**Estatística Descritiva**

**Dados:**

Categóricos (Binários, Nominal-categorias, Ordinal)

Contínuos

**Medidas de Valor Central:** ponto central ao redor do qual os dados estão distribuídos (média, mediana e moda)

Média: distribuições simétricas sem outliers.

Mediana: distribuições não simétricas ou com outliers.

Moda: valor mais frequente (dados categóricos)

**Medidas de Variabilidade:** como os dados estrão distribuídos/distante do centro

Variância: média das diferenças dos valores com relação à média.

Desvio Padrão: padronização da variância (x² para eliminar valores negativos, depois tira raiz)

|  |  |
| --- | --- |
| Distribuição de Gauss: | Probabilidade de Intervalos: |
| Diagrama  Descrição gerada automaticamente | Gráfico, Histograma  Descrição gerada automaticamente |

Covariância: avalia a variância (relação) conjunta de dois atributos.  
**Covariância positiva:** valores altos para x estão associados a valores altos para y (/)  
**Covariância negativa:** valores altos para x estão associados a valores baixos para y (\)  
Zero: ausência de relação

Uma imagem contendo objeto, relógio

Descrição gerada automaticamente

Correlação: padroniza a medida de relação entre os atributos (1 e -1)  
(covariância mostra se os atributos se relacionam positiva/negativamente, mas não mostra o grau de relação)

Texto

Descrição gerada automaticamente  
**Correlação de Pearson:** normaliza a covariância pelo desvio padrão dos atributos.  
- Variáveis seguem uma gaussiana (distribuição normal)  
- Variáveis contínuas  
- Linearidade (sequencial, ordem lógica) (ex. de estruturas não lineares: estrutura em árvore ou grafos)  
- Covariância dividida pelo desvio padrão de X vezes o desvio padrão de Y

**Correlação de Spearman:** não paramétrica (atributos não precisam ser distribuições normais)  
- Variáveis podem ser ordinais /categóricas

**Medidas Relativas:**

Quantis: intervalos com a mesma frequência (quantidade)

Uma imagem contendo Diagrama

Descrição gerada automaticamenteMediana: 2-quantil

Quartil: 4-quantil

Percentil: 100-quantil

IQR (amplitute interquartile): Q3 – Q1

**Teste de Hipótese:**

**Z-Score:** Número de desvios padrões com relação à média  
Valor 0: X = média  
Valor 1: X = 1 desvio maior que a média

**Probabilidade de Intervalos:**

Gráfico, Histograma

Descrição gerada automaticamente

**Calculando Probabilidades:**

**Gráfico

Descrição gerada automaticamente(?)**

**Teste de Hipótese:** Aceita ou Rejeita hipótese sobre uma população a partir de uma amostra de dados.  
Ex.: diferença de peso entre recém-nascidos de alta e baixa renda é significante?  
Hipótese nula (não há diferença e alternativa (há diferença)

|  |  |
| --- | --- |
| Teste Bilateral | Teste Unilateral |
|  | Ou |

**Nível de Significância:** também chamado de valor crítico.  
Probabilidade máxima para se rejeitar hipótese nula  
Usualmente 1% ou 5%  
  
P-Value: Probabilidade de a hipótese nula ser verdadeira  
Um valor pequeno (menor que valor crítico) indica evidência forte contra a hipótese nula)

**Teste t:** Verificar se há diferença significativa entra a média de uma população e uma amostra.  
Média da população conhecida, mas desvio padrão não.  
Etapas: 1 – Definir hipótese nula e alternativa  
 2 – Definir valor crítico  
 3 – Comparar o resultado com o valor crítico

|  |  |
| --- | --- |
| **Comparar 2 Grupos Pareados: t-Test Pareado** | **Comparar 2 Grupos Não-Pareados: t-Test Não-Pareado** |
| **Hipótese Nula:** Não há diferença entre a média das amostras  **Suposições:** - Diferença entre as amostras é gaussiana - Não pode conter outliers - Mesma população (pareado) | **Hipótese Nula:** Não há diferença entre a média das amostras  **Suposições:** - Observação dos grupos são independentes - Grupos seguem a curva gaussiana - Variância dos dois grupos são iguais |

|  |  |
| --- | --- |
| **Comparar 2 Grupos Pareados: Wilcoxon signed-rank Test** | **Comparar 2 Grupos Não-Pareados Mann-Whitney U Teste** |
| **Hipótese Nula:** Não há diferença entre as amostras  **Suposições:** - Não Paramétrica: diferença não corresponde a gaussiana. - Diferença segue uma distribuição simétrica - Mesma população | **Hipótese Nula:** amostras vêm de populações com mesma distribuição  **Suposições:** - Não Paramétrica: diferença não corresponde a gaussiana. - Observação dos grupos são independentes |

Tabela

Descrição gerada automaticamente

**Formas de Imputação:**

**Atributos categóricos:** Usar classificação para prever os valores ausentes de uma variável a partir das outras.

**Atributos numéricos:** Usar regressão para prever valores ausentes de uma variável a partir de outras.

**Modelos de Predição:**

**KNN (k-Nearest Neighbors):** Valores ausentes são definidos pelos valores das instâncias mais próximas**. Valores**: média dos valores vizinho (precisa de normalização)  
(PARA FAZER CÁLCULO DA DISTÂNCIA, OS ATRIBUTOS PRECISAM SER NUMÉRICOS E NORMALIZADOS)

**Discretização**: Transformas atributos contínuos em categóricos  
decidir números de intervalos e seu tamanho.  
**Supervisionada**: baseada em entropia: intervalos com mesma proporção de um rótulo  
Passo 1: Calcular entropia da classe (grau de incerteza de uma variável, segue a escala de 0 - 1)  
Ex.: Qual atributo melhor classifica uma instância.  
Passo 2: Calcular entropia condicional (entropia da classe dada uma discretização)   
Passo 3: Calcular Information gain (diferença entre a classe entropia classe e a entropia condicional)  
**Não-supervisionada**: intervalos de mesma largura/frequência  
Define intervalos iguais e separa em intervalos (discretização) ou usa a mesma frequência (n elementos)

**Normalização:** utilizada para cálculo de distância entre instâncias c/ atributos com diferentes intervalos [0,1] ou [1,-1]  
Evita que atributos com maiores valores se sobreponham aos atributos com valores menores

**Normalização por Min-Max:** valores dos atributos são ajustados para o intervalo a,b baseado no valor min e max.

**Normalização por Média e Desvio Padrão** (Z-score – número de desvios padrões que o valor está da média)

**Detecção de Outliers:**

**Métodos de Detecção Univariados:**

Interface gráfica do usuário, Texto

Descrição gerada automaticamente**1. Z-Score Robusto:** Distribuição precisa ser simétrica (usar LOG para torná-la mais simétrica, caso necessário)  
Centro = Mediana (Metade são maiores e metade são menores)  
É influenciado pelas posições dos outliers, mas não pelos seus valores.

**2. Método de Tukey:** Distribuição precisa ser simétrica.  
Baseado em quartis.  
**Outliers:** Valores menores que Q1- 1,5 \* IQR

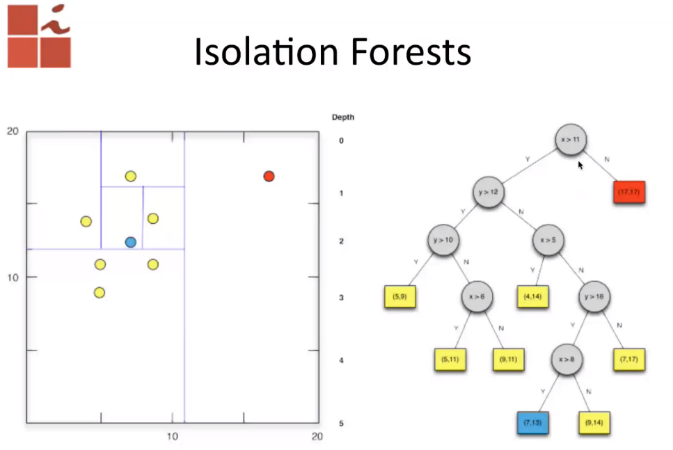
Gráfico, Gráfico de caixa estreita

Descrição gerada automaticamenteValores menores que Q3- 1,5 \* IQR

**Métodos de Detecção Bivariados:** Ex. O valor da coluna ‘preço’ e ‘área’ estão dentro de um intervalo aceitável, mas analisando os valores do preço/m², há outliers.  
Transformar a relação em univariada (razão de uma pela outra).

**Método de Detecção Multivariado:**

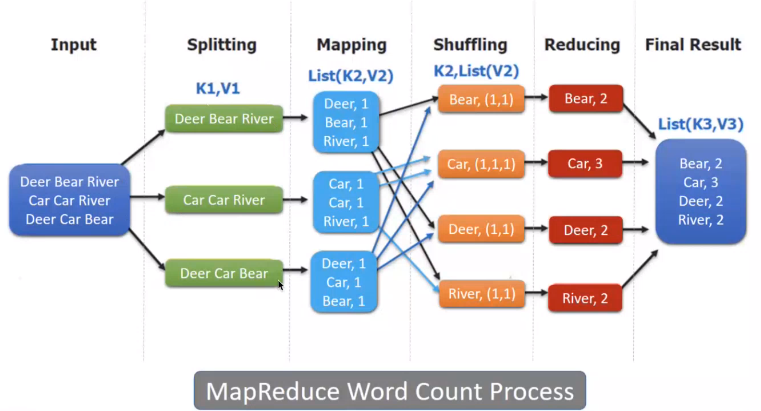
**Elliptic Envelope:** atributos seguem gaussiana (PARAMÉTRICO)  
Utiliza a distância de Mahalanobis (retira o efeito da covariância para o cálculo das distâncias)

**Isolation Forests:** atributos **NÃO** seguem gaussiana (**NÃO** PARAMÉTRICO)  
Outliers: Poucos e diferentes.  
1. Seleciona aleatoriamente um atributo.  
2. Seleciona um valor aletatório entre o máximo e mínimo  
3. Repete n vezes

Fácil de isolar outliers.

**Processamento de Dados em Larga Escala**

**Map Reduce**

**Map:** Aplica a função em todos os elementos de uma lista

**Reduce:** Realiza a computação em uma lista e retorna o resultado

**Hadoop:** Dados armazenados em arquivos armazenados (linguagem Java)  
**Apache Spark:** memória distribuída na RAM (linguagem Scala)  
**PysparK:** Extensão Python, não possui a mesma eficiência do Scala.  
 - **RDD:** Resilient Distributed Datasets – Conjuntos Resilientes de Dados Distribuidos.

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamenteDiagrama

Descrição gerada automaticamente

**Ex. Transformações:**

Map: forma um RDD distribuído a partir de um RDD (Ex. B = A.map(lambda x: x-2).  
Filter: novo RDD a partir de dados filtrados de um RDD.  
Sample: novo RDD a partir de uma amostra de um RDD.  
Union: novo RDD pela união de RDDs.  
groupByKey: (k,v),...,(k,v) -> (k,seq[v])  
Join: (k,v),...,(k,w) -> (k, <v,w>)

**Ex. Ações:**

Reduce: agregação em um RDD usando uma função (Ex. B = A.reduce(lambda x,y: x+y).  
Collect: cria um objeto com todos os itens de um RDD  
Count: número de elementos em um RDD  
countByKey: (k,v1),...,(k,vn) -> (k,n)

Diferenças entre bibliotecas Pandas e PySpark:

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente com confiança média

**Classificação e Regressão**

**Aprendizado de Máquina:** Construção de modelos preditivos e descritivos.

|  |  |
| --- | --- |
| **Supervisionado** | **Não-Supervisionado** |
| Rótulos associados a dados | Sem rótulos/categorias |
| Inferir uma função a partir de exemplos dados para predizer classes de novos exemplos  **Fases:** Treinamento e Execução |  |
| **Classificação:** categorias discretas  **Regressão:** valores contínuos | **Agrupamento:** identifica grupo de dados  **Redução de dimensionalidade:** cria representações mais suscinta dos dados |

Interface gráfica do usuário, Aplicativo

Descrição gerada automaticamente

**Treinamento:** Criar o modelo  
**Validação:** Set os hiper parâmetros  
**Teste:** Verifica a generalização do modelo

**Classificação de Modelos de Avaliação:** Precisão / Recall / F1 (F-measure) / Accuracy

Classificação de um Twitter: é positivo ou negativo?

|  |  |  |
| --- | --- | --- |
| **Matriz de Confusão** | **É positivo** | **É negativo** |
| Predizer se é positivo | Verdadeiros positivos (TP) | Falsos positivos (FP) |
| Predizer se é negativo | Falsos Negativos (FN) | Verdadeiros Negativo (TN) |

Texto

Descrição gerada automaticamente

|  |  |
| --- | --- |
| **Selecionando o melhor modelo:** | **Passos Scykit** |
| - Usar conjunto de validação/cross-validation - Buscar melhores valores dos hiper-parâmetros  - Algoritmo de ML  - Espaços de parâmetros  - Método para buscar valores candidatos  - Holdout set  - Métrica de Avaliação - Estratégias  - Grid Search  - Automl: TPOT, SMAC, auto-sklearn, optuna | Passo 1– Importar modelo escolhido  Passo 2 – Escolher os hiper-parâmetros do modelo  Passo 3 – Criar modelo com função fit  Passo 4 – Aplicar o modelo a novos dados:  - Supervisionado: função predict  - Não-supervisionado: funções transform ou predict |

**Randon Forest (Árvores de descisão) – Algoritmo de Construção**

**Uma imagem contendo Diagrama

Descrição gerada automaticamente**

**Diagnóstico de Viés e Variância do Modelo:**

**Desenho de um cachorro

Descrição gerada automaticamente com confiança médiaViés:** Desempenho do conjunto de treinamento / depende do valor alvo

**Variância:** diferença de desempenho entre treinamento e teste

**Overfitting:** bom desempenho no conjunto de treinamento. Problema em generalizar, baixo viés e alta variância  
Ex. Erro treinamento: 1% / Erro teste: 11%

**Underfitting:** Modelo não modelo bem o conjunto de treinamento.  
Alto viés e baixa variância.  
Ex. Erro treinamento: 15% / Erro teste: 16%

**Overfitting:** Adicionar dados ao conjunto treinamento, usar modelos menos complexos, selecionar outras features

**Underfitting:** Aumentar complexidade dos modelos, mais features.