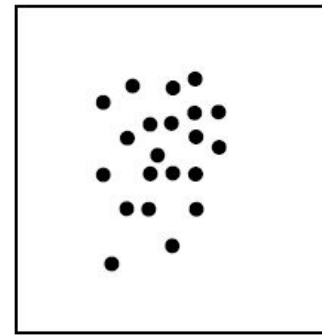
Análise bivariada: Contínua & Contínua



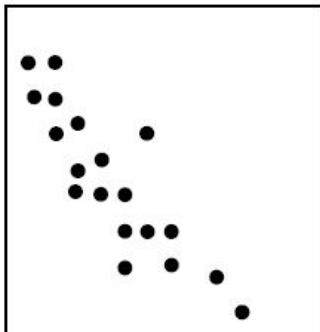
Strong positive correlation



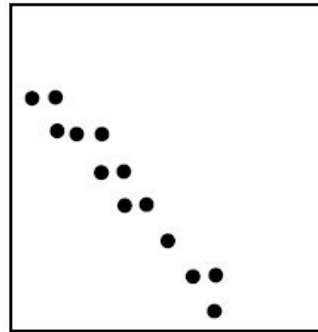
Moderate positive correlation



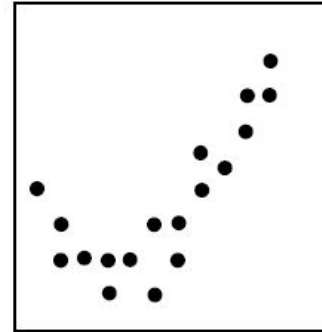
No correlation



Moderate negative correlation



Strong negative correlation

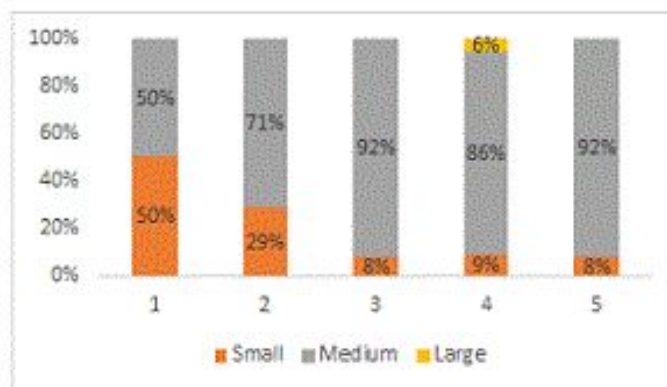
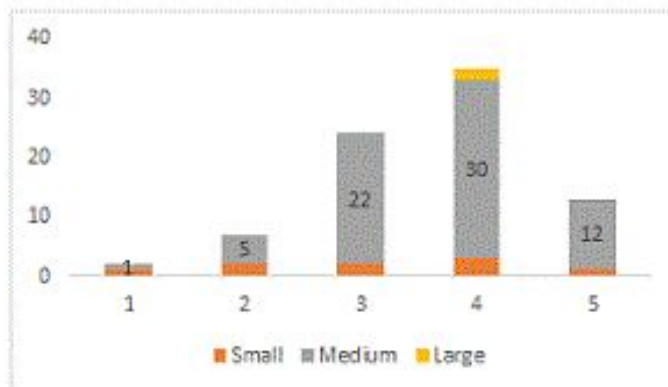


Curvilinear relationship

Análise bivariada: Categórica & Categórica

Frequency Row Pct	Product Category					Total
	1	2	3	4	5	
Small	1 11.11	2 22.22	2 22.22	3 33.33	1 11.11	9
Medium	1 1.43	5 7.14	22 31.43	30 42.86	12 17.14	70
Large	0 0.00	0 0.00	0 0.00	2 100.00	0 0.00	2
Total	2	7	24	35	13	81

Frequency Missing = 77



Análise bivariada: Categórica & Categórica



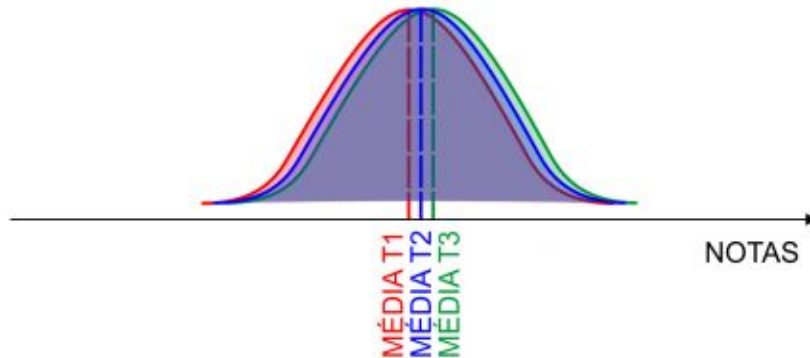
Teste X^2 : probabilidade de independência

- Probabilidade próxima de 0:
 - variáveis são dependentes.
- Probabilidade próxima de 1:
 - variáveis são independentes.

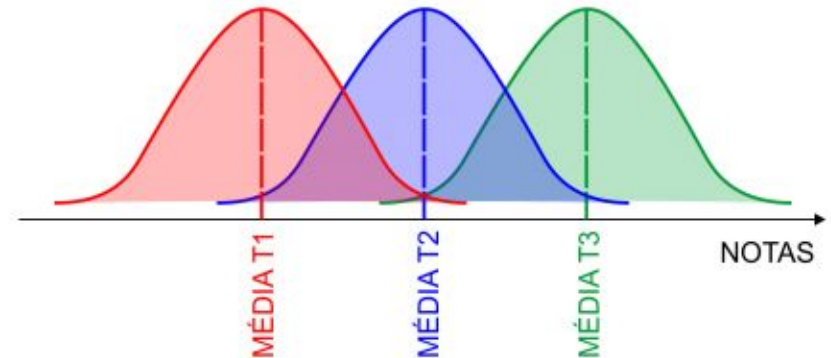
Probabilidade inferior a 0.05: isso indica que as variáveis são dependentes com 95% de confiança.

Análise bivariada: Contínua & Categórica





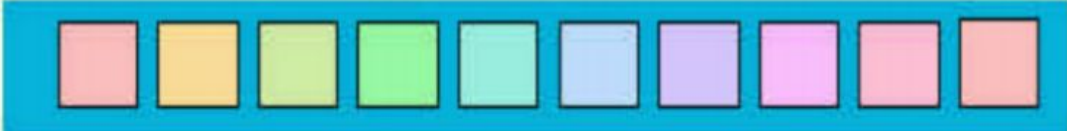

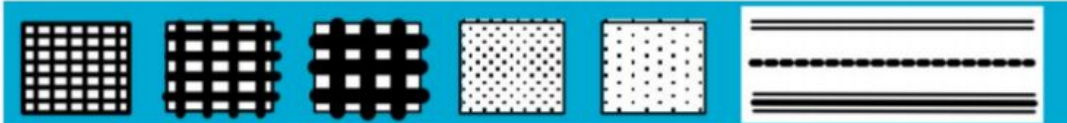
DISTRIBUIÇÃO DAS NOTAS DOS ALUNOS SUPONDO QUE
NÃO HÁ DIFERENÇA ENTRE AS TURMAS T1, T2 E T3



DISTRIBUIÇÃO DAS NOTAS DOS ALUNOS SUPONDO
DIFERENÇA ENTRE AS TURMAS T1, T2 E T3



Análise multivariada

Atributos Visuais de Bertin	
POSIÇÃO	
TAMANHO	
MARCAS	
LUMINOSIDADE	
COR	
ORIENTAÇÃO	
TEXTURA	

Análise multivariada

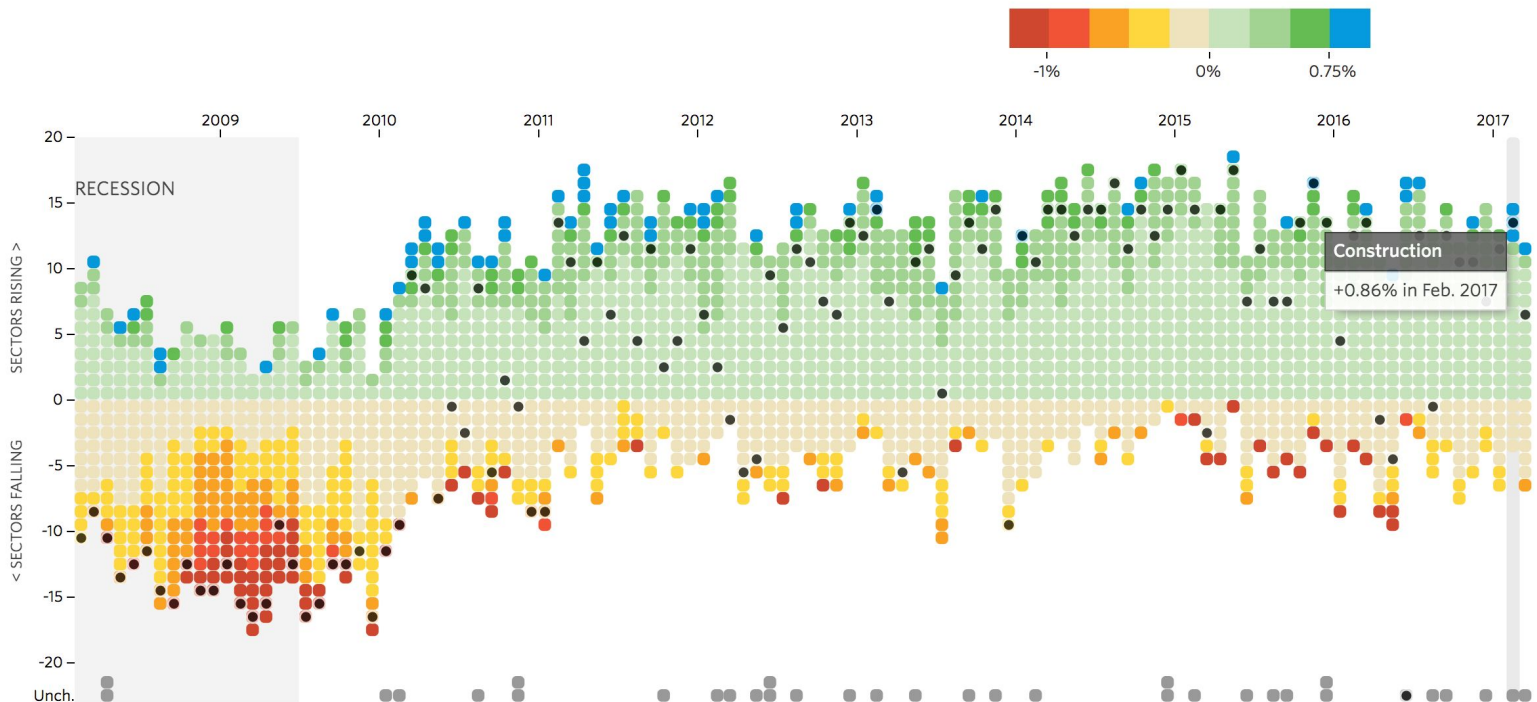
Exemplos

Track National Unemployment, Job Gains and Job Losses

By [Andrew Van Dam](#) and [Renee Lightner](#)

Winners and Losers: Job Gains and Losses [Jump to National Unemployment](#)

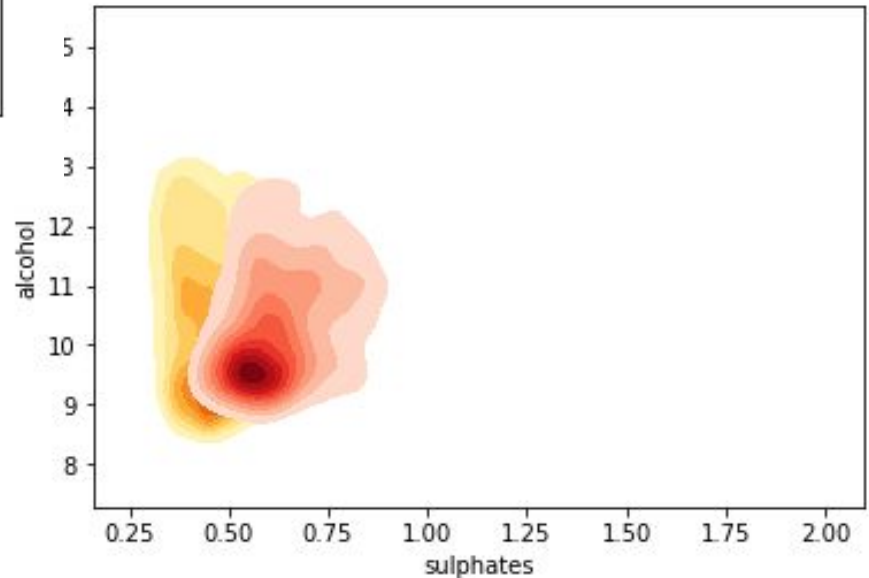
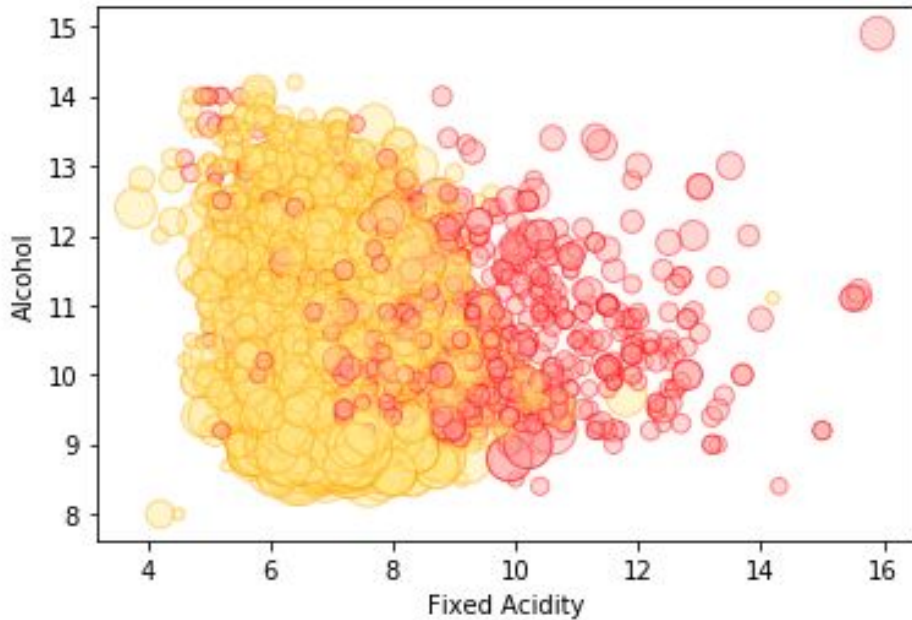
Track the number of sectors gaining or losing jobs each month. Boxes are shaded based on percentage change from the previous month in each sector's payrolls.



Análise multivariada

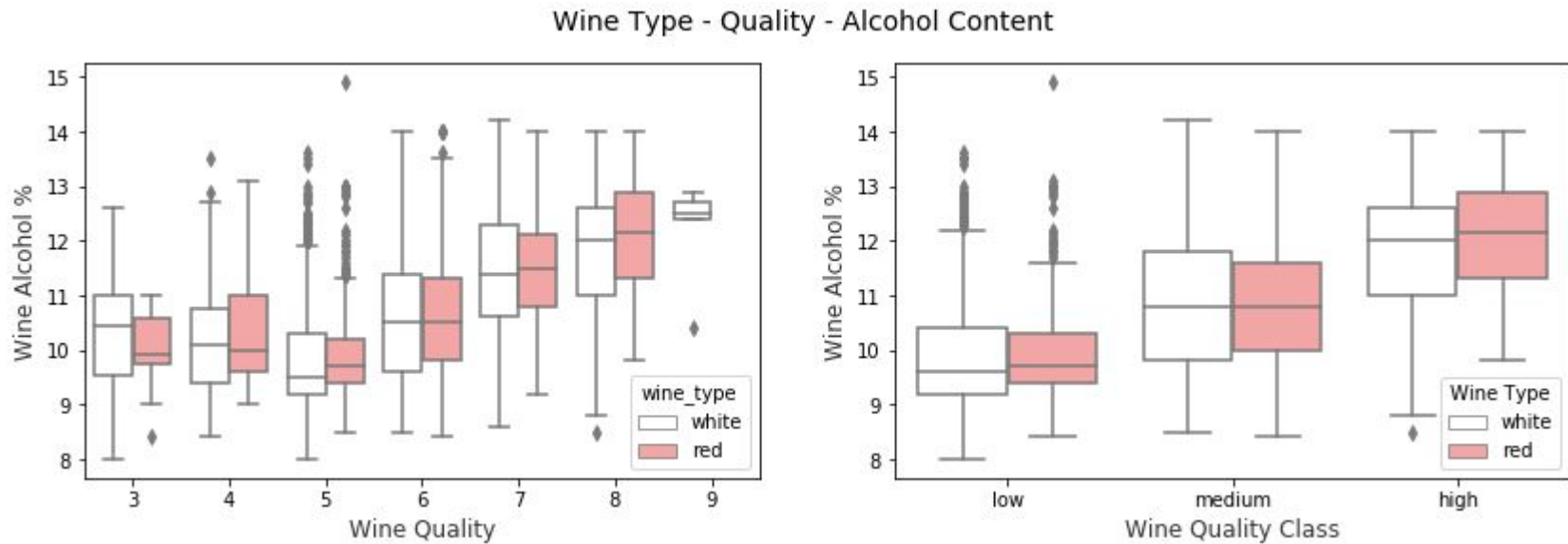
Exemplos

Wine Alcohol Content - Fixed Acidity - Residual Sugar - Type



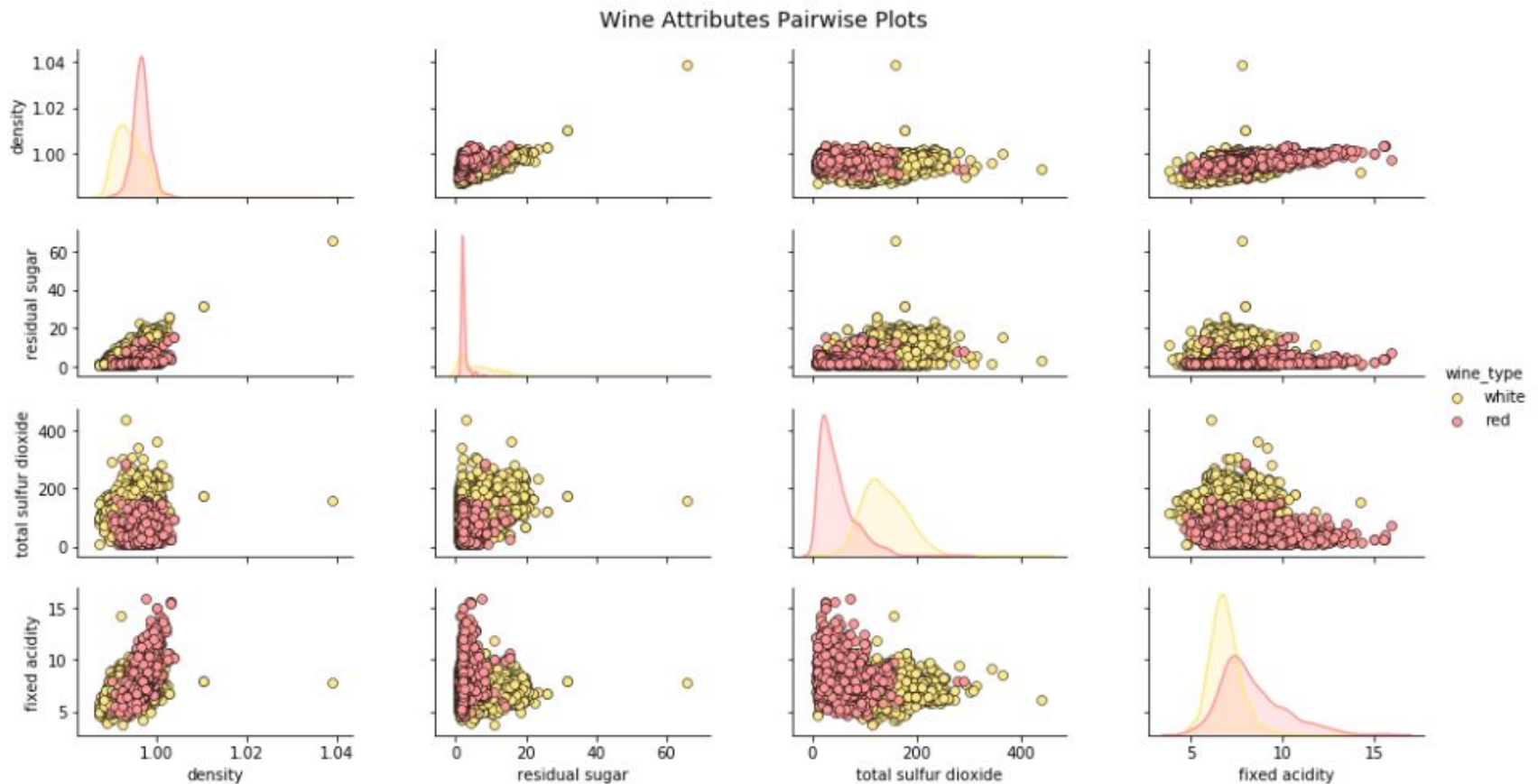
Análise multivariada

Exemplos



Análise multivariada

Exemplos

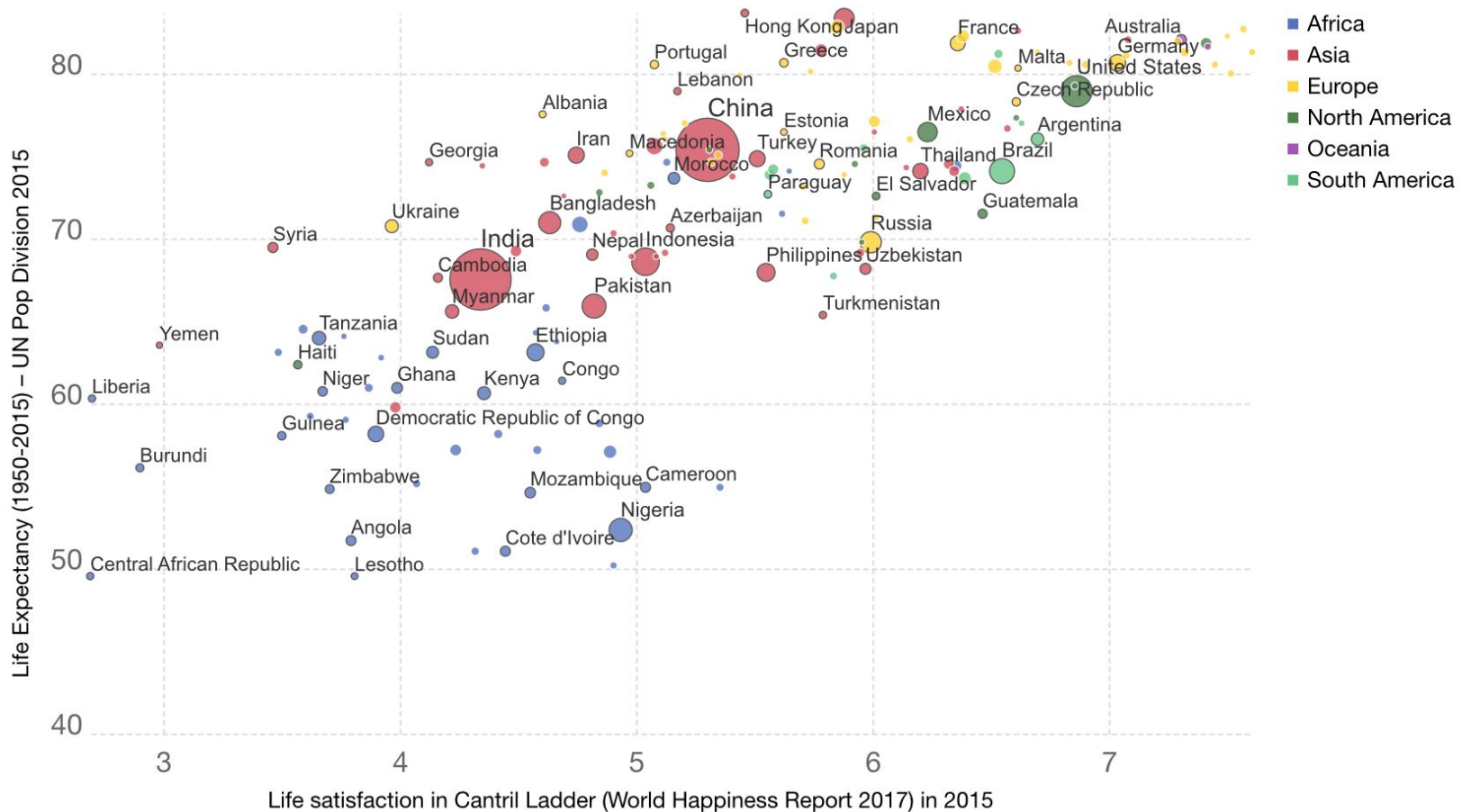


Análise multivariada

Exemplos

Life satisfaction vs Life expectancy, 2015

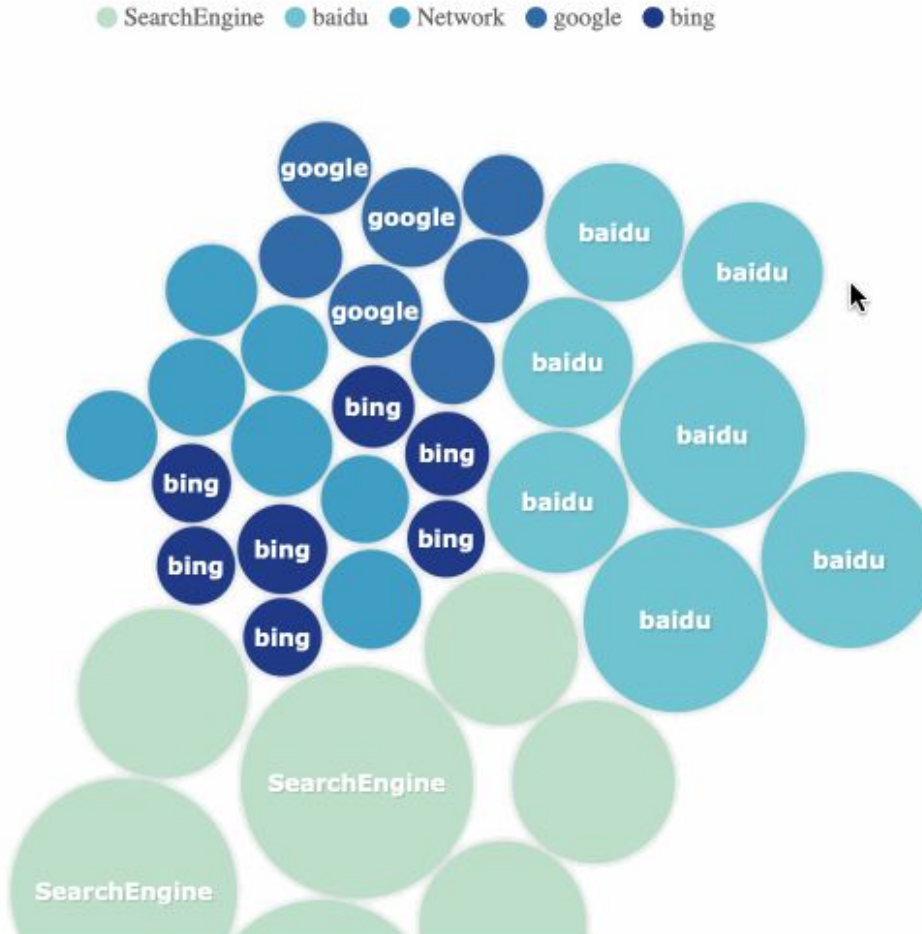
The vertical axis shows life expectancy at birth. The horizontal axis shows self-reported life satisfaction in the Cantril Ladder (0-10 point scale with higher values representing higher life satisfaction).



Análise multivariada

Exemplos

Exemplo 1: Google



Principais técnicas de pré-processamento de dados

1. Identificação de variáveis;
2. Análise de variáveis;
3. **Tratamento de dados faltantes;**
4. Tratamento de outliers;
5. Transformação de atributos;
6. Derivação de atributos;
7. Reamostragem;
8. Redução de dimensionalidade.



Tratamento de dados faltantes

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55		Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57		Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	2	1	50%
M	4	2	50%
Missing	2	2	100%

A partir dos dados, a probabilidade de jogar cricket é maior entre homens do que entre mulheres.

Tratamento de dados faltantes

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55	F	Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57	F	Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	4	3	75%
M	4	2	50%

Mas a realidade era diferente.

Por que existem dados faltantes?



- Erro na consulta que recuperou os dados;
- Erro no procedimento de raspagem;
- Atributo que não era monitorado quando os primeiros dados foram coletados;
- Distração/Recusa ao fornecer uma informação;
- Inexistência para certas instâncias;
- Dificuldade de acesso para certas instâncias;
- Erro ou limitação do equipamento de medição.
- ...

Tratamento de dados faltantes (1/6)



Descarte

Vantagem: evita introdução de erros;

Desvantagem: perda de informação.

Substituir pelo valor anterior ou posterior

Vantagem: não reduz a amostra;

Desvantagem: o valor pode continuar nulo se o valor anterior ou posterior também estiver ausente.

Tratamento de dados faltantes (2/6)




Estimação direta

- Categóricos: moda;
- Contínuos: média ou mediana;
- Temporais: splines.

Vantagem: bons resultados com poucos valores ausentes.

Desvantagem: erro de estimação pode ser acumulativo (casos com muitos valores ausentes).

Tratamento de dados faltantes (3/6)



Estimação direta

- Estimação generalizada
- Estimação com instâncias similares
 - Ex: estimar o peso de uma pessoa A com base no peso médio de outras pessoas do mesmo sexo que A.

Tratamento de dados faltantes (4/6)



Geração de modelos preditivos

Vantagem: é a estratégia mais robusta para lidar com dados ausentes;

Desvantagens:

- Os valores estimados do modelo são geralmente melhor comportados do que os valores reais;
- Os atributos do conjunto de dados pode não ter relação com o atributo com valores ausentes;
- Custo: é preciso gerar um modelo para cada atributo com valores ausentes.

Tratamento de dados faltantes (5/6)




Estimação com o kNN

Vantagens:

- O kNN pode prever atributos qualitativos e quantitativos;
- Não precisa criar um modelo preditivo para cada atributo;
- Eficiente em atributos com vários dados faltantes.

Tratamento de dados faltantes (6/6)



Estimação com o kNN

Desvantagens:

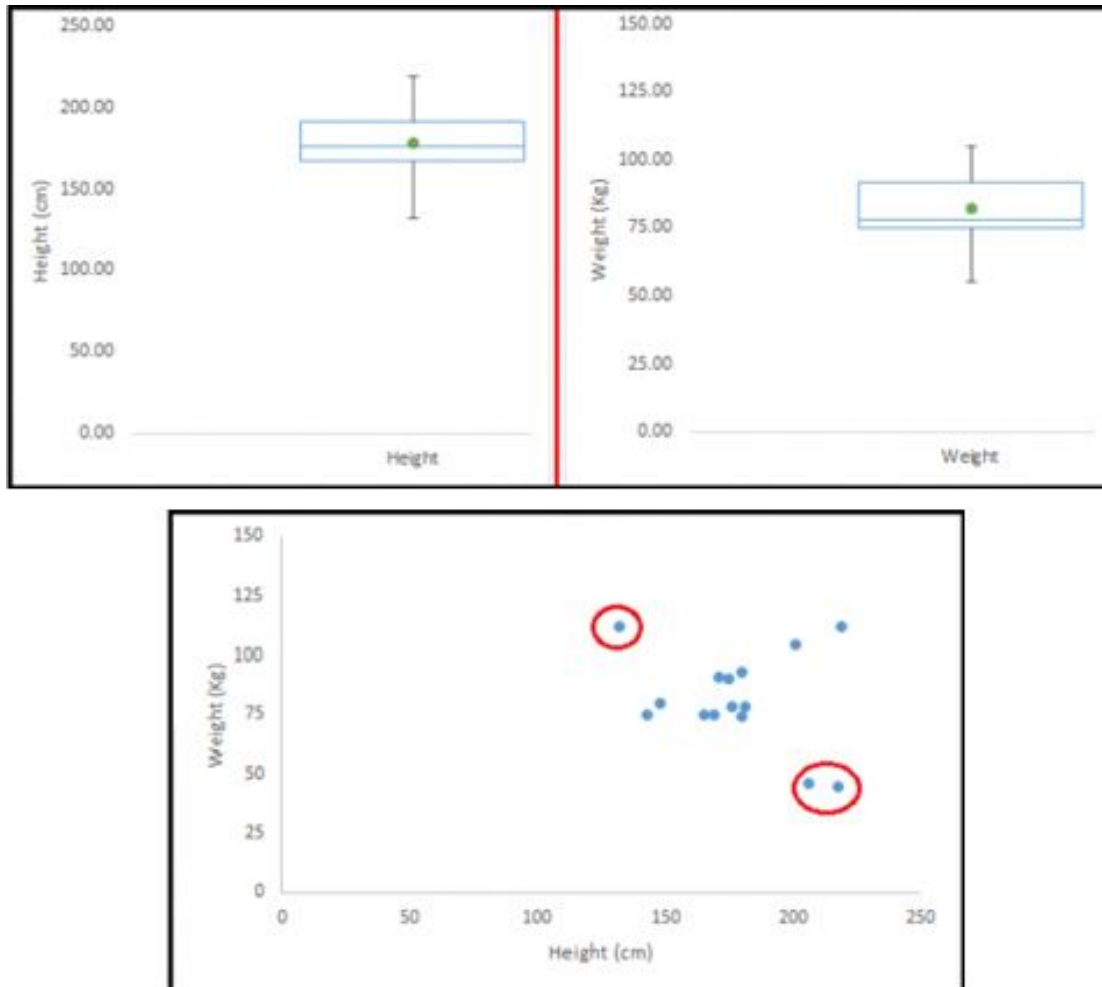
- O algoritmo kNN consome muito tempo na análise em grandes bancos de dados;
- Como definir o k?

Principais técnicas de pré-processamento de dados

1. Identificação de variáveis;
2. Análise de variáveis;
3. Tratamento de dados faltantes;
4. **Tratamento de outliers;**
5. Transformação de atributos;
6. Derivação de atributos;
7. Reamostragem;
8. Redução de dimensionalidade.

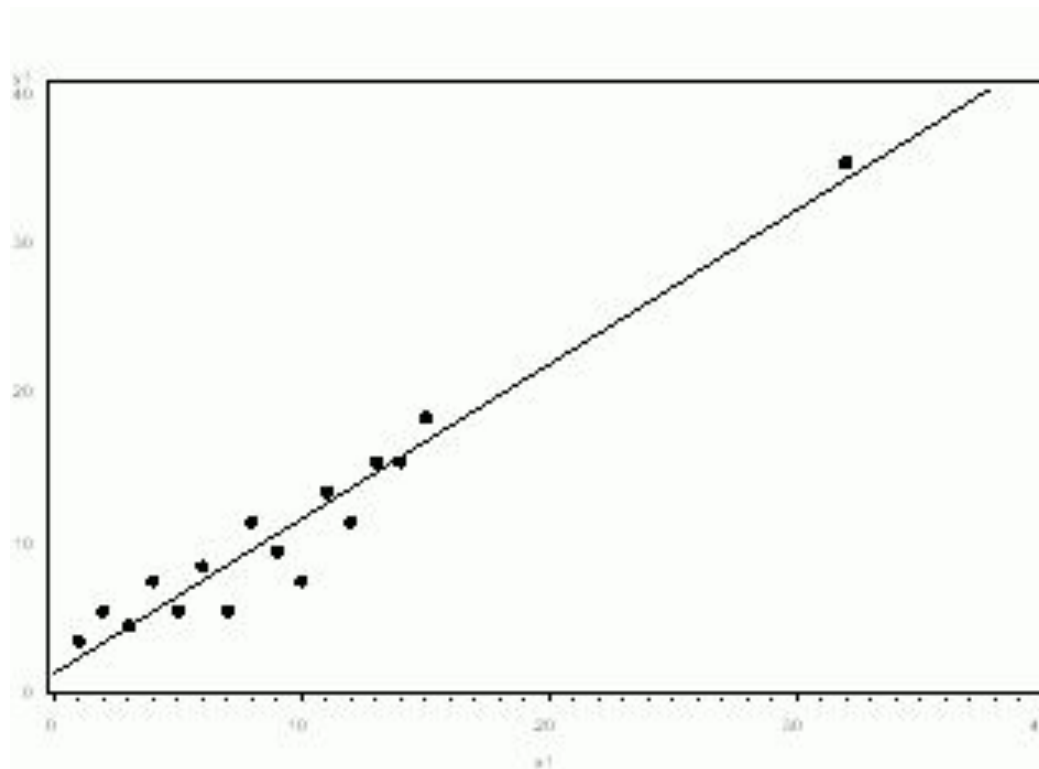


Outlier univariado e bivariado



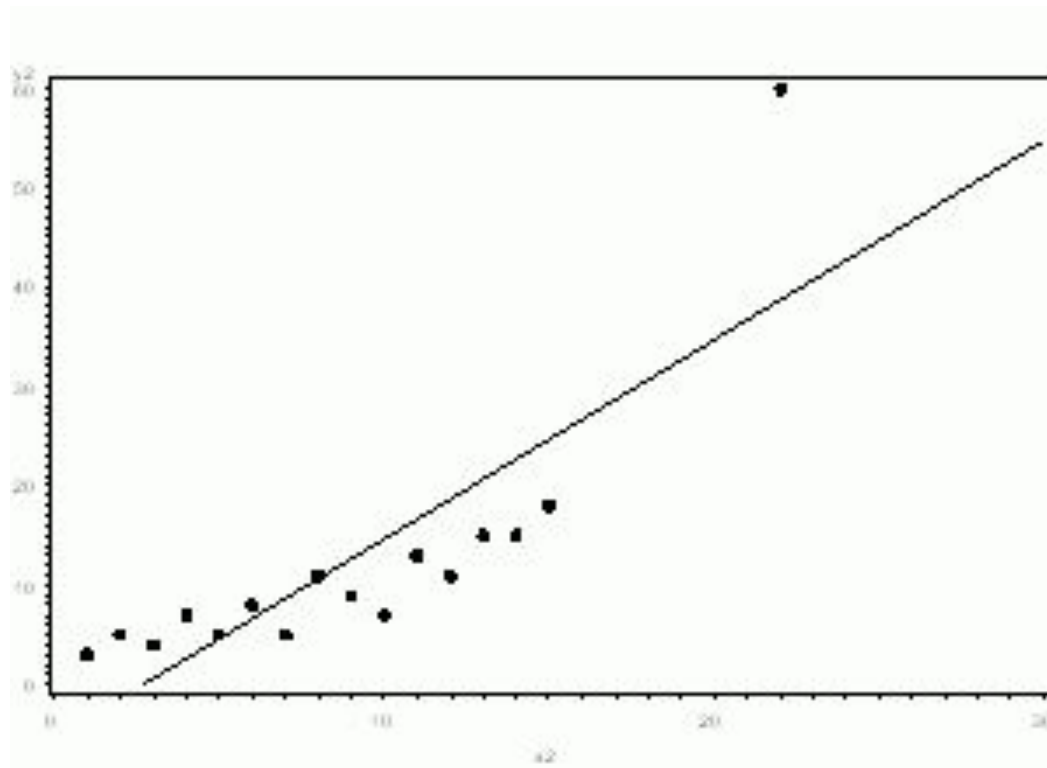
Problemas (1/3)

O outlier não afeta os padrões extraídos, mas afeta a percepção sobre a instância e o atributo.



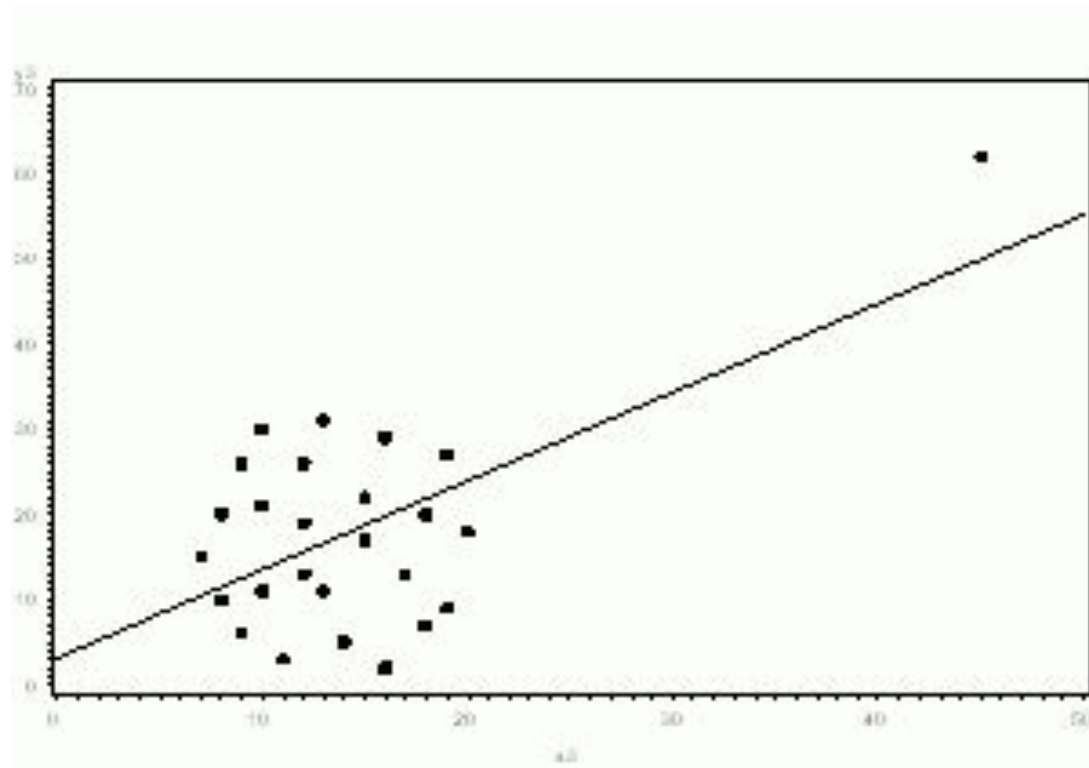
Problemas (2/3)

O outlier afeta os padrões extraídos e a percepção sobre a instância e o atributo.



Problemas (3/3)

O outlier cria um padrão inexistente.



Tipos de outliers (1/3)



Outlier natural

- Ex.: numa pesquisa com o salário de funcionários de uma empresa, é natural que uma mínima parte dos participantes possuam salários muito maiores que as demais.

Erro de amostragem

- Ex.: em um experimento relacionado a altura de atletas, apenas uma pequena parcela dos participantes eram jogadores de basquete.

Tipos de outliers (2/3)



Erros de entrada de dados

- Ex.: pessoa registrada com 680kg ao invés de 60,8kg.

Instrumento de medição com defeito

- Ex.: sensores descalibrados.

Erro experimental

- Ex: participante inicia um experimento após os demais.

Tipos de outliers (1/3)



Outlier intencional

- Ex.: pessoas que mentem num questionário.

Erro na obtenção dos dados

- Ex.: problema na rotina de consulta / integração dos dados.

Como tratar outliers?

Tratamento de outliers



Alterar o valor

Se for um erro cuja correção é trivial, basta corrigi-lo diretamente.

- Ex.: dado categórico com erro de digitação.

Quando a instância é necessária, basta estimar um outro valor (tratar como um dado faltante)

- Situações onde o valor está errado, mas não se sabe o correto;
- Situações onde o valor está correto, mas afeta negativamente o aprendizado.

Tratamento de outliers



Deletar a instância inteira (1/2)

Se a remoção não afetará a mineração

- Ex.: uma instância muito similar a diversas outras em uma coleção grande.

Se vários atributos da instância são problemáticos.

- Ex.: respondente não levou o questionário a sério.

Tratamento de outliers



Deletar a instância inteira (2/2)

Se a instância representa um exemplo que não deveria fazer parte da base

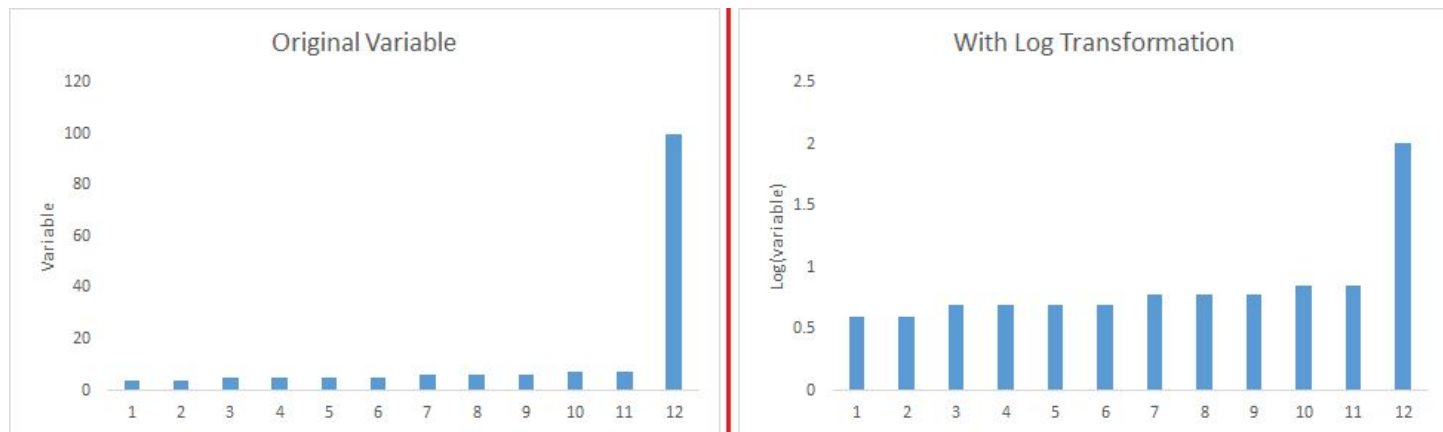
- Ex.: uma instância referente a uma criança em um experimento voltado a dados sobre adultos.

Trimming: remoção das observações com os menores e maiores valores para um determinado atributo.

Tratamento de outliers

Transformar o atributo (1/2)

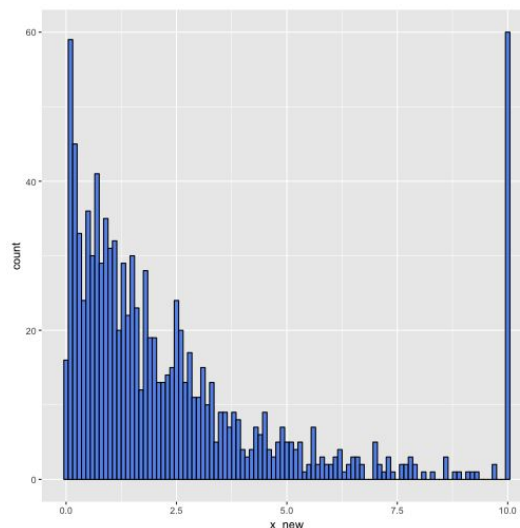
A utilização de log ou raiz quadrada / cúbica mantém a distribuição dos valores originais, mas reduz a discrepância entre eles.



Tratamento de outliers

Transformar o atributo (2/2)

Binning: ao discretizar um atributo contínuo, dados anormais podem ser agregados a valores comuns, não afetando o aprendizado.



Tratamento de outliers



Tratar separadamente

Quando o volume de outliers é suficientemente grande, uma alternativa é gerar modelos separados e depois combinar a saída.

Principais técnicas de pré-processamento de dados

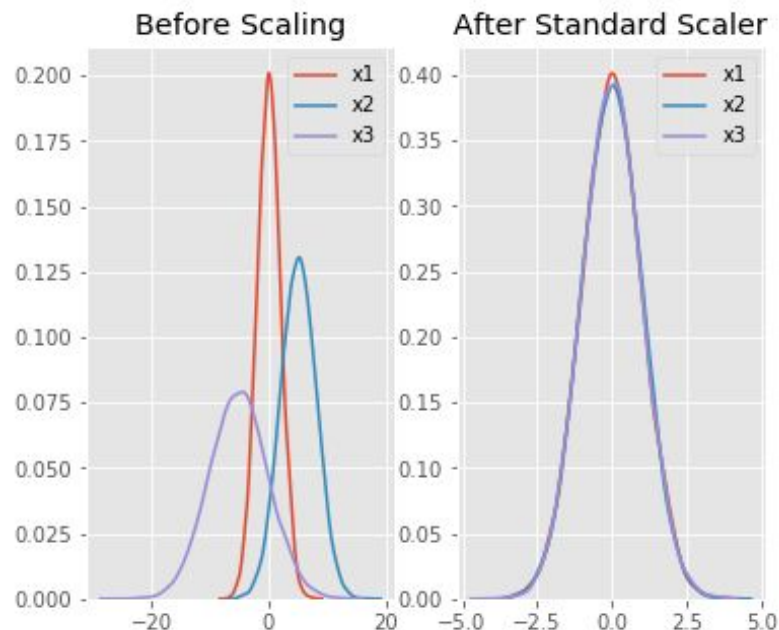
1. Identificação de variáveis;
2. Análise de variáveis;
3. Tratamento de dados faltantes;
4. Tratamento de outliers;
5. **Transformação de atributos;**
6. Derivação de atributos;
7. Reamostragem;
8. Redução de dimensionalidade.



Por que transformar atributos? (1/3)

Mudança de escala

Atributos em diferentes escalas podem prejudicar o processo de aprendizado dos modelos preditivos.

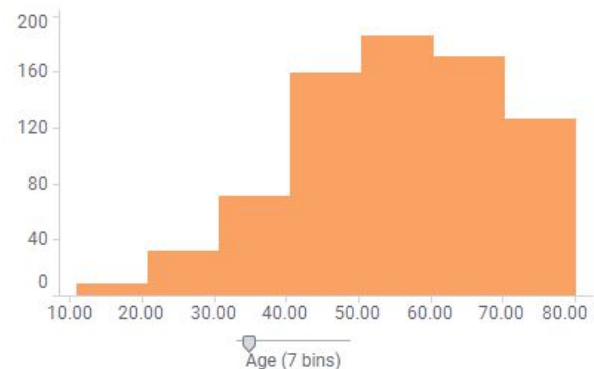
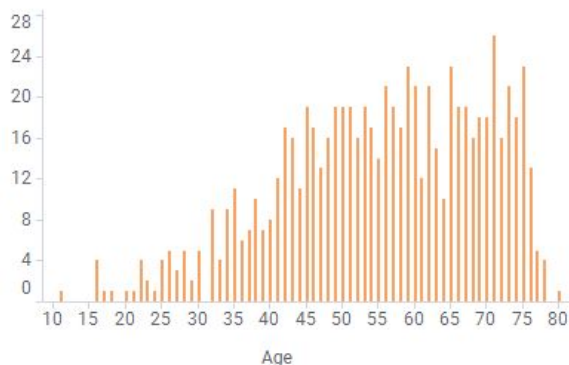


Por que transformar atributos? (2/3)


Utilidade

Muitas vezes valores próximos de um atributo possuem o mesmo valor semântico para o problema em análise.

- Ex.: Dados de idade são geralmente mais úteis quando distribuídos entre faixas (binning de dados).



Por que transformar atributos? (3/3)



Exigência do algoritmo de mineração

Transformação de simbólico para numérico

- Conversão para binário (duas categorias);
- Conversão para inteiro (categóricos ordinais);
- One hot encoding (categóricos não ordinais).

Transformação de numérico para simbólico

- Associar rótulos para cada valor;
- Discretização: média, quantiles, binning..

Principais técnicas de pré-processamento de dados

1. Identificação de variáveis;
2. Análise de variáveis;
3. Tratamento de dados faltantes;
4. Tratamento de outliers;
5. Transformação de atributos;
6. **Derivação de atributos;**
7. Reamostragem;
8. Redução de dimensionalidade.



Por que derivar atributos?



De modo geral, a derivação de atributos tem por objetivo extrair deles a informação mais útil para o problema em análise.

Alguns tipos:

- Composição:
 - Unir dois ou mais atributos para gerar um terceiro com maior utilidade. Ex: peso e altura => IMC.
- Separação:
 - Transformar um atributo em um conjunto de atributos. Ex: data => dia, mês e ano.

Principais técnicas de pré-processamento de dados

1. Identificação de variáveis;
2. Análise de variáveis;
3. Tratamento de dados faltantes;
4. Tratamento de outliers;
5. Transformação de atributos;
6. Derivação de atributos;
7. **Reamostragem;**
8. Redução de dimensionalidade.



Quando utilizar reamostragem?



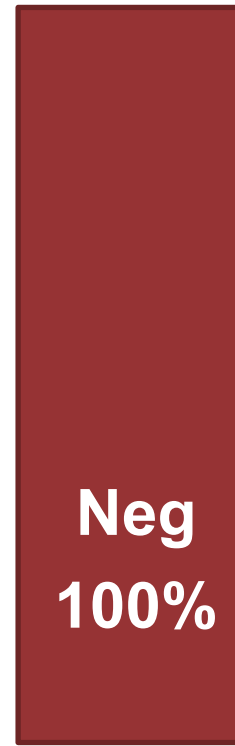
Desbalanceamento

Problemas (1/4)

Viés da acurácia



Real



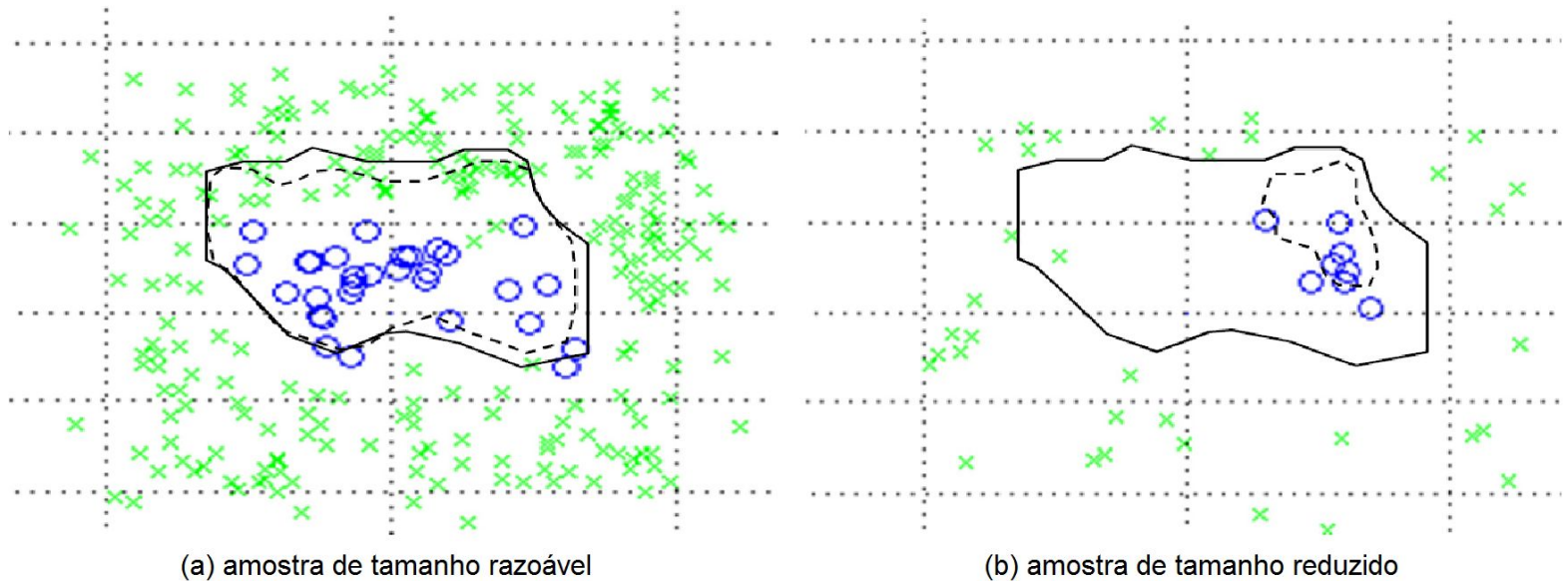
Predito

Acurácia do
modelo: 0,9

Desbalanceamento

Problemas (2/4)

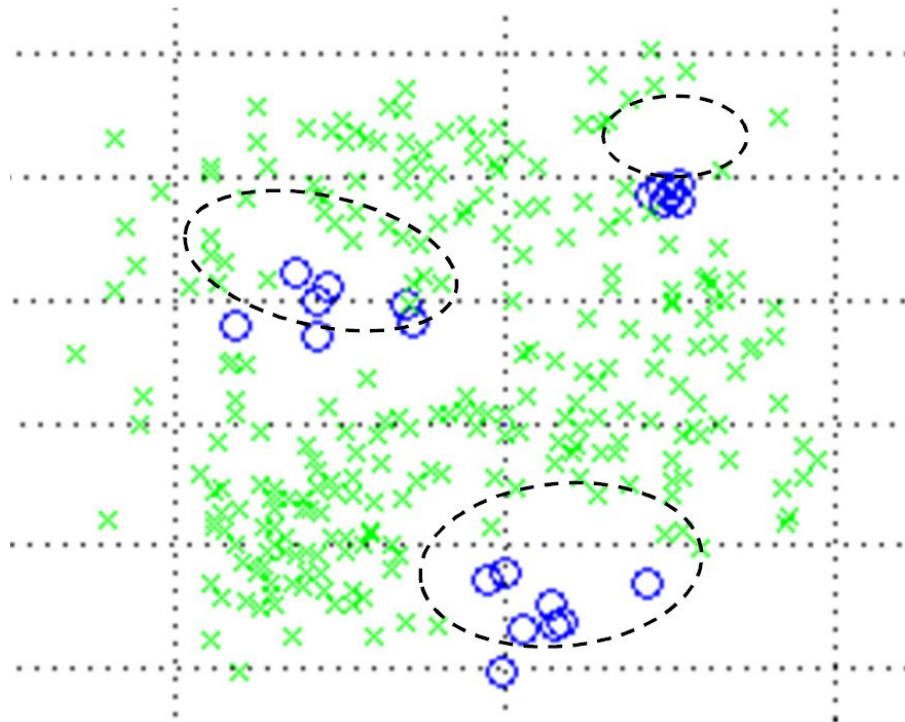
Amostras reduzidas



Desbalanceamento

Problemas (3/4)

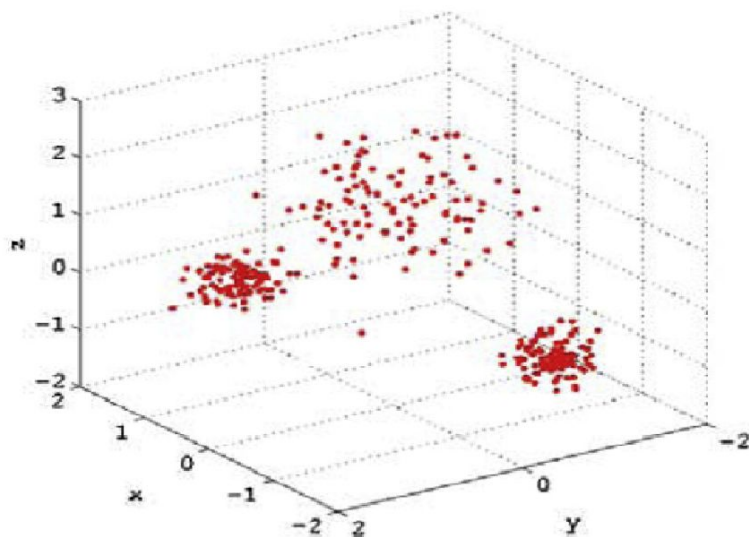
Pequenos disjuntos



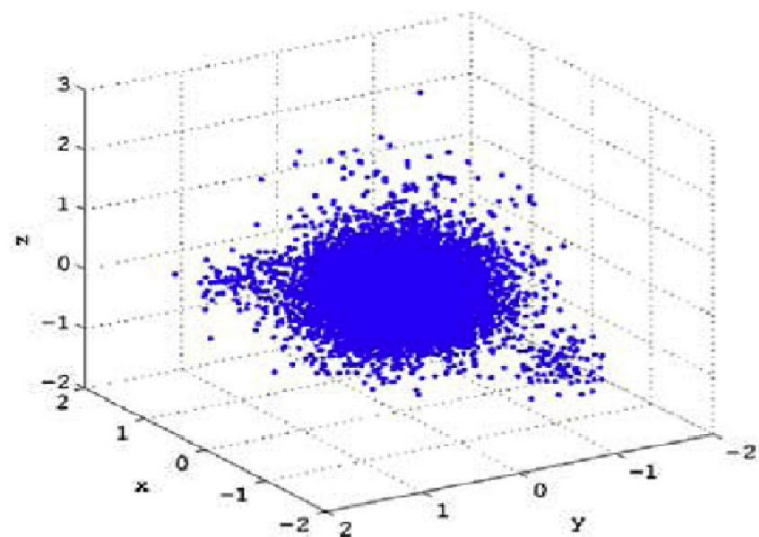
Desbalanceamento

Problemas (4/4)

Sobreposição de classes



(a) classe positiva



(b) classe negativa

Desbalanceamento

Soluções a nível de dados



Reamostragem

Equilibrar a quantidade de instâncias entre as classes

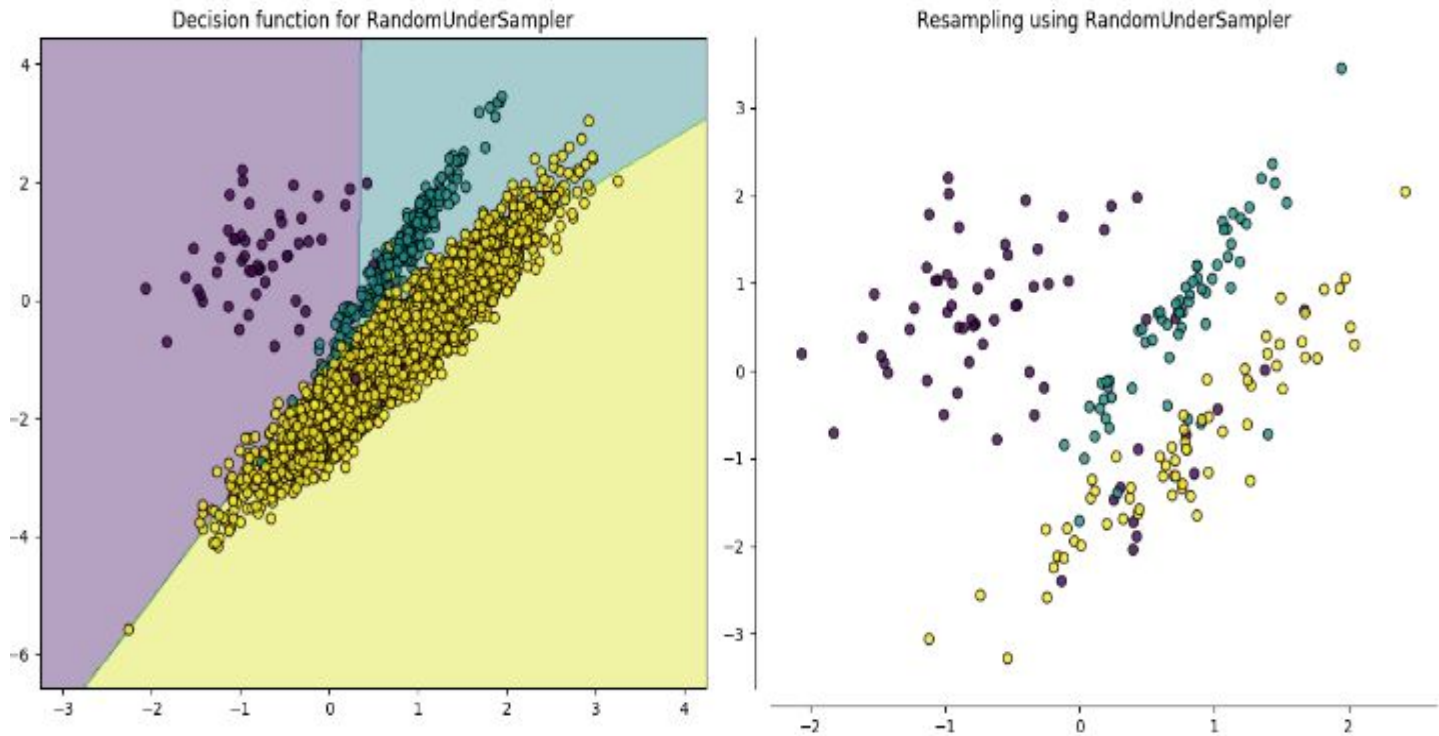
- Sobreamostragem;
- Subamostragem.

Seleção de atributos

Compor o espaço dimensional com atributos que evidenciem os limites de decisão da(s) classe(s) minoritária(s).

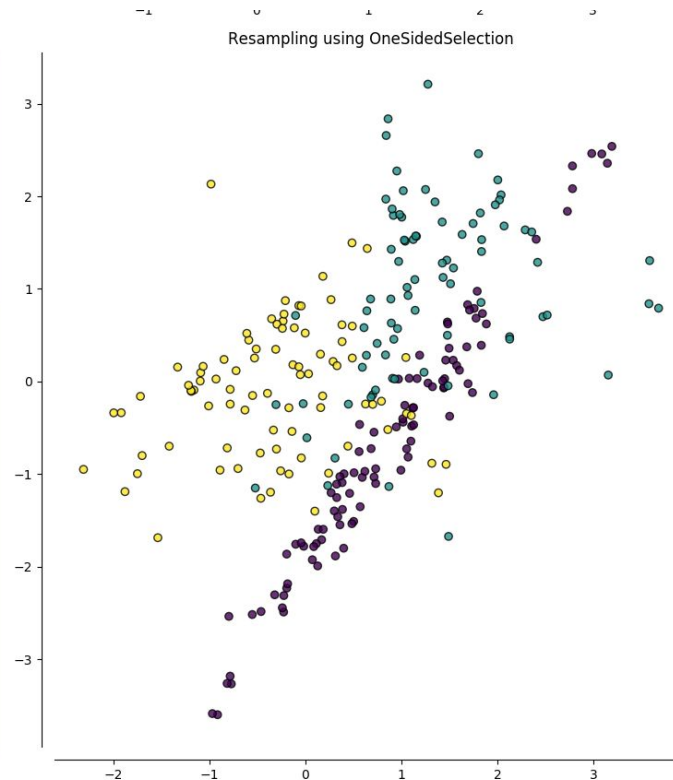
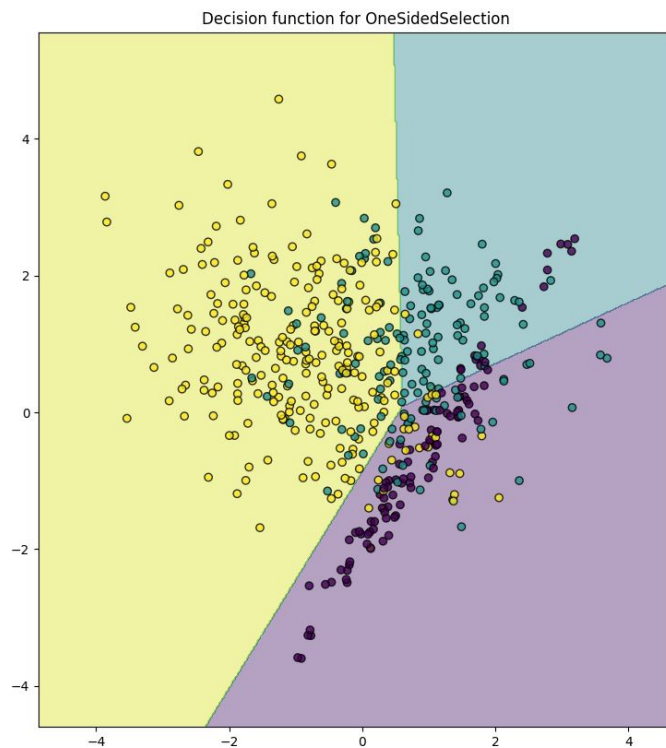
Subamostragem

Random Under Sampler



Subamostragem

One Sided Selection



Subamostragem



Vantagens

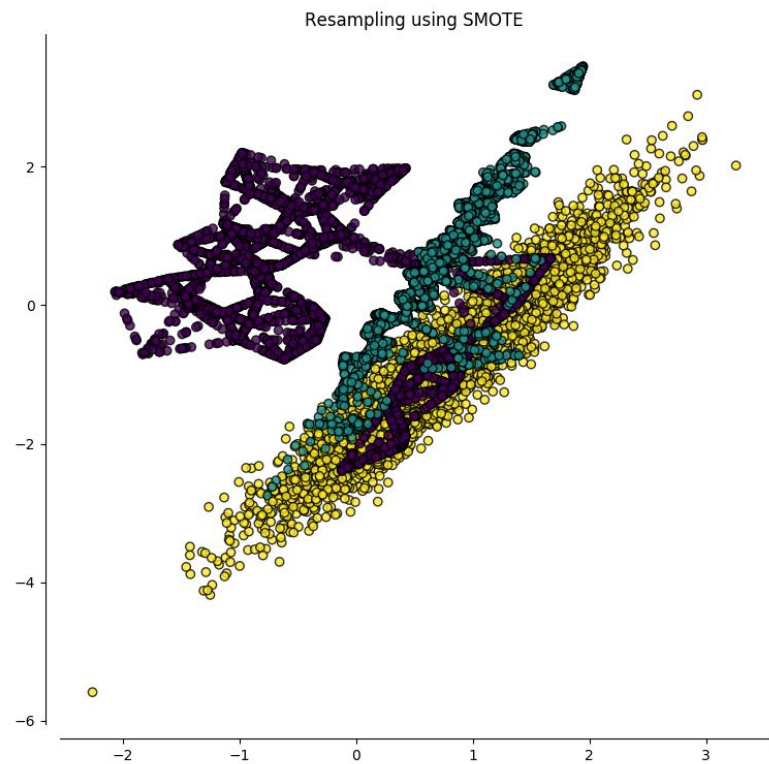
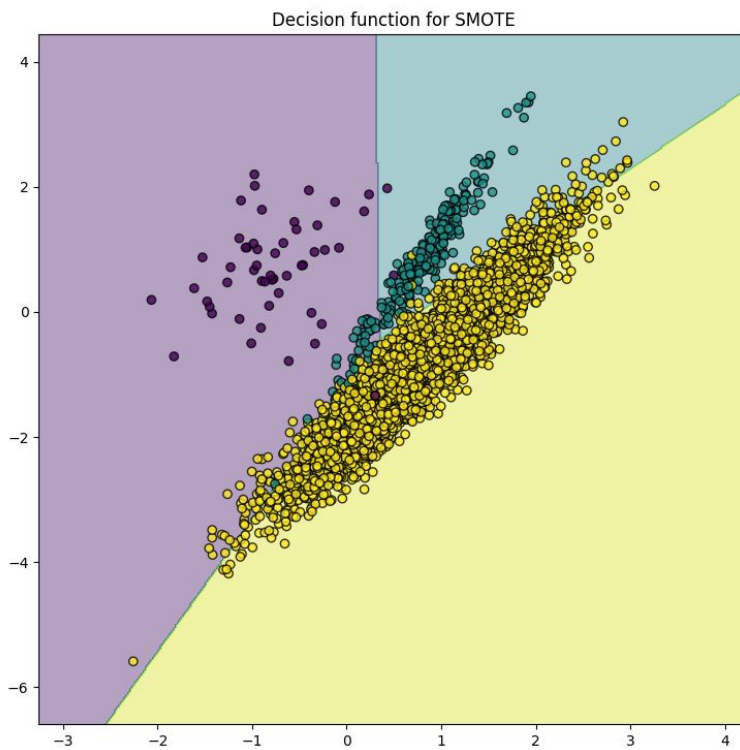
- Redução de custos de armazenamento e processamento;
- Redução dos efeitos do viés indutivo.

Desvantagem

- Perda de informação;
- Os dados da amostra podem se distanciar da distribuição real da população.

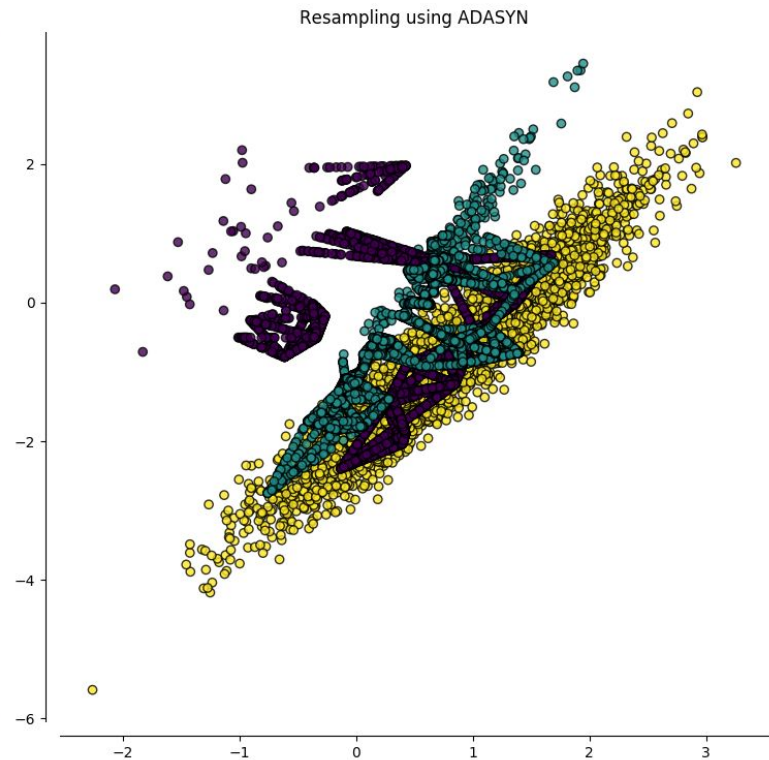
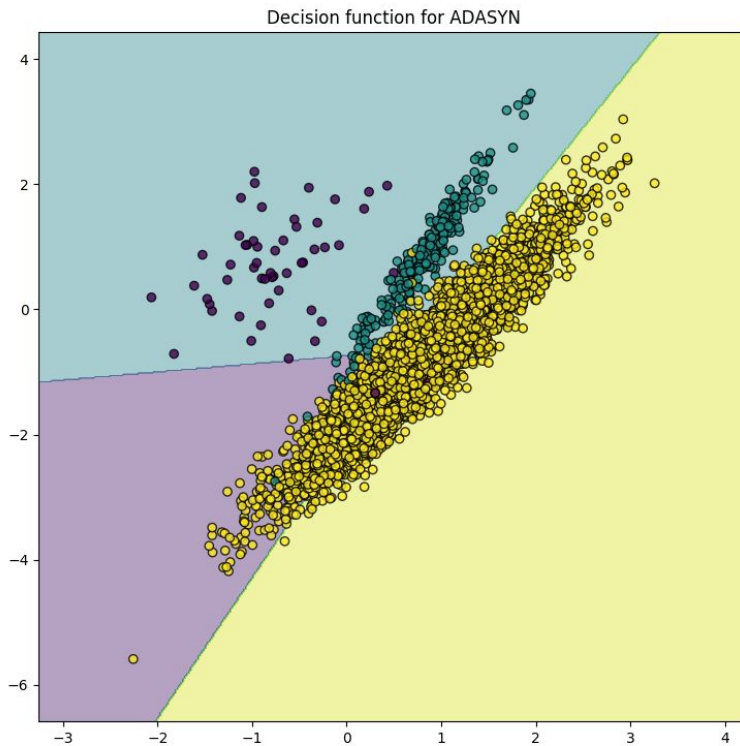
Sobreamostragem

SMOTE



Sobreamostragem

ADASYN



Sobreamostragem



Vantagens

- Redução dos efeitos da maior parte dos problemas afeitos ao desbalanceamento:
 - Viés indutivo, amostras reduzidas e Pequenos disjuntos.

Desvantagens

- Os dados obtidos não são reais e não trazem novidade sobre o comportamento real da classe;
- Aumento do custo de armazenamento e processamento.

Principais técnicas de pré-processamento de dados

1. Identificação de variáveis;
2. Análise de variáveis;
3. Tratamento de dados faltantes;
4. Tratamento de outliers;
5. Transformação de atributos;
6. Derivação de atributos;
7. Reamostragem;
8. **Redução de dimensionalidade.**

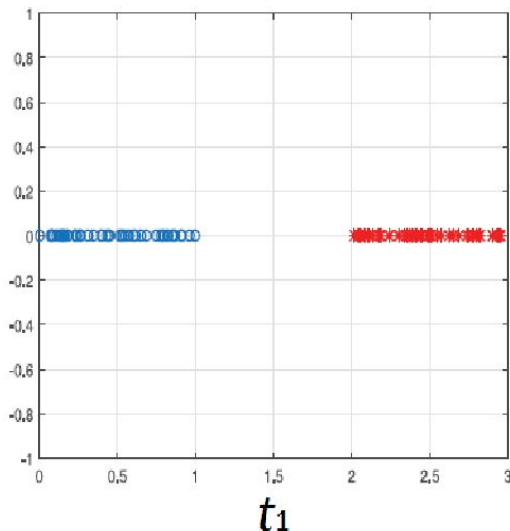


Por que reduzir a dimensionalidade?

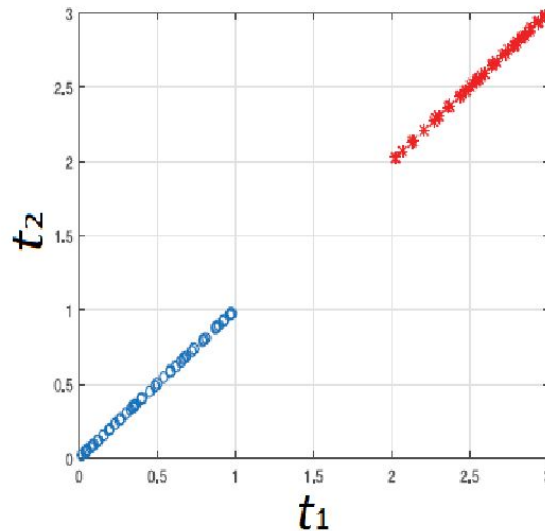
Visualização;

Custo de armazenamento / processamento;

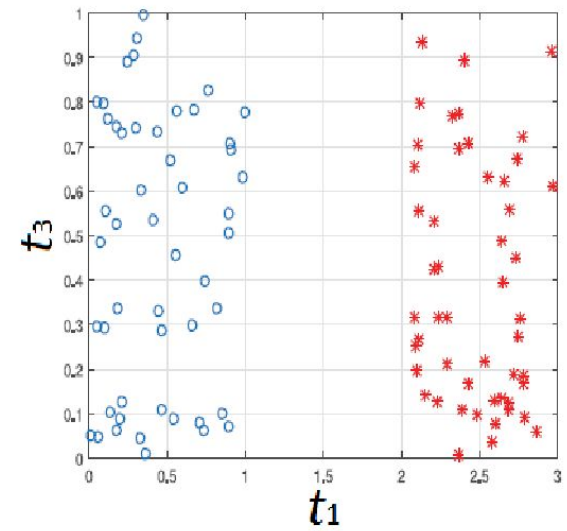
Melhorar o aprendizado.



(a) atributo relevante t_1



(b) atributo redundante t_2



(c) atributo irrelevante t_3

Redução de dimensionalidade



Seleção de atributos

Filtro

- *Term strength, chi-square, information gain...*

Wrapper

- *First Search, Forward Selection, Backward Elimination...*

Embedded

- L1 (LASSO), árvore de decisão...

Extração de atributos

PCA, LDA, LSA...

Seleção de atributos

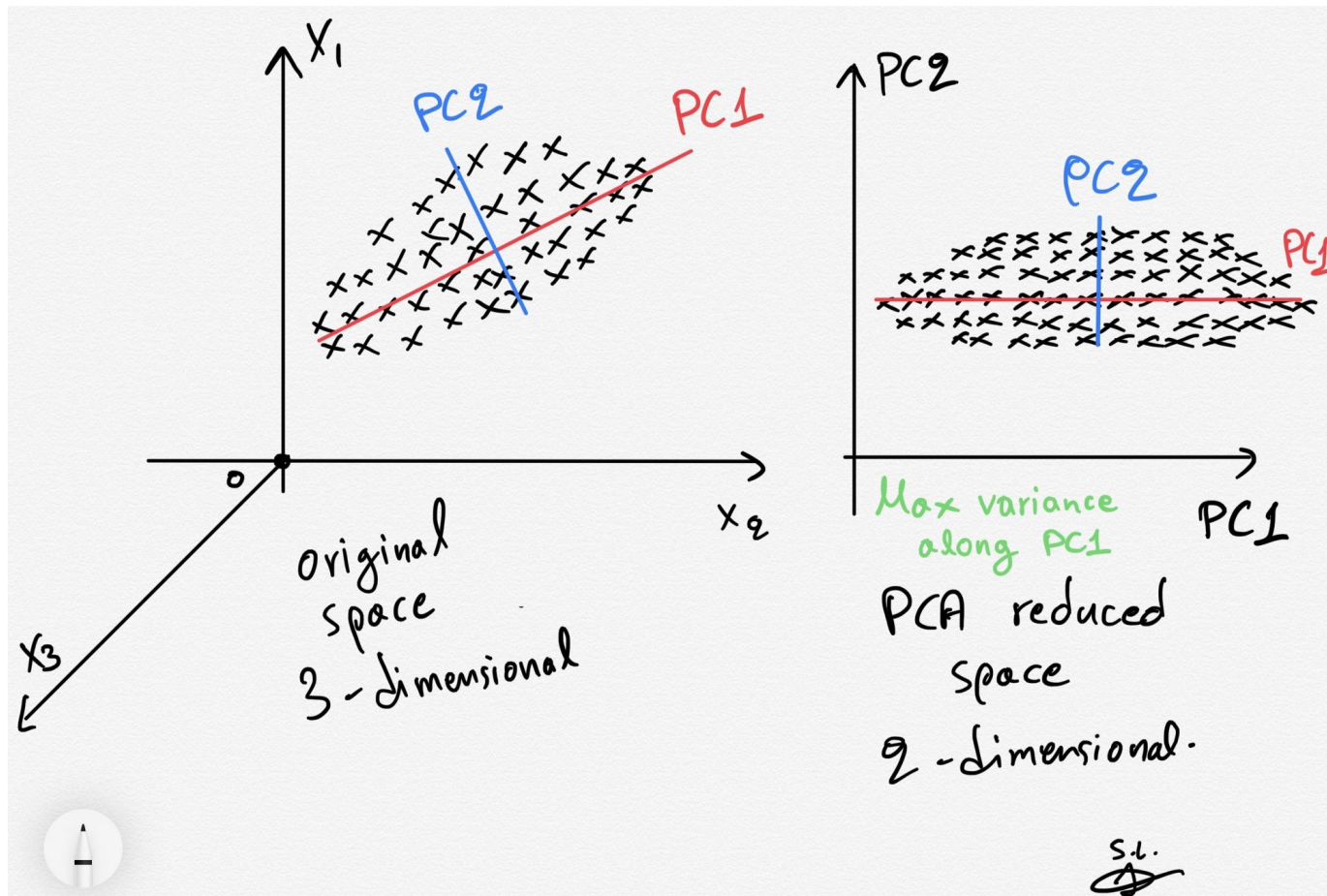
Filtragem

Atributo	Relevância
pênalti	0.92
otimização	0.87
futebol	0.83
árbitro	0.82
software	0.79
computador	0.79
performance	0.62
classificação	0.41

Seleção com N = 4
pênalti
otimização
futebol
árbitro

Extração de atributos

Principal Component Analysis (PCA)



Extração de atributos

Principal Component Analysis (PCA)

```
mtcars.pca <- prcomp(mtcars[,c(1:7,10,11)], center = TRUE,scale. = TRUE)

summary(mtcars.pca)

## Importance of components:
##              PC1      PC2      PC3      PC4      PC5      PC6
## Standard deviation  2.3782 1.4429 0.71008 0.51481 0.42797 0.35184
## Proportion of Variance 0.6284 0.2313 0.05602 0.02945 0.02035 0.01375
## Cumulative Proportion 0.6284 0.8598 0.91581 0.94525 0.96560 0.97936
##              PC7      PC8      PC9
## Standard deviation  0.32413 0.2419 0.14896
## Proportion of Variance 0.01167 0.0065 0.00247
## Cumulative Proportion 0.99103 0.9975 1.00000
```

Extração de atributos

Principal Component Analysis (PCA)

