Aprendizado de Máquina

Curso Tecnologia em Análise e Desenvolvimento de Sistemas

Professor Domingos Napolitano

Mineração de dados

Overview a Descoberta de Conhecimento em Bases de Dados

Motivação

- A informatização dos meios produtivos permitiu a geração de grandes volumes de dados:
  - Transações eletrônicas;
  - Novos equipamentos científicos e industriais para observação e controle;
  - Dispositivos de armazenamento em massa;
- Aproveitamento da informação permite ganho de competitividade: *"conhecimento é poder (e poder = $$!)"*

Motivação

- Os recursos de análise de dados tradicionais são inviáveis para acompanhar esta evolução
- *“Morrendo de sede por conhecimento em um oceano de dados”*

Motivação

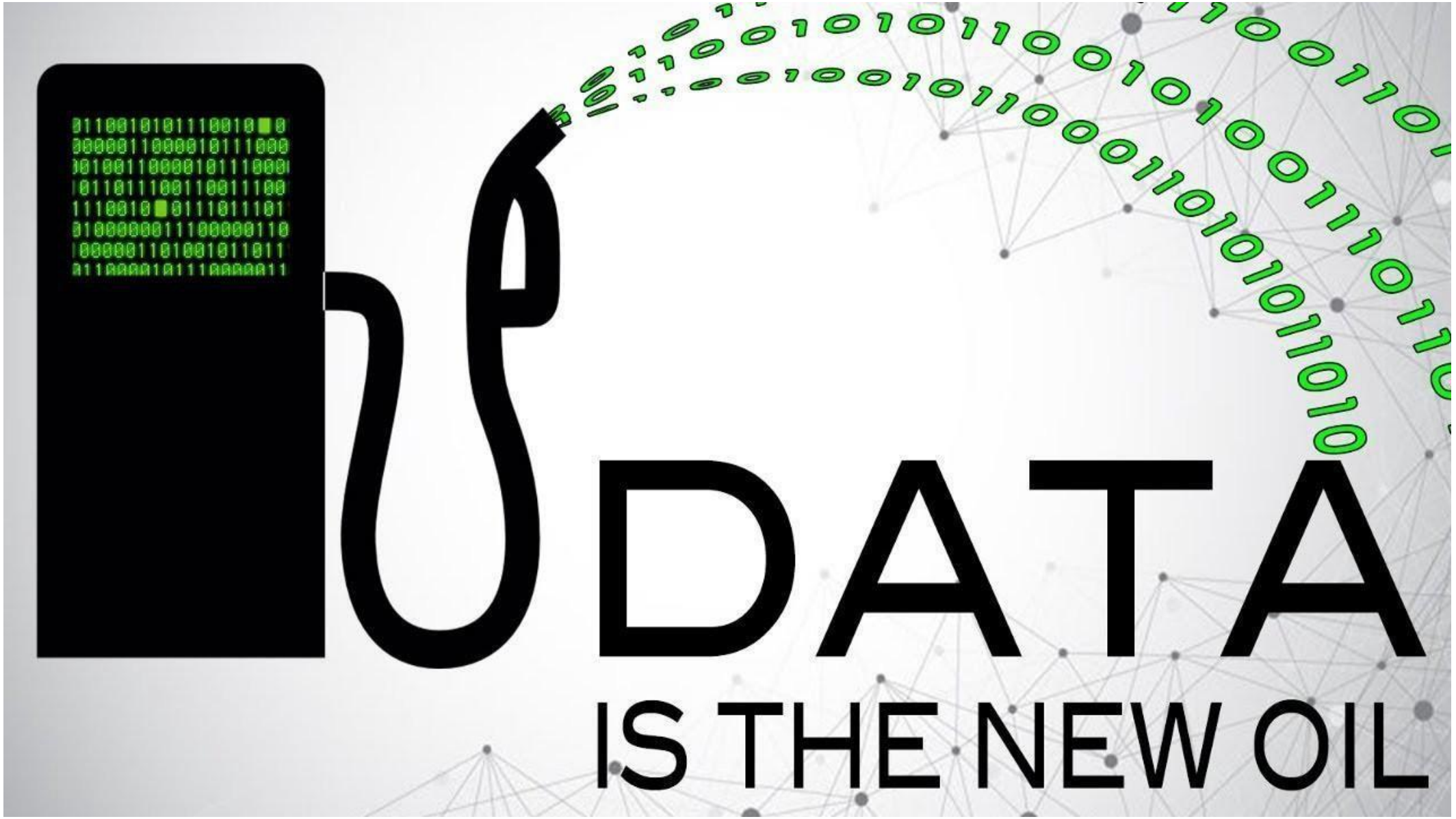
- Solução:
  - ferramentas de automatização das tarefas repetitivas e sistemática de análise de dados
  - ferramentas de auxílio para as tarefas cognitivas da análise
  - integração das ferramentas em sistemas apoiando o processo completo de descoberta de conhecimento para tomada de decisão

Exemplo Preliminar

- Um problema do mundo dos negócios: entender o perfil dos clientes
  - desenvolvimento de novos produtos;
  - controle de estoque em postos de distribuição;
  - propaganda mal direcionada gera maiores gastos e desestimula o possível interessado a procurar as ofertas adequadas;
- Quais são meus clientes típicos?

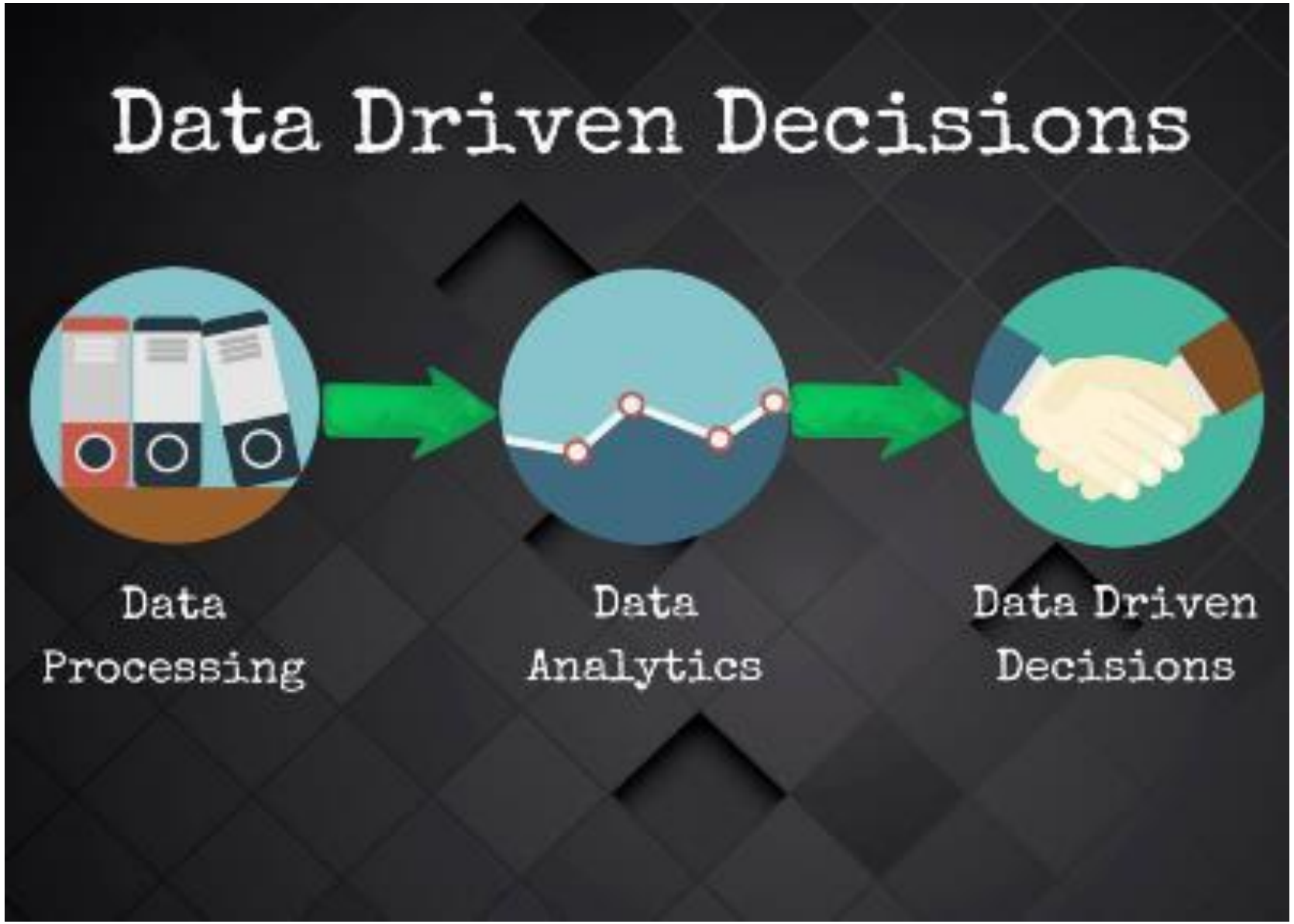
Descoberta de Conhecimento em Bancos de Dados (KDD)

- “O processo não trivial de extração de informações implícitas, anteriormente desconhecidas, e potencialmente úteis de uma fonte de dados”;
- “Torture os dados até eles confessarem”;
- O que é um padrão interessante ?

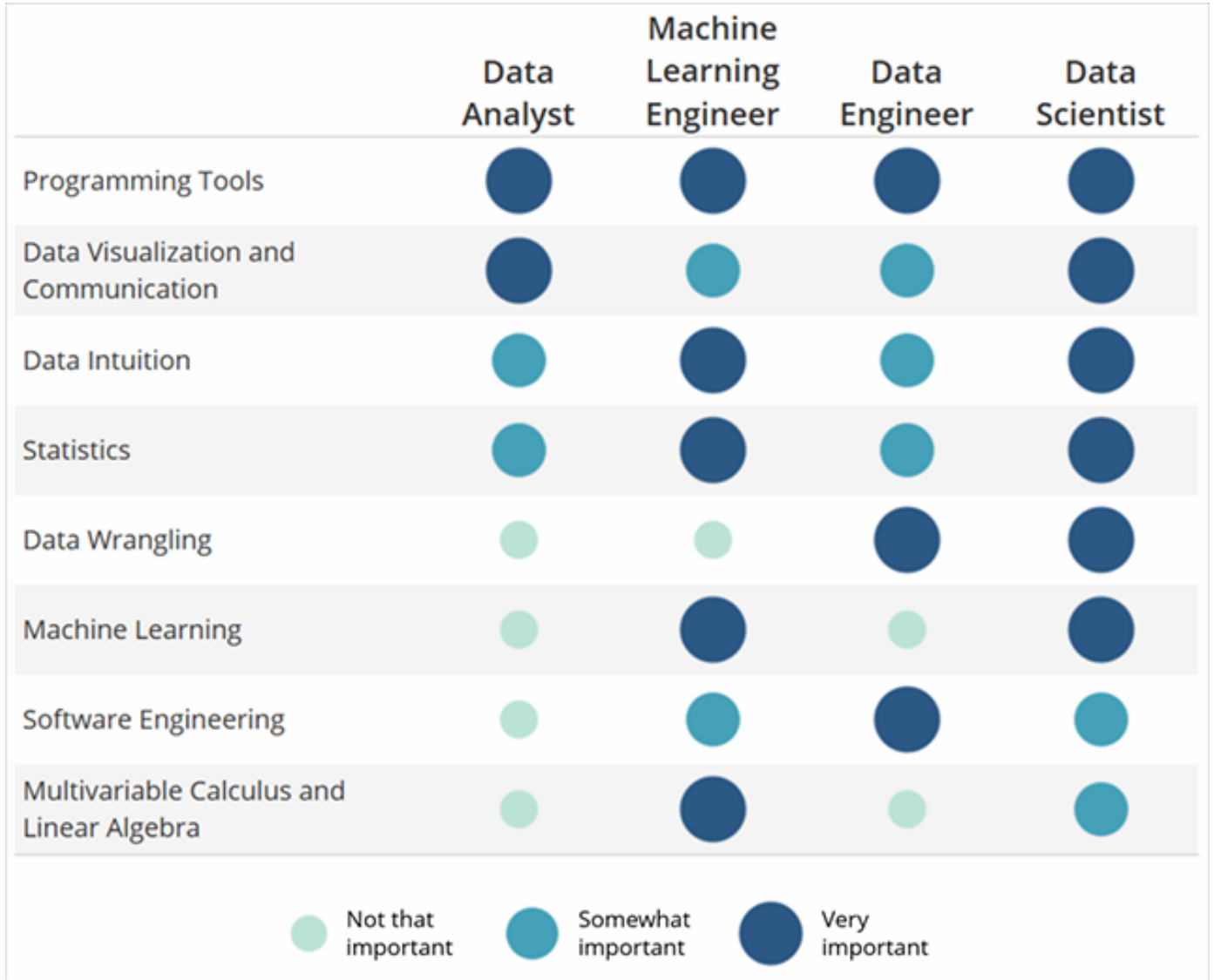


Material extraído da apresentação de Thiago Reis Guia Bolso

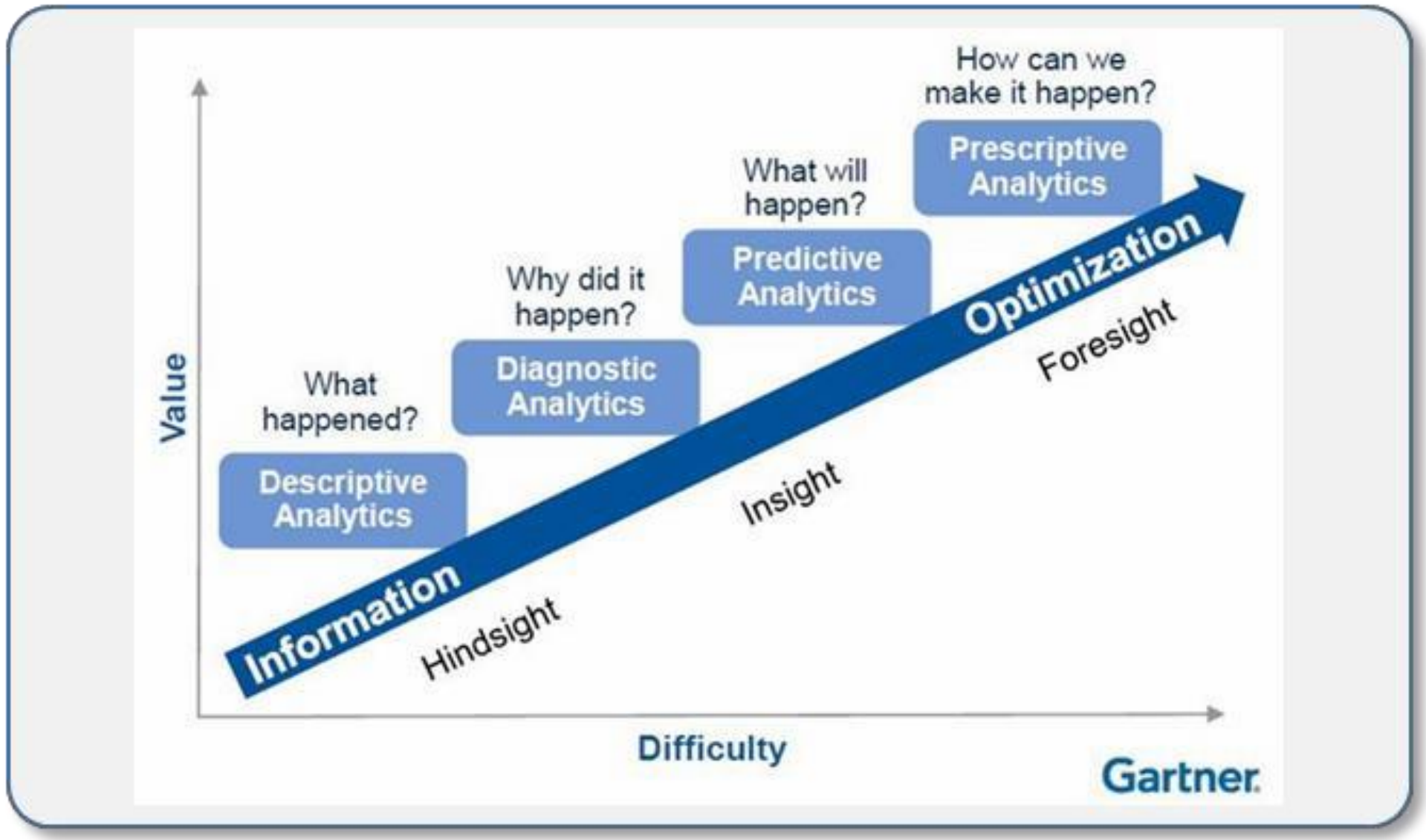
Material extraído da apresentação de Thiago Reis Guia Bolso



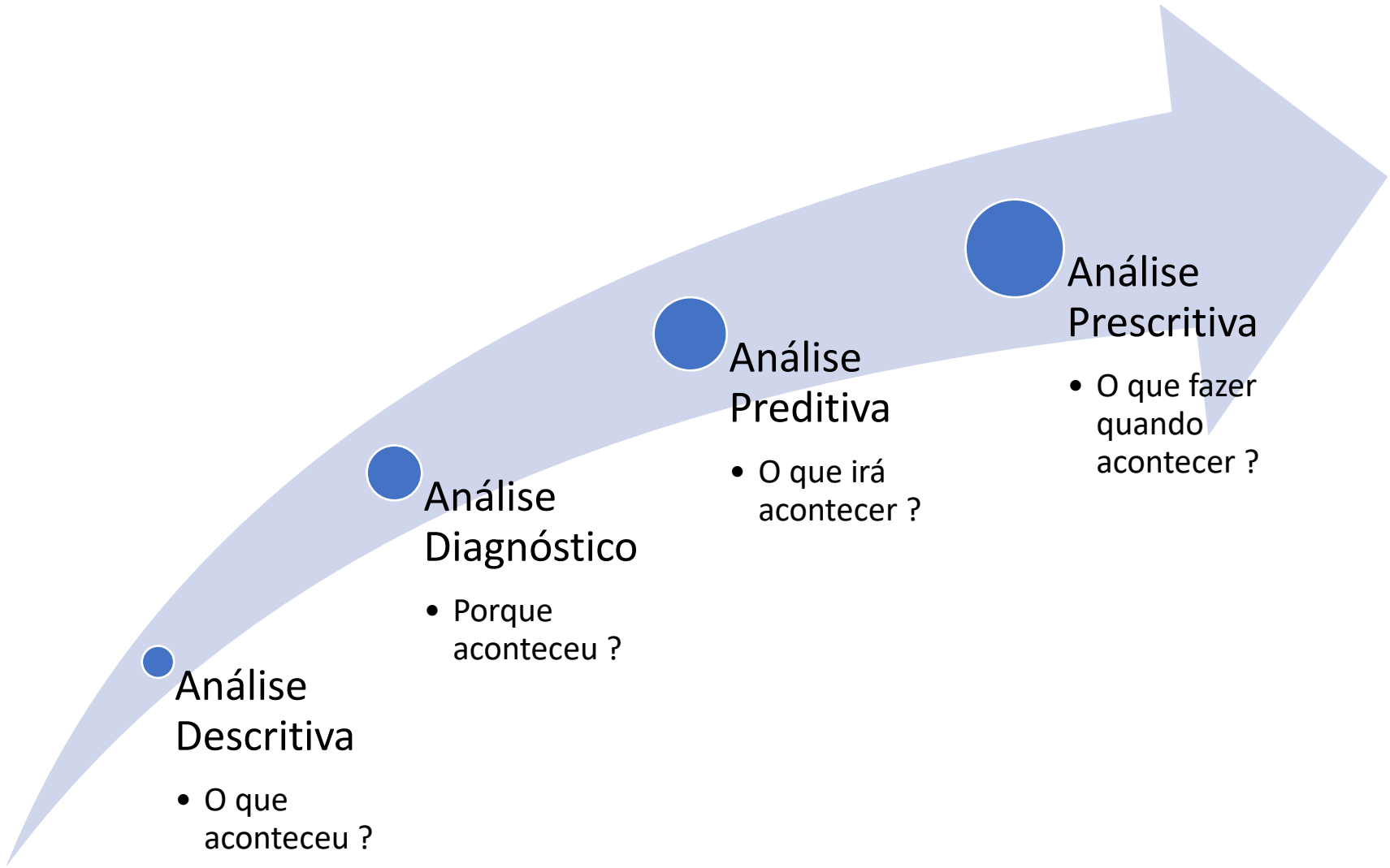
Material extraído da apresentação de Thiago Reis Guia Bolso



Material extraído da apresentação de Thiago Reis Guia Bolso



Material extraído da apresentação de Thiago Reis Guia Bolso


Análise Descritiva
• O que aconteceu ?
Análise Diagnóstico
• Porque aconteceu ?
Análise Preditiva
• O que irá acontecer ?
Análise Prescritiva
• O que fazer quando acontecer ?

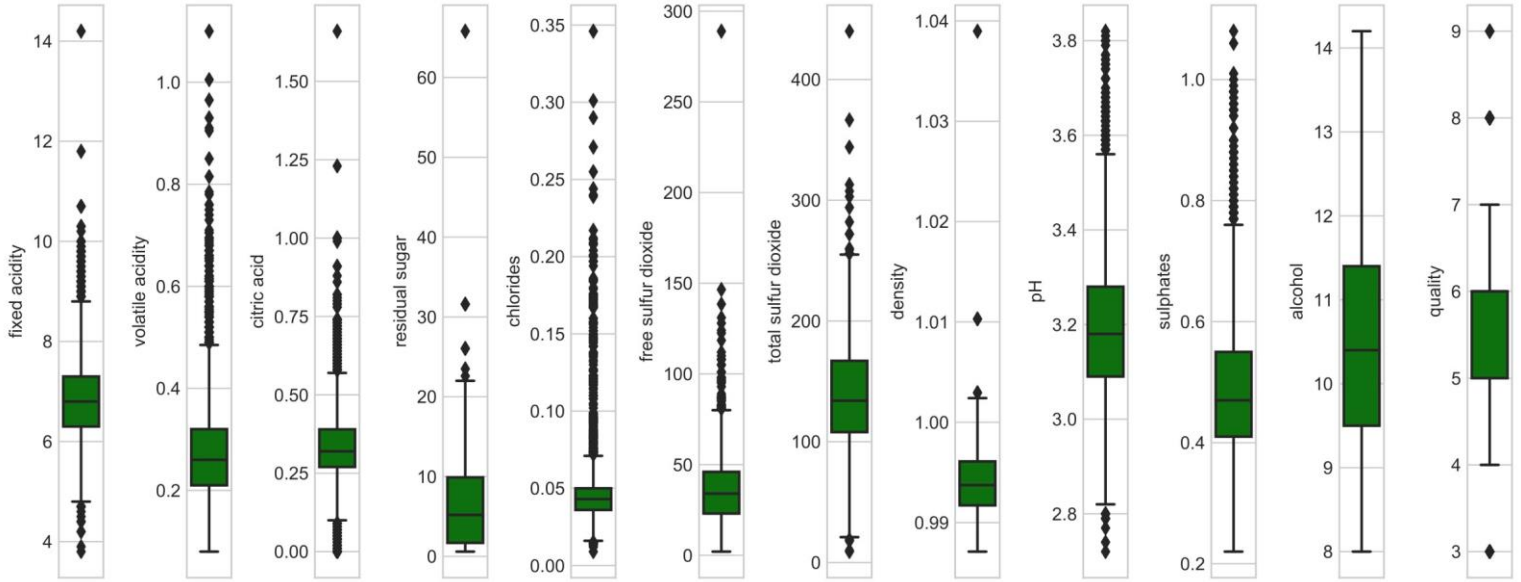
KDD x Data Mining

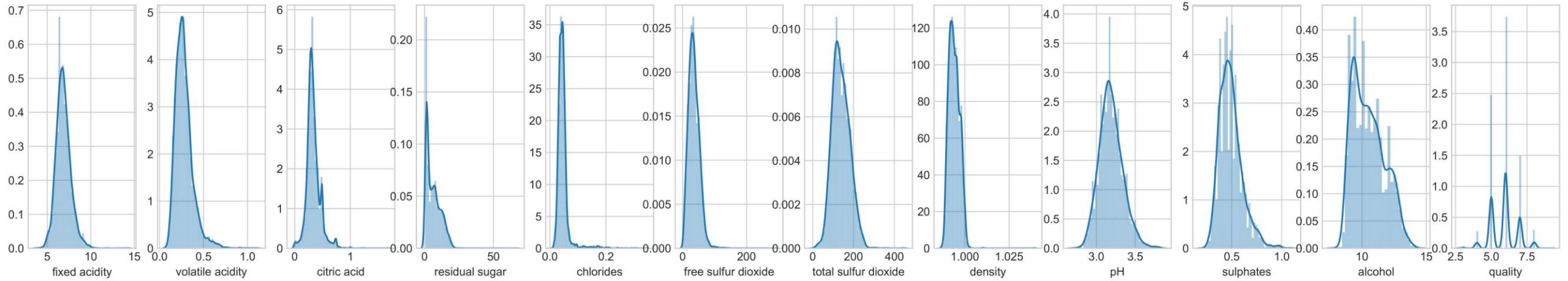
- Mineração de dados é o passo do processo de KDD que produz um conjunto de padrões sob um custo computacional aceitável;
- KDD utiliza algoritmos de *data mining* para extrair padrões classificados como “conhecimento”. Incorpora também tarefas como escolha do algoritmo adequado, processamento e amostragem de dados e interpretação de resultados;

Descrição (Description)

- É a tarefa utilizada para descrever os padrões e tendências revelados pelos dados.
- A descrição geralmente oferece uma possível interpretação para os resultados obtidos.
- A tarefa de descrição é muito utilizada em conjunto com as técnicas de análise exploratória de dados, para comprovar a influência de certas variáveis no resultado obtido.

Imagens DEA – Análise Exploratória de Dados



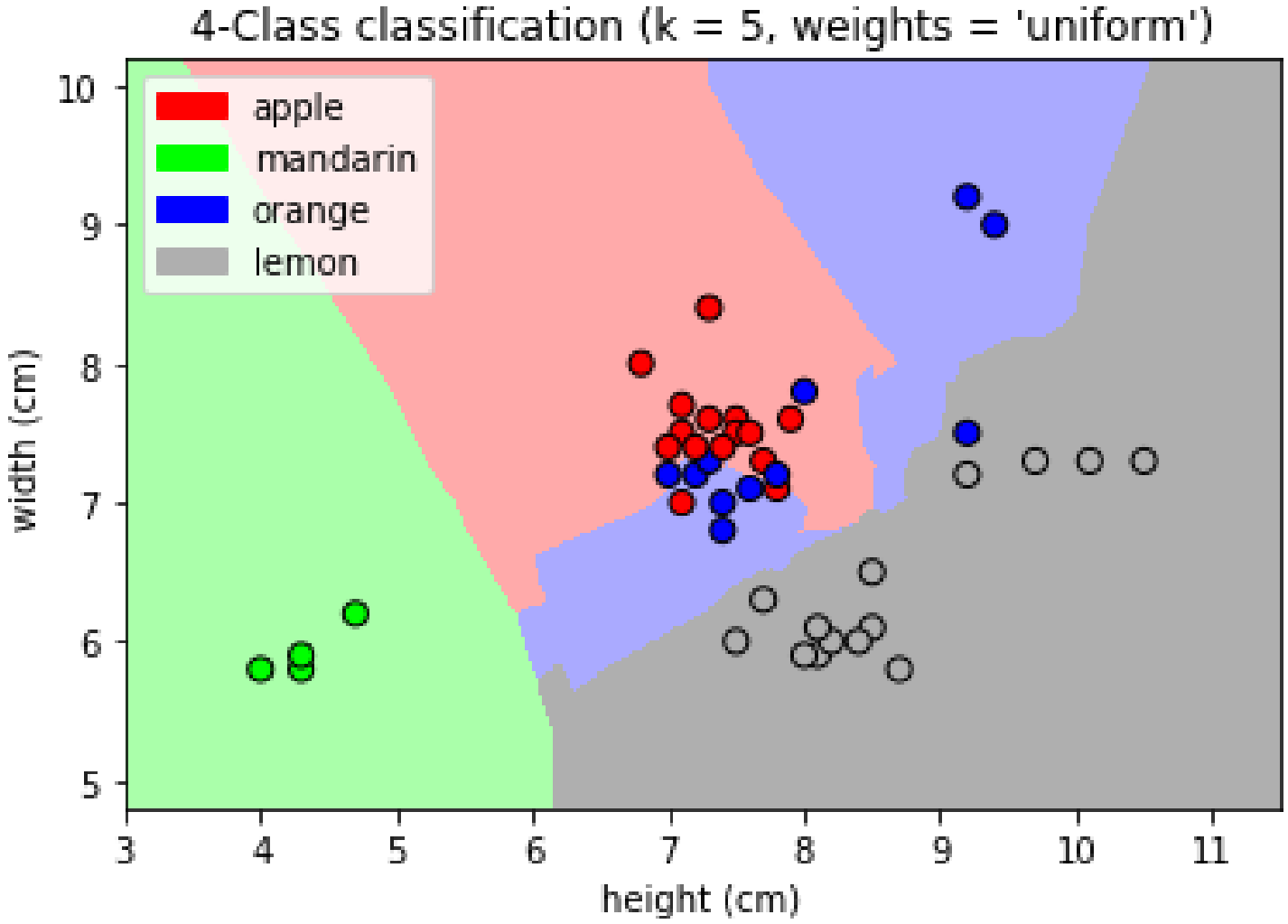


Classificação (Classification)

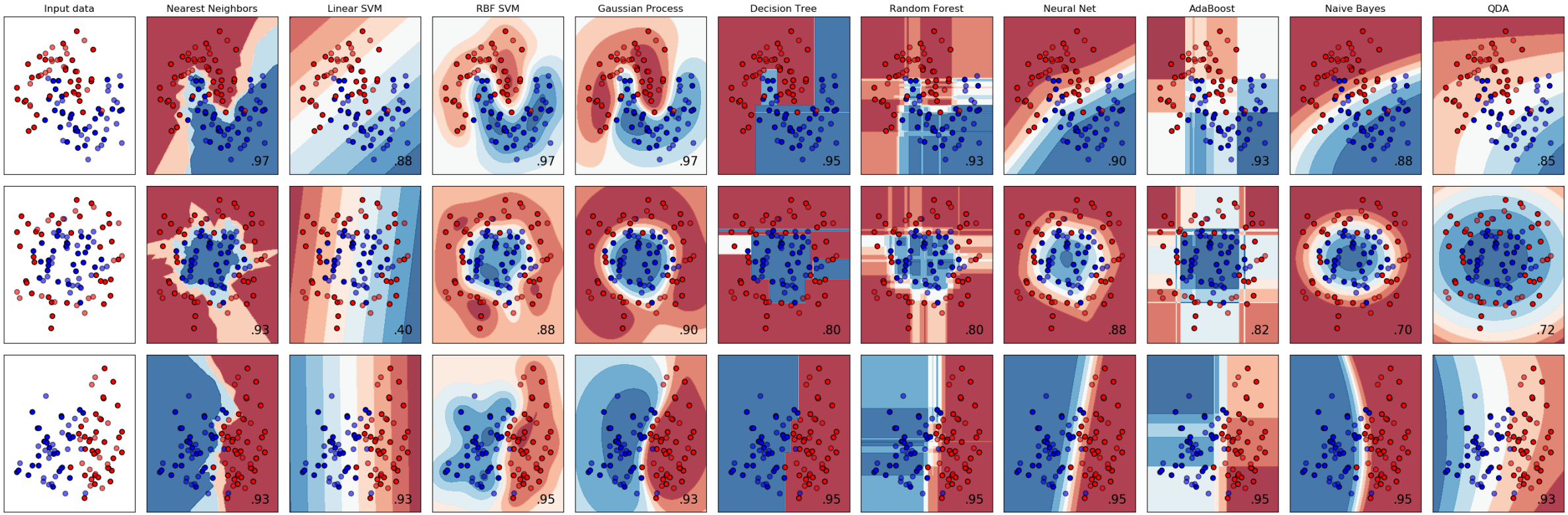
- Uma das tarefas mais comuns, a Classificação, visa identificar a qual classe um determinado registro pertence.
- Nesta tarefa, o modelo analisa o conjunto de registros fornecidos, com cada registro já contendo a indicação à qual classe pertence, a fim de ’aprender’ como classificar um novo registro (aprendizado supervisionado).

- Por exemplo, categorizamos cada registro de um conjunto de dados contendo as informações sobre os colaboradores de uma empresa:
  - Perfil Técnico,
  - Perfil Negocial e
  - Perfil Gerencial.
- O modelo analisa os registros e então é capaz de dizer em qual categoria um novo colaborador se encaixa.
- A tarefa de classificação pode ser usada por exemplo para:
  - Determinar quando uma transação de cartão de crédito pode ser uma fraude;
  - Identificar em uma escola, qual a turma mais indicada para um determinado aluno;
  - Diagnosticar onde uma determinada doença pode estar presente;
  - Identificar quando uma pessoa pode ser uma ameaça para a segurança.

Visualização Classificação



Classificação

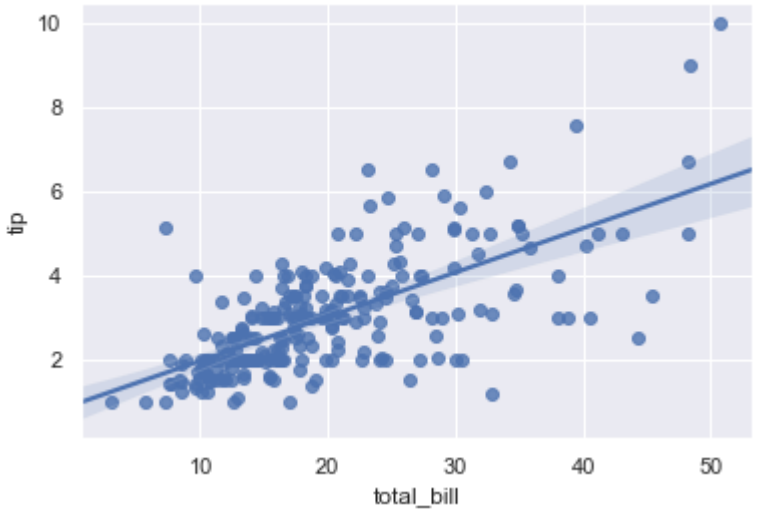


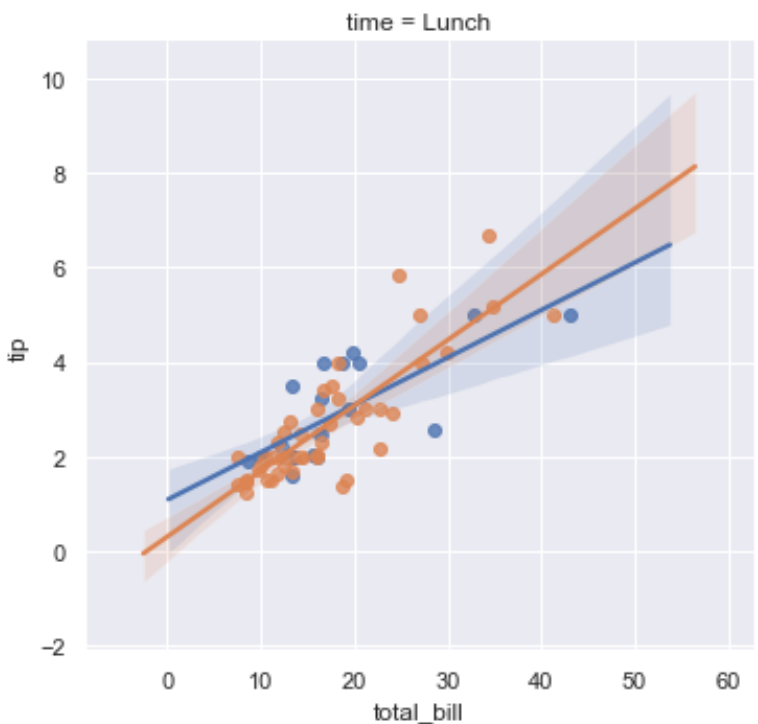
Estimação (Estimation) ou Regressão (Regression)

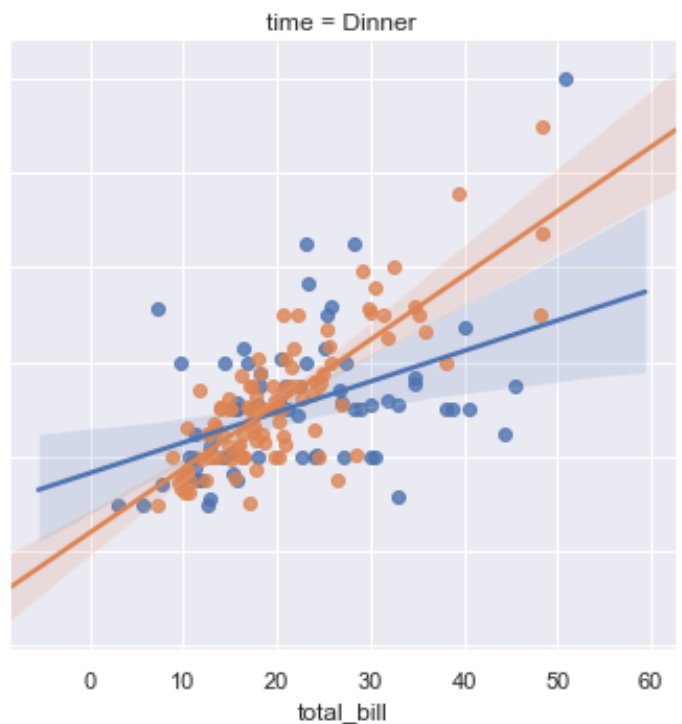
- A estimação é similar à classificação, porém é usada quando o registro é identificado por um valor numérico e não um categórico.
- Assim, pode-se estimar o valor de uma determinada variável analisando-se os valores das demais.

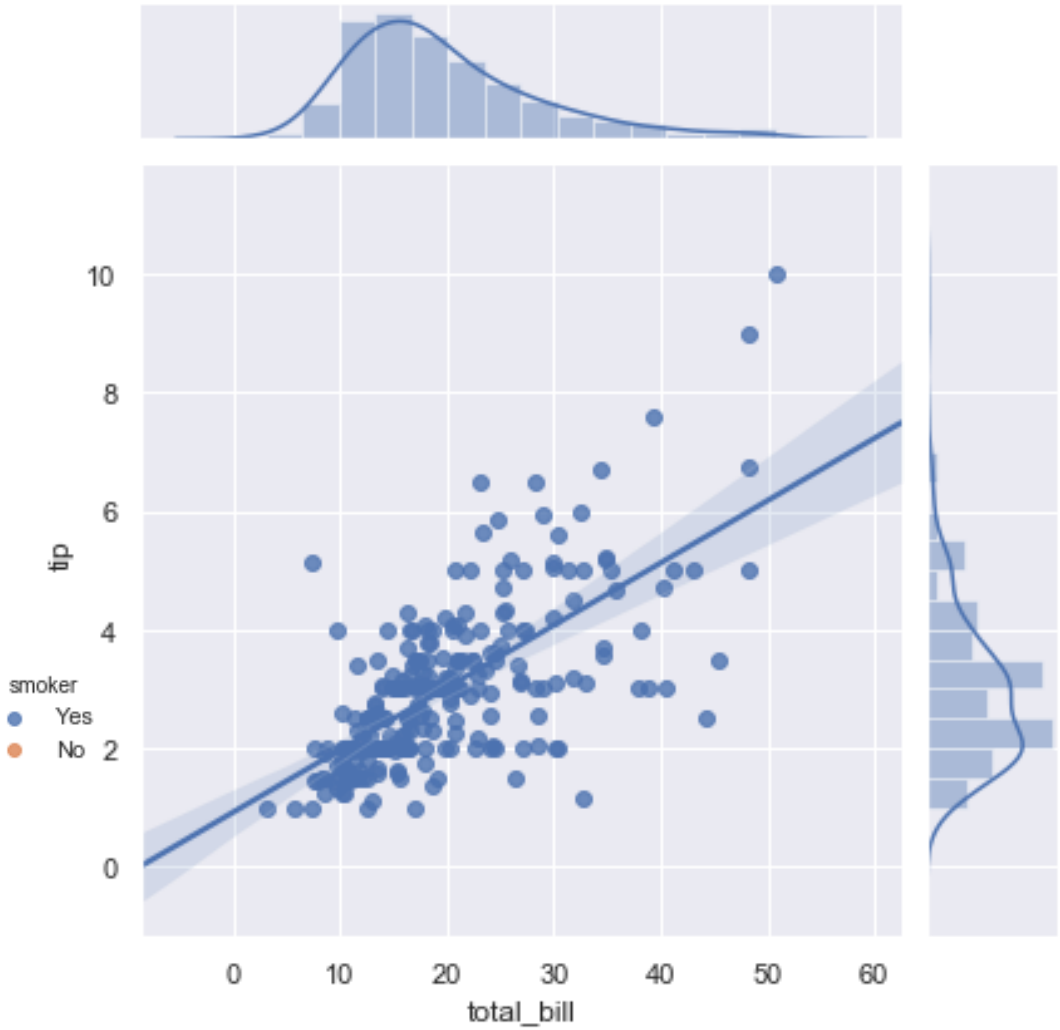
- Por exemplo, um conjunto de registros contendo os valores mensais gastos por diversos tipos de consumidores e de acordo com os hábitos de cada um.
- Após ter analisado os dados, o modelo é capaz de dizer qual será o valor gasto por um novo consumidor.
- A tarefa de estimação pode ser usada por exemplo para:
  - Estimar a quantia a ser gasta por uma família de quatro pessoas durante a volta às aulas;
  - Estimar a pressão ideal de um paciente baseando-se na idade, sexo e massa corporal.

Regressão


tip
total_bill


time = Lunch
tip
total_bill


time = Dinner
total_bill


tip
smoker
Yes
No
total_bill

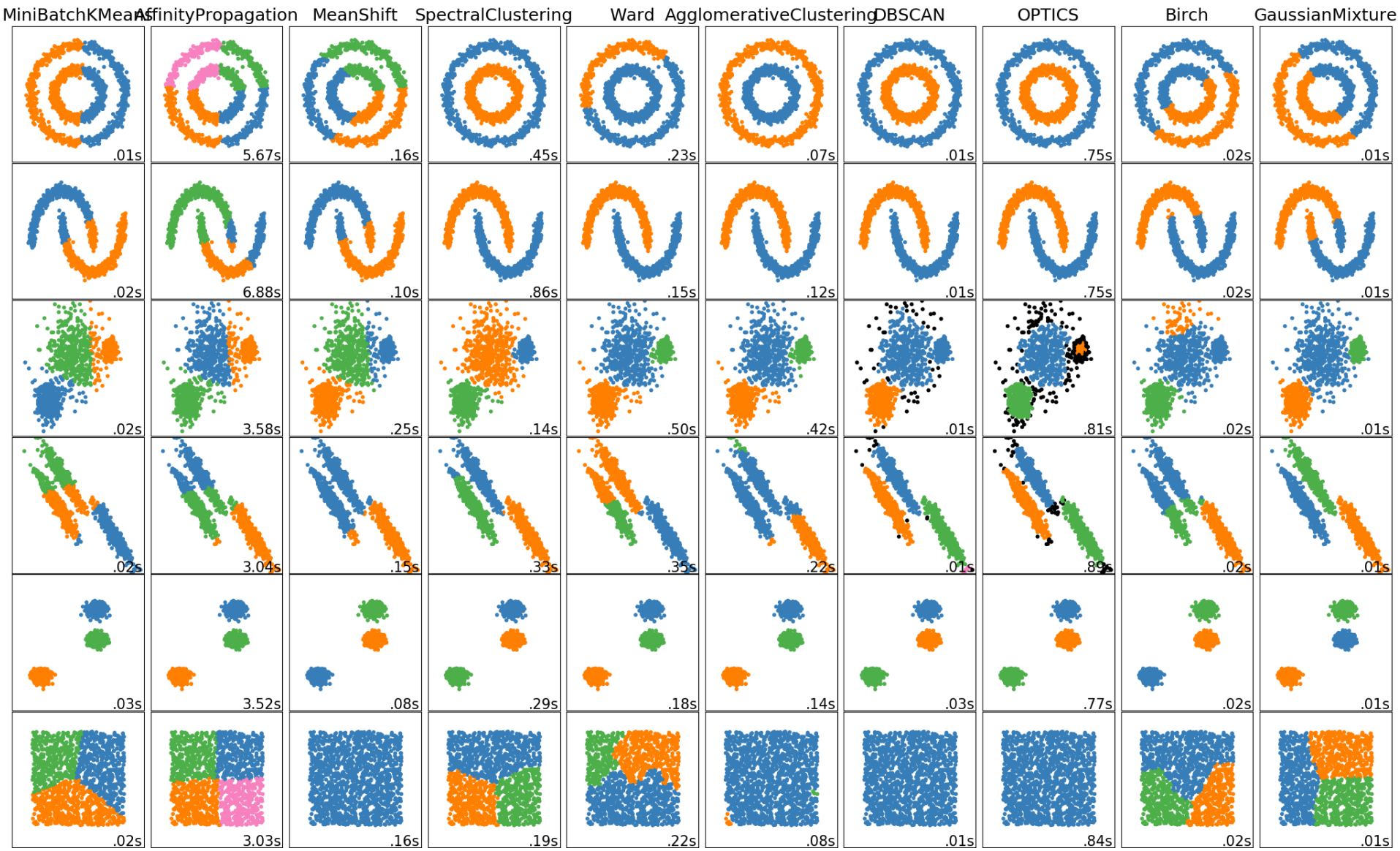
Predição (Prediction)

- A tarefa de predição é similar às tarefas de classificação e estimação, porém ela visa descobrir o valor futuro de um determinado atributo. Exemplos:
- • Predizer o valor de uma ação três meses adiante;
- • Predizer o percentual que será aumentado de tráfego na rede se a velocidade aumentar;
- • Predizer o vencedor do campeonato baseando-se na comparação das estatísticas dos times.
- Alguns métodos de classificação e regressão podem ser usados para predição, com as devidas considerações.

Agrupamento (Clustering)

- A tarefa de agrupamento visa identificar e aproximar os registros similares.
- Um agrupamento (ou cluster) é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais agrupamentos.
- Esta tarefa difere da classificação pois não necessita que os registros sejam previamente categorizados (aprendizado não-supervisionado).
- Além disso, ela não tem a pretensão de classificar, estimar ou predizer o valor de uma variável, ela apenas identifica os grupos de dados similares. Exemplos:
  - Segmentação de mercado para um nicho de produtos;
  - Para auditoria, separando comportamentos suspeitos;

Cl


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
OPTICS
Birch
GaussianMixture
.01s
5.67s
.16s
.45s
.23s
.07s
.01s
.75s
.02s
.01s
.02s
6.88s
.10s
.86s
.15s
.12s
.01s
.75s
.02s
.01s
.02s
3.58s
.25s
.14s
.50s
.42s
.01s
.81s
.02s
.01s
.02s
3.04s
.15s
.33s
.35s
.22s
.01s
.89s
.02s
.01s
.03s
3.52s
.08s
.29s
.18s
.14s
.03s
.77s
.02s
.01s
.02s
3.03s
.16s
.19s
.22s
.08s
.01s
.84s
.02s
.01s

Associação (Association)

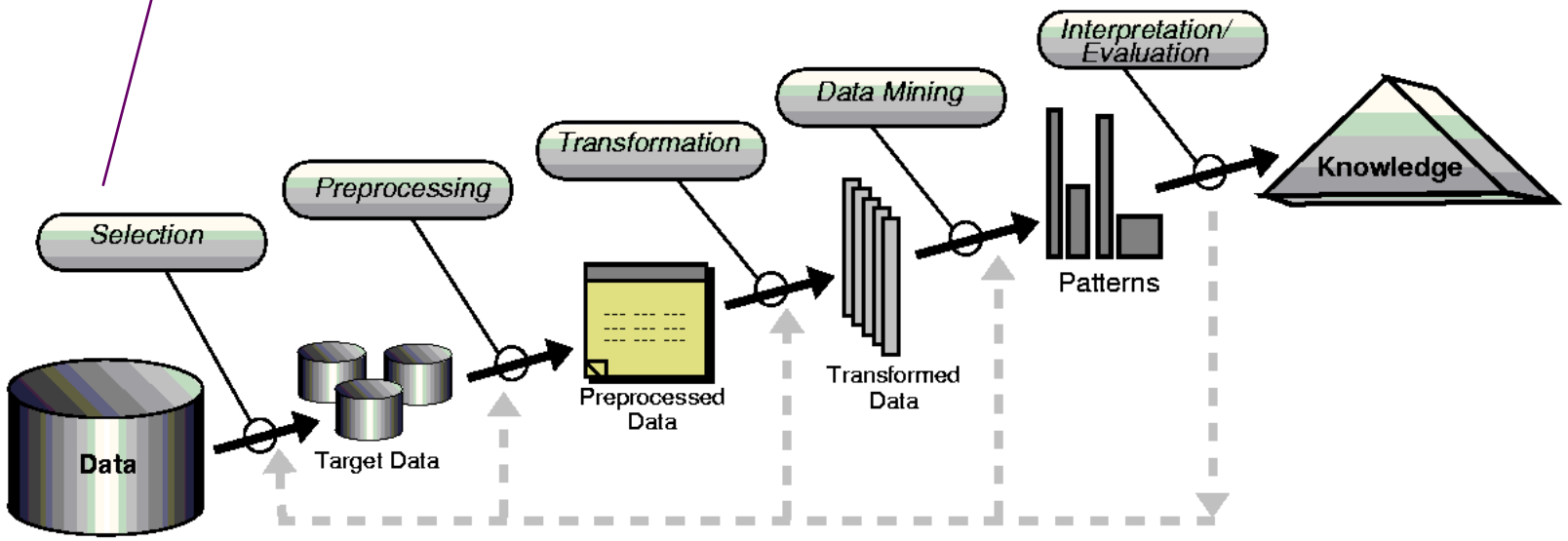
- A tarefa de associação consiste em identificar quais atributos estão relacionados. Apresentam a forma: SE atributo X ENTÃO atributo Y.
- É uma das tarefas mais conhecidas devido aos bons resultados obtidos, principalmente nas análises da "Cestas de Compras"(Market Basket), onde identificamos quais produtos são levados juntos pelos consumidores.
- Alguns exemplos:
- Determinar os casos onde um novo medicamento pode apresentar efeitos colaterais;
- Identificar os usuários de planos que respondem bem a oferta de novos serviços.

Etapas do Processo

- Seleção
- Pré-processamento
- Transformação
- Data mining (aprendizagem)
- Interpretação e Avaliação

Processo mínimo de descoberta do conhecimento

Compreensão do domínio e dos objetivos da tarefa;
Criação do conjunto de dados envolvendo as variáveis necessárias;

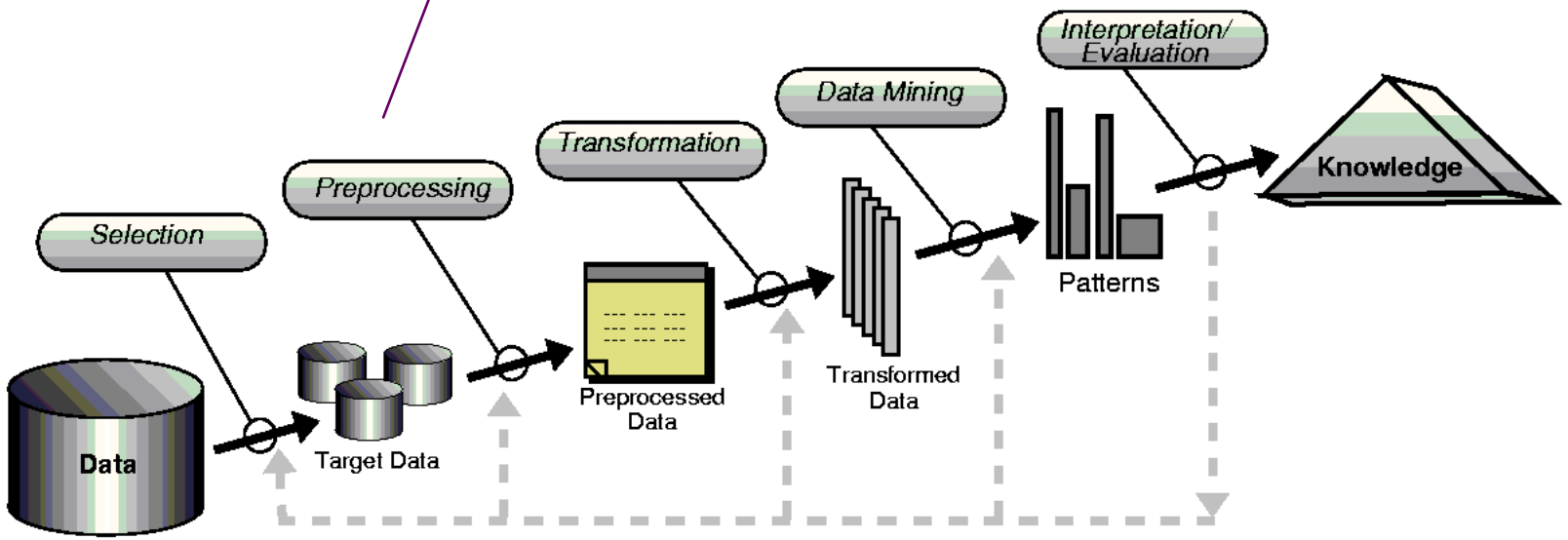


Seleção de Dados

- Selecionar ou segmentar dados de acordo com critérios definidos:
    - Ex.: Todas as pessoas que são proprietárias de carros é um subconjunto de dados determinado.

Processo mínimo

Operações como identificação de ruídos, *outliers*, como tratar falta de dados em alguns campos, etc.

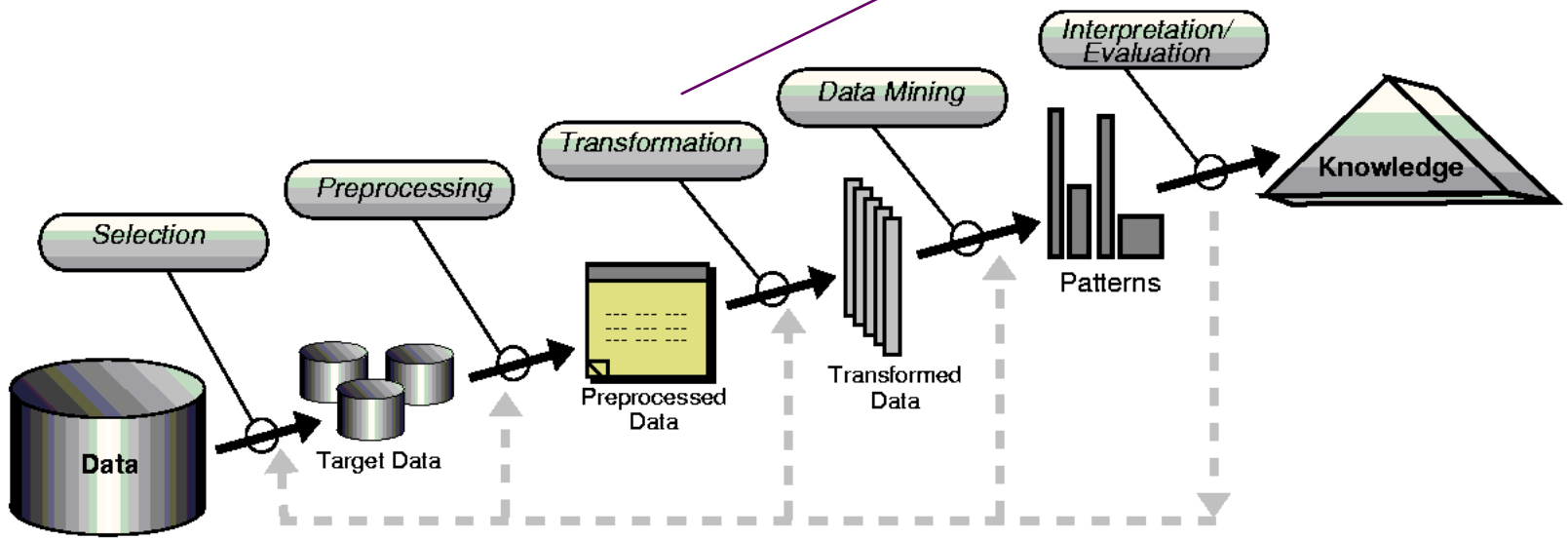


Pré-Processamento

- Estágio de limpeza dos dados, onde informações julgadas desnecessárias são removidas.
- Reconfiguração dos dados para assegurar formatos consistentes (identificação)
    - Ex. : sexo = “F” ou “M”
    - sexo = “M” ou “H”

Processo mínimo

Redução de dimensionalidade, combinação de atributos;

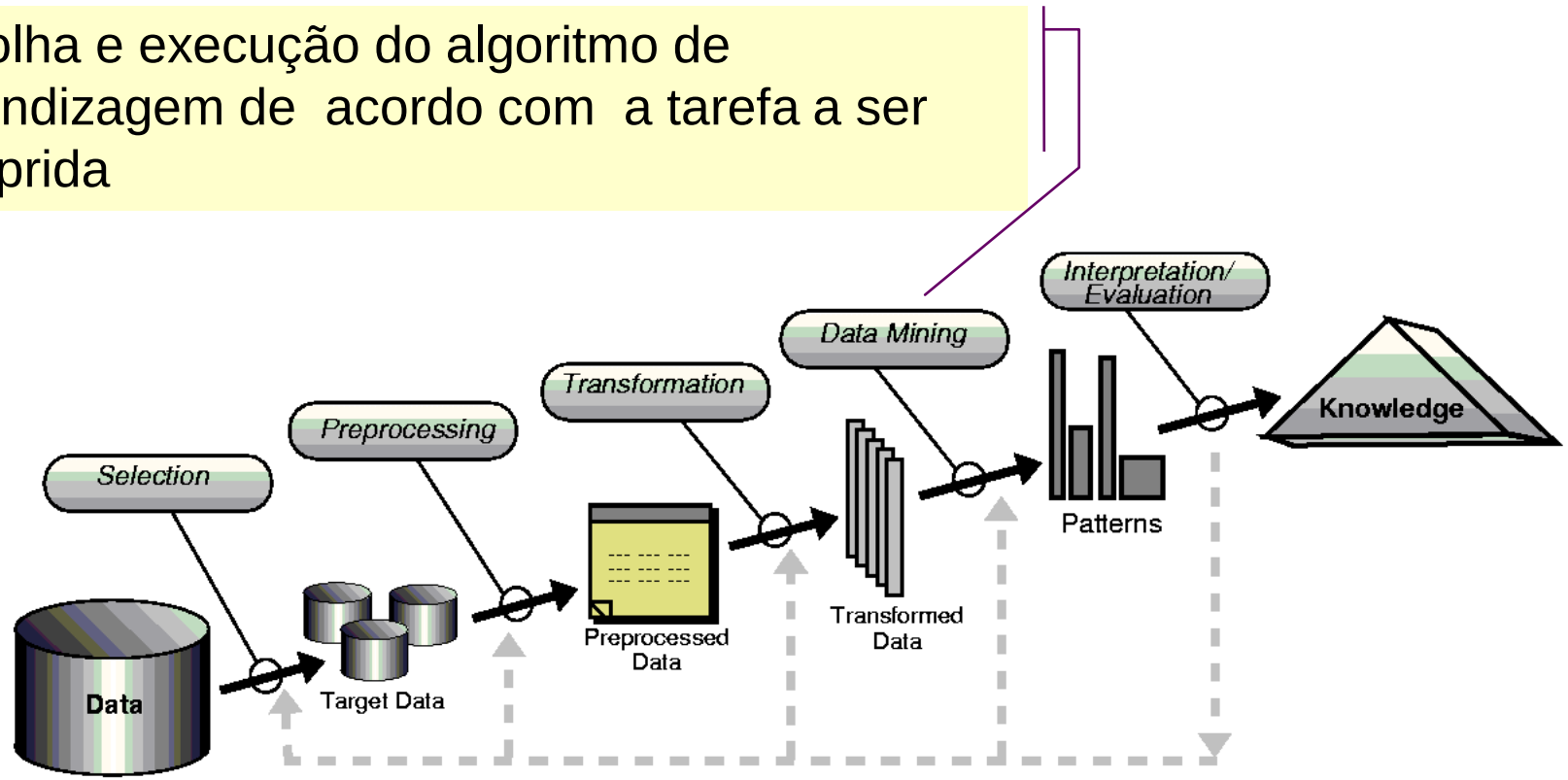


Transformação

- Transformam-se os dados em formatos utilizáveis. Esta depende da técnica data mining usada.
- Disponibilizar os dados de maneira usável e navegável.

Processo mínimo

Escolha e execução do algoritmo de aprendizagem de acordo com a tarefa a ser cumprida

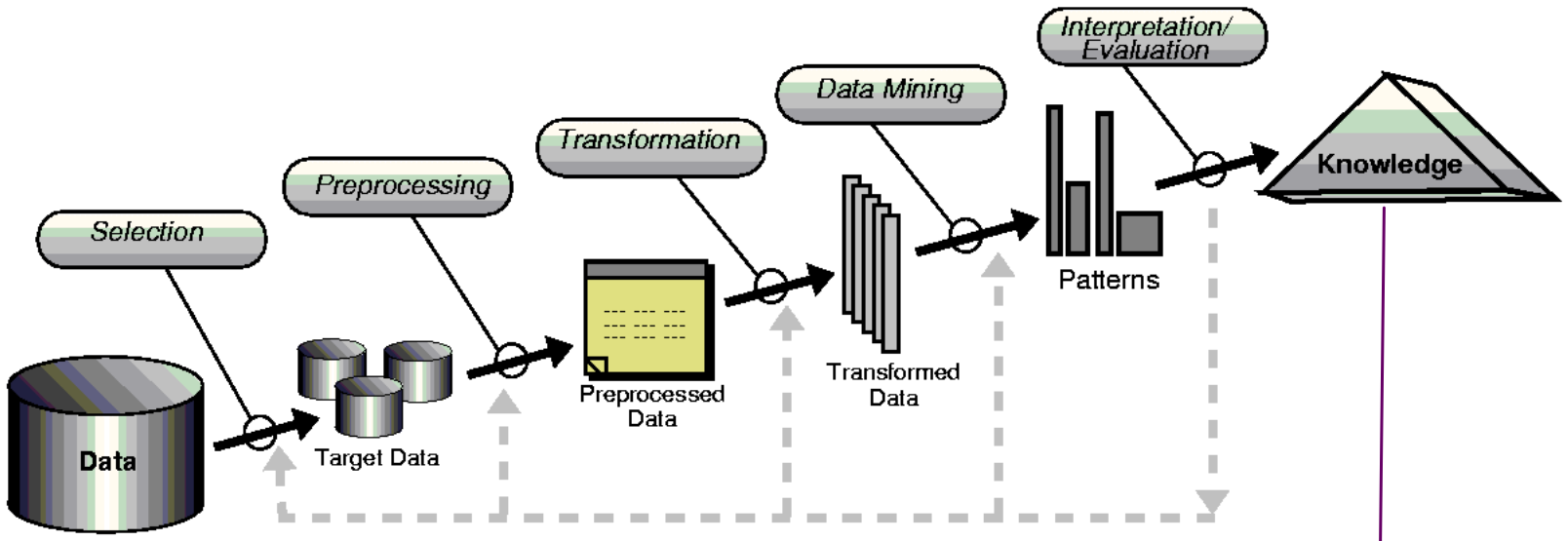


Data Mining

- É a verdadeira extração dos padrões de comportamento dos dados (exemplos)

Processo mínimo

Interpretação dos resultados, com possível retorno aos passos anteriores;



Consolidação: incorporação e documentação do conhecimento e comunicação aos interessados;

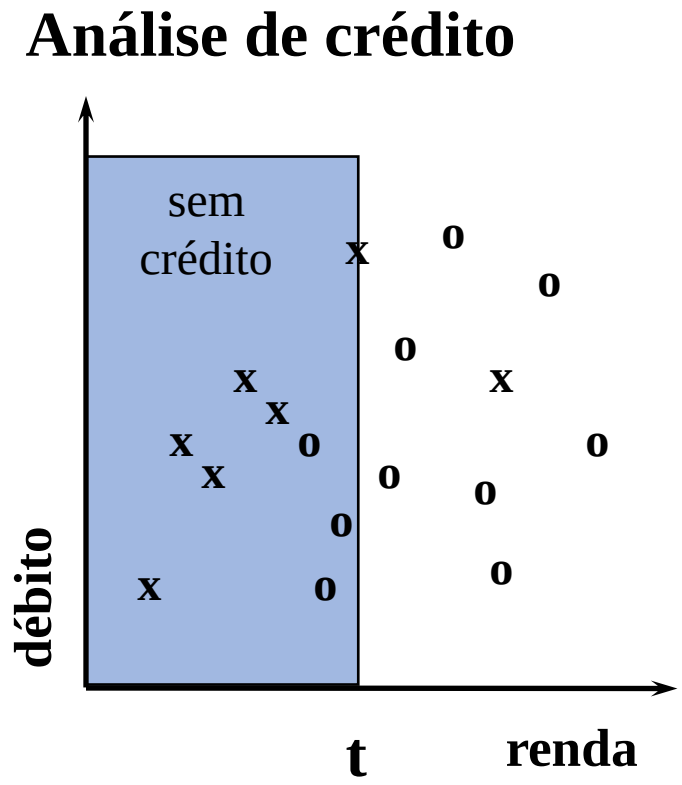
Interpretação e Avaliação

- Identificado os padrões pelo sistema, estes são interpretados em conhecimentos, os quais darão suporte a tomada de decisões humanas

Métodos de mineração de dados

- Métodos de mineração de dados normalmente são extensões ou combinações de uns poucos métodos fundamentais;
- Porém, não é viável a criação de um único método universal: cada algoritmo possui sua própria tendência indutiva;

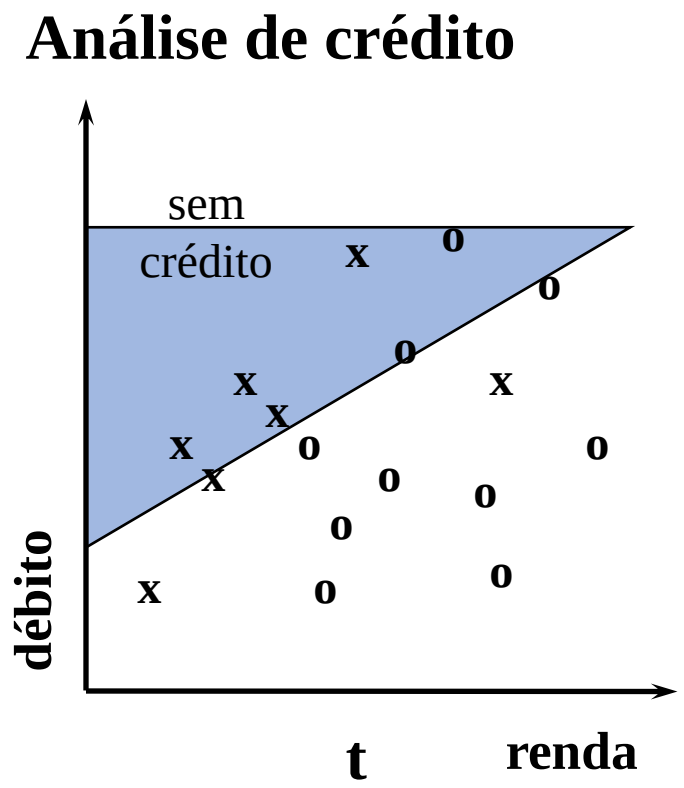
Exemplo de previsão (I)



x: exemplo recusado
o: exemplo aceito

- Um hiperplano paralelo de separação: pode ser interpretado diretamente como uma regra:
  - se a renda é menor que *t*, então o crédito não deve ser liberado
- Exemplo:
  - árvores de decisão;
  - indução de regras

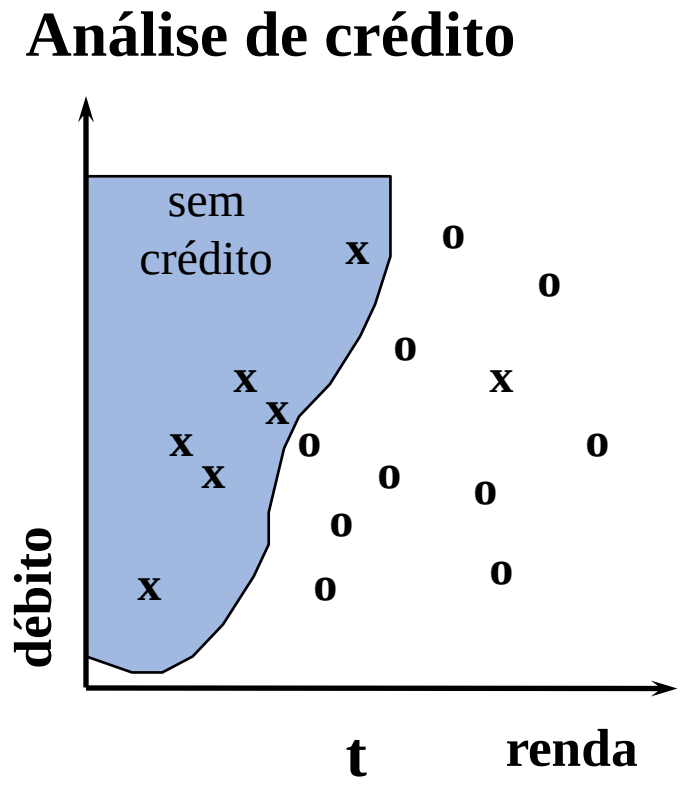
Exemplo de previsão (II)



x: exemplo recusado
o: exemplo aceito

- Hiperplano oblíquo: melhor separação:
- Exemplos:
  - regressão linear;
  - perceptron;

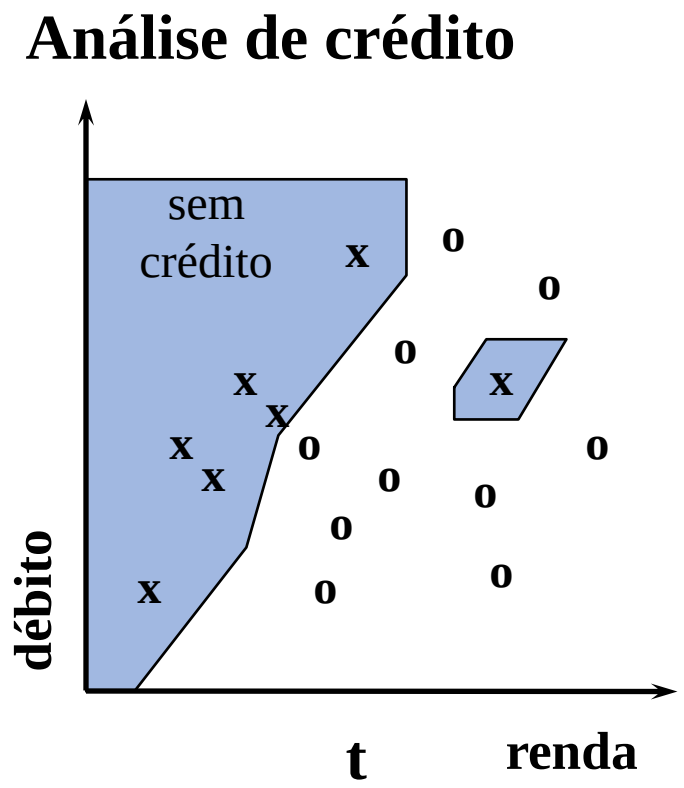
Exemplo de previsão (III)



x: exemplo recusado
o: exemplo aceito

- Superfície não linear: melhor poder de classificação, pior interpretação;
- Exemplos:
  - perceptrons multicamadas;
  - regressão não-linear;

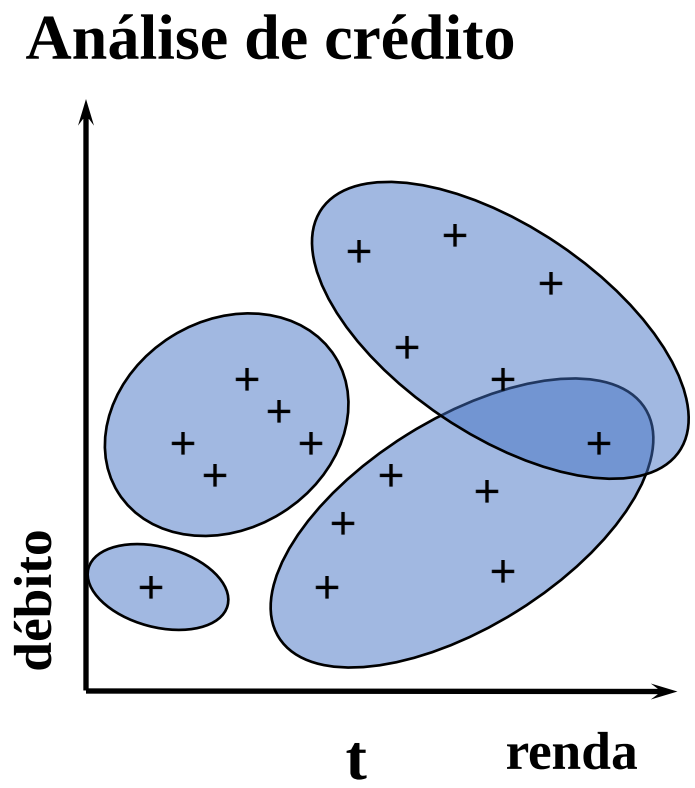
Exemplo de previsão (IV)



x: exemplo recusado
o: exemplo aceito

- Métodos baseado em exemplos;
- Exemplos:
  - k-vizinhos mais próximos;
  - raciocínio baseado em casos;

Exemplo de descrição (I)



+: exemplo

- Agrupamento
- Exemplo:
  - *vector quantization;*

Exemplo de descrição (II)

- Regras de associação
  - “98% dos consumidores que adquiriram pneus e acessórios de automóveis também se interessaram por serviços automotivos”;
  - descoberta simétrica de relações, ao contrário de métodos de classificação
    - qualquer atributo pode ser uma classe ou um atributo de discriminação;

Exemplos

Áreas de aplicações potenciais:

- Vendas e Marketing
  - *Identificar padrões de comportamento de consumidores*
  - *Associar comportamentos à características demográficas de consumidores*
  - *Campanhas de marketing direto (mailing campaigns)*
  - *Identificar consumidores “leais”*

Exemplos

Áreas de aplicações potenciais:

- Bancos
  - *Identificar padrões de fraudes (cartões de crédito)*
  - *Identificar características de correntistas*
  - *Mercado Financeiro ($$$)*

Exemplos

- Áreas de aplicações potenciais
  - Médica
    - *Comportamento de pacientes*
    - *Identificar terapias de sucessos para diferentes tratamentos*
    - *Fraudes em planos de saúdes*
    - *Comportamento de usuários de planos de saúde*

Introdução

- Exemplo (1) - Fraldas e cervejas
  - O que as cervejas tem a ver com as fraldas ?
  - homens casados, entre 25 e 30 anos;
  - compravam fraldas e/ou cervejas às sextas-feiras à tarde no caminho do trabalho para casa;
  - Wal-Mart otimizou às gôndolas nos pontos de vendas, colocando as fraldas ao lado das cervejas;
  - Resultado: o consumo cresceu 30% .

Exemplos

- Exemplo (2) - Lojas Brasileiras (Info 03/98)
  - Aplicou 1 milhão de dólares em técnicas de data mining
  - Reduziu de 51000 produtos para 14000 produtos oferecidos em suas lojas.
  - Exemplo de anomalias detectadas:
    - *Roupas de inverno e guarda chuvas encalhadas no nordeste*
    - *Batedeiras 110v a venda em SC onde a corrente elétrica é 220v*

Exemplos

Exemplo (3) - Bank of America (Info 03/98)

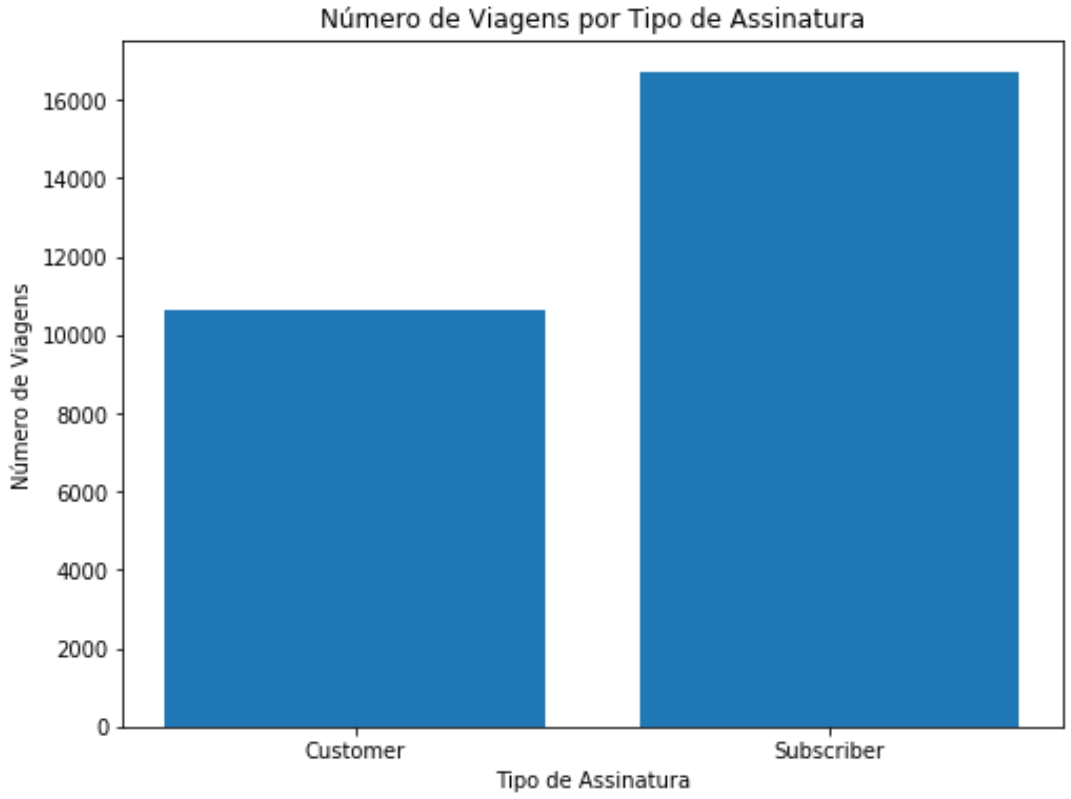
- Selecionou entre seus 36 milhões de clientes
  - Aqueles com menor risco de dar calotes
  - Tinham filhos com idades entre 18 e 21 anos
  - Resultado em três anos o banco lucrou 30 milhões de dólares com a carteira de empréstimos.

Examplo: Bay Area Bike Share

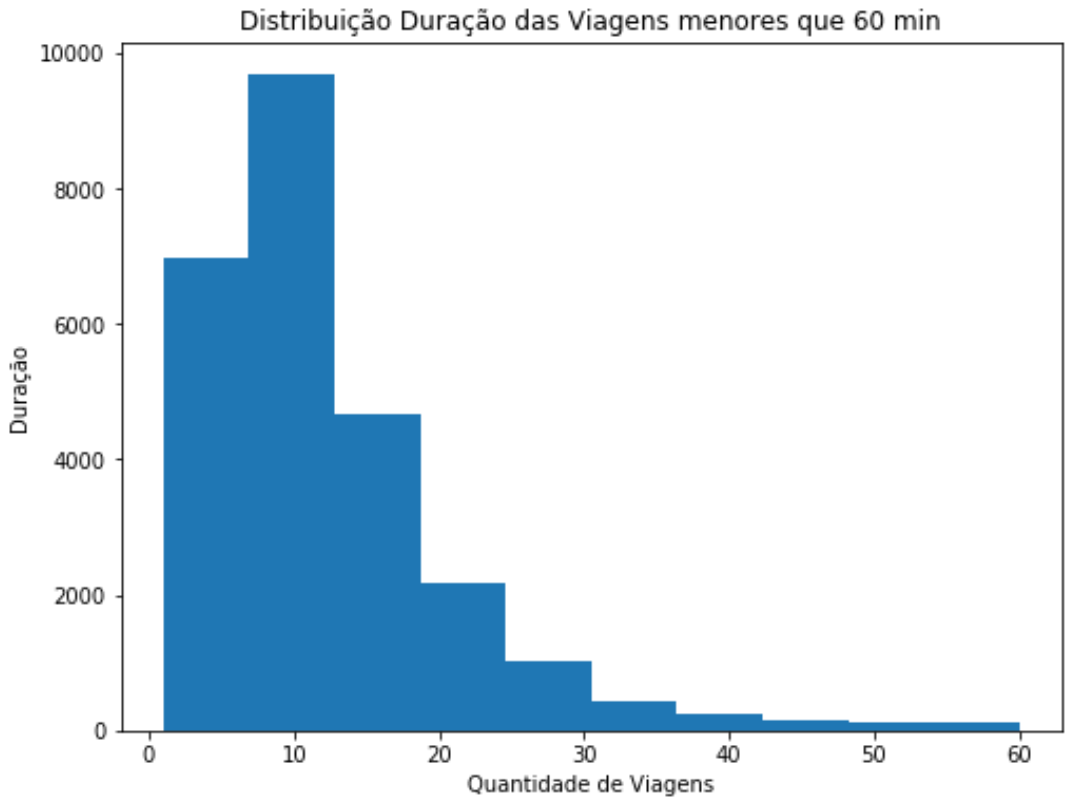
- Bay Area Bike Share é uma empresa que oferece aluguel de bicicletas on-demand para clientes em San Francisco, Redwood City, Palo Alto, Mountain View e San Jose.
- Os usuários podem desbloquear bicicletas de uma variedade de estações em cada cidade, e devolvê-las em qualquer estação dentro da mesma cidade.
- Os usuários pagam o serviço por meio de assinatura anual ou pela compra de passes de 3 dias ou 24 horas.
- Os usuários podem fazer um número ilimitado de viagens.
- Viagens com menos de trinta minutos de duração não têm custo adicional;
- Viagens mais longas incorrem em taxas de horas extras.

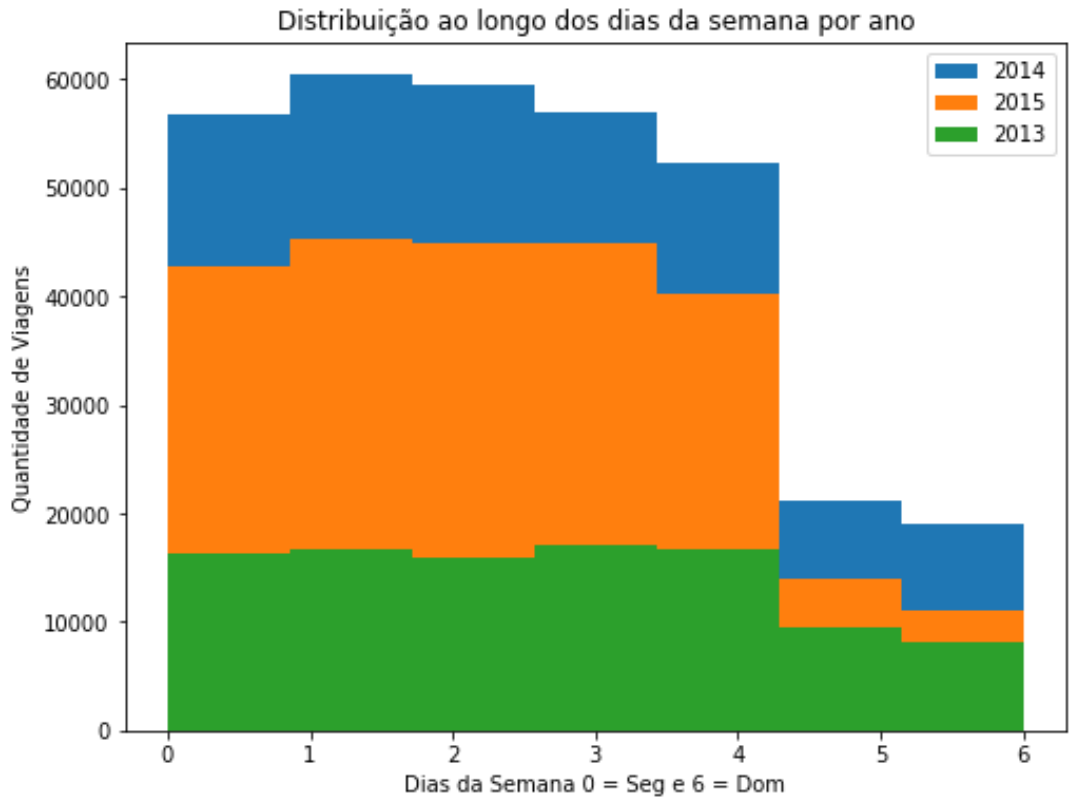
DEA – Data Exploratóry Analisys

- Existem 27345 pontos no conjunto de dados
- A duração média das viagens foi de 27 .60 minutos
- A mediana das durações das viagens f oi de 10.72 minutos
- 25% das viagens foram mais curtas do que 6.82 minutos
- 25% das viagens foram mais comprida s do que 17.28 minutos

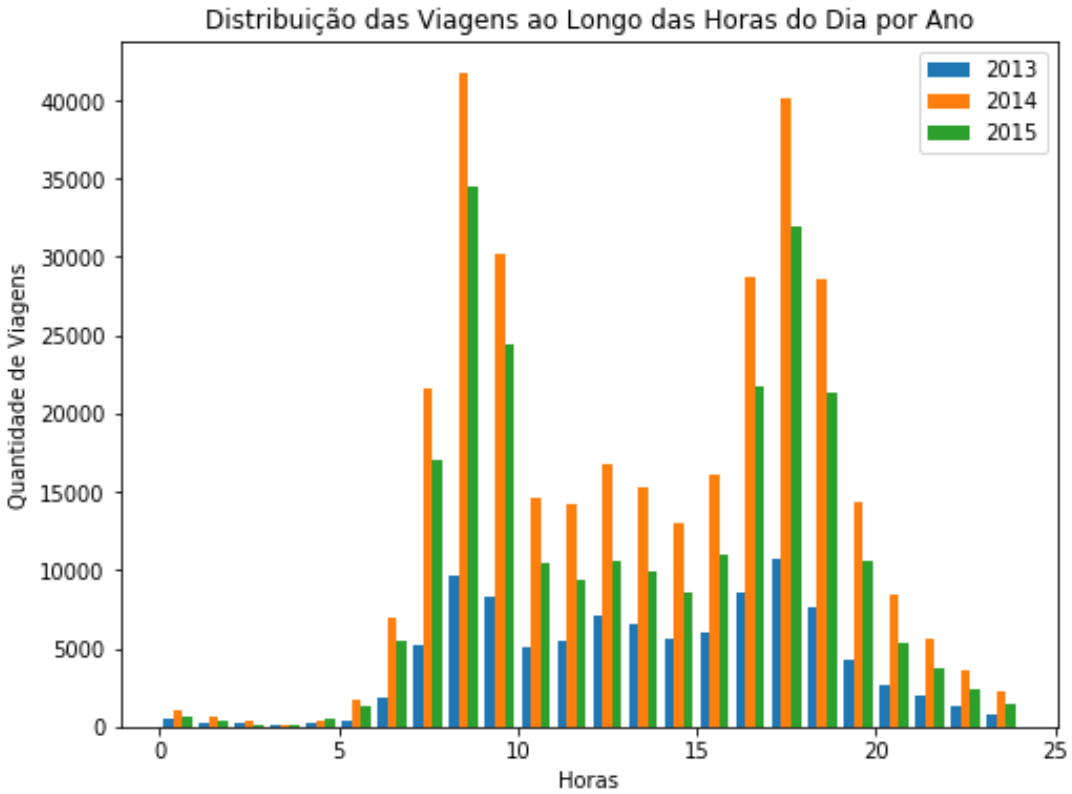


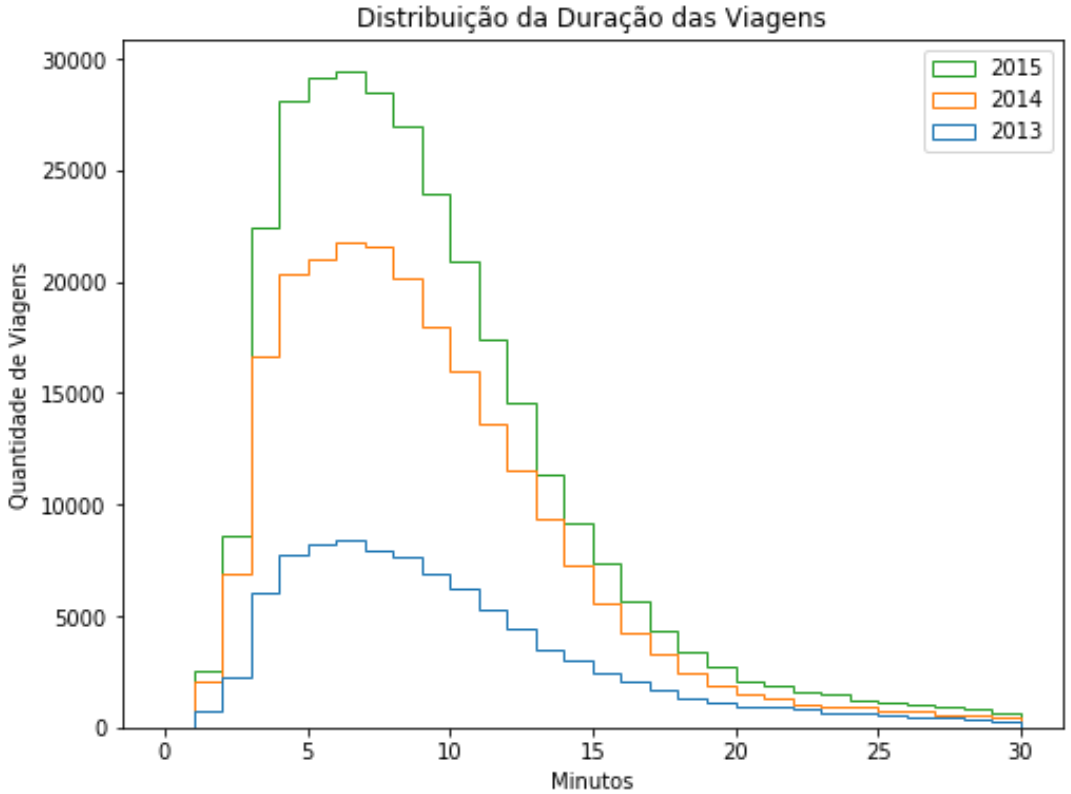
DEA Bike Share

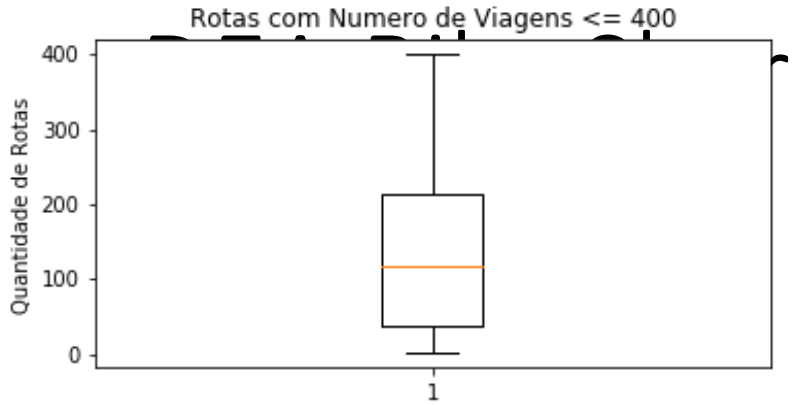


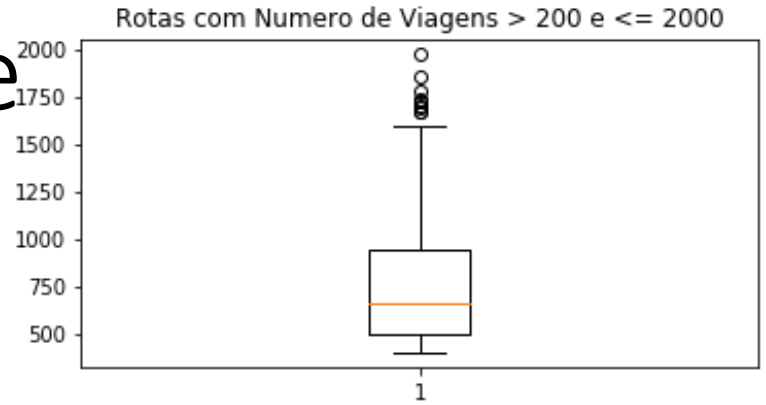


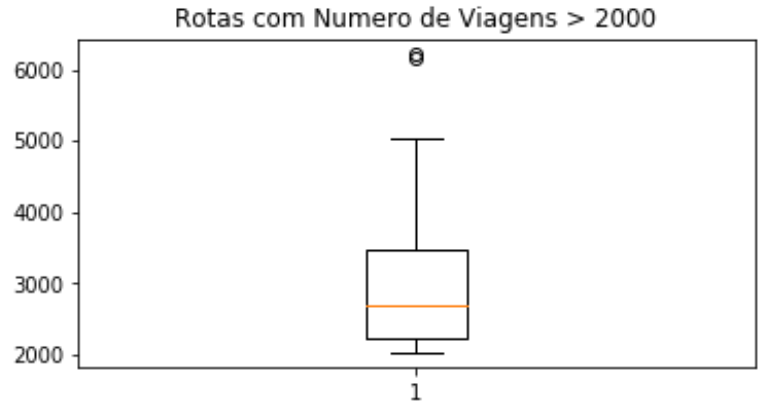
DEA – Byke Share

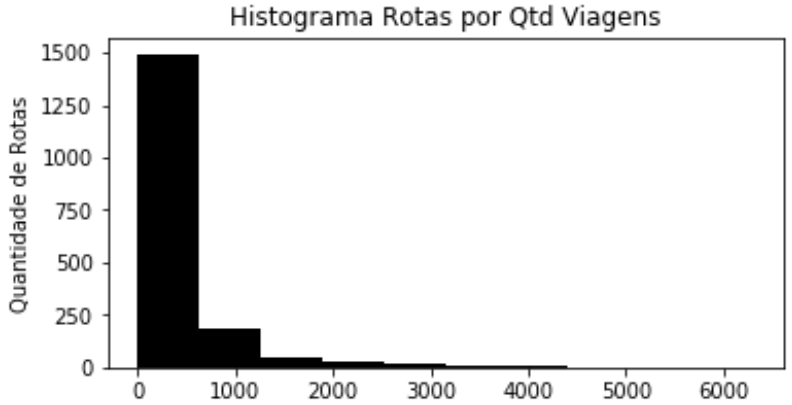


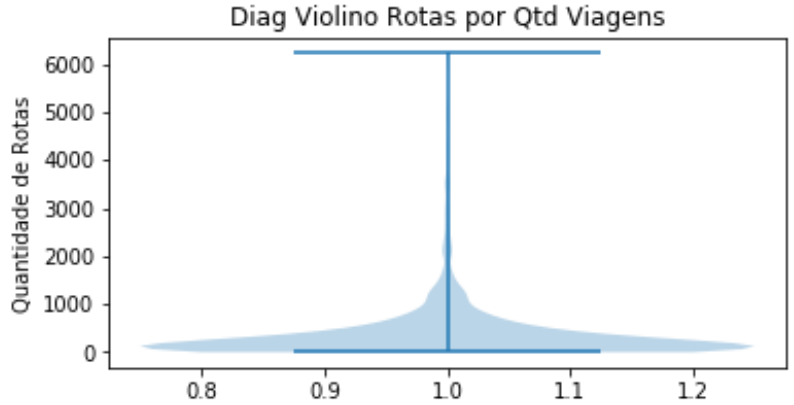


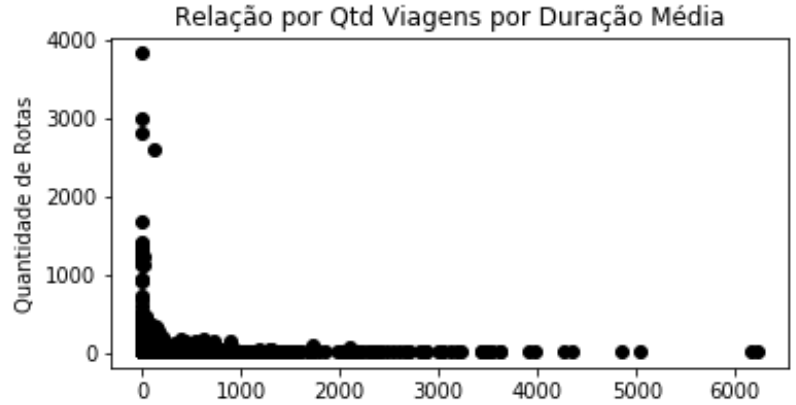

Rotas com Numero de Viagens <= 400
Quantidade de Rotas
0
100
200
300
400
1

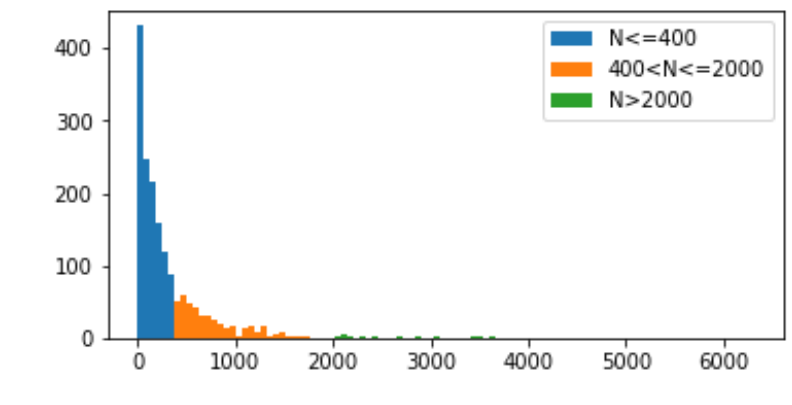

Rotas com Numero de Viagens > 200 e <= 2000
500
750
1000
1250
1500
1750
2000
1

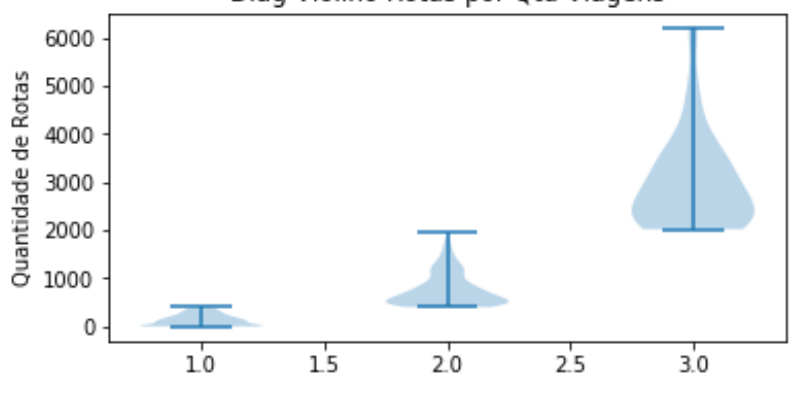

Rotas com Numero de Viagens > 2000
2000
3000
4000
5000
6000
1

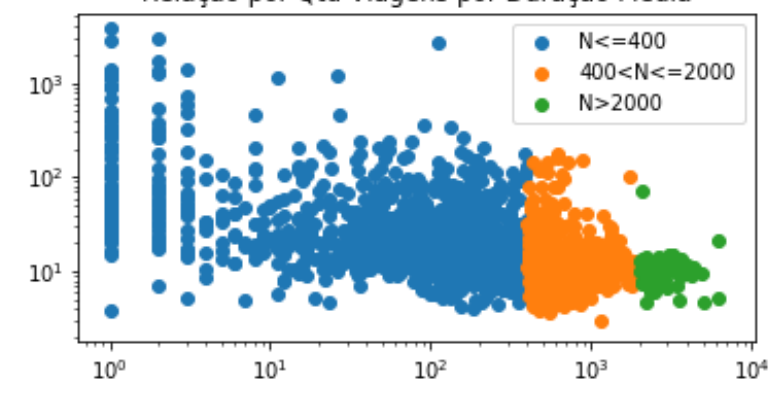

Histograma Rotas por Qtd Viagens
Quantidade de Rotas
0
250
500
750
1000
1250
1500
0
1000
2000
3000
4000
5000
6000

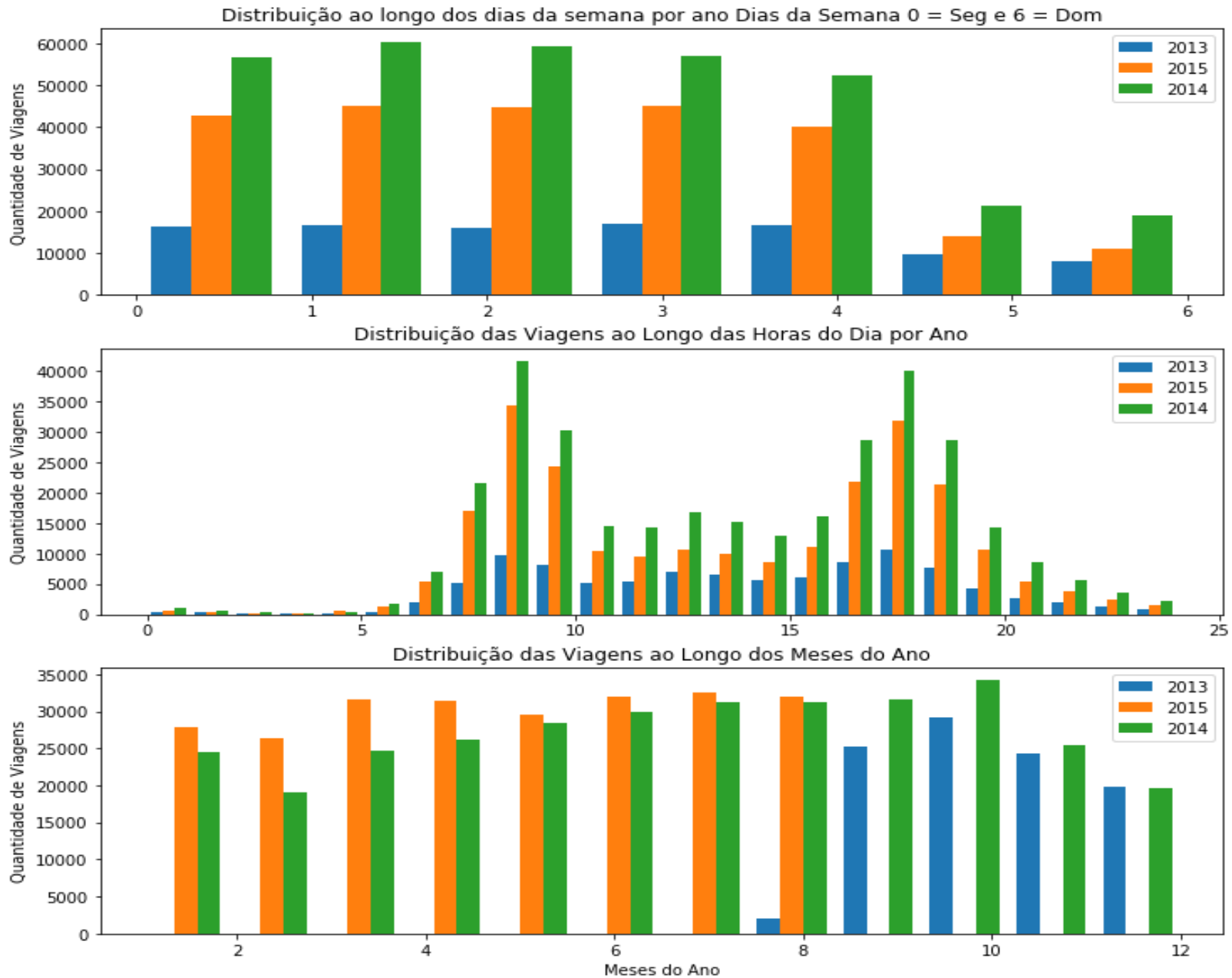

Diag Violino Rotas por Qtd Viagens
Quantidade de Rotas
0
1000
2000
3000
4000
5000
6000
0.8
0.9
1.0
1.1
1.2

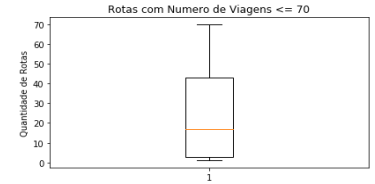

Relação por Qtd Viagens por Duração Média
Quantidade de Rotas
0
1000
2000
3000
4000
0
1000
2000
3000
4000
5000
6000

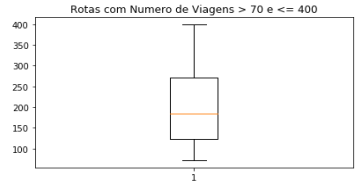

N<=400
400<N<=2000
N>2000
0
100
200
300
400
0
1000
2000
3000
4000
5000
6000

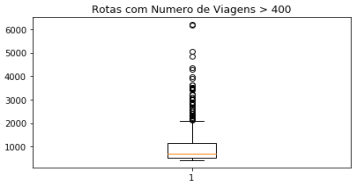

Diag Violino Rotas por Qtd Viagens
Quantidade de Rotas
0
1000
2000
3000
4000
5000
6000
1.0
1.5
2.0
2.5
3.0


Relação por Qtd Viagens por Duração Média
N<=400
400<N<=2000
N>2000
10^1
10^2
10^3
10^0
10^1
10^2
10^3
10^4

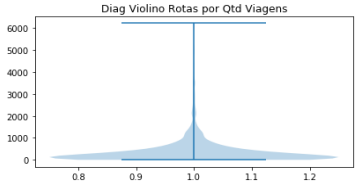

Distribuição ao longo dos dias da semana por ano Dias da Semana 0 = Seg e 6 = Dom
Quantidade de Viagens
2013
2015
2014
Distribuição das Viagens ao Longo das Horas do Dia por Ano
Quantidade de Viagens
2013
2015
2014
Distribuição das Viagens ao Longo dos Meses do Ano
Quantidade de Viagens
Meses do Ano
2013
2015
2014

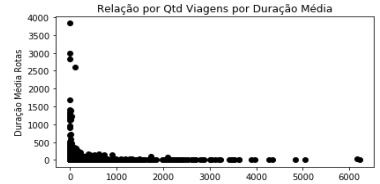

Rotas com Numero de Viagens <= 70
Quantidade de Rotas

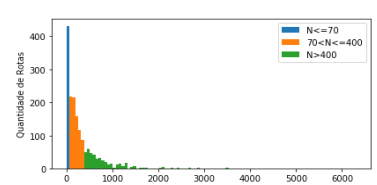

Rotas com Numero de Viagens > 70 e <= 400

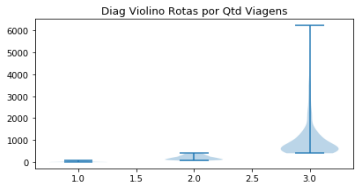

Rotas com Numero de Viagens > 400

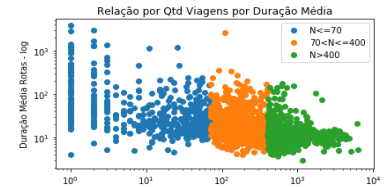

Histograma Rotas por Qtd Viagens
Quantidade de Rotas

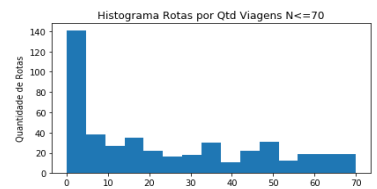

Diag Violino Rotas por Qtd Viagens

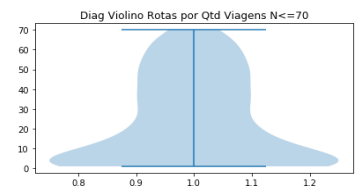

Relação por Qtd Viagens por Duração Média
Duração Média Rotas
Quatidade de Rotas

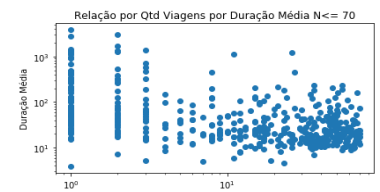

N<=70
70<N<=400
N>400
Quantidade de Rotas

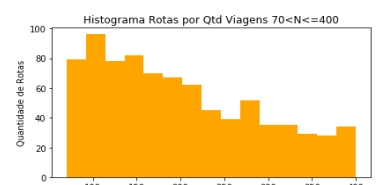

Diag Violino Rotas por Qtd Viagens

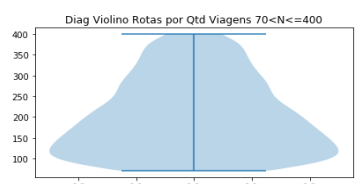

Relação por Qtd Viagens por Duração Média
N<=70
70<N<=400
N>400
Duração Média Rotas - log
Quantidade de Rotas - log

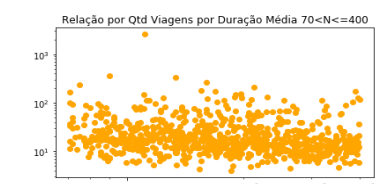

Histograma Rotas por Qtd Viagens N<=70
Quantidade de Rotas

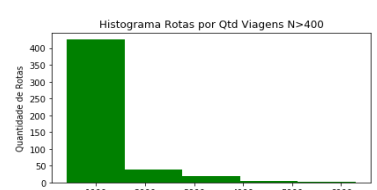

Diag Violino Rotas por Qtd Viagens N<=70

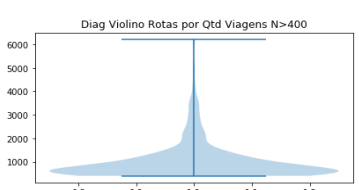

Relação por Qtd Viagens por Duração Média N<= 70
Duração Média
Quantidade de Rotas - log

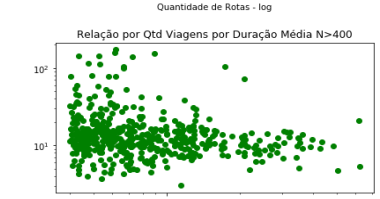

Histograma Rotas por Qtd Viagens 70<N<=400
Quantidade de Rotas


Diag Violino Rotas por Qtd Viagens 70<N<=400


Relação por Qtd Viagens por Duração Média 70<N<=400
Quantidade de Rotas - log


Histograma Rotas por Qtd Viagens N>400
Quantidade de Rotas


Diag Violino Rotas por Qtd Viagens N>400


Relação por Qtd Viagens por Duração Média N>400
Quantidade de Rotas - log

Conclusões

- Data mining é um processo que permite compreender o comportamento dos dados.
- Data mining analisa os dados usando técnicas de aprendizagem para encontrar padrões e regulariedades nestes conjuntos de dados.
- É um problema pluridisciplinar, envolve Inteligência Artificial, Estatística, Computação Gráfica, Banco de Dados.
- Pode ser bem aplicado em diversas áreas de negócios