4 ANÁLISE E CIÊNCIA DE DADOS

Este capítulo tem como objetivo a apresentação da Análise e Ciência de Dados, onde através desta área do conhecimento é possível realizar uma análise metódica e estruturada de conjuntos de dados, fundamentada em áreas como a matemática, estatística, ciências da computação, teoria de sistemas, PO e IA. Para uma melhor compreensão do leitor, subdividiremos este capítulo nos tópicos descritos a seguir:

- Apresentação, definição e importância da Análise e Ciência de Dados em Engenharia.
- Definição do ciclo de vida de projetos na Análise e Ciência de Dados.
- Apresentação de alguns fundamentos teóricos chave e ferramentas para o desenvolvimento do ciclo de vida nos projetos que envolvem a Análise e Ciência de Dados.

4.1 Definições e Importância

A Ciência e Análise de Dados é uma área interdisciplinar fundamentada nas áreas de matemática, estatística, ciências da computação, teoria de sistemas, PO, e IA (BLEI; SMYTH, 2017; CONCOLATO; CHEN, 2017; NELLI, 2018; RUNKLER, 2016), cujo objetivo principal é fornecer um significado mais fundamentado às grandes quantidades de dados e informações, produzidas e coletados (ALCÁCER; CRUZ-MACHADO, 2019; CONCOLATO; CHEN, 2017; ROSE, 2016; RUNKLER, 2016; SAUCEDO-MARTÍNEZ et al., 2018). A Ciência de Dados é um termo mais recente e abrangente que a Análise de Dados.

A Ciência de Dados é um processo de estudo dos dados orientado a descobrir problemas, resolve-los e transmitir as suas possíveis soluções (CONCOLATO; CHEN, 2017; SONG; ZHU, 2017), sendo considerado como uma disciplina empírica, onde a partir dos dados a informação é obtida utilizando um método científico (ROSE, 2016). A Análise de Dados permite analisar grandes conjuntos de dados e extrair informações para apoiar as tomada de decisões em questões específicas, onde as mesmas nem sempre são facilmente obtidas a partir da utilização de sistemas computacionais (RUNKLER, 2016).

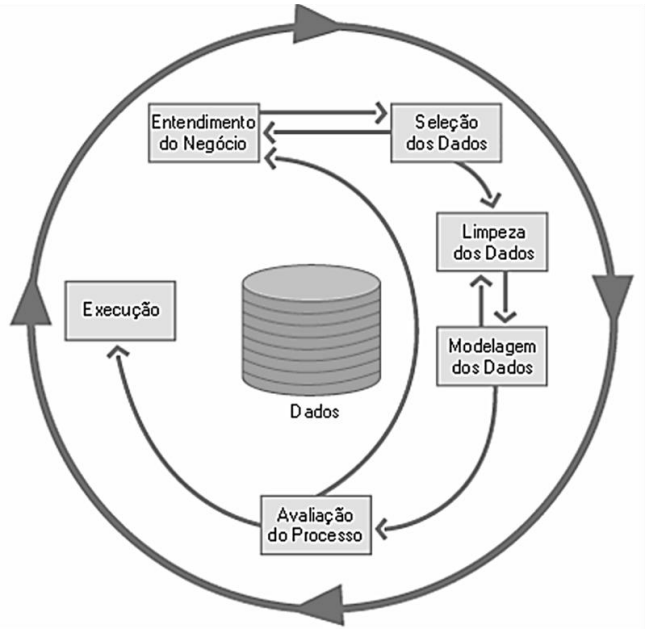
Dentro do contexto que envolve a quarta revolução industrial e dos dispositivos hiperconectados, a quantidade de dados produzidos e armazenados é muito elevada, criando assim o conceito de Big Data. Do ponto de vista da Ciência e Análise de Dados, o Big Data é a matéria-prima para a extração de conhecimento útil para a tomada de decisões, detecção de problemas e proposta de soluções (ALCÁCER; CRUZ-MACHADO, 2019; SAUCEDO-MARTÍNEZ et al., 2018).

A Ciência de Dados e a Análise de Dados estão profundamente apoiadas na utilização da IA e suas subáreas, pois agrupam algoritmos e técnicas que permitem dar solução para alguns dos problemas baseados no uso de grandes conjuntos de dados.

4.2 Ciclo de Vida na Ciência e Análise de Dados

A Ciência e Análise de Dados são áreas que precisam de um ciclo de vida composto de etapas ou fases necessárias para a realização de um projeto, devendo seguir uma metodologia estabelecida para facilitar o desenvolvimento e gerenciamento de um projeto. A mineração de dados é uma área relacionada à Ciências e Análise de Dados, como mostra a Figura 15 (ROSE, 2016), composta de um ciclo de vida do projeto conhecido como o Processo Industrial Padrão para Mineração de Dados (*Cross Industry Standard Process for Data Mining* ou CRISP-DM).

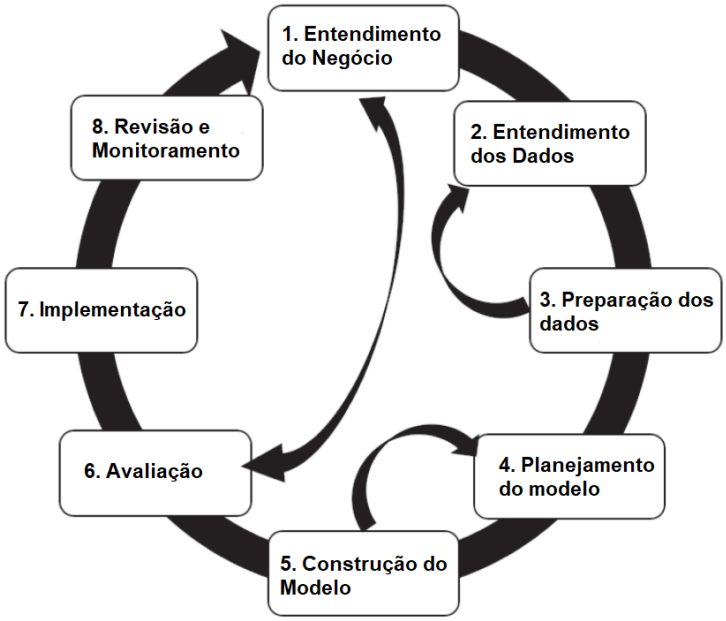
Figura 15 – Processo Industrial Padrão para Mineração de Dados (CRISP-DM).



Fonte: Amorim (2006).

O ciclo de vida de projeto CRISP-DM mostrado na Figura 15 é utilizado frequentemente e adaptado, devido à sua similaridade, com o ciclo de vida em projetos de Ciência de Dados. No trabalho de Song e Zhu (2017) é possível perceber a semelhança entre o ciclo de vida CRISP-DM e o proposto para a Ciência de Dados (Figura 16).

Figura 16 – Ciclo de vida da Ciência de Dados (SONG; ZHU, 2017).



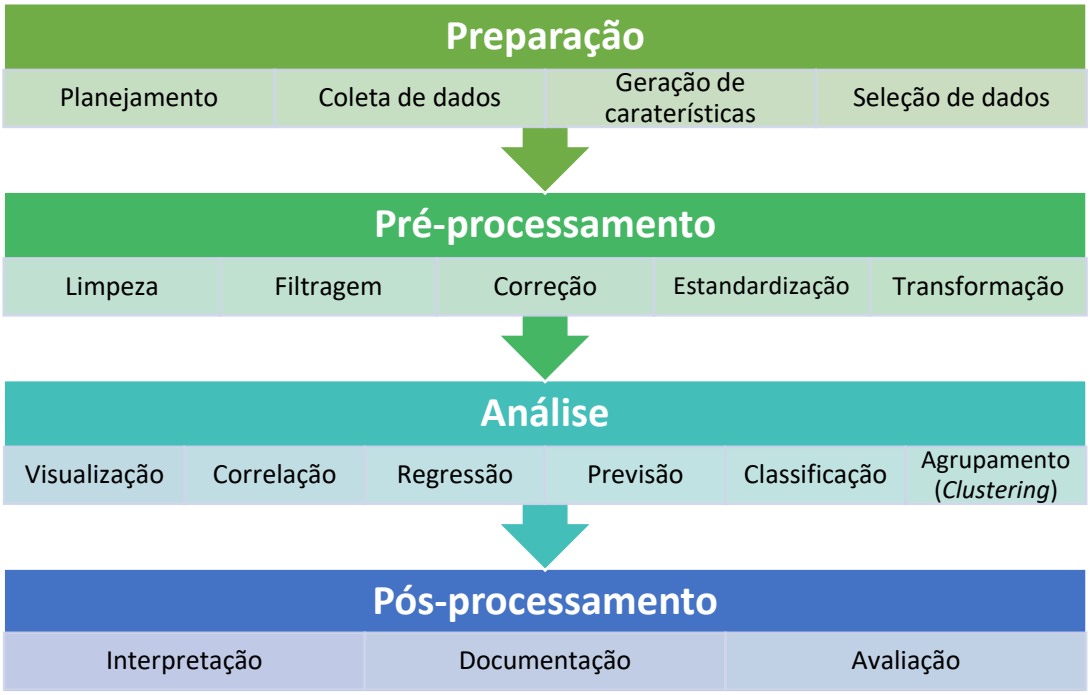
Fonte: Adaptada de Song e Zhu (2017).

A estrutura utilizada para o ciclo de vida do projeto na Análise de Dados é apresentado na Figura 17, onde são especificados os principais processos e tarefas associadas a cada fase do ciclo de vida.

Comparando-se os diagramas de ciclos de vida apresentados na Figura 15 e Figura 16 com a Figura 17, é possível concluir que o fluxo dos ciclos é diferente, pois num caso é cíclico (Figura 15 e Figura 16), enquanto no outro é linear (Figura 17).

Os diferentes comportamentos dos ciclos de vida dos projetos na Ciência e a Análise de Dados deve-se ao objetivo de cada disciplina, onde a Análise de Dados pretende resolver questões específicas (RUNKLER, 2016), enquanto que a Ciência de Dados tenta a descobrir problemas, resolve-los e comunicar as possíveis soluções (CONCOLATO; CHEN, 2017; SONG; ZHU, 2017).

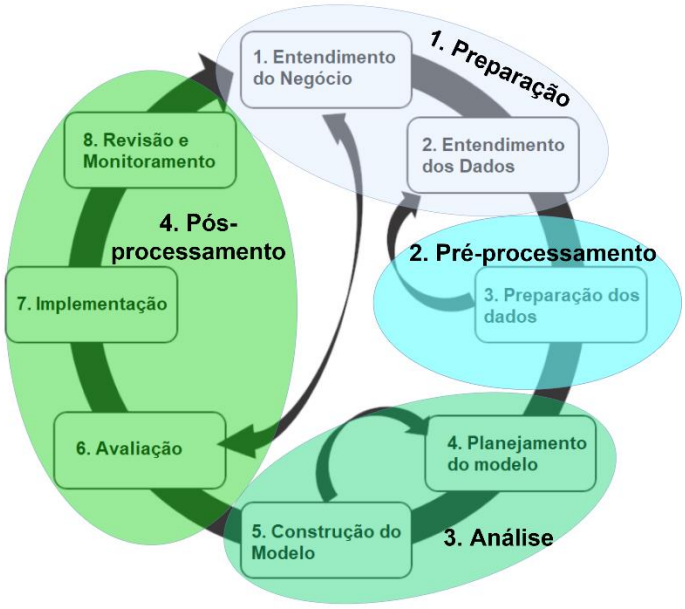
Figura 17 – Ciclo de vida de projeto na Análise de Dados.



Fonte: Elaborado pelo Autor, baseada em Runkler (2016).

Também podemos identificar algumas semelhanças nos ciclos de vida da Ciência de Dados e da Análise de Dados, pois apresentam fases com um conteúdo e forma de agir muito parecida. Na Figura 18 é apresentada uma relação das fases no ciclo de vida de projeto nas duas áreas.

Figura 18 – Comparativa do Ciclo de vida de projeto na Ciência e Análise de Dados.



Fonte: Elaborado pelo autor.

Podemos definir de forma abrangente as tarefas apresentadas na Figura 18, relacionadas a cada etapa do Ciclo de vida de projeto para a Ciência e Análise de Dados (RUNKLER, 2016; SONG; ZHU, 2017):

1. **Preparação:**
   - Definição do problema e as métricas para sua avaliação.
   - Geração de hipóteses (principalmente na Ciência de Dados).
   - Identificação das fontes dos dados.
   - Aquisição e Seleção dos dados.

2. **Pré-processamento:**
   - Limpeza, filtragem, correção, padronização e transformação de dados.

3. **Análise:**
   - Análise exploratória dos dados.
   - Escolha do tipo de modelo, de regressão, previsão, classificação ou agrupamento.
   - Seleção de métodos e técnicas para a criação do modelo.
   - Seleção das variáveis chave.
   - Criação de modelos.
   - Analisar os resultados dos modelos, até obter o modelo sob os parâmetros desejados.

4. **Pós-processamento:**
   - Avaliação dos modelos de acordo às métricas.
   - Interpretação dos modelos.
   - Documentação dos modelos.
   - Comunicação de resultados.
   - Integração de modelos em painéis de controle (principalmente na Ciência de Dados).
   - Comunicação de recomendações (principalmente na Ciência de Dados).
   - Monitoramento de desempenho e Melhora (principalmente na Ciência de Dados).

As ferramentas de software frequentemente utilizadas para a Ciência e Análise de Dados são R, Python, Matlab, SAS, KNIME, SPSS, STATISTICA, TIBCO Rapid Miner, Tableau, QlikView, oder WEKA, entre outras (RUNKLER, 2016).

Neste trabalho de doutoramento foi utilizada a linguagem Python, que é uma linguagem de programação muito utilizada na comunidade cientifica, devido a sua versatilidade em áreas como a criação de scripts e software, desenvolvimento de websites, manipulação de dados, entre outras aplicações. Além disso, oferece um amplo número de bibliotecas e pacotes que facilitam a manipulação e análise de dados, além da criação de modelos estatísticos, de aprendizado de máquina, redes neurais etc.

4.3 Ciclo de Vida na Ciência e Análise de Dados - Fundamentos Teóricos

Apresentaremos a seguir alguns fundamentos teóricos a serem considerados para o desenvolvimento das diferentes etapas do Ciclo de Vida na Ciência e Análise de Dados.

4.3.1 Etapa de Preparação

Na etapa de preparação, devemos definir inicialmente, o problema a ser resolvido, que irá permitir a correta documentação e planejamento da abordagem do problema, sendo indispensável a obtenção de dados confiáveis para realização do estudo (NELLI, 2018) resultado de pesquisas, experimentos e dados obtidos através da observação, que de acordo com a sua natureza deverão ser armazenados e tratados de diferentes formas (HEUMANN; SCHOMAKER; SHALABH, 2017; MADSEN, 2016; ROSE, 2016). Podemos destacar três formas de tipos de dados:

- **Dados Estruturados**: Seguem o formato para numa ordem específica de dados, como por exemplo ocorre nos casos dos extratos bancários, informações de voos, horários de ônibus, etc. Esses dados podem ser compartilhados através de planilhas de Excel, arquivos de variável separados através de vírgula (csv – *comma separated values*) ou separados através de tabulação (tsv - *tab separated variable*).
- **Dados Semiestruturados**: Apresentam alguma forma de estrutura, como também possuem a flexibilidade de alteração de nomes de campos e implementação de valores, como por exemplo ocorre num correio eletrônico, blogs e conteúdo das redes sociais, arquivos que podem ser compartilhados em formato XML (*Extended Markup Language*) ou JSON (*JavaScript Object Notation*).

- **Dados não estruturados**: Não seguem uma forma de estrutura específica de dados, não apresentando um modelo de dados, como acorre nos casos de uma mensagem de voz, fotos, notas digitais manuscritas, apresentações etc.

Atributos são geralmente as caraterísticas correspondentes a cada coluna de uma tabela de dados. Num banco de dados estruturados é possível identificar seus atributos e sua natureza (JOSHI, 2020). Cada atributo poderá ter uma natureza diferente, tais como uma classificação por categoria; uma representação quantitativa ou qualitativa: discreta, continua, ordinal e nominal, e poderão ser armazenados de diferentes formas: cadeias de caracteres, valores inteiros, data e hora, representação binária, etc. (HEUMANN; SCHOMAKER; SHALABH, 2017; JOSHI, 2020).

4.3.2 Pré-processamento

O objetivo da etapa de pré-processamento é reduzir possíveis erros e ruídos de uma base de dados coletada ou selecionada. Durante um processo é importante detectar o tipo de erro, manipulá-lo, e em alguns casos transformá-los e integrá-los com outros bancos de dados. Apresentaremos a seguir algumas generalidades dos processos descritos.

4.3.2.1 Tipos de erro

Numa base de dados é possível encontrar erros nos seus diferentes atributos, o que pode levar a resultados incorretos durante uma análise de dados. Existem diferentes tipos de erros. Dentre os mais relevantes (RUNKLER, 2016) podemos encontrar:

a) **Outliers ou valores atípicos:** são os dados individuais que apresentam grandes desvios ao ser comparados com os outros dados da série, e podem ocorrer devido à efeitos estocásticos ou determinísticos.

Uma das técnicas mais utilizadas na detecção de outliers é a **regra *2-sigma***, que compara a média da série com o desvio padrão como mostra a Eq. (1):

$$x_k \text{ é } outlier \iff |x_k - \mu| > 2\,\sigma \qquad (1)$$

onde x_k é o valor em estudo, μ é a média da série e σ o desvio padrão da série.

b) **Dados inválidos:** representam os valores que estão fora dos limites permitidos por um determinado atributo, como acontece em valores restringidos pelo sinal (exemplo: seu preço) ou dados em intervalos específicos (exemplo: faixa de utilização de sensores).

c) **Dados constantes:** são os valores que não alteram, e podem ser corretos ou errôneos, mas não contém nenhuma informação útil para análise, e por este motivo devem ser removidos de uma base de dados (RUNKLER, 2016).

4.3.2.2 Manipulação de erros

Após da detecção de outliers, dados inválidos e constantes podemos manipulá-los de diferentes formas (RUNKLER, 2016):

1. **Lista inválida:** Os dados permanecem inalterados, mas os índices dos dados inválidos são armazenados em uma lista separada, que é verificada em cada etapa do processamento de dados.
2. **Valores inválidos**: O outlier é substituído por um valor inválido específico, valor designado como **NaN** (*Not a Number*).
3. **Correção ou estimação:** O valor de um dado poderá ser corrigido, caso seja inválido ou estimado no caso em que estiver faltando. As técnicas mais utilizadas são as seguintes:
    a. Substituir o valor inválido pela média, mediana, ou valor mínimo ou máximo da série.
    b. Correção utilizando o valor vizinho mais próximo.
    c. Interpolação linear dos valores existentes.
    d. Interpolação não linear dos valores existentes.
    e. Estimação do valor utilizando uma regressão.
    f. Filtragem de valores.
4. **Eliminação de vetores de atributos**: Cada vetor de atributos que contenha pelo menos um valor inválido poderá ser removido.

4.3.2.3 Transformação de dados

Uma transformação de dados é exigida quando diferentes atributos possuem intervalos consideravelmente diferentes, acarretando assim, a ocorrência de um possível erro na análise, como por exemplo as grandezas podem diferir de várias ordens (RUNKLER, 2016).

Dentre as transformações mais utilizadas podemos destacar a padronização de dados, a transformação inversa, transformação de raiz, transformação logarítmica, transformação de Fisher, entre outras (RUNKLER, 2016).

Neste trabalho será utilizada a padronização dos dados, que visa a transformação dos dados para uma mesma ordem de grandeza. Essa padronização tem como objetivo deixar as variáveis com uma média igual a 0 e um desvio padrão igual a 1, como mostra a Eq. (2):

$$z = \frac{x_k - \mu}{\sigma} \tag{2}$$

onde z é o valor padronizado, x_k é o valor a padronizar, μ a média e σ o desvio padrão dos dados da variável em estudo.

4.3.2.4 Integração de dados

As bases de dados podem fornecer diversas informações consideradas relevantes, e por esse motivo, torna-se muito importante a integração de múltiplas bases de dados de diferentes fontes (RUNKLER, 2016). A Figura 19 exemplifica a integração de diferentes bases de dados, onde cada banco de dados apresenta um atributo, que é integrado utilizando a data.

Figura 19 – Exemplo de Integração de bancos de dados baseados na Data.

Data	Dado 1
...	...
...	...

+

Data	Dado 2
...	...
...	...

=

Data	Dado 1	Dado 2
...	...	...
...	...	...

Fonte: Elaborado pelo Autor, baseado em Runkler (2016).

4.3.3 Análise de dados

A fase de análise de dados compreende as diferentes atividades, cujo objetivo é a criação de um modelo útil para o problema e o tipo de dados em estudo. Inicialmente, utilizando conceitos estatísticos, é realizado uma aproximação dos dados, através de sua exploração e representação gráfica, sendo definido o tipo de modelo a ser utilizado, para posterior desenvolvimento do modelo de interesse.

4.3.3.1 Análise Estatística

Um conjunto de dados pode conter muitas variáveis e observações, sendo importante o uso de funções estatísticas que consigam expressar esses os dados de uma forma significativa. Dentre as funções estatísticas mais relevantes para análise dados podemos destacar as medidas de tendência central, medidas de dispersão e outras caraterísticas no estudo de séries temporais.

4.3.3.1.1 Medidas de Tendência Central

Dentre as medidas de tendência central mais importantes estão a média, mediana, os quartis, a moda. É importante destacar que no caso da distribuição dos dados em estudo for uma distribuição normal, a média e a mediana apresentarão os mesmos valores, e em qualquer outro tipo de distribuição dos dados os valores serão diferentes.

- **Média:** é um valor que representa o centro da série de dados em estudo e é calculada como mostra a Eq. (3):

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3}$$

  onde μ é a média, x_i é cada dado da série de dados e n é o número de elementos da série

- **Mediana:** é o valor médio localizado na lista ordenada de valores da série em estudo. Quando a lista for ímpar esse valor estará localizado na metade da lista, e quando a lista for par esse valor será representado pela média dos dois valores localizados na metade da lista.

- **Quartil:** generalização da ideia de mediana, onde enquanto a mediana divide uma lista de valores ordenados em duas partes iguais, o quartil divide os dados de uma lista em quatro porções iguais.
- **Percentil:** generalização da ideia de quartil, que divide os dados em cem partes iguais. É possível afirmar que o quartil inferior, primeiro quartil ou $Q_{1/4}$ equivale ao 25° percentil, assim mesmo o segundo quartil ou $Q_{2/4}$ equivale à mediana e ao 50° percentil, finalmente o terceiro quartil, quartil superior ou $Q_{3/4}$ ao 75° percentil.
- **Moda:** é o valor que mais se repete numa lista de dados. Para determiná-lo basta realizar a contagem de cada item de uma lista, e encontrar a frequência de cada um, selecionando o mais frequente.

4.3.3.1.2 Medidas de Dispersão

- **Variância:** medida de dispersão representada pelo quadrado da média das distâncias entre os valores dos dados e a média (não está na mesma unidade de medida dos dados), como mostra a Eq. (4):

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{4}$$

onde μ é a média, x_i é cada dado da série de dados, n é o número de elementos da série e σ^2 a variância.

- **Desvio padrão:** é a raiz quadrada da variância, sendo uma medida de dispersão que está na mesma unidade de medida que os dados, podendo assim, ser interpretada como "a distância média" entre os valores dos dados e a média como mostra a Eq. (5),

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \tag{5}$$

onde μ é a média, x_i é cada dado da série de dados, n é o número de elementos da série e σ o desvio padrão.

4.3.3.1.3 Medidas para análise de Séries Temporais

Séries temporais podem ser analisadas a partir da utilização de ferramentas estatísticas para a decomposição de uma série temporal, permitindo assim, a obtenção de diferentes parâmetros, tais como o ciclo de tendência através da simplificação, a componente sazonal, e o ruído (ou erro), e fatores dependentes do tempo. A partir destes resultados é possível analisarmos o comportamento, e os possíveis componentes da série temporal em estudo, e esta informação será importante para o desenvolvimento e escolha de parâmetros em modelos de previsão.

A biblioteca utilizada neste trabalho para a decomposição das séries temporais foi a *statsmodels* (SEABOLD; PERKTOLD, 2010), que contém diversas ferramentas estatísticas disponibilizadas em linguagem de programação Python. A decomposição das séries temporais foi realizada utilizando duas estruturas diferentes, conforme a natureza da variável em estudo:

- **Decomposição aditiva:** mais apropriada de ser utilizada quando a tendência e sazonalidade são aproximadamente constantes com ocorrência de poucas variações.
- **Decomposição multiplicativa:** mais apropriada de ser utilizada, no caso em que a tendência e sazonalidade apresentem mudanças de forma não linear (HYNDMAN; ATHANASOPOULOS, 2018).

O modelo matemático da decomposição aditiva, é mostrado através da Eq. (6), enquanto que a Eq. (7), mostra a decomposição multiplicativa:

$$y_t = S_t + T_t + R_t \tag{6}$$

$$y_t = S_t \times T_t \times R_t \tag{7}$$

onde

y_t: Dados da série temporal,

S_t: Componente sazonal,

T_t: Tendência e

R_t: Ruído.

4.3.3.2 Representação Gráfica de Dados

Uma representação gráfica de dados permitirá examinar os mesmos, possibilitando encontrar valores improváveis ou errados, ajudando assim à detecção de possíveis erros nos bancos de dados. Ela também é importante para ter uma ideia de padrões, estruturas, tendências e relacionamentos entre os diferentes atributos, representando uma importante ajuda complementar para a análise estatística (MADSEN, 2016).

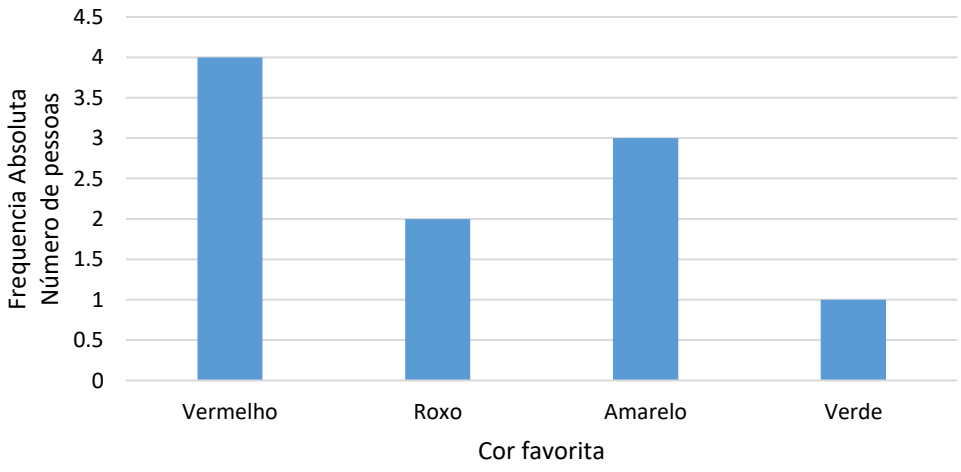
Existem múltiplas representações gráficas para os dados, entre os mais destacados encontramos os gráficos de barras, histogramas, gráfico de setores, gráficos de dispersão, gráficos de linhas e diagrama de caixa, que serão apresentados a seguir.

4.3.3.2.1 Gráficos de barras

Um gráfico de barras é uma ferramenta simples para a visualização de frequências relativas ou absolutas dos valores observados para uma determinada variável. O uso deste tipo de representação é bem abrangente, e permitirá a representação de variáveis nominais e ordinárias.

Usualmente na implementação deste gráfico, a altura de cada barra (eixo *y*) é determinada pela frequência absoluta ou relativa da respectiva categoria que é representada no eixo *x*. Na Figura 20 é mostrado um exemplo de representação através de gráfico de barras.

Figura 20 – Exemplo de Gráfico de Barras.

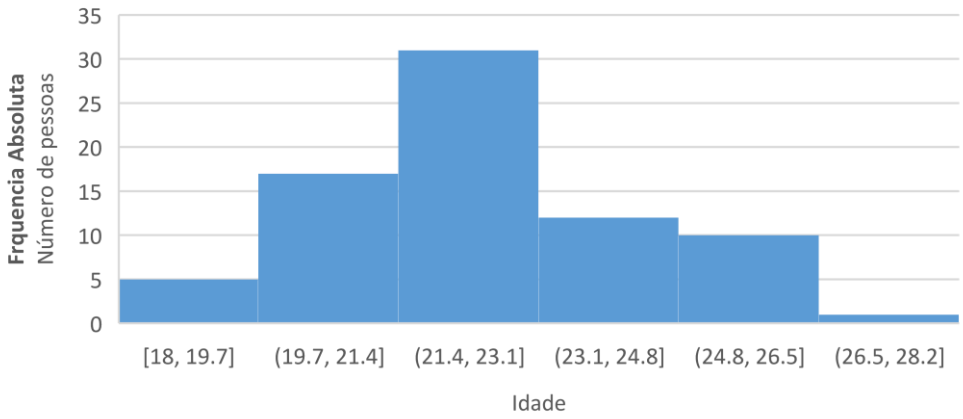


Fonte: Elaborado pelo Autor.

4.3.3.2.2 Histogramas

Um Histogramas é uma representação da distribuição de valores de um conjunto de dados, onde são apresentados os dados em diferentes grupos, mostrando sob a forma de barras cada categoria (eixo x), e o valor da frequência absoluta na vertical (eixo y). Na Figura 21 é possível observar um exemplo de um histograma que apresenta número de alunos numa sala de aula em função de sua idade.

Figura 21 – Exemplo de um Histograma.

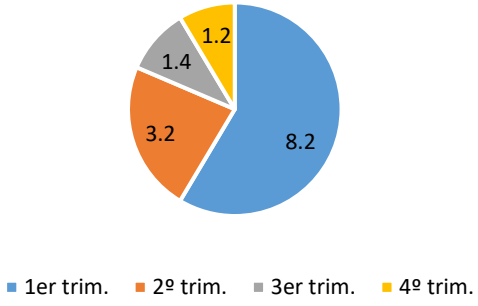


Fonte: Elaborado pelo Autor.

4.3.3.2.3 Gráfico de setores

O gráfico de setores (ou de pizza) é um círculo particionado em segmentos, onde cada segmento representa uma categoria, permitindo assim, a visualização de frequências absolutas ou relativas de variáveis nominais e ordinárias. Cada segmento representará a frequência relativa da categoria, como mostra a Figura 22.

Figura 22 – Exemplo de Gráfico de Setores.

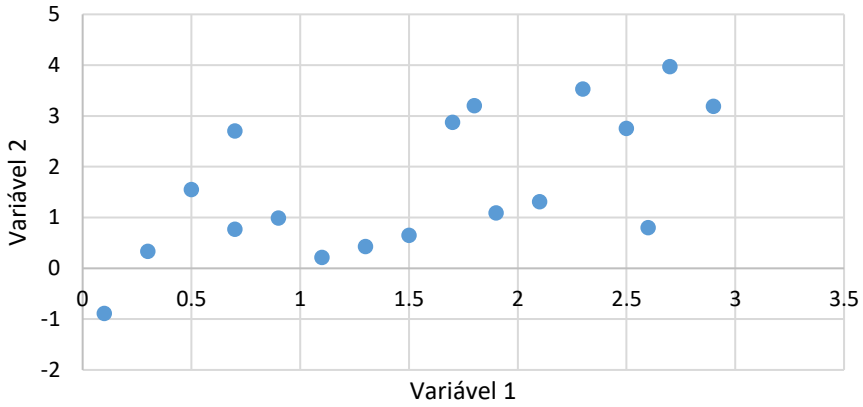


Fonte: Elaborado pelo Autor.

4.3.3.2.4 Gráfico de Dispersão

Um gráfico de dispersão permite visualizar graficamente a relação entre duas variáveis contínuas, sendo obtido através da representação de observações emparelhadas de duas variáveis num sistema de coordenadas bidimensional, como apresenta a Figura 23.

Figura 23 – Exemplo de Gráfico de Dispersão.

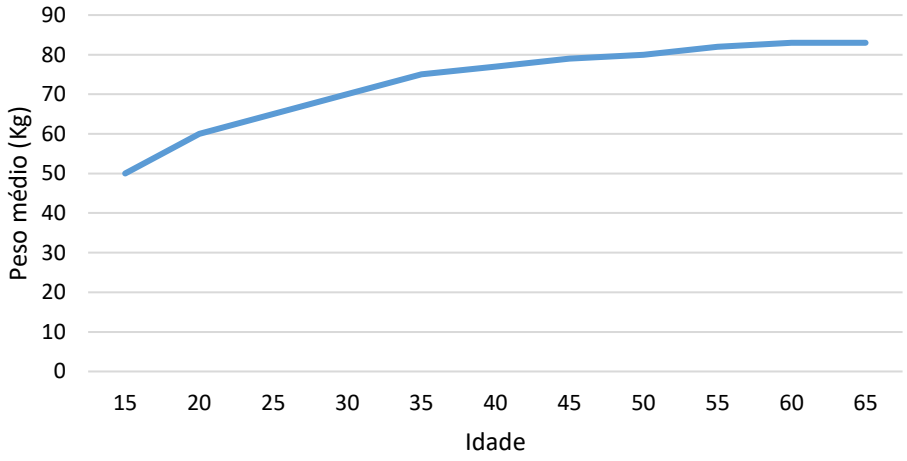


Fonte: Elaborado pelo Autor.

4.3.3.2.5 Gráficos de Linhas

Os gráficos de linhas são utilizados para representar uma tendência, permitindo distinguir o comportamento de uma variável (considerada dependente) em termos de uma outra variável (independente). A Figura 24 mostra a representação do peso médio de uma pessoa em função de sua idade.

Figura 24 – Exemplo de Gráfico de Linhas.

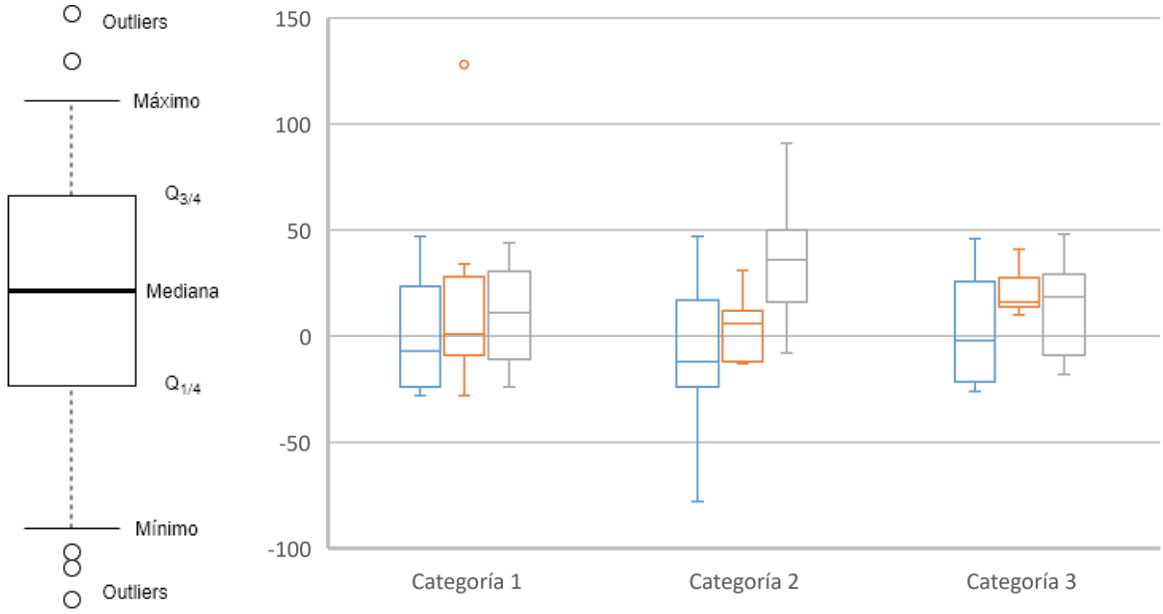


Fonte: Elaborado pelo Autor.

4.3.3.2.6 Diagrama de Caixas

O diagrama de caixa é uma representação gráfica que resume várias medidas de tendência central e dispersão, permitindo assim observar a distribuição de uma variável usando sua mediana, quartis, mínimos, máximos e outliers (Figura 25).

Figura 25 – Componentes do Gráfico de Caixa e exemplo.

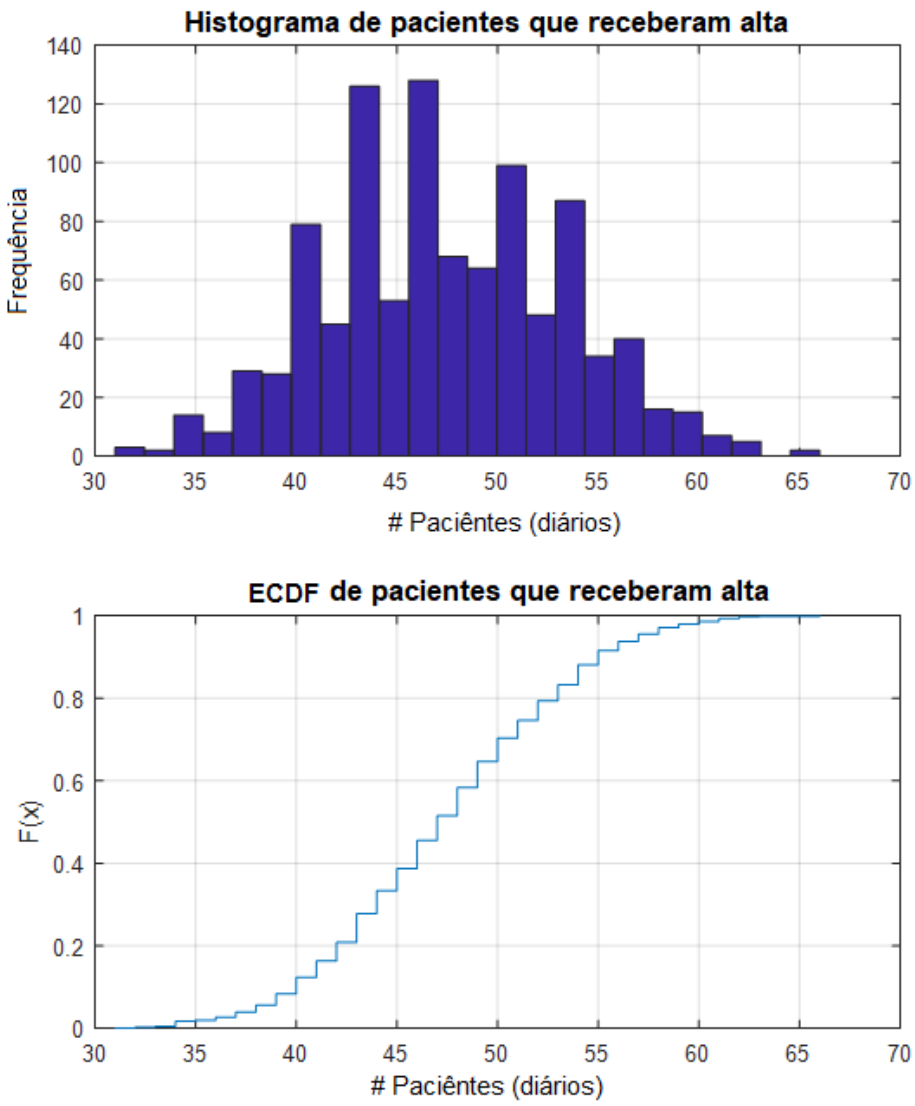


Fonte: Elaborado pelo Autor.

4.3.3.2.7 Função de distribuição acumulada empírica

A Função de distribuição acumulada empírica (ECDF – *Empirical Cumulative Distribution Function*) é uma abordagem utilizada para a visualização simultânea da frequência de variáveis e seu histograma, focado na representação empírica da frequência relativa da distribuição das variáveis (Figura 26).

Figura 26 – Exemplo da ECDF e comparativa com o histograma.



Fonte: Elaborado pelo Autor.

4.3.3.3 Tipos de Modelos

Para a modelagem dos dados utilizamos geralmente modelos preditivos e modelos descritivos. Modelos preditivos têm como objetivo a previsão do comportamento do sistema ao que os dados foram coletados, não importando a capacidade de interpretação do modelo, enquanto que os modelos descritivos permitem dar uma interpretação aos fenômenos apresentados através de um conjunto de dados, permitindo assim a detecção das variáveis que conseguem influir na dinâmica do sistema do qual os dados foram coletados.

Os tipos de modelos deverão ser selecionados de acordo com seus objetivos (NELLI, 2018), podendo ser definidos como modelos de:

- **Classificação:** modelos com resultado de tipo de categoria.
- **Regressão:** modelos com resultado de tipo numérico.
- **Agrupamento:** modelos com resultado de tipo descritivo.

Para esses diferentes modelos podemos encontrar técnicas clássicas de modelagem baseadas no aprendizado baseado no uso de estatística, tais como a regressão linear, regressão logística etc. Também é possível encontrar modelos fundamentados em IA, e particularmente no aprendizado de máquina como as redes neurais, redes neurais profundas etc.

Neste trabalho de doutoramento serão aprofundadas as técnicas relacionadas com a modelagem de séries temporais, utilizado nos estudos de caso propostos neste trabalho. No próximo capítulo, aprofundaremos algumas técnicas fundamentadas na utilização de Redes Neurais para regressão e previsão. As técnicas de regressão clássicas, baseadas no aprendizado estatístico não serão utilizadas neste trabalho, e muitas referências nesta área poderão ser encontradas na literatura (JAMES et al., 2013; RUNKLER, 2016).

Para a previsão de séries temporais abordadas neste trabalho serão utilizados os modelos de previsão: ingênua, os métodos de suavização exponencial, e os Modelo Auto Regressivo Integrado de Médias Móveis (ARIMA), descritos a seguir.

4.3.3.3.1 Previsão Ingênua

O modelo de previsão ingênua é um método simples de previsão de séries temporais baseado em dados históricos, como mostra a Eq. (8):

$$y_{t+h} = y_t \tag{8}$$

onde y_t é o dado histórico, e irá predizer o valor de y_{t+h}, e h indica o horizonte de previsão selecionado para o modelo, onde esse valor é usualmente é 1, pois a previsão desejada é um passo no futuro. Da mesma forma, em alguns modelos sazonais pode ser utilizado $h = k + 1$, onde k é o valor da sazonalidade da série temporal.

Este modelo é um modelo aproximado, matematicamente muito simples, possibilitando com ser validado com bom desempenho em alguns casos, tais como em previsão de modelos econômicos e financeiros (HYNDMAN; ATHANASOPOULOS, 2018).

4.3.3.3.2 Previsão baseada nos Métodos de Suavização Exponencial

Este modelo de previsão é baseado na realização de médias ponderadas do histórico de dados, com um decaimento exponencial dos pesos, sendo muito utilizado em ampla gama de aplicações industriais (HYNDMAN; ATHANASOPOULOS, 2018).

Neste trabalho de pesquisa será utilizado o Método de Suavização Exponencial Simples e Dupla. No método simples os valores do histórico são multiplicados por uma constante que decresce exponencialmente, como mostra a Eq. (9):

$$y_{t+1} = \alpha\, y_t + (1-\alpha)\, y_{t-1} + \ (1-\alpha)^2 y_{t-2} + \cdots + (1-\alpha)^n y_{t-n} \qquad (9)$$

onde, y é a variável de interesse para diferentes instantes de tempo. O valor α é a constante de suavização, que corresponde a um valor $0 \leq \alpha \leq 1$, fazendo decrescer a suavização exponencialmente a razão de *n*.

A Eq. (10) define o Método de Suavização Exponencial Dupla:

$$\begin{gathered} l_t = \ \alpha\, y_t + (1-\alpha)\, (l_{t-1} + b_{t-1}) \\ b_t = \beta(l_t - l_{t-1}) + (1-\beta)\, b_{t-1} \\ y_{t+h} = l_t + h\, b_t \end{gathered} \qquad (10)$$

onde l_t corresponde a uma estimativa do nível da série no momento t, b_t indica o nível da tendência ou bias da série temporal no tempo t. Os parâmetros das equações α, a constante de suavização; e β, a constante de suavização da tendência, onde $0 \leq \alpha \leq 1$ e $0 \leq \beta \leq 1$.

Para a implementação e otimização das constantes dos Métodos de Suavização Exponencial Simples e Dupla neste trabalho foi utilizada a biblioteca *statsmodels*.

4.3.3.3.3 Previsão baseada no Modelo Auto regressivo Integrado de Médias Móveis (ARIMA)

O modelo de previsão ARIMA e de suavização exponencial são métodos muito utilizados para a previsão de séries temporais. Ao contrário dos métodos de suavização exponencial, o método ARIMA oferece uma abordagem que tem como objetivo a descrição dos valores de autocorrelação da série temporal em estudo, conseguindo assim realizar uma previsão baseada nesta informação (HYNDMAN; ATHANASOPOULOS, 2018).

A notação deste modelo obedece a forma ARIMA (*p, d, q*), onde *p* é a ordem do processo autorregressivo (*AR*), *d* corresponde ao grau de diferenciação envolvido (*I*), e *q* indica a ordem do processo de Média Móvel (*MA*).

O modelo ARIMA sazonal permite a modelagem de uma série temporal sazonal e sua notação usual é ARIMA $(p,d,q)(P,D,Q)^s$, onde *P, D* e *Q* têm o mesmo significado que *p, d* e *q*, mas para a parte sazonal do modelo, e *s* é o número de períodos em um ciclo sazonal. O modelo ARIMA sazonal implementado neste trabalho segue inicialmente a metodologia de Box-Jenkins descrita em (BOX et al., 2015).

Para conseguir a implementação indicada do modelo ARIMA é indispensável que a série temporal em estudo apresente a propriedade da estacionariedade.

Uma série temporal estacionária apresenta propriedades como uma média, variância e autocorrelação constante no tempo. Se uma série não for estacionaria, deverá ser utilizado alguns métodos que irão permitir a estacionariedade da mesma. Dentre os métodos utilizados para atingir a estacionariedade de uma série temporal, podemos destacar os métodos da diferenciação, o da eliminação de tendência, e a aplicação da função logarítmica ou a raiz quadrada para a estabilização da variância.

O método mais utilizado é o método da diferenciação, que é apresentado na Eq. (11), onde y_i' é o resultado da primeira diferença da série, e y_i a série temporal não estacionária.

$$y_i' = y_i - y_{i-1} \tag{11}$$

O método da diferenciação pode ser aplicado múltiplas vezes, e de diferentes formas dependendo da natureza da série temporal não estacionaria. As séries estacionarias sazonais utilizam o valor da constante de sazonalidade (*k*) como termo de diferenciação, como apresenta a Eq. (12).

$$y_i' = y_i - y_{i-k} \tag{12}$$

Posteriormente, é possível aplicar um teste de estacionariedade ou de raiz unitária, como são os casos dos testes de Dicker-Fuller, o Aumentado de Dickey-Fuller, o de Phillips-Perron, dentre outros. O teste vai indicar se a série temporal obtida após de hipoteticamente atingir a estacionariedade, e no caso de não seja atingida é importante aplicar novamente as técnicas para atingir esta propriedade. A seguir, deverá ser realizado a aproximação das constantes *p* e *q* mediante o uso de gráficas da função de autocorrelação (ACF) e autocorrelação parcial (PACF).

A ACF é a correlação cruzada de uma série temporal com ela mesma, sendo importante para a identificação de padrões repetitivos no sinal, enquanto, a PACF é a aplicação da ACF a uma série temporal estacionária com seus próprios valores desfasados. Seguindo as recomendações de Box et al. (2015) é possível indicar um modelo de acordo com os tipos de gráficas obtidas na ACF e PACF.

Nos dias atuais, devido a elevada capacidade computacional disponibilizada, é possível utilizar métodos para obter as constantes *p*, *d* e *q* utilizando uma busca em grade (*Grid Search*), que é a otimização de hiper-parâmetros, como *p, d* e *q.* A busca em grade faz uma busca exaustiva da melhor combinação de hiper-parâmetros, onde um intervalo de busca é indicado para cada hiper-parâmetro.

No caso da busca em grade do modelo ARIMA é utilizada a função *auto-arima* do pacote para Python *pmdarima* (SMITH, 2017). O critério de escolha utilizado para o modelo ARIMA é baseado no critério de informação de Akaike (AIC), que indica a qualidade dos modelos estatísticos para um determinado conjunto de dados (TADDY, 2019).

4.3.4 Pós-processamento

Na etapa de pós-processamento podemos destacar a avaliação dos modelos obtidos utilizando métricas, pois é de vital importância comparar os resultados dos diferentes métodos. Nenhum método consegue dominar em todos os conjuntos de dados, como mostra a teoria do *no free lunch*. Logo, um método específico pode funcionar melhor, mas algum outro método pode funcionar melhor em um conjunto de dados semelhante, mas diferente.

Existem diferentes métricas para a análise e avaliação dos modelos, permitindo assim uma comparativa quantitativa do desempenho de cada modelo. A seguir serão abordadas algumas métricas tipicamente utilizadas para analisar modelos previsão, e em alguns casos modelos de regressão, onde estas métricas são baseadas no erro. As métricas mais utilizadas para tal finalidade são o Erro Quadrático Médio (EQM), a Raiz do EQM (REQM), Erro Médio Absoluto (EMA) e Coeficiente de Determinação ou R^2.

4.3.4.1 Erro Quadrático Médio (EQM)

O Erro Quadrático Médio (EQM) é uma métrica que mostra a média da diferença entre o valor esperado e o obtido ao quadrado, como mostra a Eq. (13).

$$EQM(y, \hat{y}) = \frac{1}{s} \sum_{i=0}^{s-1} (y_i - \hat{y}_i)^2 \tag{13}$$

onde y é um vetor com o valor esperado ou real, $\hat{y}$ é um vetor com os valores da previsão feita pelo modelo em estudo, e s o número de amostras.

O EQM é um valor que não está perto da escala das bases de dados ou dos dados padronizados, pois penaliza ao quadrado os erros muito grandes, ou seja, um menor EQM significa um melhor modelo.

4.3.4.2 Raiz do EQM (REQM)

A Raiz do Erro Quadrático Médio (REQM) é uma métrica que avalia o erro, como mostra a Eq. (14).

$$REQM(y, \hat{y}) = \sqrt{\frac{1}{s} \sum_{i=0}^{s-1} (y_i - \hat{y}_i)^2} \tag{14}$$

onde y é um vetor com o valor esperado ou real, $\hat{y}$ é um vetor com os valores da previsão feita pelo modelo em estudo, e s o número de amostras.

O REQM é um valor que está na mesma escala dos valores das bases de dados ou dos dados padronizados, obtendo a raiz quadrada do EQM, ou seja, um menor REQM significa um melhor modelo.

4.3.4.3 Erro Médio Absoluto (EMA)

O Erro Médio Absoluto (EMA) é uma métrica que mostra a média do valor absoluto da diferença entre o valor esperado e o obtido, como mostra a Eq. (15).

$$EMA(y, \hat{y}) = \frac{1}{s} \sum_{i=0}^{s-1} |y_i - \hat{y}_i| \tag{15}$$

onde y é um vetor com o valor esperado ou real, $\hat{y}$ é um vetor com os valores da previsão feita pelo modelo em estudo, e s o número de amostras.

O EMA é um valor que está na mesma escala dos valores das bases de dados ou dos dados padronizados, mas a diferença com o REQM é que não penaliza os erros quadráticos, ou seja, um menor EMA significa um melhor modelo.

4.3.4.4 Coeficiente de Determinação ou R^2

O Coeficiente de Determinação ou R^2 é uma proporção que é utilizada para determinar a qualidade de ajuste de um modelo, indicando a quantidade de amostras não vistas e provavelmente serão previstas pelo modelo. A Eq. (16) mostra matematicamente o Coeficiente de Determinação ou R^2.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{s-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{s-1}(y_i - \mu)^2} \tag{16}$$

onde μ é a média aritmética de y, y é um vetor com o valor esperado ou real, $\hat{y}$ é um vetor com os valores da previsão feita pelo modelo em estudo, e s o número de amostras.

O melhor valor possível de R^2 é 1.0, que indica que o modelo de previsão está perfeitamente ajustado com os dados reais. Um resultado de -1.0 é possível e indica que a relação é inversa entre o modelo de previsão e a saída esperada. Um resultado de 0.0 mostra a incapacidade de previsão do modelo.

4.4 Conclusões do Capítulo e próximas etapas

Neste capítulo concluímos que a Ciência de Dados e a Análise de Dados são áreas correlacionadas, com um ciclo de vida do projeto muito semelhante e bem estruturado, fundamentada particularmente no tratamento e interpretação de dados através de técnicas e modelos estatísticos. A interpretação e entendimento dos diferentes conjuntos de dados permite extrair informação e conclusões chave para o entendimento dos sistemas em estudo e a possível proposta de soluções de natureza descritiva ou preditiva.

Por outro lado, a Inteligência Artificial é uma área muito ampla que complementa e permite modelos mais complexos e robustos na Ciência de Dados e Análise de Dados, através do uso de aprendizado de máquina. Além da modelagem, através da IA podemos encontrar algoritmos de otimização, que são importantes para a solução de problemas e melhoria de sistemas complexos. A IA é uma área fundamental que permite autonomia e robustez no contexto da I4.0 e a PO, mostrando a grande capacidade do Big Data no desenvolvimento de modelos inteligentes.

O próximo capítulo desta tese de doutoramento corresponde à Inteligência Artificial (IA), onde serão definidos alguns fundamentos teóricos, aprofundando particularmente as áreas da computação evolutiva e a aprendizagem de máquina. Serão detalhados os algoritmos genéticos, as redes neurais e o aprendizado por reforço, descrevendo sua metodologia e aplicações, orientando a abordagem para os casos de estudo presentes neste trabalho.

Referência bibliográfica

https://repositorio.unicamp.br/acervo/detalhe/1149374?guid=1693595816980&returnUrl=%2fresultado%2flistar%3fguid%3d1693595816980%26quantidadePaginas%3d1%26codigoRegistro%3d1149374%231149374&i=1