

**UNIVERSIDADE ESTADUAL PAULISTA “JÚLIO DE MESQUITA FILHO”
FACULDADE DE ENGENHARIA
CAMPUS DE ILHA SOLTEIRA**

EDUARDO GARCEZ DE SOUZA

SI-CHRONOS (Sistema Integrado de Análises Cronológicas)

Orientadora: Prof.^a Dra. Christiane Marie Schweitzer

Ilha Solteira
Ano 2024

EDUARDO GARCEZ DE SOUZA

SI-CHRONOS (Sistema Integrado de Análises Cronológicas)

Trabalho de conclusão de curso
apresentado à Universidade Estadual
Paulista (UNESP), para obtenção do título
de Bacharel em Engenharia Mecânica.

Orientadora: Prof.^a Dra. Christiane Marie
Schweitzer

Ilha Solteira

2024

FICHA CATALOGRÁFICA

Desenvolvido pelo Serviço Técnico de Biblioteca e Documentação

S729s Souza, Eduardo Garcez de.
SI-CHRONOS (sistema integrado de análises cronológicas) / Eduardo Garcez de Souza. -- Ilha Solteira: [s.n.], 2024
62 f. : il.

Trabalho de conclusão de curso (Graduação em Engenharia Mecânica) - Universidade Estadual Paulista (UNESP), Faculdade de Engenharia, Ilha Solteira, 2024

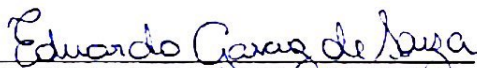
Orientador(a): Christiane Marie Schweitzer

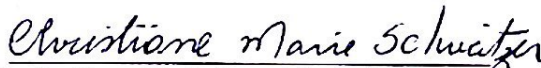
Inclui bibliografia

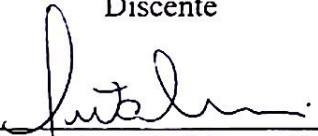
1. Análise de dados. 2. Python. 3. Interface gráfica. 4. Estatísticas. 5. Ferramenta de análise. 6. Visualização de dados.

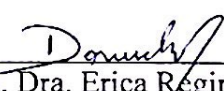
ATA DE DEFESA DE TRABALHO DE GRADUAÇÃO

Aos dezessete dias do mês de junho do ano de dois mil e vinte e quatro, o discente **EDUARDO GARCEZ DE SOUZA**, matriculada sob o RA: 182054721, tendo como banca examinadora sua orientadora a Profa. Dra. Christiane Marie Schweitzer, o Prof. Dr. Antonio Marcos Cossi e a Profa. Dra. Erica Regina Marani Daruichi Machado, apresentou o Trabalho de Graduação do Curso de Engenharia Mecânica (FEIS/UNESP), intitulado "SI-CHRONOS (Sistema Integrado de Análises Cronológicas)", obtendo a nota 10 (Dez - APROVADO).


EDUARDO GARCEZ DE SOUZA
Discente


Profa. Dra. Christiane Marie Schweitzer
Orientadora


Prof. Dr. Antonio Marcos Cossi
Membro da banca


Profa. Dra. Erica Regina Marani
Daruichi Machado
Membro da banca

RESUMO

Este trabalho apresenta o desenvolvimento de um programa em Python denominado SI-CHRONOS (Sistema Integrado de Análises Cronológicas), que visa facilitar as análises de dados de diversas áreas. O programa proporciona uma análise dos dados, permitindo ao usuário visualizar informações estatísticas, gráficos de distribuição de normalidade e gráficos de dispersão em relação ao tempo. O sistema possui uma interface gráfica amigável desenvolvida com a biblioteca Tkinter, possibilitando ao usuário interagir de forma intuitiva. O SI-CHRONOS se trata de uma contribuição significativa para a comunidade acadêmica e científica, fornecendo uma solução prática e eficiente para a análise de dados em pesquisas. O trabalho busca oferecer uma ferramenta acessível e fácil de usar, permitindo aos usuários sem experiência em programação realizar análises estatísticas de forma precisa, rápida e confiável.

Palavras-chave: Análise de dados, Python, Interface Gráfica, Estatísticas, Ferramenta de Análise, Visualização de dados.

ABSTRACT

This work presents the development of a Python program called SI-CHRONOS (Integrated Chronological Analysis System), which aims to facilitate data analysis in various fields. The program provides data analysis, allowing users to visualize statistical information, normality distribution graphs, and scatter plots over time. The system features a user-friendly graphical interface developed with the Tkinter library, enabling intuitive user interaction. SI-CHRONOS is a significant contribution to the academic and scientific community, offering a practical and efficient solution for data analysis in research. The work aims to provide an accessible and easy-to-use tool, allowing users without programming experience to perform statistical analyses accurately, quickly, and reliably.

Keywords Data Analysis, Python, Graphical Interface, Statistics, Analysis Tool, Data Visualization.

Sumário

1 – INTRODUÇÃO	10
1.1 – OBJETIVO GERAL	11
1.2 – OBJETIVO ESPECÍFICO	11
1.3. Estrutura do Trabalho	13
2 - REVISÃO DE LITERATURA	14
2.1 – ESTATÍSTICA DESCRITIVA	14
2.1.1 – Média	15
2.1.2 – Mediana	16
2.1.3 – Moda	17
2.1.4 – Variância e Desvio Padrão	17
2.2 – VARIÁVEIS ALEATÓRIAS	20
2.2.1 - Tipos de Variáveis	20
2.2.2 - Distribuição de Variáveis Aleatórias	21
2.2.3 - Frequência de Variáveis Aleatórias	25
2.3 – CORRELAÇÃO DE DADOS	26
2.4 – <i>POKA YOE</i>	27
2.5 – APLICAÇÕES DA ESTATÍSTICA NA CIÊNCIA	27
2.5.1 - Uma Pesquisa sobre Informática de Biodiversidade: Conceitos, Práticas e Desafios	27
2.5.2 - Abordagens de Aprendizado de Máquina para Previsão de Eventos Climáticos Extremos em Áreas Urbanas: Desafios e Resultados Iniciais...	29
2.5.3 - Compreensão da Mobilidade Humana e Dinâmica de Carga de Trabalho devido a Diferentes Eventos de Grande Escala Utilizando Dados de Telefones Móveis	30
3 - MATERIAL E MÉTODOS	31
3.1 - REQUISITOS	32
3.1.1 - Requisitos Funcionais	32

3.1.2 - Requisitos não funcionais	32
3.2 – AMBIENTE DE DESENVOLVIMENTO	33
3.2.1 - Ferramenta de Desenvolvimento	33
3.2.2 - Criação e Estrutura do Projeto	34
3.3 – MODELAGEM – DIAGRAMAS UML	35
3.3.1 - Casos de Uso	35
3.3.2 - Diagramas de Sequência	37
3.4 – FASE DE DESENVOLVIMENTO DO SOFTWARE	39
3.4.1 – Escolha da Linguagem de Programação	39
3.4.2 – Módulo de Interface Gráfica (UI) e Experiência do Usuário (UX)	40
3.4.3 – Módulo de Manipulação de Dados	43
3.4.4 – Módulo de Visualização de Dados	43
3.4.5 – Módulo de Controle e Lógica de Funcionamento	44
3.4.6 – Módulo de Testes e Depuração	45
3.4.7 – Módulo de Documentação	45
3.4.8 – Melhorias e Expansões Futuras	46
3.4.9 – Confiabilidade	46
4 - RESULTADOS E DISCUSSÃO	47
4.1 – Validação do sistema	47
4.1.1 – Histograma e curva de normalidade	47
4.1.2 – Melhor curva entre os pontos amostrais	49
4.1.3 – Correlação de Variáveis	51
4.1.4 – Comparativo final	53
4.2 – Aplicação	54
5 – CONCLUSÃO	58
5.1 – TRABALHOS FUTUROS	59
5.1.1 – Suporte a mais formatos	59

5.1.2 – Expansão de Funcionalidades Analíticas.....	59
5.1.3 – Transformação em um Executável.....	60
5.1.4 – Otimização de Desempenho e Escalabilidade	60
6 - REFERÊNCIAS BIBLIOGRÁFICAS.....	61

1 – INTRODUÇÃO

O uso de sistemas estatísticos para análise de dados é de suma importância em uma variedade de campos, desde a ciência e a pesquisa até o mundo dos negócios e a tomada de decisões governamentais.

Os sistemas estatísticos permitem identificar padrões e tendências nos dados que podem não ser óbvios à primeira vista. Eles fornecem ferramentas para visualizar e resumir grandes conjuntos de dados, facilitando a interpretação e a extração de visões significativas. São usados, também, para testar hipóteses e validar teorias em uma variedade de disciplinas. Fornecem métodos para determinar se os resultados observados são estatisticamente significativos e se podem ser generalizados para uma população maior.

Usados para analisar dados experimentais, avaliar a validade das descobertas e comunicar os resultados de forma clara e precisa, os sistemas estatísticos desempenham um papel crucial em todas as etapas do processo científico, desde a coleta de dados até a publicação dos resultados.

A análise estatística fornece uma base objetiva para a tomada de decisões. Em lugar de depender apenas da intuição ou da experiência pessoal, os dados estatísticos permitem que as decisões sejam fundamentadas em evidências sólidas, o que pode reduzir o viés e aumentar a eficácia das decisões. Prever eventos futuros e construir modelos que descrevam o comportamento de sistemas complexos são técnicas especialmente importantes em áreas como finanças, meteorologia, saúde, engenharia e marketing.

Atualmente, os sistemas estatísticos são eficazes na detecção de anomalias e fraudes em conjuntos de dados, ajudando a identificar padrões suspeitos ou comportamentos incomuns que podem indicar atividades fraudulentas ou problemas nos dados, fornecendo as ferramentas necessárias para extrair visões significativas, tomar decisões informadas e avançar o conhecimento em diversas áreas do conhecimento humano.

Nesse contexto, o desenvolvimento de sistemas estatísticos torna-se um poderoso aliado na manipulação e visualização de dados de forma ágil e precisa.

O presente trabalho tem como objetivo desenvolver um programa em Python, denominado "SI-CHRONOS" (Sistema Integrado de Análises Cronológicas), que visa facilitar as análises de dados de áreas diversas, como uma solução flexível e acessível para pesquisadores, cientistas de dados e profissionais que necessitam realizar análises estatísticas, visualizações gráficas e interpretações de dados de maneira eficaz.

A finalidade do *software* é proporcionar uma plataforma amigável e de fácil utilização, que permita aos usuários importar seus conjuntos de dados, explorar suas características e obter diferentes e importantes visões para tomada de decisões embasadas em evidências. O programa permite que o usuário selecione o arquivo de texto contendo os dados, especifique os nomes das colunas, seus valores mínimos e máximos, além de definir o intervalo de tempo entre a coleta das amostras.

O sistema integrado de análises cronológicas abrange uma série de funcionalidades, tais como: análise estatística descritiva, geração de gráficos de distribuição de normalidade para cada variável, exibição de matriz de correlação entre as variáveis por meio de mapa de calor, e criação de gráficos de dispersão para análise de tendências e relações entre os dados.

1.1 – OBJETIVO GERAL

O objetivo geral deste trabalho se trata do desenvolvimento de um programa em linguagem Python, intitulado "SI-CHRONOS", que permita que os usuários sejam capazes de realizar análises informativas e rápidas de conjuntos de dados contidos em arquivos de texto (.txt), organizados em colunas. O programa visa otimizar o tempo de análise e minimizar os erros de natureza humana, além de oferecer uma plataforma amigável e de fácil utilização para a interpretação e visualização de dados, possibilitando a execução de novas abordagens para tomada de decisões embasadas em evidências, sem que o usuário tenha o trabalho de recorrer a múltiplos *softwares* em busca de transformar seus dados em visibilidades.

1.2 – OBJETIVO ESPECÍFICO

Esse trabalho visa garantir que o sistema seja uma ferramenta completa, e de alto valor agregado para a análise de dados, permitindo que usuários de diversas

áreas consigam realizar análises detalhadas de forma simples e rápida. Para alcançar esse objetivo, serão necessárias as realizações das seguintes etapas:

- I.** Implementar um algoritmo eficiente para importação e manipulação de dados provenientes de arquivos de texto com estrutura de colunas;
- II.** Desenvolver funcionalidades para a realização de análises estatísticas descritivas, incluindo média, moda, mediana e variância, para cada coluna do conjunto de dados;
- III.** Criar gráficos de distribuição de frequência e curvas de normalidade para auxiliar na avaliação da aderência dos dados a uma distribuição Gaussiana em cada variável analisada;
- IV.** Gerar uma matriz de correlação entre as variáveis, permitindo identificar possíveis relações e padrões existentes nos dados;
- V.** Construir gráficos de dispersão para análise de tendências e comportamentos das variáveis ao longo do tempo ou em relação a outras variáveis, possibilitando uma visualização mais abrangente dos dados;
- VI.** Incluir recursos para que o usuário possa definir os valores mínimos e máximos de cada coluna, a fim de personalizar as análises de acordo com as necessidades específicas de cada conjunto de dados;
- VII.** Implementar um mecanismo de geração de curvas de tendência para auxiliar na identificação de padrões e comportamentos das variáveis ao longo do tempo;
- VIII.** Disponibilizar uma interface interativa para que o usuário possa interagir facilmente com o programa, selecionar os arquivos de texto, ajustar parâmetros e visualizar os resultados das análises;
- IX.** Realizar testes comparativos com ferramentas já validadas para garantir a precisão e eficácia das análises realizadas pelo programa;
- X.** Validar o programa por meio da aplicação em diferentes conjuntos de dados, demonstrando sua eficiência e utilidade na interpretação e visualização de informações.

1.3. Estrutura do Trabalho

Além da Seção 1, o trabalho segue a seguinte estrutura: na Seção 2, encontra-se uma revisão da literatura, sendo essa fundamental para o entendimento do trabalho. Na Seção 3, serão detalhadas as etapas de desenvolvimento do programa, os requisitos funcionais e não funcionais, e a modelagem do sistema utilizando diagramas UML. A Seção 4 se concentra na apresentação dos resultados e sua análise. Por fim, na Seção 5, serão apresentadas as considerações finais, e, também, uma discussão das possibilidades de projetos futuros.

2 - REVISÃO DE LITERATURA

Nesta seção, será apresentada uma revisão pontual da literatura que fundamenta os conceitos e métodos utilizados no desenvolvimento do sistema. A revisão abordará os tópicos relacionados à estatística descritiva, variáveis aleatórias e correlação de dados, proporcionando uma base sólida para a compreensão das análises realizadas pelo programa.

2.1 – ESTATÍSTICA DESCRITIVA

A estatística descritiva se trata de uma área fundamental da estatística, sendo ela responsável por resumir, organizar e apresentar os dados de forma compreensível, descrevendo características dos dados e assim permitindo uma compreensão mais clara das informações presentes em um conjunto de dados.

Dentre os principais objetivos da estatística descritiva estão a identificação de padrões, tendências e variações nos dados, além da fundamental importância para o fornecimento de novas abordagens para a tomada de decisões informadas. Essa disciplina é amplamente utilizada em diversas áreas, como ciências sociais, negócios, ciência de dados, pesquisas científicas e muito mais. [8]

A estatística descritiva abrange várias medidas e técnicas, dentre elas podemos citar brevemente:

Medidas de Tendência Central:

- Média: Calculada como a soma de todos os valores dividida pelo número de observações. Ela fornece uma ideia geral sobre o valor típico do conjunto de dados.
- Mediana: Valor que separa o conjunto de dados ao meio, com metade dos valores acima dela e metade abaixo. Ela é menos sensível a valores extremos do que a média.
- Moda: Representa o valor mais frequente no conjunto de dados. O conjunto pode apresentar somente uma moda (unimodal) ou várias modas (multimodal).

Medidas de Dispersão:

- Variância: Mensura a dispersão dos valores em relação à média.
- Desvio Padrão: Trata-se da raiz quadrada da variância e fornece uma medida de dispersão mais facilmente interpretável, na mesma unidade dos dados originais.
- Amplitude: É a diferença entre o maior e o menor valor do conjunto de dados.

Gráficos e Visualizações:

- Histogramas: Gráficos que demonstram a frequência de ocorrência de valores em intervalos específicos.
- Gráficos de Barras: Usados para representar a distribuição de valores em categorias distintas.
- Gráficos de Dispersão: Mostram a relação entre duas variáveis, permitindo identificar padrões e tendências.

A estatística descritiva possui um papel de extrema importância na interpretação inicial de dados, ajudando a detectar padrões, anomalias e informações relevantes. No contexto do programa desenvolvido, essas técnicas de estatística descritiva serão aplicadas automaticamente para fornecer novas ideias aos usuários sobre seus conjuntos de dados, de forma a minimizar a incidência de erros humanos, permitindo uma análise mais informada e eficaz.

Dentre as medidas citadas, iremos nos aprofundar em quatro delas em específico: média, moda, mediana e variância.

2.1.1 – Média

A média se trata de uma medida estatística de tendência central, responsável por representar o valor médio de um conjunto de dados. Esta medida é calculada por meio da soma de todos os valores do conjunto e divisão pelo número total de observações [8].

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (1)$$

onde:

- x_i : observação número i da variável x ;
- $\bar{x}$: média da variável x ;
- n : número de observações.

Esta medida possui a vantagem de ser igualmente sensível a todas as observações do conjunto de dados. É amplamente utilizada com o intuito de encontrar o valor típico ou representativo do conjunto.

Contudo, devido a essa sensibilidade, a média é facilmente alterada em casos de presença de valores extremos, sendo esses também chamados de “*outliers*”. Tornando-se distorcida e não representativa do conjunto como um todo, nesses casos sendo necessário optar por outras formas de análise, como a mediana, que é menos afetada por valores extremos.

2.1.2 – Mediana

A mediana se trata de uma medida que representa o valor central de um conjunto de dados ordenados. Ao contrário da média, que é a soma de todos os valores dividida pelo número de observações, a mediana é o valor que divide o conjunto de dados ao meio, com metade dos valores acima dela e metade abaixo. Sendo amplamente utilizada em conjunto de dados com a presença de muitos *outliers*, visto que a mediana é definida pela ordem em que os dados estão dispostos e não pela magnitude dos valores. [8]

Para se determinar a mediana, deve-se seguir os seguintes passos:

- i. Ordene os dados em ordem crescente ou decrescente;
- ii. Se o número de observações for ímpar, a mediana será o valor do meio da sequência ordenada;
- iii. Se o número de observações for par, a mediana será a média dos dois valores do meio.

Para fins de cálculo temos:

$$md(x) = \begin{cases} x_{(\frac{n+1}{2})}, & \text{se } n \text{ ímpar;} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, & \text{se } n \text{ par.} \end{cases} \quad (2)$$

2.1.3 – Moda

A moda representa o valor mais frequente de um conjunto de dados. Enquanto a média se baseia na soma de todos os valores dividida pelo número total de observações, e a mediana se concentra no valor central quando os dados estão ordenados, a moda identifica o valor que ocorre com maior repetição. Sendo especialmente útil quando se deseja determinar quais os agrupamentos nas distribuições.

Essa medida é muito utilizada em variáveis categóricas (qualitativas) ou discretas (quantitativas), como categorias de produtos, notas de uma avaliação, idades em grupos etários, gênero de uma população, entre outros. Sendo responsável por fornecer informações relevantes sobre as características dominantes de um conjunto de dados, principalmente quando a ocorrência é um fator a ser considerado.

Os conjuntos de dados podem ser classificados em três tipos em relação a moda, são eles:

1. Unimodal: Possui uma única moda.
2. Bimodal: Possui duas modas distintas. Podendo indicar a presença de subgrupos diferentes ou diferentes picos na distribuição.
3. Multimodal: Possui mais de duas modas. Indicando uma variedade de agrupamentos.

A análise da moda é de fundamental importância tanto em conjuntos de dados quantitativos quanto qualitativos. Para dados quantitativos, ela determina os valores predominantes de uma amostra, permitindo uma compreensão mais clara da distribuição dos dados. Em relação aos dados qualitativos, a moda auxilia na tomada de decisões ao identificar a preferência ou escolha mais comum em um conjunto de opções, o que é amplamente aplicável a negócios, além de permitir a identificação de características comuns em um grupo ou população.

2.1.4 – Variância e Desvio Padrão

A variância é uma medida estatística que quantifica a dispersão ou a variabilidade dos dados em relação à média. Enquanto a média, a mediana e a moda estão focadas na tendência central dos dados, a variância avalia o quão distantes os valores individuais estão da média geral.

Essa medida é particularmente útil para entender o grau de dispersão dos dados e como eles se espalham em torno da média [5]. Quanto maior a variância, maior é a dispersão dos dados, indicando que eles estão mais distantes da média. Por outro lado, uma variância menor sugere que os dados estão mais próximos da média.

A variância é calculada seguindo os seguintes passos:

- I. Calcule a média dos dados;
- II. Subtraia cada valor da média e eleve o resultado ao quadrado;
- III. Some os resultados dos quadrados obtidos;
- IV. Divida essa soma pelo número total de observações.

Resumida pela seguinte fórmula:

$$\sigma^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n} \quad (3)$$

onde:

- x_i : observação número i da variável x ;
- $\bar{x}$: média da variável x ;
- n : número de observações;
- σ^2 : variância.

A análise da variância é fundamental para avaliar a dispersão, pois ajuda a entender o quão distantes os valores individuais estão da média, fornecendo uma ideia da dispersão geral dos dados. Além disso, permite comparar a variabilidade entre diferentes conjuntos de dados, identificando qual conjunto tem maior dispersão, e detectar *outliers*, ou valores extremamente distantes da média, que podem impactar as análises.

A variância se trata de uma medida crucial em estatística e análise de dados, proporcionando informações sobre a variabilidade e a distribuição dos valores, e é amplamente utilizada para entender a amplitude dos dados e a consistência das observações.

Em paralelo a medida, temos o desvio padrão, calculado através da raiz quadrada da variância:

$$\sigma = \sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{n}} \quad (4)$$

O desvio padrão é um grande facilitador na análise de dispersão dos dados, visto que se difere da variância no quesito de unidades. Enquanto a variância é apresentada em unidades ao quadrado dos dados originais, o desvio padrão é expresso nas mesmas unidades dos dados originais. O fator “quadrado” também eleva a sensibilidade da variância a *outliers*, o que já não ocorre com o desvio padrão [8].

Na tabela a seguir, podemos notar as principais diferenças entre esses dois tipos de análise de dispersão:

Tabela 1 – Variância x Desvio Padrão

Característica	Variância	Desvio Padrão
Definição	Média dos quadrados das diferenças entre cada valor e a média.	Raiz Quadrada da Variância.
Unidade de Medida	Expressa em unidades ao quadrado dos dados originais.	Expressa na mesma unidade dos dados originais.
Sensibilidade a Outliers	Sensível a valores extremos devido à elevação ao quadrado.	Menos sensível a valores extremos.
Interpretação	Mais difícil de interpretar devido às unidades ao quadrado.	Mais fácil de interpretar e comparar.

Fonte: Elaborado pelo próprio autor.

2.2 – VARIÁVEIS ALEATÓRIAS

No âmbito das análises estatísticas e da modelagem de fenômenos complexos, o conceito de variáveis aleatórias desempenha um papel fundamental. A Seção 2.2 deste trabalho se dedica a uma exploração aprofundada das variáveis aleatórias, abordando sua classificação, distribuição e frequência, fatores de extrema importância no fornecimento de informações relevantes para a compreensão e interpretação de eventos sujeitos a incertezas.

Pode-se resumir o termo “variáveis aleatórias” como qualquer regra que associe um valor a cada resultado possível de um espaço amostral de um experimento. Afirmação que, traduzida a um viés matemático, implica que uma variável aleatória é uma função definida no espaço amostral com contradomínio composto por um conjunto de número reais [8].

2.2.1 - Tipos de Variáveis

A compreensão dos diferentes tipos de variáveis é de suma importância para a seleção adequada de técnicas em uma análise de conjunto de dados. Ao classificar as variáveis em categorias qualitativas ou quantitativas e em seus subtipos, estamos equipados para tomar decisões informadas sobre como realizar análises estatísticas, interpretar resultados e tirar conclusões significativas dos dados em questão. Isso permite que sejam escolhidas abordagens estatísticas que se alinhem precisamente à natureza das variáveis presentes, aumentando assim a precisão e relevância das conclusões obtidas a partir da análise.

As variáveis podem ser classificadas com base na natureza dos dados que representam e nas operações estatísticas que podem ser aplicadas a elas. Existem dois tipos principais de classificação: variáveis qualitativas (categóricas) e variáveis quantitativas (numéricas).

As variáveis qualitativas, também conhecidas como categóricas, representam características que podem ser categorizadas em grupos ou até mesmo distintas. Normalmente são utilizadas em situações em que se deseja determinar a frequência de determinada variável, como por exemplo gênero, cor dos olhos, estado civil, tipos de produtos, entre outros.

Esse agrupamento pode ser dividido em mais dois subgrupos:

- Nominais: As categorias não se relacionam em qualquer tipo de ordem. Por exemplo: cor dos olhos, gênero, língua falada.
- Ordinais: As categorias possuem uma relação de ordem, não quantificável, entre uma e outra. Por exemplo: níveis de satisfação do cliente (baixo, médio, alto).

Por outro lado, tem-se as variáveis quantitativas, também conhecidas como numéricas, essas que são capazes de representar quantidades numéricas, logo, mensuráveis. Elas podem ser separadas em mais dois subgrupos:

- Discretas: Representam valores que são contáveis e inteiros, constituindo um conjunto finito e não permitindo intermediários. Exemplos: Número de pessoas em uma sala, número de vezes que um dado cai em um determinado número, entre outros.
- Contínuas: Representam quaisquer valores presentes em um intervalo contínuo, incluindo fracionários e decimais (conjunto dos reais), assim, abrangendo o determinado intervalo de forma completa e infinita. Exemplos: Temperatura de um fluido, a qual pode ser tanto $1,03^{\circ}\text{C}$ quanto $1,03045^{\circ}\text{C}$, altura, comprimento, tempo, peso.

2.2.2 - Distribuição de Variáveis Aleatórias

Como retratado na seção 2.2.1, as variáveis podem ser tanto discretas, assumindo um conjunto finito de possíveis valores, quanto contínuas, abrangendo um intervalo de valores infinitos. A distribuição de variáveis aleatórias descreve como os possíveis valores de uma variável aleatória estão distribuídos, fornecendo uma estrutura para entender a probabilidade de diferentes resultados e padrões em experimentos ou situações com algum grau de incerteza.

Existem várias distribuições de variáveis aleatórias amplamente reconhecidas, cada uma com suas características e aplicações específicas.

2.2.2.1 – Distribuição Binomial

A Distribuição Binomial se trata de uma das distribuições de probabilidade mais utilizadas na estatística, sendo aplicada em situações em que um processo possui apenas dois resultados possíveis, geralmente denominados "sucesso" e "fracasso", com o intuito de modelar eventos binários ou dicotômicos, onde estamos interessados em quantificar a probabilidade de sucesso em um certo número de tentativas independentes. Como por exemplo, quantas vezes um dado irá cair com a face que contenha um número par para cima.

Esse tipo de distribuição é caracterizado por dois principais parâmetros:

- Número de tentativas (n): O número total de experimentos independentes ou tentativas.
- Probabilidade de sucesso (p): A probabilidade de ocorrer o resultado desejado em uma única tentativa.

E sua fórmula é dada por:

$$P(x = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad (5)$$

onde:

- $P(x = k)$: probabilidade de obter k sucessos em n tentativas;
- $\binom{n}{k}$: coeficiente binominal, o qual representa o número de combinações de n elementos tomados a cada sucesso;
- p^k : probabilidade de k sucessos;
- $(1 - p)^{n-k}$: probabilidade de $n - k$ fracassos.

Sua média e desvio padrão são calculados das seguintes formas, respectivamente:

$$\mu = np \quad (6)$$

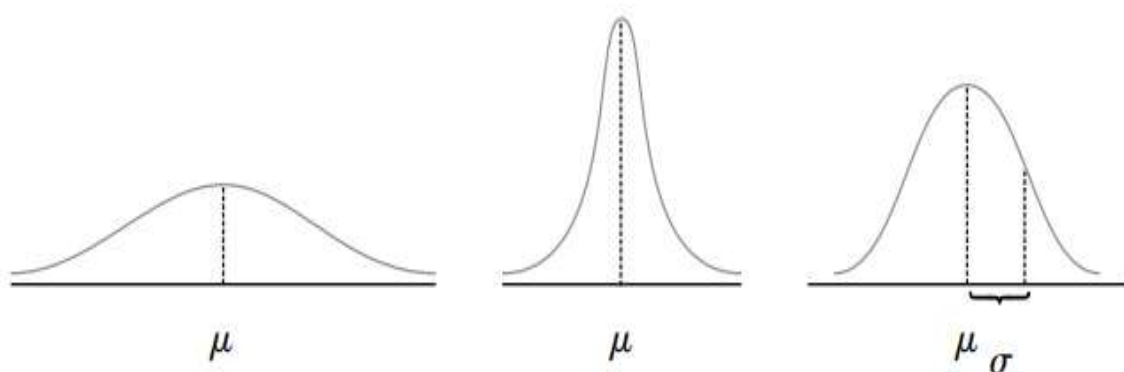
$$\sigma = \sqrt{np(1 - p)} \quad (7)$$

A distribuição binomial, uma distribuição discreta, trabalha com um número finito de possibilidades, e uma de suas propriedades importantes é o fato de que, quando o número de tentativas é muito superior à probabilidade de sucesso, ela pode ser aproximada pela distribuição de Poisson.

2.2.2.2 – Distribuição Normal (Gaussiana):

A Distribuição Normal, também conhecida como Distribuição Gaussiana, desempenha um papel essencial na estatística, sendo empregada em diversas análises. Ela é relevante para modelar diversos conjuntos de dados do mundo real, caracterizando-se por uma curva de formato de sino e simetria em torno da média, como pode ser visualizado na Figura 1.

Figura 1 – Curvas de Distribuição Normal



Fonte: [8].

Características principais da Distribuição Normal:

- Possível de ser utilizada tanto em casos de variáveis discretas quanto contínuas;
- Na distribuição normal, a média, mediana e moda são todas iguais e coincidem no centro da distribuição;
- Simetria: A Distribuição Normal exibe simetria perante a média, indicando que os valores à esquerda e à direita da média são praticamente idênticos;

- Regra Empírica (também chamada de regra “68-95-99.7”): Aproximadamente 68% dos dados estão dentro de um desvio padrão da média, 95% estão dentro de dois desvios padrões e 99.7% estão dentro de três desvios padrões;
- Parâmetros Fundamentais: A Distribuição Normal é definida pelos parâmetros média (μ) e desvio padrão (σ).

Para que possamos determinar a forma da curva de distribuição da probabilidade de ocorrência de cada valor da variável aleatória, faz-se necessário o cálculo da função densidade de probabilidade:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (8)$$

onde, $f(x)$ é a densidade de probabilidade no ponto, μ denota a média e σ representa o desvio padrão.

A Distribuição Normal possui grande aplicabilidade em diversos campos, como a estatística e ciência, por conta de sua presença em inúmeros fenômenos naturais e processos aleatórios que englobam uma infinidade de variáveis aleatórias potenciais. Ela se posiciona como uma ferramenta central na modelagem de variáveis contínuas, como por exemplo, peso, altura e temperatura. Além disso, desempenha um papel de destaque em testes estatísticos, análises de regressão e uma série de outros métodos estatísticos de natureza avançada [8].

2.2.2.3 – Distribuição de Poisson

A Distribuição de Poisson é uma distribuição de probabilidade discreta, visto que foca em metrificar a quantidade de eventos inteiros, de acordo com Bussab e Morettin (2017), a distribuição é frequentemente utilizada para modelar eventos raros e discretos, que ocorrem em um determinado intervalo de tempo. Esse tipo de distribuição é amplamente utilizada em diversas áreas da estatística e ciência, como em áreas da biologia, engenharia, finanças e epidemiologia. Ela é particularmente aplicável a situações em que os eventos são independentes uns dos outros e ocorrem a uma taxa constante.

Características principais da Distribuição de Poisson:

- Eventos Raros: Essa distribuição é adequada para descrever eventos que são considerados raros e independentes. Exemplos incluem: acidentes (de quaisquer gêneros) ocorridos em um determinado intervalo de tempo, defeitos em uma linha de produção, chegada de clientes em um estabelecimento, chamadas telefônicas recebidas por uma central em um determinado intervalo de tempo, entre outros.
- Parâmetro Lambda (λ): representa a taxa média de ocorrência dos eventos num intervalo de tempo específico.
- Variabilidade: A variância da distribuição de Poisson é igual ao seu parâmetro λ . Ou seja, quanto maior o valor de λ , maior será a dispersão dos eventos ao redor da média.

O cálculo da probabilidade de ocorrência de determinado evento em um intervalo especificado é dado pela fórmula de distribuição de Poisson:

$$P(X = k) = \frac{(e^{-\lambda} \lambda^k)}{k!} \quad (9)$$

onde, k é o número de eventos desejados e λ a taxa média de ocorrência do evento.

2.2.3 - Frequência de Variáveis Aleatórias

A frequência das variáveis aleatórias pode ser muito bem analisada por meio do uso de gráficos de dispersão, estes que consistem em uma representação bidimensional na qual cada ponto no gráfico representa uma observação individual que possui valores para duas variáveis diferentes [8]. O eixo horizontal (eixo x) representa uma variável e o eixo vertical (eixo y) representa outra. Ao associar pares de valores das duas variáveis, o gráfico de dispersão revela a relação entre elas, permitindo uma análise visual direta das interações.

No caso do atual projeto, os gráficos de dispersão serão utilizados com o intuito de monitorar o comportamento das variáveis em relação ao tempo decorrido de experimento, dessa forma, a respectiva variável aleatória sendo disposta no eixo y,

enquanto a frequência de observações, dada em unidades de tempo, sendo disposta no eixo x. Assim, os gráficos de dispersão que relacionam variável-tempo serão capazes de apontar o comportamento e a tendência daquela variável em específico com o decorrer do experimento.

2.3 – CORRELAÇÃO DE DADOS

Um dos aspectos-chave da análise estatística é denominado correlação de dados, este que é responsável por quantificar o grau de associação entre variáveis, assim, permitindo compreender como as mudanças em uma variável podem afetar as demais, fornecendo melhores abordagens sobre a direção e força dessa relação.

Neste trabalho, teremos enfoque no estudo das variáveis quantitativas, visto que o programa SI-CHRONOS, na sua atual versão, possui a capacidade de avaliar somente variáveis numéricas.

A medida de correlação mais comumente utilizada para variáveis quantitativas é o denominado coeficiente de correlação de Pearson, numérico, variante entre -1 e +1, no qual, quanto mais próximo a zero, mais indicará uma falta de associação (ou dependência) entre as variáveis. Um valor positivo próximo a 1 indica uma correlação positiva forte, ou seja, à medida que uma variável aumenta, a outra também tende a aumentar. Por outro lado, um valor negativo próximo a -1 indica uma correlação negativa forte, onde o aumento de uma variável está associado à redução da outra [8].

O coeficiente de Pearson é calculado pela seguinte relação:

$$r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}} \quad (10)$$

Para uma melhor visibilidade da relação entre as variáveis, também são utilizados os gráficos de dispersão previamente mencionados na Seção 2.2.3.

2.4 – POKA YOKE

O conceito “*Poka Yoke*” se trata de uma metodologia fundamentada na prevenção de erros, desenvolvida por Shigeo Shingo no contexto do Sistema Toyota de Produção. O termo *Poka Yoke* significa “à prova de erros” em japonês e refere-se a qualquer mecanismo ou dispositivo que ajude a evitar erros humanos durante o processo de produção [15].

A filosofia por trás do *Poka Yoke* é fundamentada em que, embora os erros humanos sejam inevitáveis, suas consequências podem ser minimizadas ou eliminadas por meio de projetos inteligentes e controles automáticos. Isso é alcançado através da implementação de dispositivos físicos ou sistemas de controle que detectam e previnem erros antes que eles possam causar danos ou defeitos. Exemplos de *Poka Yoke* incluem guias visuais, sensores, alarmes e outras formas de automação que sinalizam ou impedem ações incorretas. Ao incorporar o *Poka Yoke* no desenvolvimento do SI-CHRONOS, buscou-se criar um sistema que reduza a probabilidade de erros na análise de dados temporais, garantindo a precisão e a confiabilidade dos resultados. Esta abordagem não só melhora a eficiência do processo de análise, mas também contribui para a minimização de custos e a melhoria contínua da qualidade.

2.5 – APLICAÇÕES DA ESTATÍSTICA NA CIÊNCIA

A estatística é uma ferramenta poderosa que desempenha um papel fundamental em uma ampla variedade de campos científicos. Nesta seção, exploraremos três estudos científicos recentes que demonstram como a análise estatística é aplicada para abordar questões complexas e relevantes em diferentes domínios.

2.5.1 - Uma Pesquisa sobre Informática de Biodiversidade: Conceitos, Práticas e Desafios

Ao abordar a área de informática em biodiversidade, é relevante citar o estudo intitulado “*A Survey of Biodiversity Informatics: Concepts, Practices, and*

Challenges", conduzido por Luiz M. R. Gadelha Jr. e colaboradores [11]. Este estudo se debruça sobre um tema de importância crucial no cenário global atual, relacionado à biodiversidade e seu impacto diante do crescimento populacional e atividades econômicas.

No resumo apresentado, os autores destacam a magnitude sem precedentes da população humana e suas atividades econômicas, que estão exercendo um impacto cada vez maior nos ambientes globais. A crescente preocupação dos países em relação ao consumo de recursos naturais e à capacidade dos ecossistemas de supri-los é evidenciada. Para a conservação eficaz da biodiversidade, torna-se essencial disponibilizar indicadores e conhecimento de maneira aberta para tomadores de decisão, de forma que possam utilizá-los de maneira eficiente.

O estudo enfoca a transformação de dados brutos em informações sintetizadas, prontas para uso, o que requer múltiplas etapas de refinamento. As metodologias e técnicas empregadas para gerenciar e analisar esses dados constituem a área conhecida como informática em biodiversidade (ou e-Biodiversidade). O ciclo de vida dos dados de biodiversidade envolve planejamento, coleta, certificação, descrição, preservação, descoberta, integração e análise. Pesquisadores, sejam eles produtores ou consumidores de dados de biodiversidade, provavelmente se envolverão em atividades relacionadas a pelo menos uma dessas etapas.

O trabalho utiliza uma variedade de fontes de dados relacionadas à biodiversidade, incluindo dados de ocorrência de espécies, listas de espécies, dados baseados em amostras e observações, dados moleculares, literatura acadêmica, imagens, vídeos, e sensoriamento remoto. Esses dados são coletados de diversas maneiras, como expedições de campo, observações de cidadãos cientistas (entusiastas e voluntários que participam ativamente na coleta de dados científicos), redes de biossensores e muito mais.

Para garantir a qualidade e adequação desses dados, são adotadas várias medidas, incluindo a padronização dos termos e o uso de ontologias para garantir a consistência, a verificação da precisão taxonômica, a redução de erros de georreferenciamento e o uso de ferramentas como o BioGeomancer e o Geolocate para validar as coordenadas geográficas. Além disso, são propostos *frameworks* para avaliação e gerenciamento da qualidade dos dados, como o desenvolvido por Veiga

et al. (2017), que permite aos usuários definir os requisitos de qualidade dos dados e implementar soluções para melhorar a adequação dos mesmos para usos específicos.

Embora o estudo não se concentre em uma única ferramenta estatística, ele adota uma abordagem abrangente para gerenciar e garantir a qualidade dos dados de biodiversidade, envolvendo várias etapas e técnicas de coleta, padronização e avaliação de dados. Isso inclui análises estatísticas que utilizam técnicas de ciência de redes, genômica e gestão de fluxos de trabalho científicos para entender a estrutura e dinâmica das interações ecológicas.

2.5.2 - Abordagens de Aprendizado de Máquina para Previsão de Eventos Climáticos Extremos em Áreas Urbanas: Desafios e Resultados Iniciais

Ao explorar as abordagens para previsão de eventos climáticos extremos em áreas urbanas, podemos citar o estudo conduzido por Fabio Porto e colaboradores, intitulado "*Machine Learning Approaches to Extreme Weather Events Forecast in Urban Areas: Challenges and Initial Results*" [12]. Este trabalho assume uma relevância crítica, considerando a crescente importância da previsão de chuvas intensas e seus impactos nas áreas urbanas, como evidenciado no contexto da cidade do Rio de Janeiro.

Nesse estudo, é examinada a necessidade de previsões precisas para eventos climáticos extremos, ressaltando a influência das mudanças climáticas e seus efeitos em eventos como chuvas intensas. Os autores investigam as limitações dos métodos de previsão atuais, em especial os resultados do Centro de Operações do Rio (COR), que emitiu alertas para eventos climáticos extremos, porém com desafios em termos de acurácia.

Uma parte crucial do estudo está na análise da abordagem atual adotada pelo COR, que envolve meteorologistas interpretando as previsões numéricas do tempo e monitorando uma variedade de fontes de dados, incluindo imagens de radar e sensores meteorológicos. No entanto, o estudo reconhece a lacuna significativa em relação à disponibilidade de dados relacionados a eventos climáticos extremos.

Para lidar com essa limitação, o estudo propõe a exploração de modelos de Aprendizado de Máquina (AM) como alternativa para melhorar as previsões de eventos climáticos extremos. Três abordagens distintas são apresentadas, todas

envolvendo o uso de modelos de aprendizado profundo, capazes de aprender padrões espaço-temporais a partir de observações de precipitação de estações meteorológicas e dados de previsão do tempo.

O estudo ilustra a integração entre aprendizado de máquina e análise estatística em várias etapas do processo de previsão de eventos climáticos extremos. No processo de pré-processamento dos dados, técnicas estatísticas como o algoritmo *k-means clustering* e o cálculo de distâncias utilizando o algoritmo *Dynamic Time Warping* são empregadas para identificar padrões e agrupar regiões do espaço de dados com comportamentos semelhantes. Além disso, na construção do conjunto de dados rotulado, o modelo de redes neurais auto-organizáveis (SOM) é utilizado como uma ferramenta semi-automática de rotulagem, enquanto a avaliação da qualidade das anotações é realizada por meio de um modelo de árvore de decisão (DT). Essa abordagem demonstra como o aprendizado de máquina se beneficia da análise estatística para processar e interpretar os dados, fornecendo *insights* valiosos para a modelagem e previsão de eventos climáticos extremos.

Em suma, o estudo se trata de uma contribuição significativa para a compreensão das abordagens de previsão de eventos climáticos extremos em áreas urbanas, focando especialmente nos desafios enfrentados na cidade do Rio de Janeiro. A exploração de técnicas de aprendizado de máquina para previsão, juntamente com a análise crítica das limitações atuais, faz deste trabalho uma referência valiosa para aqueles interessados na interseção entre ciência climática, tecnologias de previsão e tomada de decisões urbanas diante de eventos climáticos extremos.

2.5.3 - Compreensão da Mobilidade Humana e Dinâmica de Carga de Trabalho devido a Diferentes Eventos de Grande Escala Utilizando Dados de Telefones Móveis

Nesta pesquisa, destaco o estudo intitulado "*Understanding Human Mobility and Workload Dynamics Due to Different Large-Scale Events Using Mobile Phone Data*", conduzido por Humberto T. Marques-Neto e colaboradores [13]. Este estudo oferece novas abordagens sobre como a análise de dados de telefones móveis pode

contribuir para aprimorar a maneira como as operadoras lidam com cargas de trabalho incomuns causadas por eventos de grande escala.

Os autores exploram a análise da mobilidade humana e a dinâmica resultante na carga de trabalho da rede devido a três tipos diferentes de eventos de grande escala: uma partida de futebol importante, um concerto de rock e uma celebração de ano novo, realizados em uma grande cidade brasileira. A análise é baseada na caracterização dos registros de chamadas de telefones móveis feitas em torno do tempo e local de cada evento. Ou seja, a mobilidade humana e a carga de trabalho da rede são analisadas em termos do número de chamadas de telefone móvel, seus tempos de chegada e partida, e suas durações. Para a análise dos padrões de movimento dos participantes dos eventos de grande escala, são utilizados mapas de calor.

A pesquisa empregou uma variedade de métodos estatísticos, incluindo análise de séries temporais, análise de dados geoespaciais, análise de regressão, análise de dados longitudinais e análise descritiva. Essas metodologias foram utilizadas para compreender os padrões de mobilidade humana e a dinâmica da carga de trabalho da rede durante os eventos de grande escala, fornecendo *insights* valiosos para o planejamento e gerenciamento da infraestrutura de telecomunicações.

Os resultados obtidos são valiosos para a compreensão da mobilidade humana induzida por eventos de grande escala. Além disso, esses resultados podem fornecer informações críticas para os gerentes de rede em relação a estratégias eficazes de gerenciamento de capacidade e planejamento. Em paralelo, os autores apresentam o aplicativo "PrediTraf", desenvolvido para auxiliar as operadoras de telefonia móvel no planejamento de infraestrutura em eventos de grande escala.

3 - MATERIAL E MÉTODOS

Nesta seção, apresentaremos uma descrição detalhada do material e dos métodos utilizados no desenvolvimento do SI-CHRONOS, uma aplicação que possibilita a análise estatística de dados provenientes de arquivos de texto (.txt). O SI-CHRONOS foi construído em Python, fazendo uso da biblioteca "Tkinter" para criação uma interface gráfica intuitiva e acessível aos usuários.

3.1 - REQUISITOS

Antes de entrar em detalhes sobre os procedimentos utilizados no desenvolvimento, é crucial compreender os requisitos funcionais e não funcionais que direcionaram a criação do SI-CHRONOS.

3.1.1 - Requisitos Funcionais

Os requisitos funcionais representam as principais funcionalidades que a aplicação deve oferecer aos seus usuários. Para o SI-CHRONOS, os requisitos funcionais incluem:

- **Upload de Dados:** O SI-CHRONOS permite aos usuários carregar dados a partir de arquivos de texto (.txt) que contenham séries temporais. Isso facilita a entrada de dados de forma flexível e versátil.
- **Análise Estatística:** A aplicação é capaz de realizar análises estatísticas básicas, incluindo a criação de histogramas, gráficos de dispersão e cálculos de estatísticas descritivas. Essas análises ajudam os usuários a compreender a distribuição e o comportamento dos dados.
- **Visualização de Resultados:** O SI-CHRONOS fornece uma interface gráfica que permite aos usuários visualizar os resultados das análises estatísticas de forma clara e intuitiva.
- **Exportação de Gráficos:** Os usuários têm a capacidade de salvar os gráficos gerados em um arquivo PDF para referência futura. Funcionalidade essa que facilita o compartilhamento e a documentação dos resultados.

3.1.2 - Requisitos não funcionais

Os requisitos não funcionais estabelecem critérios de desempenho, usabilidade e confiabilidade que a aplicação deve atender. Para o SI-CHRONOS, os requisitos não funcionais incluem:

- **Usabilidade:** A interface foi projetada para ser intuitiva e amigável. Isso garante que usuários com diferentes níveis de experiência em análise estatística possam utilizar a aplicação de forma eficaz.
- **Desempenho:** O SI-CHRONOS é otimizado para processar grandes conjuntos de dados de forma eficiente. Isso garante que a análise seja rápida e responsiva, para uma quantidade considerável de volume de dados.
- **Confiabilidade de dados:** O *software* prioriza a confiabilidade dos dados dos usuários. As informações carregadas não são armazenadas em nenhum tipo de banco de dados, assim garantindo que não sejam acessíveis ou compartilhadas de forma não autorizada.
- **Portabilidade:** A aplicação é compatível com diferentes sistemas operacionais, garantindo que os usuários possam acessá-la em uma variedade de ambientes.

3.2 – AMBIENTE DE DESENVOLVIMENTO

Nesta seção, detalharemos o ambiente de desenvolvimento específico que foi utilizado na criação do SI-CHRONOS, incluindo a ferramenta principal e a estrutura do projeto.

3.2.1 - Ferramenta de Desenvolvimento

O ambiente de desenvolvimento para o projeto SI-CHRONOS baseou-se principalmente no Jupyter Notebook, que faz parte da distribuição Anaconda. As principais ferramentas utilizadas são:

- **Ambiente de Desenvolvimento - Jupyter Notebook (Anaconda):** O Jupyter Notebook é uma aplicação web de código aberto que permite criar e compartilhar documentos que contêm código interativo, equações, visualizações e texto narrativo. Ele é utilizado em ciência de dados e análise numérica devido à sua capacidade de fornecer um ambiente interativo para o desenvolvimento.
- **Linguagem de Programação - Python:** O projeto foi desenvolvido Python, por ser uma linguagem de programação adotada na análise de dados e em muitos

campos científicos e, também, de fácil codificação. Além disso, outro fator determinante para sua escolha está relacionado ao conjunto de bibliotecas disponíveis para Python, juntamente com sua sintaxe amigável.

3.2.2 - Criação e Estrutura do Projeto

A seguir, será detalhado como o projeto foi criado, destacando as estruturas e bibliotecas utilizadas em cada etapa:

1. **Criação da Interface Gráfica:** A interface gráfica foi criada utilizando a biblioteca “Tkinter” para Python. Esta biblioteca fornece as ferramentas necessárias para criar interfaces de usuário simples e interativas. Foi utilizada para elaborar a janela principal, rótulos, botões e campos de entrada.
2. **Manipulação de Dados:** A manipulação de dados, incluindo a leitura de arquivos de texto e a criação de estruturas de dados, foi realizada com a ajuda das bibliotecas “Pandas” e “NumPy”. A Pandas se trata de uma biblioteca de análise de dados que fornece estruturas de dados flexíveis e ferramentas de análise de dados. Ela foi utilizada para leitura e manipulação dos dados carregados pelo usuário. Enquanto a NumPy, que se trata de uma biblioteca fundamental para computação científica em Python, foi utilizada para as operações numéricas e cálculos estatísticos.
3. **Visualização de Dados:** Para criar gráficos e visualizações, foi empregada a biblioteca “Matplotlib”. Essa é uma biblioteca de visualização de dados que permite criar gráficos, histogramas e visualizações personalizadas, no nosso caso, sendo utilizada para criação de gráficos de dispersão, histogramas e matrizes de correlação.
4. **Análise Estatística:** A biblioteca “SciPy” foi utilizada para realizar testes estatísticos, como o teste de normalidade. A SciPy se trata de uma biblioteca baseada na NumPy, e fornece funções adicionais para otimização, estatísticas e processamento de sinais.
5. **Salvando Gráficos:** Para salvar e organizar os gráficos gerados em um arquivo PDF, foi empregada a biblioteca a biblioteca Matplotlib, mais especificamente sua função “Matplotlib.backends.backend_pdf”.

3.3 – MODELAGEM – DIAGRAMAS UML

A modelagem desempenha um papel crucial na compreensão e representação de sistemas complexos. Nesta seção, serão abordados os principais aspectos da modelagem utilizando a Linguagem de Modelagem Unificada (UML), uma ferramenta para visualização, especificação, construção e documentação de sistemas orientados a objetos. Os diagramas UML selecionados para análise são os de Casos de Uso e Diagrama de Sequência.

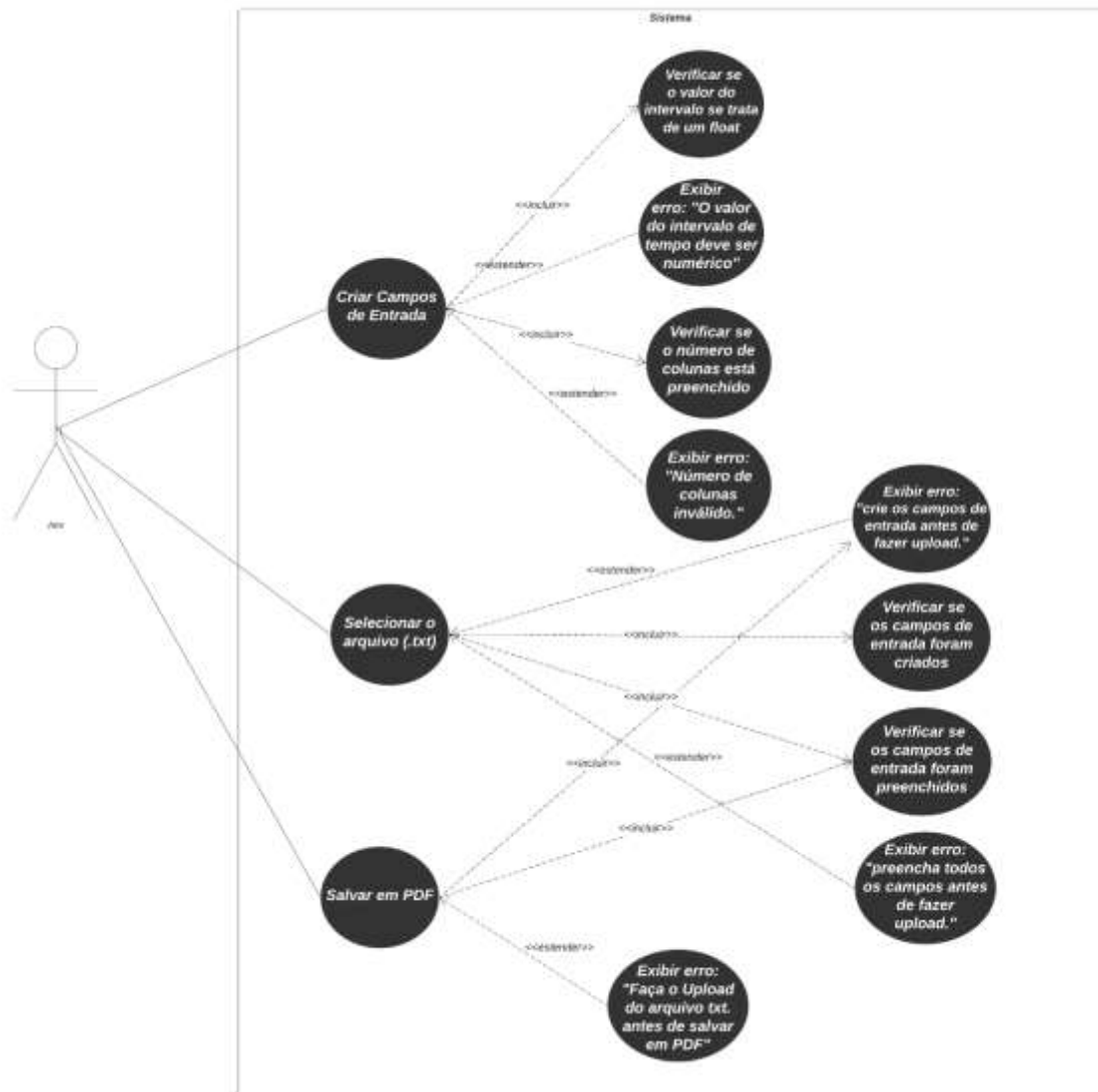
3.3.1 - Casos de Uso

Primeiramente temos o diagrama de casos de uso, esse que desempenha um papel central na identificação e especificação de requisitos funcionais do sistema. Esse tipo de diagrama fornece uma visão holística das interações entre atores e o sistema, representando o sistema como um conjunto coeso de interações e destacando as principais funcionalidades e comportamentos.

Esse tipo de diagrama não apenas identifica requisitos funcionais, mas também fornece uma linguagem visual intuitiva para comunicar esses requisitos. Na presente pesquisa, os Casos de Uso serão explorados como uma ferramenta fundamental para capturar, analisar e documentar os requisitos do sistema. Ao adotar uma abordagem holística, buscamos não apenas listar funcionalidades, mas entender como essas funcionalidades se encaixam no contexto mais amplo do sistema.

A seguir, apresentamos o Diagrama de Casos de Uso que resume visualmente a interação entre atores e o sistema, destacando não apenas as ações que os usuários podem tomar, mas também as respostas do sistema por meio das extensões e inclusões.

Figura 2 – Diagrama de Casos de Uso



Fonte: Elaborado pelo próprio autor.

No diagrama (Figura 2), as ações do usuário estão representadas de maneira clara, evidenciando, pelas setas, as possíveis interações com o sistema. Além disso, as respostas do sistema, indicadas pelas notações <<incluir>> e <<estender>>, fornecem uma compreensão mais aprofundada de como o sistema reage a diferentes cenários e requisitos específicos.

As notações “<<incluir>>” e “<<estender>>” são utilizadas para descrever relacionamentos entre casos de uso, indicando como diferentes funcionalidades interagem e se relacionam em um sistema.

O relacionamento “<<incluir>>” indica que um caso de uso (incluso) é uma parte necessária ou obrigatória de outro caso de uso (includente). Isso significa que a funcionalidade do caso de uso incluído é sempre executada como parte do caso de uso includente. Por exemplo, toda vez que o usuário clicar no botão “Seleciona o .txt”, o sistema irá verificar se os campos de entrada já foram previamente criados.

Enquanto o relacionamento “<<estender>>”, indica que um caso de uso (estende) pode ser expandido com funcionalidades adicionais em circunstâncias específicas, sem afetar o fluxo principal do caso de uso. Seguindo o mesmo exemplo, caso o usuário pressione o botão “Seleciona o .txt”, sem ter previamente criado os campos de entrada, o sistema irá retornar a mensagem “Crie os campos de entrada antes de fazer upload”, impedindo que o usuário continue o uso do programa sem seguir a sequência de uso correta.

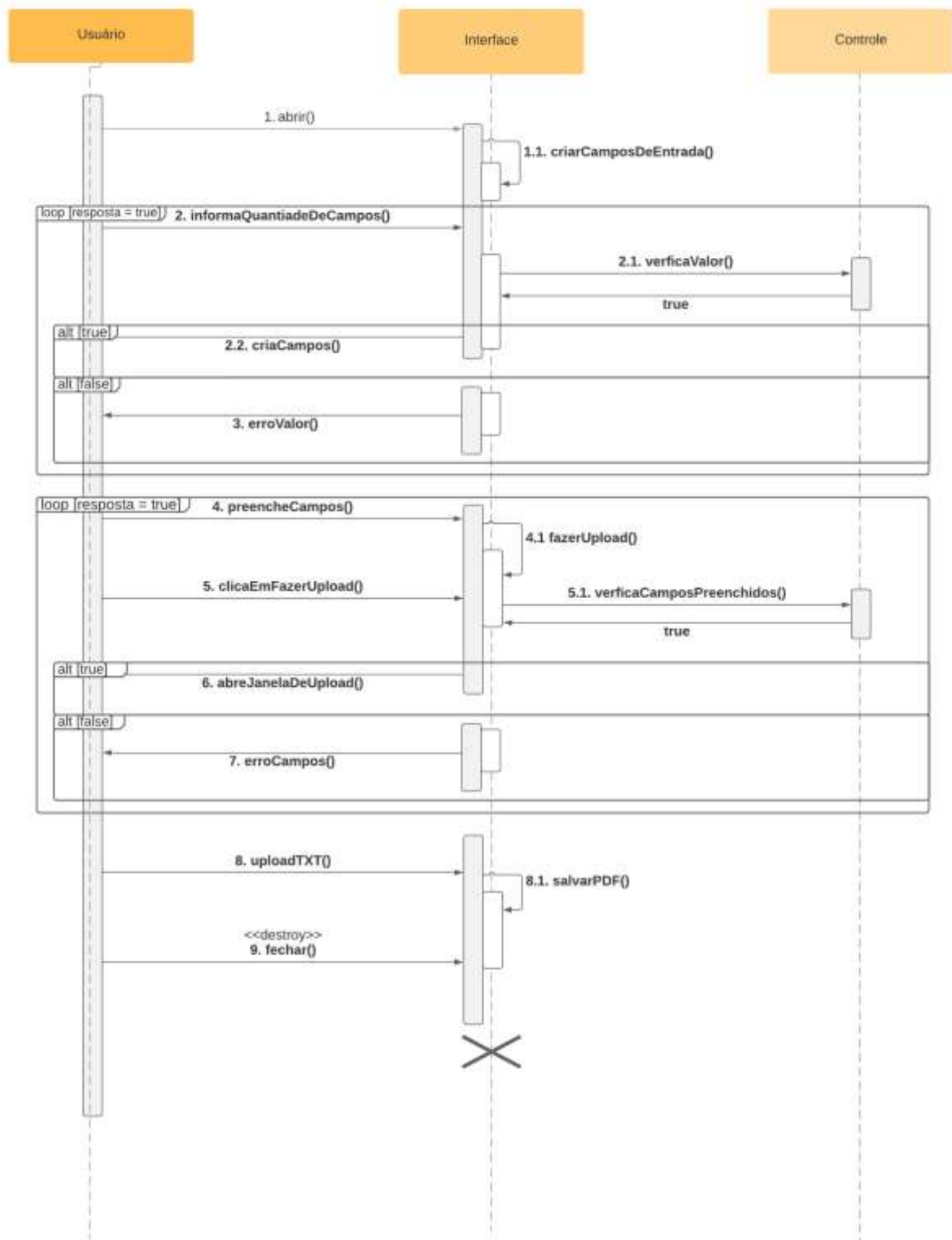
3.3.2 - Diagramas de Sequência

A abordagem para a modelagem de sistemas avança para além dos Casos de Uso, explorando agora os Diagramas de Sequência. Estes desempenham um papel crucial na visualização das interações dinâmicas entre diferentes partes do sistema, proporcionando uma compreensão clara do fluxo temporal das ações.

Estes diagramas oferecem uma representação gráfica das interações entre objetos, permitindo que se compreenda como as diferentes partes do sistema se comunicam e colaboram para alcançar um objetivo comum. Em contraste com os Casos de Uso, que fornecem uma visão mais estática, os Diagramas de Sequência nos conduzem por um caminho temporal, revelando a dinâmica subjacente ao funcionamento do sistema.

Na Figura 3, é apresentado o Diagrama de Sequência que destaca as interações entre diferentes objetos durante a utilização do sistema “SI-CHRONOS”.

Figura 3 – Diagrama de Sequência



Fonte: Elaborado pelo próprio autor.

No diagrama, as setas indicam a ordem temporal das mensagens trocadas entre os objetos (Usuário, Interface e Controle). As linhas de vida (pontilhadas) representam a existência dos objetos ao longo do tempo.

3.4 – FASE DE DESENVOLVIMENTO DO SOFTWARE

A etapa de desenvolvimento e implementação do sistema envolveu a transformação dos requisitos e especificações definidos nos ideais do projeto em código executável. A seguir, estão destacados os principais aspectos dessa implementação, os quais podem ser divididos em uma estrutura modular, abordando todas etapas principais, desde a escolha de linguagem, arquitetura, lógica de funcionamento, interface do usuário, confiabilidade e testes.

3.4.1 – Escolha da Linguagem de Programação

A escolha da linguagem de programação Python para o desenvolvimento do SI-CHRONOS foi fundamentada em diversas vantagens que a tornam uma opção sólida para projetos de ciência de dados e interface gráfica.

Python é conhecido por sua sintaxe simples e clara, tornando o código mais legível e fácil de entender. Isso acelera o desenvolvimento, facilita a manutenção do software e contribui para uma curva de aprendizado suave.

Além disso, Python possui uma comunidade ativa e uma vasta gama de bibliotecas de código aberto. Para o “SI-CHRONOS”, a utilização de bibliotecas como “tkinter”, “pandas”, e “matplotlib” permitiu uma implementação eficiente de funcionalidades complexas, sem a necessidade de desenvolvimento exaustivo. Como por exemplo a integração suave com a biblioteca “pandas”, a qual facilita a manipulação de conjuntos de dados, enquanto “matplotlib” oferece recursos poderosos para a criação de visualizações gráficas. [1]

Ao considerar esses fatores, a escolha de Python como a linguagem principal para o desenvolvimento do “SI-CHRONOS” não apenas proporcionou eficiência no desenvolvimento, mas também garantiu a acessibilidade, robustez e flexibilidade necessárias para atender aos requisitos do projeto, também sendo de mais fácil

compreensão para os demais estudantes que desejem continuar/modificar o projeto, visto que terá código aberto.

3.4.2 – Módulo de Interface Gráfica (UI) e Experiência do Usuário (UX)

O módulo de Interface Gráfica (UI) desempenha um papel crucial no “SI-CHRONOS”, sendo responsável por criar uma interface amigável que facilita a interação dos usuários com o software. Desenvolvido utilizando a biblioteca “tkinter” do Python, esse módulo oferece uma variedade de elementos de interface, proporcionando uma experiência do usuário intuitiva e eficiente.

Dentre os elementos de interface estão:

- **Rótulos (*Labels*):** Utilizados para exibir informações descritivas na interface do usuário. No “SI-CHRONOS”, rótulos são empregados para indicar instruções, mensagens informativas e identificação de elementos visuais.
- **Botões (*Buttons*):** Permitem aos usuários realizar ações específicas, como seleção de arquivos, execução de análises e salvamento de resultados. A disposição estratégica dos botões no layout visa otimizar a usabilidade.
- **Campos de Entrada (*Entry Fields*):** Facilitam a inserção de informações pelo usuário, como o número de colunas desejado ou intervalo de tempo para análise. Esses campos são projetados para serem claros e intuitivos.
- **Janelas de Upload (*File Upload Windows*):** Oferecem uma maneira eficiente de os usuários selecionarem arquivos necessários para a análise. A integração de janelas de upload simplifica o processo de entrada de dados.

Figura 4 – Interface SI-CHRONOS

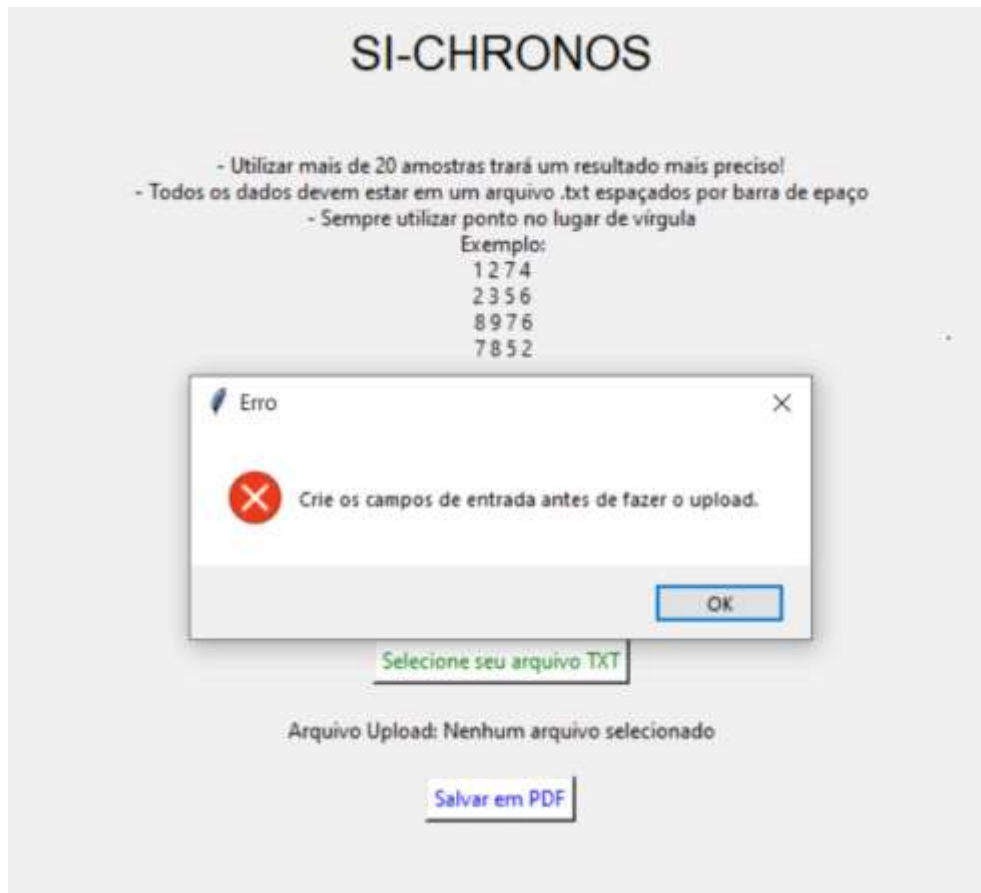
The screenshot displays the SI-CHRONOS interface. At the top, the title 'SI-CHRONOS' is centered. Below it, three instructions are listed: '- Utilizar mais de 20 amostras trará um resultado mais preciso!', '- Todos os dados devem estar em um arquivo .txt espaçados por barra de espaço', and '- Sempre utilizar ponto no lugar de vírgula'. An example of data input is shown: 'Exemplo: 1 2 7 4, 2 3 5 6, 8 9 7 6, 7 8 5 2'. Below the instructions, there are two input fields: 'Número de colunas:' and 'Intervalo de tempo (s):'. A button labeled 'Criar campos de entrada' is positioned below the second input field. A green button labeled 'Selecione seu arquivo TXT' is located below the 'Criar campos de entrada' button. Below this, the text 'Arquivo Upload: Nenhum arquivo selecionado' is displayed. At the bottom, a blue button labeled 'Salvar em PDF' is visible.

Fonte: Elaborado pelo próprio autor.

Por outro lado, tem-se a experiência do Usuário (UX), a qual foi meticulosamente planejada para garantir que cada interação seja marcada por um design intuitivo. A disposição cuidadosa dos elementos na interface foi pensada para proporcionar uma experiência natural e eficaz, permitindo que os usuários se envolvam com o software de maneira instintiva.

O *Feedback* Visual desempenha um papel fundamental, com mensagens informativas e indicadores visuais incorporados para oferecer um retorno imediato ao usuário. Essa abordagem visa assegurar uma compreensão clara das ações realizadas, promovendo uma interação fluida e transparente.

Figura 5 – *Feedback* visual do Sistema



Fonte: Elaborado pelo próprio autor.

A busca pela simplicidade e clareza é evidente na apresentação das opções e nas instruções fornecidas pelo "SI-CHRONOS". Essa abordagem visa minimizar possíveis confusões, tornando o software acessível a usuários com diferentes níveis de familiaridade com a ferramenta.

A facilidade de navegação é uma prioridade na estrutura da interface, que foi projetada de maneira lógica. Isso facilita a navegação pelos diversos recursos do software, garantindo uma experiência coesa e agradável para os usuários, independentemente de sua expertise técnica.

A resiliência a erros é incorporada por meio da implementação de mecanismos que lidam com situações adversas de maneira amigável. Mensagens claras e orientações são fornecidas para auxiliar os usuários na correção de possíveis problemas, contribuindo para uma experiência de uso confiável e eficiente.

3.4.3 – Módulo de Manipulação de Dados

O Módulo de Manipulação de Dados desempenha um papel crucial no “SI-CHRONOS”, sendo responsável por processar e analisar os dados fornecidos pelos usuários. Este componente utiliza a biblioteca "pandas" em Python para estruturar e manipular conjuntos de dados, assegurando eficiência na execução de operações como leitura de arquivos e análise estatística.

Ao empregar a biblioteca "pandas", o *software* se beneficia de estruturas de dados flexíveis, como *DataFrames*, que facilitam a organização e manipulação dos dados. A leitura de arquivos, um passo essencial na interação com o usuário, é realizada de maneira eficiente, permitindo ao programa processar informações provenientes de diferentes fontes de maneira consistente e rápida.

A análise estatística, outro aspecto crucial do módulo, é otimizada pela rica funcionalidade oferecida pelo "pandas". Essa capacidade de análise abrange desde operações simples, como a identificação de estatísticas descritivas, até análises mais avançadas, contribuindo para a geração de conclusões a partir dos dados fornecidos.

3.4.4 – Módulo de Visualização de Dados

O módulo de visualização de dados exerce um papel crucial no contexto do sistema, sendo encarregado de transformar os dados processados em representações visuais significativas. Este componente faz uso da consagrada biblioteca "matplotlib" em Python, permitindo a geração de uma variedade de gráficos, tais como gráficos de dispersão, evidenciando padrões e tendências nos dados, histogramas que ilustram a distribuição de variáveis, e matrizes de correlação que revelam relações entre diferentes conjuntos de dados.

O Módulo de visualização de dados, ao utilizar a biblioteca "matplotlib", reforça a abordagem do sistema em fornecer não apenas análises estatísticas robustas, mas também uma representação visual marcante, promovendo uma compreensão mais completa e acessível dos resultados obtidos.

3.4.5 – Módulo de Controle e Lógica de Funcionamento

O Módulo de Controle e Lógica de Funcionamento desempenha um papel central na orquestração do sistema, sendo responsável por coordenar a interação harmoniosa entre os diversos componentes. Este componente atua como o cérebro operacional do *software*, implementando as funções essenciais que sustentam a estrutura e operação do sistema como um todo.

No âmbito do controle, o módulo supervisiona a execução de tarefas fundamentais, tais como a criação dinâmica de campos de entrada, a gestão do processo de upload de arquivos, a coordenação do processamento eficiente de dados, a orquestração de uma sequência coesa de interações e a instigação da geração de gráficos visuais.

A criação dinâmica de campos de entrada é um aspecto distintivo desse módulo, permitindo que o sistema se adapte dinamicamente às necessidades específicas dos usuários. A habilidade de criar campos de entrada de forma programática confere ao sistema uma flexibilidade notável, garantindo que a interface permaneça intuitiva e amigável, independentemente das variações nos requisitos dos usuários. A gestão do processo de *upload* de arquivos é outra faceta crítica, proporcionando aos usuários a capacidade de fornecer dados ao sistema de maneira eficiente e direta. O módulo implementa um fluxo de trabalho coeso que permite a importação de conjuntos de dados externos, viabilizando análises e visualizações subsequentes.

Ao coordenar o processamento de dados, o módulo assegura uma manipulação eficaz das informações fornecidas pelos usuários. Essa funcionalidade é fundamental para a condução de análises estatísticas e a subsequente criação de visualizações coerentes, contribuindo para uma compreensão mais profunda dos dados.

A condução de uma sequência coesa de interações é um elemento crucial para a experiência do usuário, garantindo uma navegação suave e lógica dentro do sistema. O módulo de controle gerencia os eventos e ações do usuário, assegurando uma resposta adequada e coerente em todas as etapas da interação.

3.4.6 – Módulo de Testes e Depuração

A fase de testes foi integral na implementação, envolvendo testes unitários, de integração e de sistema. Esses testes foram conduzidos para identificar e corrigir possíveis erros, garantindo o desempenho estável do sistema em diversas situações. Para assegurar a confiabilidade dos resultados do SI-CHRONOS, utilizamos o Excel como uma ferramenta de validação e comparação.

O Excel é amplamente reconhecido e validado como uma ferramenta eficaz para análise de dados, oferecendo funcionalidades robustas e amplamente testadas ao longo dos anos. Utilizamos o Excel para conferir os resultados obtidos pelo SI-CHRONOS, verificando a precisão e a integridade das análises realizadas pelo nosso software. Além disso, o Excel foi empregado para comparar o tempo de análise, permitindo uma avaliação objetiva da eficiência do SI-CHRONOS em relação a uma ferramenta já estabelecida.

3.4.7 – Módulo de Documentação

Uma documentação completa do código foi elaborada para auxiliar na compreensão e manutenção futura. Cada função, variável e componente significativo do sistema é explicado, proporcionando uma referência detalhada para desenvolvedores presentes e futuros.

Figura 6 – Código comentado.

```
249 # obter o intervalo de tempo
250 intervalo = float(intervalo_entry.get())
251
252 # imprime o conteúdo do arquivo na tela
253 print(df)
254
255 # cria gráfico de distribuição de normalidade para cada coluna
256 for col in df.columns:
257     # histograma
258     plt.figure()
259     # Histograma com cores personalizadas, bordas e transparência
260     plt.hist(df[col], bins=20, color='skyblue', edgecolor='black', alpha=0.7)
261     # cria a curva de distribuição normal
262     mu, std = stats.norm.fit(df[col])
263     xmin, xmax = plt.xlim()
264     x = np.linspace(xmin, xmax, 100)
265     p = stats.norm.pdf(x, mu, std) * (xmax - xmin) # Escala ajustada
266     plt.plot(x, p, 'k', linewidth=2)
267     plt.title(f"Distribuição de Frequência e Curva de Normalidade - {col}")
268     plt.xlabel(col)
269     plt.ylabel("Frequência")
270     pdf_plots.append(plt.gcf()) # Use plt.gcf() to get the current figure
271
272
273 # Atualiza o rótulo com o nome do arquivo
274 file_label.config(text=f"Arquivo Upload: {os.path.basename(file_path)}")
275 file_label.pack(pady=5)
276 upload_button.config(state='disabled') # Desabilita o botão após o upload
```

Fonte: Elaborado pelo próprio autor.

3.4.8 – Melhorias e Expansões Futuras

A arquitetura foi desenvolvida considerando a possibilidade de futuras melhorias e expansões. A estrutura modular permite a adição de novas funcionalidades ou aprimoramentos sem comprometer a estabilidade do sistema.

Essa flexibilidade arquitetônica não apenas simplifica o processo de implementação de melhorias, mas também agiliza a expansão do escopo do *software*. Novas funcionalidades podem ser integradas de forma incremental, proporcionando aos usuários uma experiência contínua e aprimorada sem a necessidade de reestruturações massivas.

3.4.9 – Confiabilidade

Na implementação do *software* em questão, uma escolha estratégica foi feita em relação ao armazenamento de informações do usuário. Diferentemente de muitas aplicações que fazem uso de bancos de dados persistentes para reter dados entre sessões, nosso programa optou por uma abordagem singular para garantir a confiabilidade dos dados e a privacidade do usuário.

A característica distintiva do “SI-CHRONOS” reside na sua capacidade de reiniciar todas as informações a cada execução. Em outras palavras, a cada vez que o programa é iniciado, ele inicia um estado limpo, sem qualquer histórico residual das atividades ou dados dos usuários.

Esta abordagem não apenas fortalece a confiabilidade de dados, mas também simplifica a experiência do usuário. Não há necessidade de preocupações com exclusão manual de dados ou com potenciais exposições de informações confidenciais, pois o sistema é automaticamente restaurado a um estado inicial a cada uso.

Em suma, a decisão de não implementar um banco de dados persistente no *software* visa proporcionar aos usuários uma ferramenta eficiente, na qual, a cada execução, o programa irá oferecer um ambiente de trabalho limpo para as informações do usuário.

4 - RESULTADOS E DISCUSSÃO

Nesta seção, serão apresentados resultados das análises realizadas utilizando o programa SI-CHRONOS, comparando-os com os resultados obtidos através do uso do Microsoft Excel para o mesmo banco de dados, afim de se aferir a eficiência da análise. Justifica-se o uso do *software* Excel por se tratar de uma ferramenta de análise validada e amplamente utilizada. A comparação entre SI-CHRONOS e Excel se baseia em diferentes parâmetros de desempenho, incluindo tempo de processamento e precisão das análises. O objetivo é demonstrar as vantagens do SI-CHRONOS em termos de economia de tempo e simplicidade de uso, o que pode proporcionar benefícios significativos para pesquisadores e profissionais que necessitam realizar análises de dados de forma eficiente e rápida.

4.1 – Validação do sistema

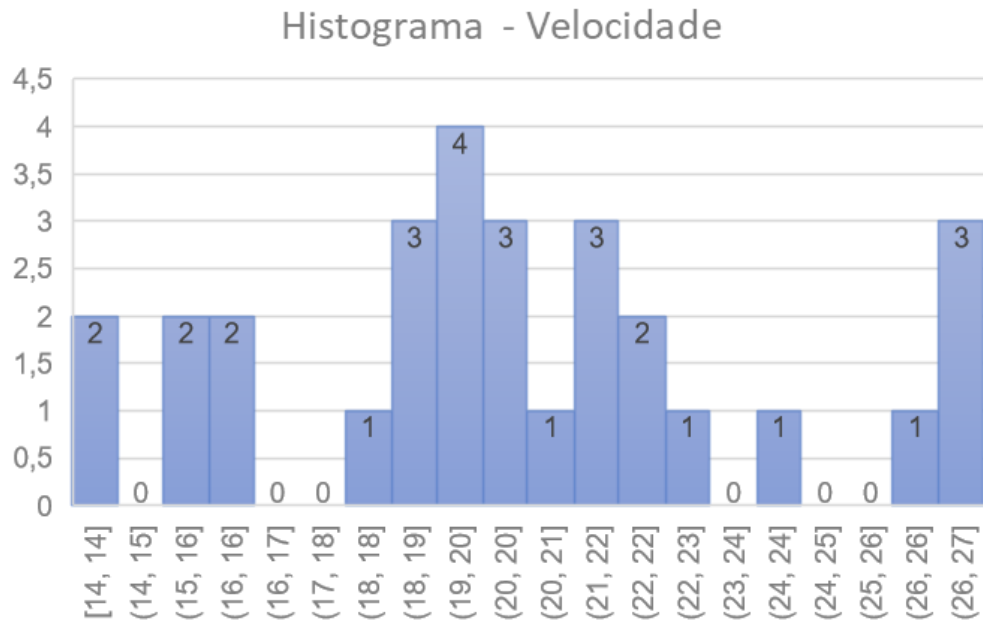
Para avaliar a eficiência do programa SI-CHRONOS em comparação com o Microsoft Excel, foram realizados testes específicos medindo o tempo necessário para executar várias análises de dados. Os resultados demonstram uma significativa economia de tempo ao utilizar o SI-CHRONOS, conforme detalhado a seguir.

Os dados de velocidade utilizados para as análises foram gerados de forma aleatória para simular diferentes cenários de desempenho. Essa abordagem permite uma avaliação imparcial das capacidades do SI-CHRONOS em comparação com o Excel, destacando as diferenças de tempo de processamento e a simplicidade de uso entre as duas ferramentas.

4.1.1 – Histograma e curva de normalidade

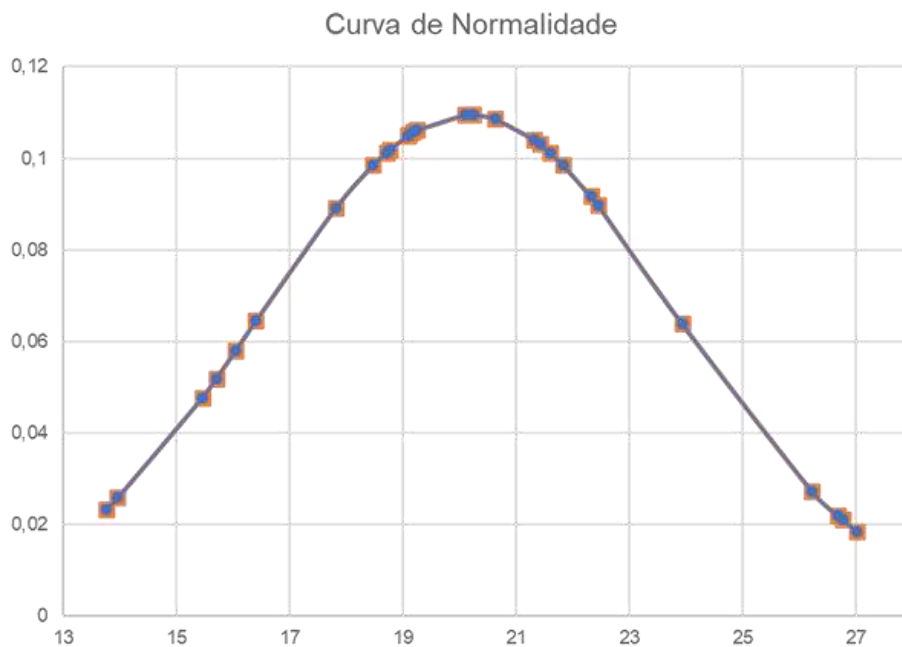
Por meio da utilização do Excel, o processo levou cerca de 16 minutos e 30 segundos. Esse tempo inclui a formatação dos dados, a criação dos gráficos, visto que foi necessária a criação de duas visualizações separadas, uma para a visualização da curva de normalidade e outra para a distribuição de frequência.

Figura 7 – Distribuição de frequência (Excel).



Fonte: Elaborado pelo próprio autor por meio do *Software* Excel.

Figura 8 – Curva de normalidade (Excel).

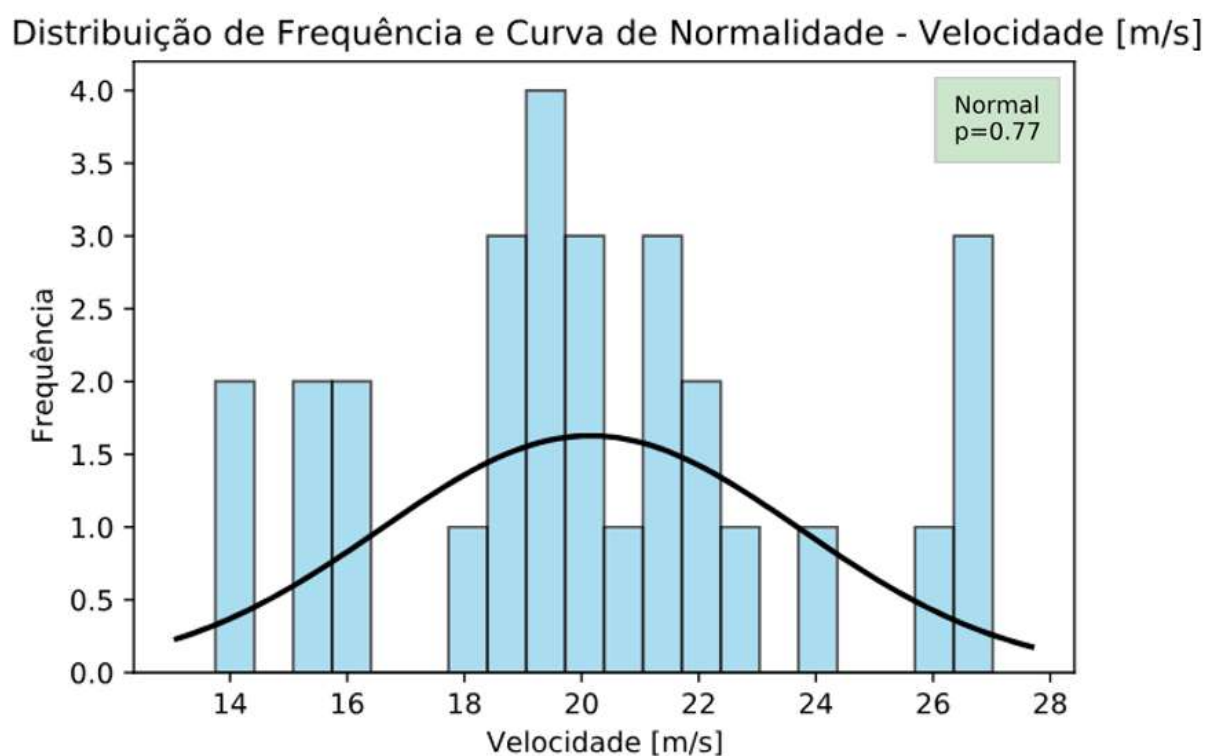


Fonte: Elaborado pelo próprio autor por meio do *Software* Excel.

Com o SI-CHRONOS, o processo é concluído em menos de um minuto, visto que o sistema responde praticamente instantaneamente, somente sendo levado em consideração o tempo de interação do usuário com a interface e de inserção dos

dados, que foram cerca de 50 segundos. O sistema automatiza a criação do histograma, adiciona a curva de normalidade e calcula o 'p-valor' para verificar a normalidade da distribuição. Isso reduz drasticamente o tempo necessário para concluir a tarefa, fornecendo os dados de forma consolidada em uma única visualização

Figura 9 – Histograma e curva de normalidade (SI-CHRONOS).

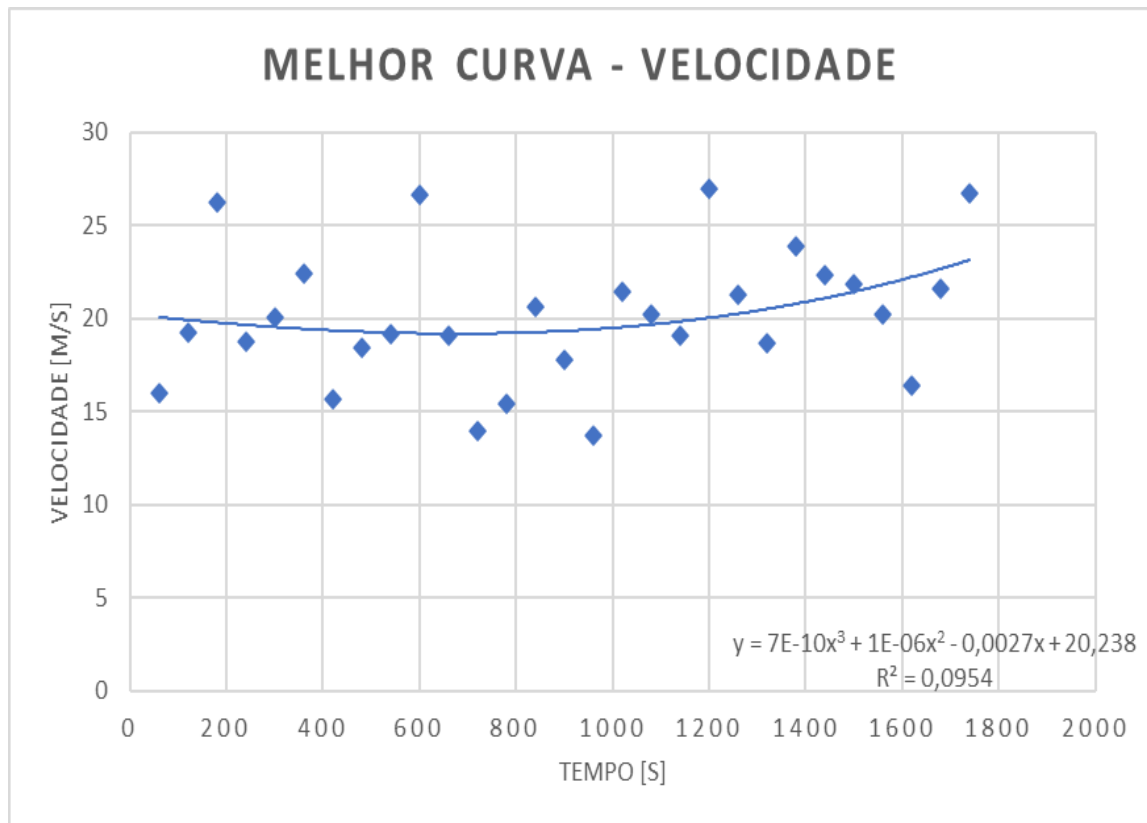


Fonte: Elaborado pelo próprio autor por meio do sistema SI-CHRONOS.

4.1.2 – Melhor curva entre os pontos amostrais

No Excel, o ajuste demandou 18 minutos, incluindo o tempo de inserção dos dados, a escolha do tipo de curva e o ajuste manual, além do tempo de aprendizado do usuário, que pode variar de pessoa para pessoa.

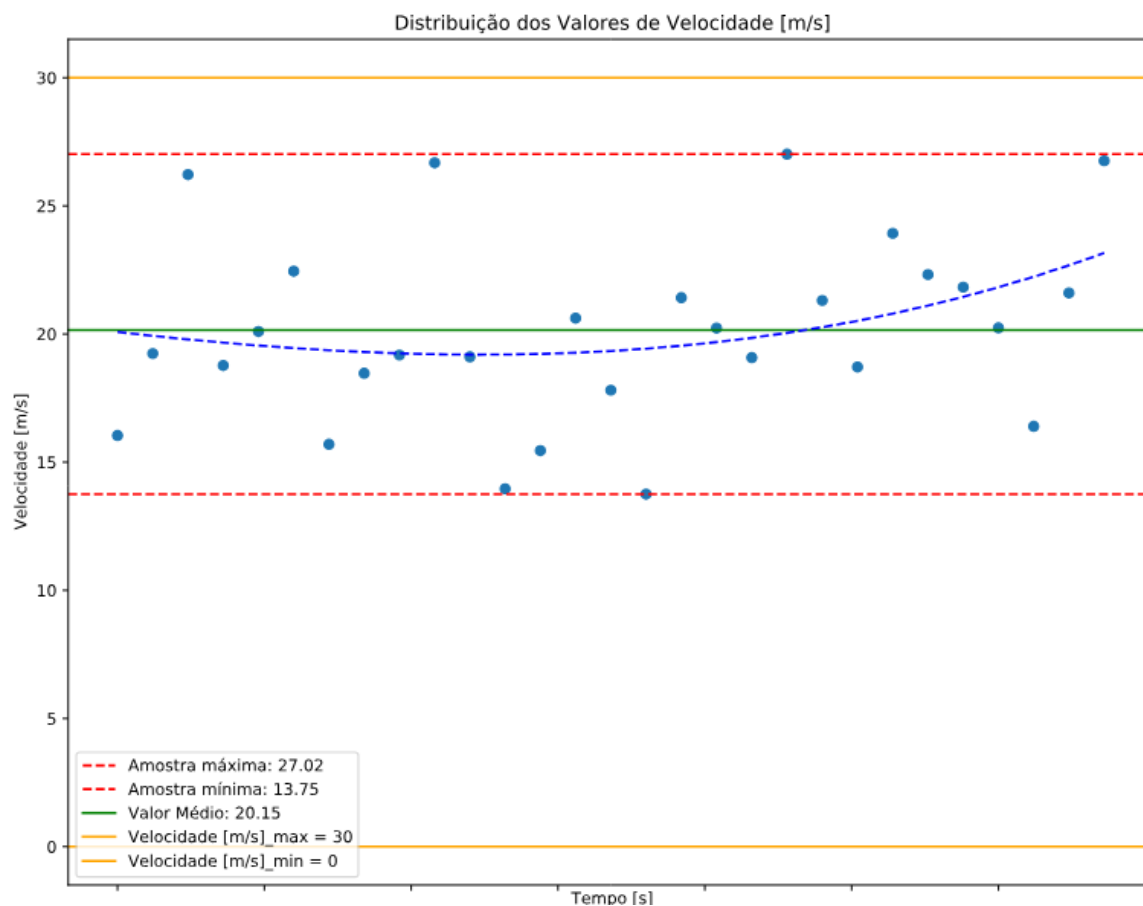
Figura 10 – Melhor curva entre os dados (Excel).



Fonte: Elaborado pelo próprio autor por meio do *Software* Excel.

Já com o SI-CHRONOS, o ajuste é concluído em menos de 1 minuto. Visto que o sistema emprega algoritmos automatizados que selecionam rapidamente a melhor curva de ajuste, eliminando a necessidade de intervenção manual extensa e, consequentemente, reduzindo o tempo total necessário.

Figura 11 – Melhor curva entre os dados (SI-CHRONOS).



Fonte: Elaborado pelo próprio autor por meio do sistema SI-CHRONOS.

4.1.3 – Correlação de Variáveis

Com o uso do Excel, foram necessários 26 minutos para realizar a correlação entre as variáveis selecionadas. O processo incluiu a instalação do complemento *Tool Kit* de Análise de Dados, a preparação dos dados e a execução da análise, além do tempo de aprendizado do usuário, que pode variar.

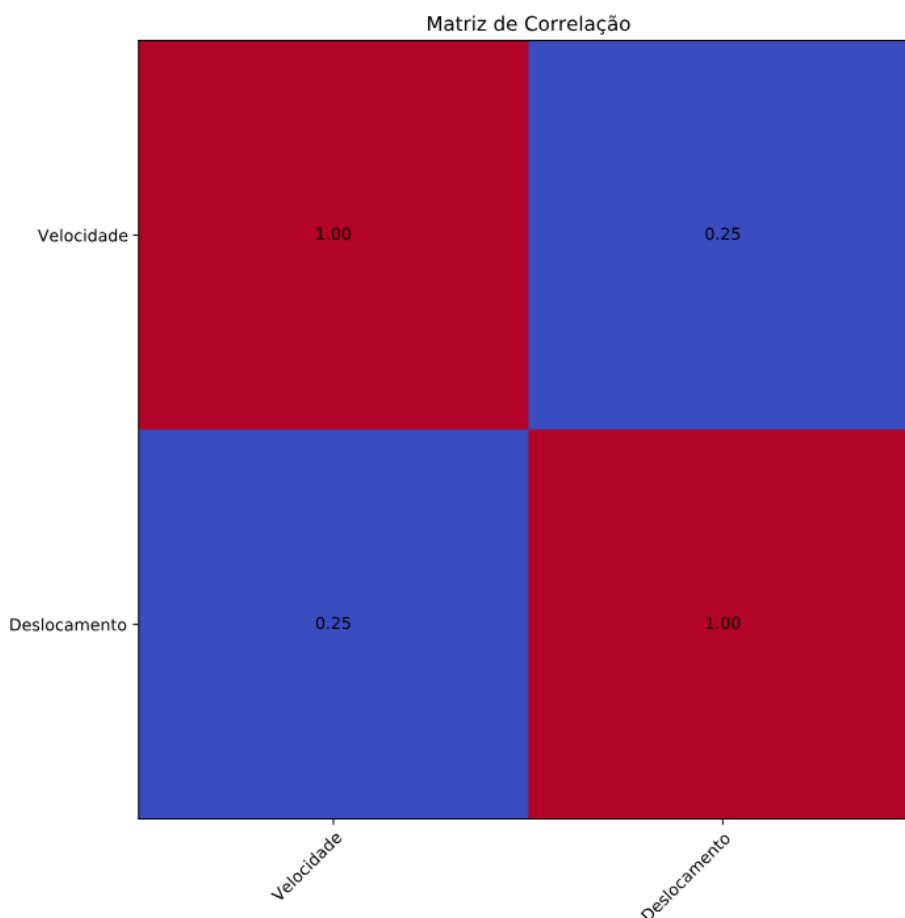
Figura 12 – Correlação de Variáveis (Excel).

	<i>Velocidade</i>	<i>Deslocamento</i>
<i>Velocidade</i>	1	0,245
<i>Deslocamento</i>	0,245	1

Fonte: Elaborado pelo próprio autor por meio do *Software Excel*.

Já com o SI-CHRONOS, a correlação foi realizada em menos de 1 minuto já em conjunto com as análises anteriores. O programa integra ferramentas de análise estatística que permitem a execução de correlações de forma quase instantânea, sem a necessidade de complementos adicionais. A análise da correlação de variáveis pelo SI-CHRONOS não apenas reduziu o tempo de processamento, mas também simplificou o fluxo de trabalho, tornando o processo mais intuitivo e menos propenso a erros.

Figura 13 – Correlação de variáveis (SI-CHRONOS).



Fonte: Elaborado pelo próprio autor por meio do sistema SI-CHRONOS.

4.1.4 – Comparativo final

Ao comparar a utilização do SI-CHRONOS com o Microsoft Excel, é evidente uma discrepância marcante em termos de eficiência e praticidade na análise de dados. Considerando todas as análises realizadas em conjunto, por meio do Excel foi necessário um tempo total de 65 minutos para a finalização de todas análises, levando em conta todos os processos envolvidos, incluindo formatação dos dados, inserção de complementos, ajustes manuais, e o tempo de aprendizado do usuário, variável entre diferentes indivíduos.

Por outro lado, o SI-CHRONOS executou todas as análises simultaneamente, sendo essa uma das vantagens do sistema, a capacidade de realizar todas as análises de uma só vez em um tempo praticamente instantâneo, sendo considerado apenas o tempo de interação do usuário com a interface do sistema e a inserção dos dados,

que no caso analisado foi de 50 segundos. Isso representa uma redução percentual de aproximadamente 98,7% no tempo total necessário em comparação com o Excel.

É importante ressaltar que os dados gerados pelo Excel foram utilizados como ferramenta de validação para os resultados obtidos pelo SI-CHRONOS, os quais se mostraram compatíveis com os obtidos pelo Excel, demonstrando a confiabilidade e precisão do sistema como uma alternativa eficaz para análise de dados.

Além da economia de tempo, o SI-CHRONOS simplificou o fluxo de trabalho ao consolidar todas as etapas de análise em uma única plataforma, proporcionando uma experiência mais intuitiva para o usuário. A capacidade do sistema de realizar análises complexas de forma rápida e precisa torna-o uma ferramenta valiosa para pesquisadores e profissionais que necessitam de resultados confiáveis em um curto espaço de tempo.

Em suma, os resultados desta comparação evidenciam claramente as vantagens do SI-CHRONOS em termos de eficiência, simplicidade de uso e precisão das análises, destacando-o como uma ferramenta indispensável para análise de dados em diversas áreas de pesquisa e prática profissional.

4.2 – Aplicação

Para exemplo de aplicação prática do sistema foram utilizados dados provenientes da Agência Nacional de Águas e Saneamento Básico (ANA) [14]. O programa foi aplicado com o objetivo de analisar a vazão Q95 do rio Pardo, localizado próximo a cidade de Barretos-SP, durante todos os meses do ano de 2021 e fornecer informações sobre as características hidrológicas do rio e suas implicações para a gestão dos recursos hídricos.

Os dados fornecidos pela ANA são de suma importância, pois representam uma fonte confiável e abrangente de informações sobre os recursos hídricos do Brasil.

Tabela 2 – Vazão Q95 do Rio Pardo

MÊS	Q95 [m³/s]
JAN	301,212
FEV	267,004
MAR	302,676

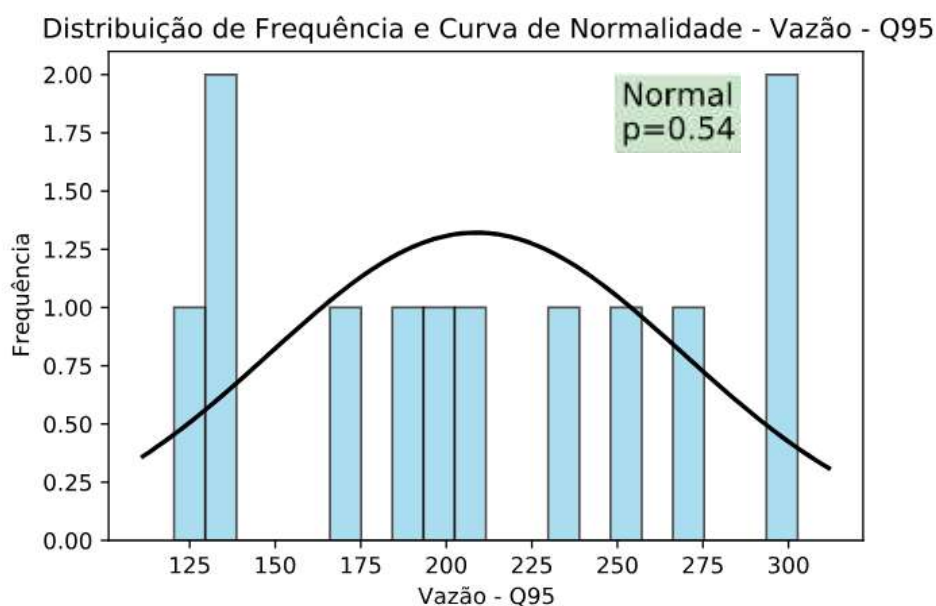
ABR	250,731
MAI	209,483
JUN	193,611
JUL	172,137
AGO	133,228
SET	135,106
OUT	120,439
NOV	188,618
DEZ	231,334

Fonte: [1]

A vazão Q95 se trata um parâmetro fundamental na avaliação da disponibilidade hídrica e na compreensão dos padrões de fluxo do rio ao longo do tempo. Por meio da análise da vazão, buscamos identificar padrões sazonais, tendências de longo prazo e possíveis anomalias que possam influenciar as condições hidrológicas do rio.

O programa desenvolvido permitiu a criação de gráficos de dispersão que representam a relação entre a vazão do rio e o tempo, além de fornecer informações sobre a distribuição dos valores das vazões. Esses gráficos foram complementados com análises estatísticas, incluindo testes de normalidade e correlação, que contribuíram para uma compreensão mais abrangente das características hidrológicas do rio Pardo.

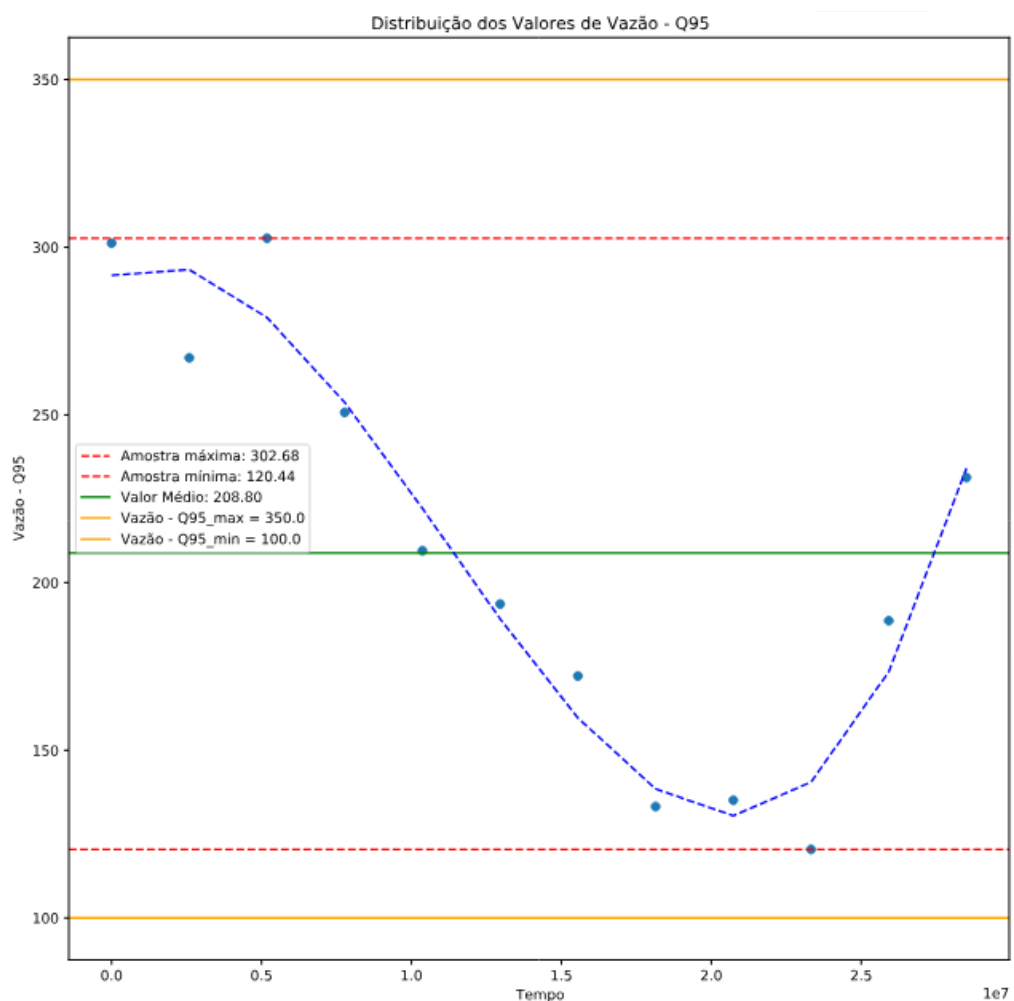
Figura 14 – Distribuição de Frequência e Curva de Normalidade (Q95)



Fonte: Elaborado pelo próprio autor através do aplicativo SI-CHRONOS.

A distribuição se mostrou estatisticamente normal, com um p-valor de 0,54, identificando a normalidade dos dados. A curva de sino é claramente delineada no gráfico de distribuição de normalidade. Esse resultado era esperado, já que na natureza grande parte das situações tende a seguir uma distribuição normal. No entanto, em amostras pequenas (menos de 20 amostras), os testes de normalidade podem ter baixo poder estatístico, o que significa que podem não ser sensíveis o suficiente para detectar desvios significativos da normalidade, mesmo que existam. Portanto, embora o resultado do teste sugira normalidade, é essencial reconhecer as limitações inerentes à análise de dados com amostras tão pequenas e considerar outras evidências e métodos de análise para uma interpretação mais abrangente e precisa dos dados.

Figura 15 – Distribuição de Valores de Q95



Fonte: Elaborado pelo próprio autor através do aplicativo SI-CHRONOS.

Através da distribuição, pode-se chegar à conclusão de que os dados de vazão Q95 ao longo dos meses revela padrões sazonais distintos e sugere uma correlação entre os meses e os valores de vazão. Observa-se um aumento notável nos valores de vazão nos meses de outubro a janeiro, com os picos mais altos registrados em janeiro (301,212 m³/s), março (302,676 m³/s) e dezembro (231,334 m³/s). Esse aumento pode ser atribuído a uma variedade de fatores sazonais, como precipitação aumentada devido a estações chuvosas. Por outro lado, os valores de vazão tendem a diminuir de fevereiro a setembro, atingindo os valores mais baixos em agosto (133,228 m³/s) e setembro (135,106 m³/s). Essa queda pode ser associada a períodos de estiagem ou menor precipitação durante esses meses. Esses padrões sazonais na vazão Q95 destacam a importância de considerar as variações temporais na disponibilidade de recursos hídricos ao planejar e gerenciar projetos de uso da água e para a compreensão dos ecossistemas aquáticos.

E também, caso necessária uma análise mais detalhada do caso, em questão de tempo, o *software* nos fornece a melhor equação da curva, a qual pode ser utilizada para a determinação de Q95 em qualquer período de tempo.

Figura 16 – Melhor Equação entre os pontos

Equação: $y = 4.456e-20 x^3 - 1.506e-12 x^2 + 4.29e-06 x + 291.6$ $R^2 = 0.95$
--

Fonte: Elaborado pelo próprio autor através do aplicativo SI-CHRONOS.

No caso estudado, foi obtida uma equação de polinomial de terceiro grau, além de um valor de 0,95 para R^2 , ou seja, o modelo é capaz de explicar uma grande parte da variabilidade dos dados observados. Isso geralmente é considerado um valor muito alto para o R^2 e sugere que o modelo de regressão se ajusta muito bem aos dados.

Ao examinarmos os resultados obtidos pela análise da vazão Q95, buscamos extrair informações relevantes que possam orientar políticas e práticas de manejo de recursos hídricos na região de Barretos. Essa análise é de suma importância para a gestão sustentável dos recursos hídricos, bem como possíveis estratégias para enfrentar os desafios associados à variabilidade hidrológica e às mudanças climáticas.

5 – CONCLUSÃO

Este trabalho teve como objetivo principal o desenvolvimento de um programa em Python denominado SI-CHRONOS (Sistema Integrado de Análises Cronológicas) para facilitar a análise de dados em diferentes áreas, reduzindo o tempo de análise e minimizando a incidência de riscos de erros de natureza humana. Com a implementação deste software, possibilitou-se alcançar os objetivos propostos de maneira eficaz, conforme demonstrado pelos resultados obtidos.

O programa foi projetado para minimizar a intervenção do usuário em processos críticos, garantindo que as análises fossem realizadas de forma automática e sem erros, seguindo assim, o conceito *Poka Yoke*, que se refere à prevenção de erros humanos através de mecanismos e processos que evitam a ocorrência de falhas. Esse enfoque preventivo foi essencial para assegurar a confiabilidade dos resultados e aumentar a eficiência do *software*.

Em paralelo, o SI-CHRONOS apresentou uma excelente capacidade de reduzir significativamente o tempo necessário para a análise de dados. Testes comparativos realizados com o Microsoft Excel mostraram que o sistema conseguiu diminuir o tempo de análise em 98,72%. Essa economia de tempo representa uma grande vantagem para pesquisadores e profissionais que dependem de análises rápidas e precisas para tomar decisões embasadas.

Além do mencionado, o programa apresenta uma interface gráfica amigável, elaborada com a biblioteca “Tkinter”. Essa abordagem permite que os usuários interajam de maneira intuitiva, tornando-o acessível a qualquer tipo de público. A criação deste sistema representa uma valiosa contribuição para a comunidade acadêmica e científica, fornecendo uma solução prática e eficaz para a análise de dados em pesquisas.

Em suma, o desenvolvimento do SI-CHRONOS foi realizado com êxito, demonstrando eficácia na redução do tempo de análise e na minimização de erros através da aplicação do conceito de *Poka Yoke*. A ferramenta mostrou-se útil e eficiente, proporcionando uma plataforma acessível e de fácil utilização para análises estatísticas, contribuindo assim para o avanço de futuras pesquisas e estudos em diversas áreas do conhecimento.

5.1 – TRABALHOS FUTUROS

A versão inicial do "SI-CHRONOS" representa um marco significativo em nossa jornada para fornecer aos usuários uma plataforma acessível e intuitiva para explorar e compreender dados temporais de forma mais eficiente. No entanto, é possível reconhecer que há espaço para melhorias e expansões substanciais que podem enriquecer ainda mais a funcionalidade e utilidade do programa.

Este aplicativo não é apenas uma ferramenta, mas uma expressão da colaboração e engajamento da comunidade em torno de um objetivo comum: capacitar os usuários a explorar e compreender dados temporais de forma mais eficiente e intuitiva. Criado por membros da comunidade acadêmica, para a comunidade, este programa é um testemunho do poder da colaboração e da filosofia do código aberto, onde qualquer indivíduo interessado pode se sentir livre a modificar, adaptar e dar continuidade ao programa caso for de sua vontade.

Seguindo esse viés, são apresentados pontos nos quais trabalhos futuros podem se concentrar para impulsionar ainda mais a evolução do programa.

5.1.1 – Suporte a mais formatos

Expandir o suporte a uma variedade de formatos de arquivo pode melhorar a interoperabilidade do programa com outras ferramentas e sistemas. Isso pode incluir a capacidade de importar e exportar dados em formatos populares como CSV, Excel, JSON, entre outros.

5.1.2 – Expansão de Funcionalidades Analíticas

Adicionar novas funcionalidades analíticas pode aumentar a utilidade do programa para uma variedade de aplicações. Por exemplo, a inclusão de métodos de análise estatística avançada, técnicas de aprendizado de máquina e ferramentas de visualização de dados interativas pode permitir uma análise mais profunda e abrangente dos conjuntos de dados temporais.

5.1.3 – Transformação em um Executável

Uma maneira de tornar o "SI-CHRONOS" mais acessível e fácil de usar para uma ampla gama de usuários é transformar o código-fonte em um executável independente. Isso permitirá que os usuários executem o programa sem a necessidade de instalar ou configurar bibliotecas adicionais, simplificando o processo de uso. Além disso, a distribuição de um executável pode facilitar a adoção do programa por usuários que não estejam familiarizados com o ambiente de desenvolvimento Python, ampliando assim o alcance e o impacto da aplicação. A criação de um executável também pode contribuir para uma melhor experiência do usuário, eliminando potenciais barreiras técnicas e simplificando o processo de instalação e execução do programa. Para isso, será necessário explorar ferramentas e técnicas de empacotamento e distribuição de *software*, como PyInstaller, cx_Freeze ou py2exe, para converter o código Python em um executável que possa ser utilizado em diferentes plataformas, como Windows, macOS e Linux. Dessa forma, estaremos dando mais um passo em direção à democratização do acesso e à promoção da análise de dados temporais para um público mais amplo e diversificado.

5.1.4 – Otimização de Desempenho e Escalabilidade

Investir em otimizações de desempenho e escalabilidade pode garantir que o programa seja capaz de lidar com grandes volumes de dados e executar análises complexas de forma eficiente. Isso pode incluir a implementação de técnicas de paralelização, otimização de algoritmos e gerenciamento eficaz de recursos do sistema.

6 - REFERÊNCIAS BIBLIOGRÁFICAS

1. McKinney, W. Data Analysis and Visualization with Python. O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.
2. Amadebai, E. 13 Reasons Why Data Is Important in Decision Making. Analytics For Decisions. Disponível em: <<https://www.analyticsfordecisions.com/data-is-important-in-decision-making/>>. Acesso em: 2 fev. 2024.
3. Idris, I. Python Data Analysis. Packt Publishing Ltd. Birmingham B3 2PB, October 2014.
4. GUPTA, B.C.; GUTTMAN, I. Estatística e probabilidade com aplicações para engenheiros e cientistas, LTC, 2017.
5. MENDENHALL, W.; BEAVER, R.J., BEAVER, B. M. Introduction to Probability and Statistics, Metric Version, Cengage, 2020.
6. BALAKRISHNAN, N.; KOUTRAS, M.V.; POLITIS, K.G. Introduction to Probability Models and Applications, John Wiley & Sons, 2020.
7. BUSSAB, W.O.; MORETTIN, P.A. Estatística Básica. Editora Saraiva, 2017.
8. DEVORE, J.L. Probabilidade e estatística para engenharia e ciências. 9 ed. São Paulo: CENGAGE Learning, 2018.
9. Pressman, R. et. Al. Engenharia de Software – uma abordagem profissional, McGraw Hill Education, 2021.
10. Oliveira, D. et.al. Inovação pela abordagem design thinking: um estudo de projetos desenvolvidos em uma instituição de ensino superior. CNEC Educação, v. 2, n. 1, 2019. Disponível em: <<http://sys.facos.edu.br/ojs/index.php/cneceduacao/article/viewFile/320/292>>. Acesso em: 5 mai. 2024.
11. Gadelha Jr, L. M. et al. A survey of biodiversity informatics: Concepts, practices, and challenges. 1 out. 2020. Disponível em: <<https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1394>>. Acesso em: 7 mai. 2024.
12. Porto, F. et al. Machine Learning Approaches to Extreme Weather Events Forecast in Urban Areas: Challenges and Initial Results. Supercomputing Frontiers and Innovations, 9(1), 49–73 (2022). Disponível em: <<https://doi.org/10.14529/jsfi220104>>. Acesso em: 7 mai. 2024.

13. Marques-Neto, H.T., Xavier, F.H.Z., Xavier, W.Z. et al. Understanding Human Mobility and Workload Dynamics Due to Different Large-Scale Events Using Mobile Phone Data. *J Netw Syst Manage* 26, 1079–1100 (2018). Disponível em: <<https://doi.org/10.1007/s10922-018-9454-3>>. Acesso em: 7 mai. 2024.
14. AGÊNCIA NACIONAL DE ÁGUAS (ANA). Índices e Estatísticas Hidrometeorológicas - Estação Fluviométrica (Dados Consistidos e Brutos) - número de anos da série. 11 de mar. 2022. Disponível em: https://dadosabertos.ana.gov.br/datasets/2a045c4bb2aa4c599fcca73724cae786_3/explore?location=-20.774870%2C-48.545690%2C9.00. Acesso em: 1 maio 2024.
15. Shingo, S. *Zero Quality Control: Source Inspection and the Poka-Yoke System*. CRC Press, 1986.