CENTRO UNIVERSITÁRIO FEI

BRUNO ARTHUR BASSO SILVA

USO DE ÁRVORES DE DECISÃO PARA ANÁLISE DE DADOS NAS DISCIPLINAS DE LABORATÓRIO DE FÍSICA

São Bernardo do Campo

2024

BRUNO ARTHUR BASSO SILVA

USO DE ÁRVORES DE DECISÃO PARA ANÁLISE DE DADOS NAS DISCIPLINAS DE LABORATÓRIO DE FÍSICA

Relatório Parcial de Iniciação Didática apresentado ao Centro Universitário FEI, como parte dos requisitos do Programa PIBIC-FEI. Orientado pela Prof. Eliane de Fátima Chinaglia

São Bernardo do Campo

2024

RESUMO

Existe uma necessidade de mudança no curso de engenharia, sendo necessário tornar o aluno mais atualizado no cenário da tecnologia e crescimento exponencial de Machine Learning. Com o auxílio de ferramentas e algoritmos. Esses algoritmos podem contribuir para o Laboratório de Física. Dessa maneira os alunos poderão trabalhar de forma mais prática e eficaz alguns tópicos que exigem maior tempo para compreensão. Utilizando Python como linguagem de programação e o Scikit-Learn como principal biblioteca de Árvore de Decisão, além de bibliotecas como Matplotlib para esboçar gráficos e dendrogramas, com a finalidade de ajudar o aluno de engenharia a entender melhor os conceitos e os dados. Ademais, construímos um trilho de ar, capaz de retirar o atrito, ligado a uma polia, na qual existem duas massas que compõem o projeto, uma após a polia e outra no começo do trilho, juntamente com sensores que captam a variação da velocidade e tempo, utilizando o Data Studio. Dessa forma, auxilia o aluno na extração de dados para intensificar a acurácia da sua Inteligência Artificial. Após a aquisição, o foco será voltado totalmente para teoria e conceitos de Árvore de Decisão e sua precisão, gerando uma resposta com resultados bem próximos dos resultados experimentais.

Palavras-chave: Árvore de Decisão. Python. Machine Learning. Scikit-Learn. Física Básica.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Com o aumento exponencial da Inteligência Artificial (IA), sua utilização no cotidiano das pessoas se tornou comum, sendo em sua maioria para facilitar e auxiliar em tarefas básicas, com o aprimoramento dos conceitos gerais. Visando a formação do curso de Engenharia, a utilização da IA é indiscutível. Para isso, o Laboratório de Física irá contribuir a fim de mostrar os conceitos de uma máquina inteligente para capacitar os engenheiros. Essas tecnologias têm o potencial de aprimorar processos, otimizar a tomada de decisões e catalisar a inovação em uma vasta gama de domínios (IBM, 2023).

O Machine Learning (ML) é o subconjunto da IA que se concentra na construção de sistemas que aprendem, ou melhoram o desempenho, sem ser explicitamente programado, evolvendo algoritmos que podem identificar padrões e tomar decisões com base nos dados.

Consequentemente, a compreensão dos princípios e das aplicações da IA e ML, tornou-se não apenas fundamental para profissionais de tecnologia, mas também para aqueles que buscam se aprimorar em suas respectivas áreas de atuação e explorar as oportunidades proporcionadas por essas tecnologias.

À medida que a relevância dessas tecnologias cresce tanto em nossa sociedade quanto no mercado de trabalho, torna-se essencial que os futuros engenheiros adquiram, pelo menos, uma compreensão básica de como os algoritmos de IA funcionam, a fim de empregá-los de maneira consciente, positiva e construtiva. No contexto do ensino de física básica para estudantes de engenharia, especialmente em ambientes de laboratório experimental, a coleta e análise de dados desempenham um papel crucial (MEI/CNI, 2018).

O objetivo do projeto é apresentar a um aluno de engenharia os conceitos básicos de uma IA atrelado aos conceitos de física, como a segunda lei de newton, utilizando também os conceitos básicos de programação.

Nesse contexto, este projeto visa aproveitar a oportunidade de coletar e analisar dados durante as aulas de laboratório para proporcionar aos alunos ingressantes a interação com algoritmos básicos de ML. Tal abordagem não só

contribui para o entendimento da física, mas também para a familiarização dos alunos com conceitos de IA destacando a importância dessa habilidade em sua formação, juntamente explorando a prática em linguagens de programação desde o início da graduação (MITCHELL, 1997)

Neste projeto serão explorados algoritmos de ML tanto não supervisionados (agrupamento) quanto supervisionados (Árvore de Decisão) para analisar dados obtidos em experimentos realizados no laboratório de física. O método de análise de agrupamentos (cluster) é um exemplo de algoritmo não supervisionado que visa identificar correlações significativas entre observações com base em variáveis numéricas, resultando na formação de grupos homogêneos internamente, mas heterogêneos entre si. Por outro lado, a Árvore de Decisão, um algoritmo supervisionado, segmenta conjuntos de dados com base em características específicas, com foco em minimizar o erro de classificação e permitir a classificação automática de novas observações.

A fim de assegurar a fidelidade do projeto, a aquisição de dados se torna crucial para fazer o devido treinamento da IA permitindo aumentar a acurácia de saída do programa.

2 REVISÃO BIBLIOGRÁFICA

Dado que a meta do projeto consiste em enriquecer o ensino introdutório de Física através da tecnologia e aprofundamento no ramo da IA, é essencial refinarmos as abordagens dos temas tratados neste projeto. O tópico selecionado para a criação de uma simulação foi a cinemática e a dinâmica das partículas.

2.1 ANÁLISE DE AGRUPAMENTO HIERÁRQUICO

A análise de agrupamentos hierárquico é um método de ML cujo objetivo é verificar se há uma correlação significativa entre as diversas observações em função de variáveis numéricas. Como resultado, obtém-se uma estrutura de grupos (cluster) que não possuem correlação significativa entre si, mas as observações contidas em cada grupo carregam uma relação de similaridade. Ou seja, obtém-se grupos heterogêneos em si, mas homogêneos internamente.

No agrupamento hierárquico, o resultado que mostra como os dados podem ser agrupados em diferentes níveis de granularidade através de cálculos de distância (FÁVERO, 2022).

Quando se trata de clusters, é obrigatório entender os conceitos de métodos e métricas a fim de calcular a distância entre diferentes pontos da dispersão dos dados em função de uma variável. Os métodos e métricas criam uma hierarquia de clusters, agrupando os dados de forma gradual, nível por nível. Os métodos determinam como os clusters são formados, enquanto as métricas avaliam a similaridade entre os pontos de dados (SCIPY, 2008).

Existem vários métodos e métricas possíveis para agrupar os dados, como “Single Linkage”, que é a distância mínima entre as observações de um grupo A em relação a um grupo B, “Complete Linkage”, é a distância máxima entre as observações de um grupo A em relação a um grupo B (SCIPY, 2008).

Os melhores métodos e métricas para o desenvolvimento do projeto, se trata dos métodos *single linkage* e *ward*, ambos usando a métrica *euclidean*. Escolhemos esses porque, posteriormente, na seção 4.2, terá uma análise mais

clara visualizando os gráficos dos respectivos métodos e métricas. Abaixo será apresentado o cálculo de cada um dos métodos que serão utilizados:

Método: *single.*

$$d(u,v) = \min\big(dist(u[i], v[j])\big) \quad (1)$$

Calculando a distância mínima $d(u,v)$ entre os pontos dos cluster *u* e *v.*

Métrica*: euclidean.*

Essa métrica faz a distância euclidiana entre os pontos pelo seguinte cálculo:

$$d(u,v) = \sqrt{((u1-u2)^2 + (v1-v2)^2)} \quad (2)$$

Método: *ward.*

$$d(u,v) = \sqrt{\frac{|v|+|s|}{T}d(v,s)^2 + \frac{|v|+|t|}{T}d(v,t)^2 - \frac{|v|}{T}d(s,t)^2} \quad (3)$$

Onde:

u, v, t: clusters que são usados para calcular a distância pela equação (3);

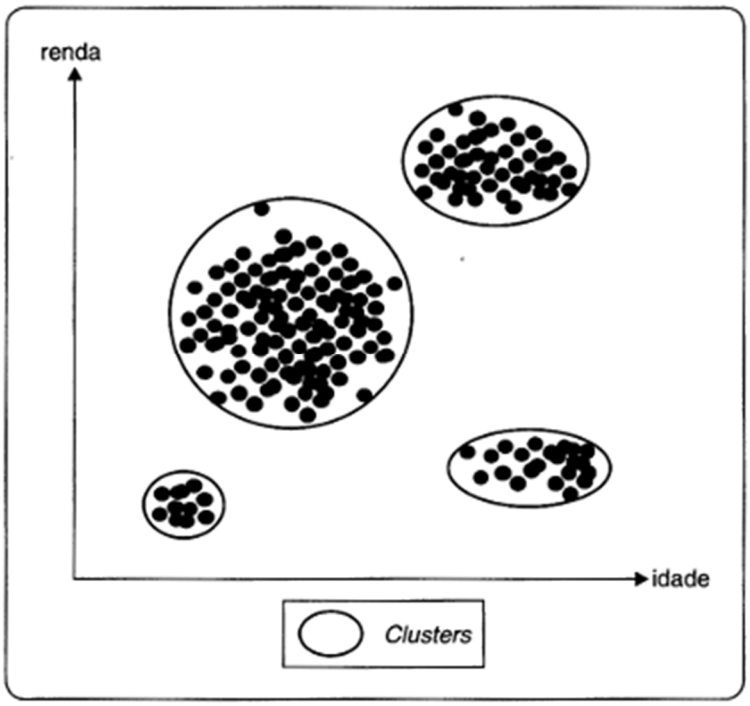
T: Representa o tamanho total do novo cluster "u", que é a soma dos tamanhos dos clusters *s* e *t* (| *s* |+ | *t* |).

Posteriormente, será feita a devida análise desses 2 casos presentes nesse tópico, totalmente detalhado.

A figura 1 representa o funcionamento da distribuição de um agrupamento hierárquico com clusters e o exemplo é referente a renda pela idade da população.

A partir da formação desses clusters, percebe-se que as pessoas de idade e mais novas tem uma renda menor em relação a uma pessoa que está inserida no mercado de trabalho (FÁVERO, 2022).

Figura 1 - Agrupamento de 4 clusters



Fonte: 'Manual de Análise de Dados'. v.9.1. p – 301.

Basicamente os clusters são agrupamentos feitos para otimizar os resultados e a dispersão dos dados.

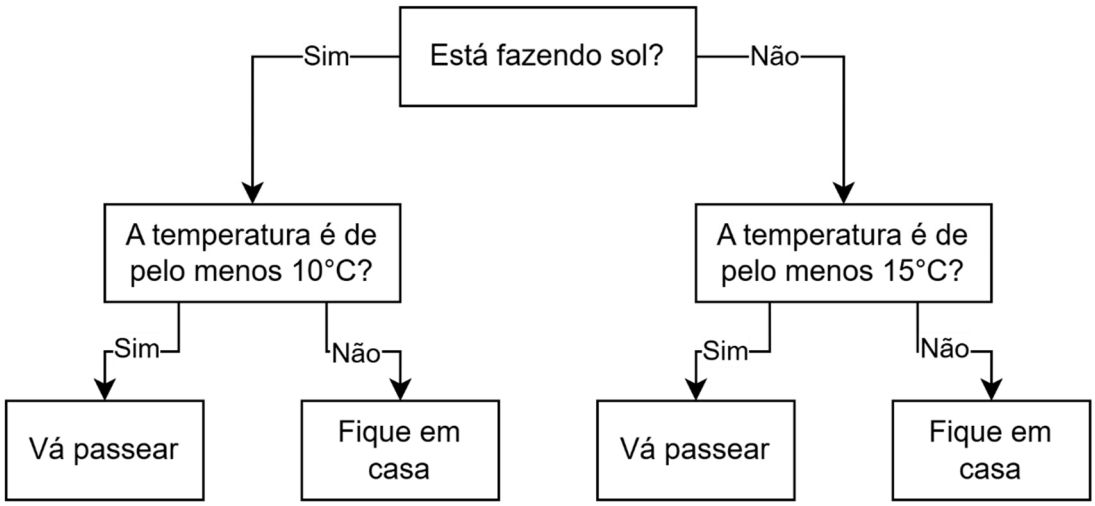
2.2 ÁRVORE DE DECISÃO

As árvores de decisão são ferramentas poderosas no mundo do aprendizado de máquina, utilizadas tanto para classificação quanto para regressão. Imagine-as como diagramas que simulam o processo de tomada de decisão, guiando-o por diferentes caminhos até alcançar o resultado. Tem como objetivo segmentar um conjunto de dados com base em características ou atributos, dividindo-o em grupos distintos ou categorias. Cada nó interno da árvore representa uma decisão baseada em uma condição sobre os atributos

dos dados, e cada folha da árvore corresponde a uma classe ou categoria de classificação. Em sua essência, uma árvore de decisão se assemelha a uma árvore invertida, crescendo a partir de um nó raiz e se ramificando em diversos outros nós. Cada nó representa uma decisão, e cada ramificação, um possível resultado dessa decisão (KLOSTERMAN, 2020).

A figura 2, nos mostra exatamente o funcionamento lógico de uma árvore se baseando nas decisões tomadas (KLOSTERMAN, 2020).

Figura 2 - Árvore de decisão de se é possível sair de acordo com o clima



Fonte: Bruno Arthur Basso Silva “adaptado de” ‘Projetos de Ciência de Dados com Python’, v.5. p – 207.

2.2.1 Índice de Gini

Podemos analisar essa árvore pelas variáveis presentes, porém quando se trata de decisões que a IA deve tomar, é imprescindível saber o posicionamento de cada pergunta e para isso, devemos usar um índice, mais conhecido como *Índice de Gini,* que verifica a impureza de cada variável em si a fim de entender a respectiva colocação (DANIYA, 2020).

O *Gini* consegue avaliar a impureza dos dados, ou seja, quanto menor a dispersão dos valores graficamente, mais puro ele é. Tendo isso em vista, temos que calcular a impureza da resposta tanto para “sim” quanto para “não.

$$1 - \left(\frac{atr1}{atr1 + atr2}\right)^2 - (\frac{atr2}{atr2 + atr1})^2 \quad (4)$$

- *atr1:* refere-se à primeira resposta se caso for “sim” ou “não”;
- *atr2:* refere-se à segunda resposta se caso for “sim” ou “não”.

Após esse cálculo, sabemos o grau de impureza para *sim* e *não* da variável escolhida para fazer a pergunta, porém não sabemos a impureza da variável em si. Para isso, basta fazer um outro cálculo usando as impurezas feitas anteriormente juntamente com o total de respostas dos atributos.

$$(\left(\frac{Tsim}{T}\right) * IMPsim) + (\left(\frac{Tnão}{T}\right) * IMPnão) \quad (5)$$

- $Tsim$: quantidade de respostas para *sim*;
- $Tnão$: quantidade de respostas para *não*;
- T: total de respostas para ambos;
- $IMPsim$: impureza para *sim*;
- $IMPnão$: impureza para *não*;

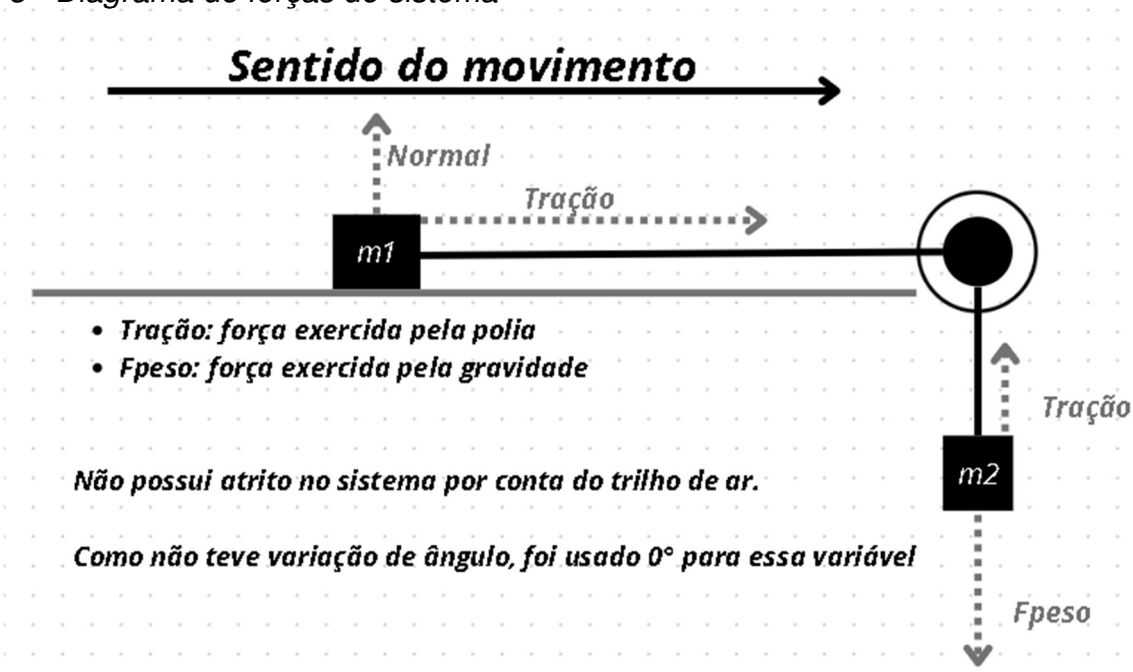
Somente neste momento, temos noção da impureza daquela certa variável. Se o índice for mais perto do zero, significa que ela é bastante pura, caso contrário, pode considerar impura.

A raiz segue o mesmo conceito apresentado, porque uma variável somente se torna raiz, se, dentre todas as possíveis variáveis, for a mais pura (DANIYA, 2020).

2.3 CINEMÁTICA E DINÂMICA DE DOIS CORPOS ACOPLADOS

Os tópicos de física básica abordados nesse projeto que são “cinemática” e “dinâmica” e está totalmente relacionado com os corpos do sistema conforme a figura 3 (HALLIDAY, 2016).

Figura 3 - Diagrama de forças do sistema



Fonte: Bruno Arthur Basso Silva

Neste caso especificamente, os dados extraídos são numéricos e classificados de tal forma:

m1: massa do corpo colocado em um plano horizontal sem atrito;
m2: massa do corpo suspenso;

Assim, a análise da parte teórica é essencial para visar o resultado dos dados experimentais, logo, calcular a aceleração teórica ajuda a compreender se está tentando um resultado próximo do medido.

Neste sentido, existe a necessidade de usar a 2° lei de newton para descobrir a aceleração teórica para comparar com a aceleração retirada do experimento.

Como trata-se da 2° lei de newton, também conhecida como princípio fundamental da dinâmica, é indispensável não mencionar a relação crucial que

essa lei estabelece, entre a força aplicada a um corpo, sua massa e aceleração, que tem como sua principal fórmula:

$$\vec{F} = m \times \vec{a} \tag{6}$$

$\vec{F}$: resultante que atua sobre o corpo, em N;
m: massa do corpo, em kg;
$\vec{a}$: aceleração do corpo, em m/s².

Depois da menção da lei de newton, é notório o uso desta no projeto, porém é preciso fazer algumas alterações por se tratar de dois corpos e sem a presença do atrito.

$$a = \frac{m_2 \times g}{m_1 + m_2} \tag{7}$$

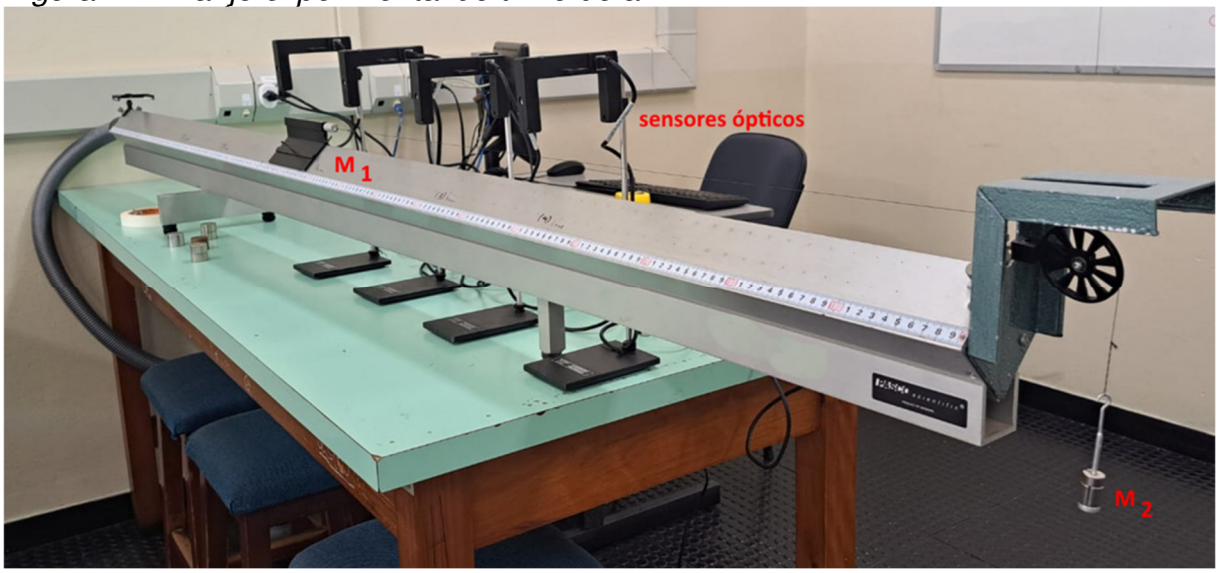
Com a aceleração devidamente isolada para fins de calculá-la, agora é possível fazer a comparação com a aceleração medida após o experimento.

3 METODOLOGIA

A primeira etapa para o desenvolvimento desse projeto consistiu na aquisição de dados para obter um banco de dados eficiente e confiável, com o intuito de treinar a IA. Conforme ilustrado na figura 4, foi utilizado um trilho de ar para a obtenção de dados a fim de minimizar as forças de atrito no movimento de dois corpos de massas diferentes e acoplados por um fio. Sensores ópticos posicionados estrategicamente permitiram a medida da velocidade de um dos corpos (m_1). Com esse arranjo experimental conseguimos determinar a aceleração do sistema a partir das medidas de velocidade em função do tempo.

O trilho pode ser horizontal ou inclinado de um ângulo. Nesta primeira etapa, mantivemos o trilho horizontal.

Figura 4 - Arranjo experimental do trilho de ar



Fonte: Bruno Arthur Basso Silva

Na tabela 1, está apresentada a variação de valores para a aquisição de dados referentes as massas utilizadas.

Tabela 1 - Variação das massas utilizadas no experimento

m_1	m_2
0.209	0.004 a 0.013
0.309	0.014 a 0.019
0.409	0.02 a 0.03

Fonte: Bruno Arthur Basso Silva

3.1 AQUISIÇÃO DOS DADOS

Para a aquisição de dados utilizamos o programa Data Studio, que foi uma ferramenta essencial para visualização e análise de dados quando se trata da obtenção de informações de velocidade e tempo por meio desse software específico. Durante o experimento, com o movimento de um objeto ao longo de

um trilho de ar, o aplicativo registra meticulosamente em intervalos regulares de tempo através de sensores com o intuito de melhorar a precisão dos resultados. Em seguida, as informações coletadas são armazenadas e processadas no banco de dados associado ao software. Para este trabalho, foram utilizados quatro sensores com distâncias simétricas (figura 4). Após obtermos cada dado, armazenamos em uma planilha no Excel.

3.1.1 Dados e definição dos parâmetros

Em relação aos dados de entrada, eles podem ser numéricos ou categóricos, e o algoritmo de classificação busca criar uma estrutura hierárquica de decisões que minimize o erro de classificação, permitindo que novas observações sejam classificadas automaticamente com base nas decisões tomadas nos nós da árvore.

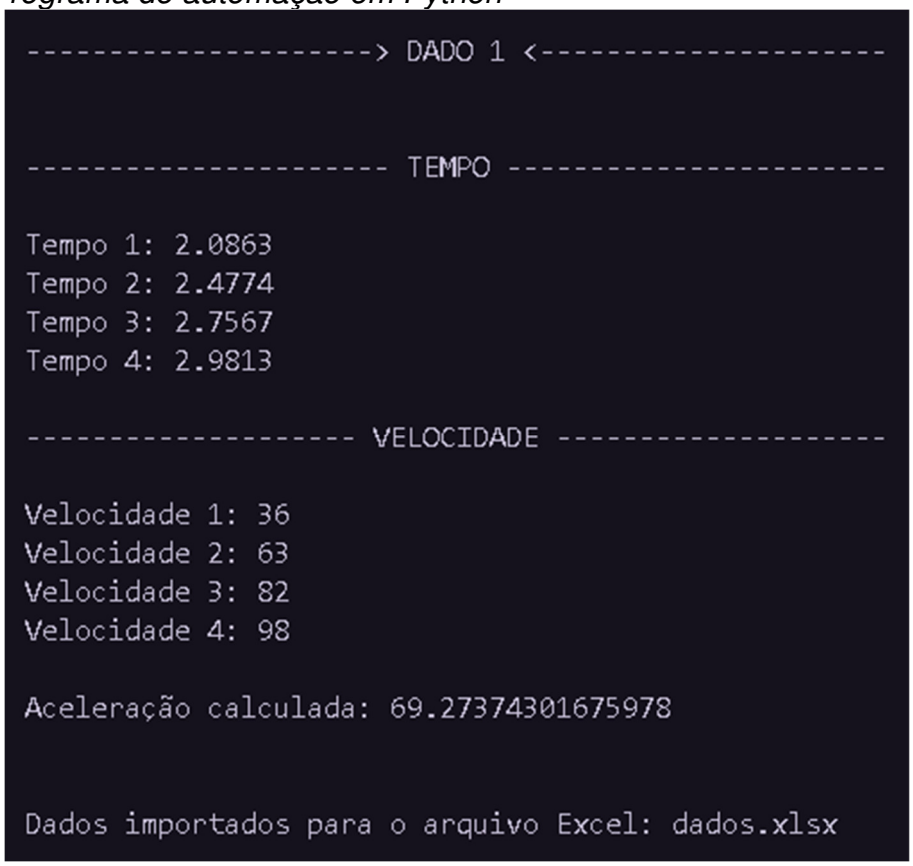
Quando se trata de Árvore, é necessário saber que os parâmetros são indispensáveis para sair o resultado esperado e entender quais serão utilizados. Depois de compreender melhor o uso dos parâmetros, foi escolhido as seguintes variáveis para melhor execução do algoritmo de classificação: a massa *m1* e aceleração extraída, pois o foco da IA consiste em fornecer o resultado de um grupo de massas *m2.* Não foi colocado o ângulo como parâmetro ainda pois precisamos fazer mais variação desse parâmetro para ser útil na árvore.

3.2 DESENVOLVIMENTO DE PROGRAMAS PARA A AQUISIÇÃO

Durante a aquisição, foi feito um programa para otimizar e facilitar o trabalho experimental, utilizando a linguagem Python e a bibliotecas Pandas para colocar os dados diretamente na planilha “dados.xlsx” no Excel. A figura 5 mostra o *input,* os valores de tempo e velocidade obtidos do Data Studio, e o *output,* que é a aceleração calculada. A aceleração calculada serve somente como referência para verificar a qualidade dos dados obtidos. A aceleração

experimental é obtida por meio da análise gráfica da velocidade em função do tempo.

Figura 5 - Programa de automação em Python



Fonte: Bruno Arthur Basso Silva

Após uma coleta razoável de dados, foi possível a criação do primeiro código teste para analisarmos o cluster e uma primeira versão da arvore de decisão.

3.3 UTILIZAÇÃO DE BIBLIOTECAS

Neste contexto, foi preciso utilizar algumas bibliotecas necessárias para o desenvolvimento da IA e parte de análise do projeto (PYTHON, 2001). São elas:

- Scikit-Learn: também conhecido como sklearn, é uma das bibliotecas mais populares em Python para aprendizado de máquina. Ela oferece uma ampla variedade de algoritmos de aprendizado de máquina, tanto para tarefas supervisionadas quanto não supervisionadas (SCIKIT-LEARN, 2007).

- Pandas: é amplamente utilizada para manipulação e análise de dados. Ela oferece estruturas de dados poderosas, como o Data Frame, que permite armazenar e manipular conjuntos de dados de forma eficiente. Além disso, é possível realizar operações de limpeza, transformação e análise estatística em dados tabulares de maneira intuitiva e eficaz (PANDAS, 2008).
- Matplotlib.pyplot: é uma sub-biblioteca do Matplotlib, que fornece uma interface para criar gráficos de alta qualidade em Python. Com ele, é possível criar uma ampla variedade de gráficos, como gráficos de linha, histogramas, gráficos de dispersão e muito mais (MATPLOTLIB, 1991).
- Seaborn: é uma biblioteca de visualização de dados que fornece uma interface de alto nível para criar gráficos estatísticos atraentes e informativos. E simplifica a criação de gráficos complexos, como diagramas de dispersão com regressão linear, box plots e mapas de calor (W3SCHOOL, 2012).
- SciPy: é uma biblioteca usada para computação científica e técnica. Ela é construída sobre o NumPy e fornece funcionalidades adicionais para operações matemáticas, otimização, processamento de sinais, álgebra linear, integração numérica, interpolação, estatísticas, entre outros. (SCIPY, 2008).

4 RESULTADOS OBTIDOS

Após a análise e estudos sobre o entendimento do projeto e os conceitos básicos apresentados, é essencial expor todos os resultados obtidos até então para entender os pontos positivos e negativos, visando a continuação do projeto.

4.1 DADOS ADQUIRIDOS

Os dados que foram obtidos estão no caminho planejado, pelo fato da aceleração experimental ser compatível com a aceleração teórica esperada. Isso significa que a aquisição dos dados está sendo realizada de forma adequada, assim fazendo com que o banco de dados para o treino da IA seja mais eficaz. Até o momento, tivemos a aquisição de mais de 250 dados para compor o banco de dados.

4.2 ANÁLISE DOS DENDROGRAMAS

Para realizarmos uma análise exploratória do banco de dados, foi criado um dendrograma para compreender melhor os intervalos de valores de massa e aceleração medida.

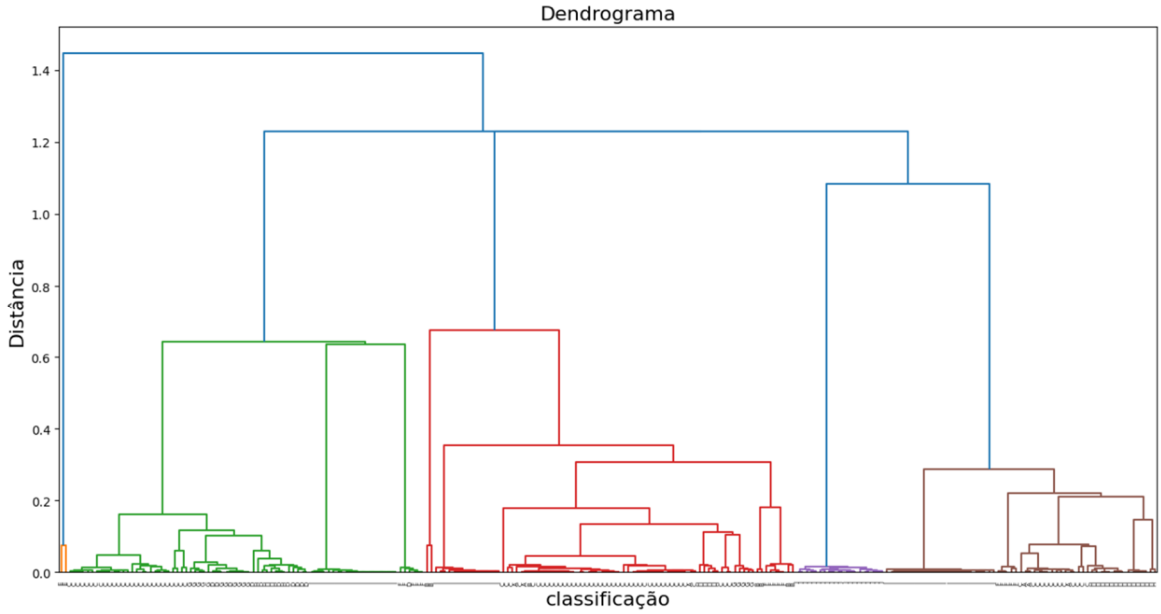
Um dendrograma é um tipo de diagrama que mostra como os dados estão agrupados ou divididos em *clusters*. Quando se tem uma grande quantidade de dados, ele ajuda a identificar padrões, agrupando os dados mais semelhantes entre si (PUIGBÒ, 2009).

Existem algumas maneiras de construir um dendrograma, usando diferentes métodos e métricas como explicados na seção 2.1. Ao analisá-lo, podemos identificar quais grupos de dados são mais coesos e quais são mais heterogêneos de acordo com as informações conhecidas sobre o experimento. Isso nos ajuda a entender a estrutura dos nossos dados e a tomada de decisão baseadas nessa compreensão. O objetivo neste momento é verificar se é possível agrupar as massas *m2, em função das massas m1 e do valor da aceleração.*

Portanto, na Figura 6 e 7, mostra exatamente dois métodos que foram eficazes na distribuição dos dados facilitando a visualização. Método *single linkage* com a métrica *euclidean*, e o método *ward* com a métrica *euclidean,* respectivamente. Foram testados vários outros métodos e outras métricas, porém esses foram o que melhor se ajustaram à nossa realidade.

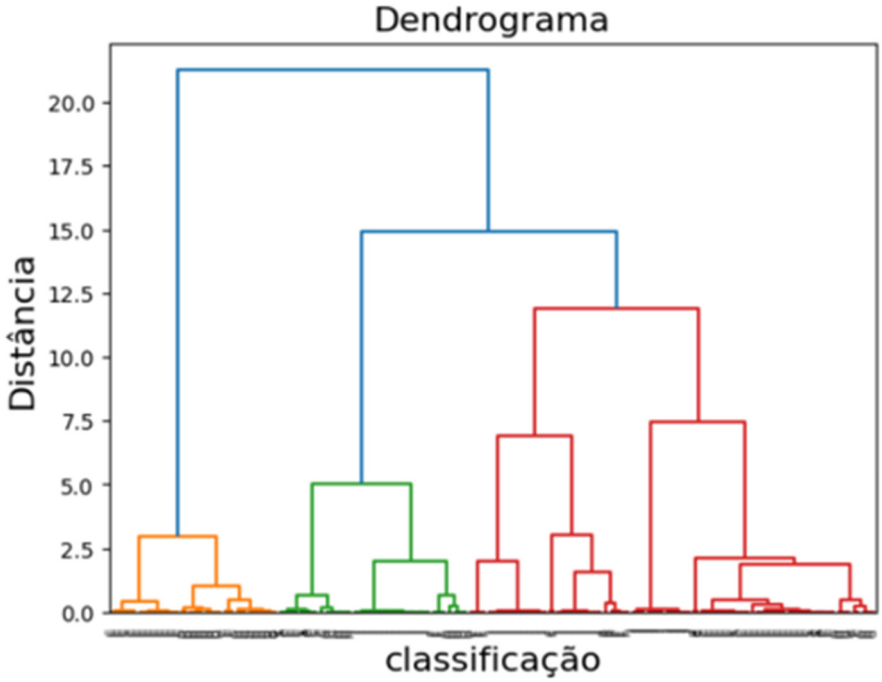
O eixo x se trata dos valores de *m2* do banco de dados, que foram classificados como letras de “A” a “J”, com o intuito de facilitar a visualização da análise, e o eixo y se refere aos valores dos cálculos da distância referente ao método escolhido. Devido ao grande número de dados não é possível observar de modo adequado as informações da classificação mostradas nas figuras. Essa classificação será melhor discutida mais adiante.

Figura 6 - Dendrograma do método single e métrica euclidean



Fonte: Bruno Arthur Basso Silva

Figura 7 - Dendrograma do método ward e métrica euclidean



Fonte: Bruno Arthur Basso Silva

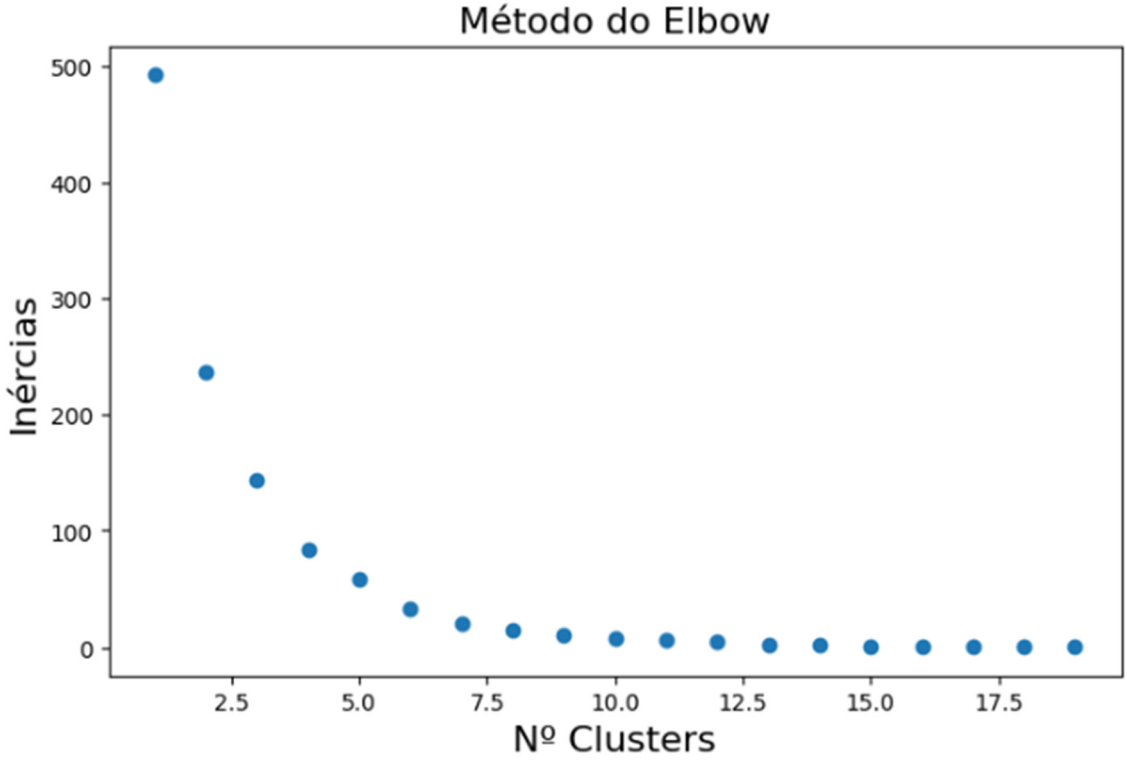
Para definir quantos clusters iremos usar na classificação para agrupar as observações é usual utilizar o método Elbow.

4.3 MÉTODO DE ELBOW

Este método consiste em ajudar os cientistas de dados a selecionarem o número ideal de clusters, ajustado por um ponto de inflexão da curva, também conhecido como “cotovelo” (YELLOWBRICK, 2016).

Na figura 8, podemos visualizar a mudança na taxa de variação da inercia exatamente em 5 clusters, ou seja, utilizaremos 5 agrupamentos para analisar os dados.

Figura 8 - Gráfico do método de Elbow



Fonte: Bruno Arthur Basso Silva

4.4 FORMAÇÃO DOS CLUSTERS

Na tabela 2, é apresentada a quantidade de cada massa m2, classificadas de A à J, em cada um dos 5 clusters.

Tabela 2 - Quantidade de cluster (y) pela quantidade de grupos (x)

m2 (kg)	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Grupo	0.309	0.309	0.309	0.209	0.209	0.209	0.409	0.409	0.409	0.409
1°	3	7	43	0	0	5	5	5	15	0
2°	3	0	13	0	0	5	0	15	25	0
3°	0	0	26	12	0	5	11	5	20	0
4°	0	0	0	0	2	0	0	0	0	0
5°	0	0	0	0	0	0	0	0	0	20

Fonte: Bruno Arthur Basso Silva

É notório que os clusters estão totalmente mal distribuídos. Por exemplo, somente o grupo 1 contêm quase todas as massas m2 utilizadas. Nossa expectativa era que os clusters fossem capazes de identificar de modo mais adequado as massas m2.

Assim, para analisarmos de uma forma diferente a variabilidade dos dados obtidos, partimos para uma análise estatística usando box plot, como será apresentado a seguir.

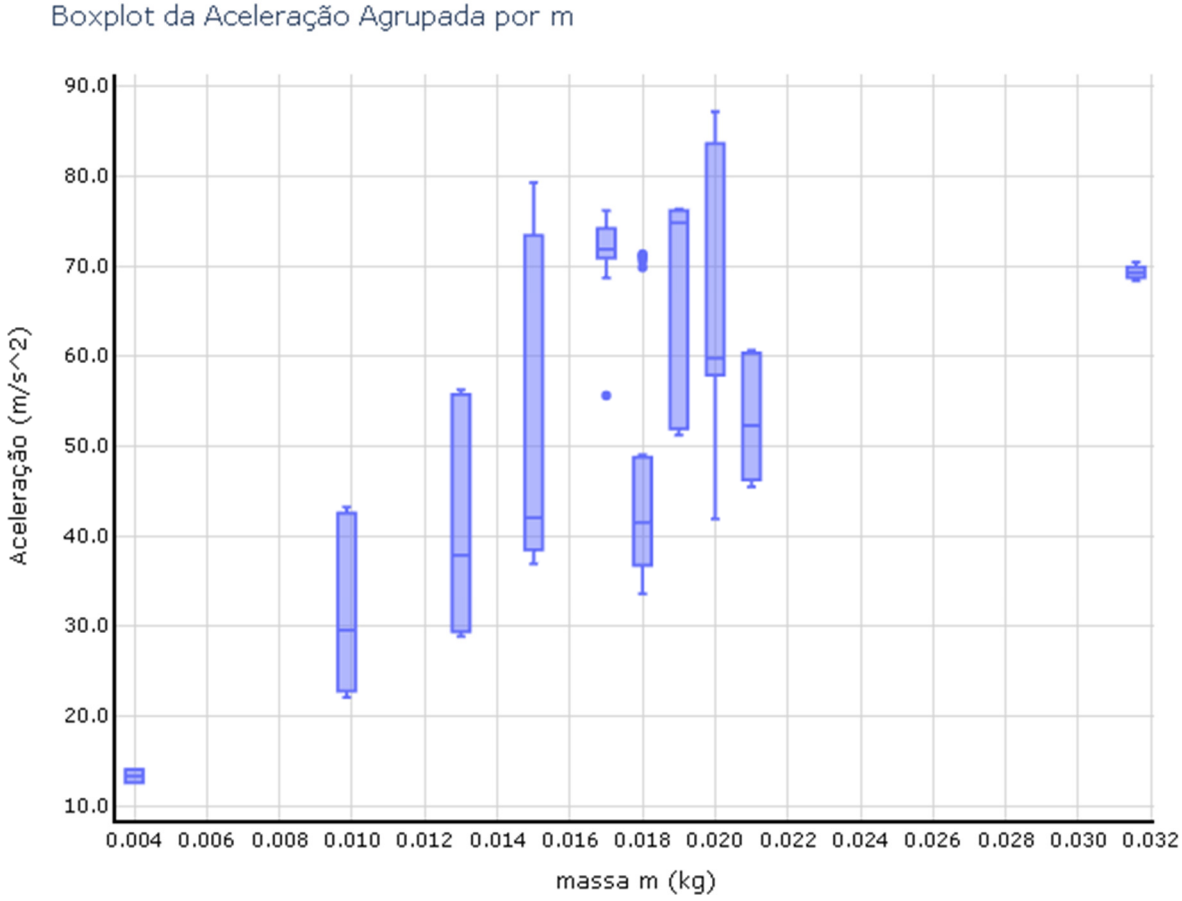
4.5 ANÁLISE DADOS USANDO UM BOX PLOT

O Box Plot, ou Diagrama de Caixa, é uma ferramenta gráfica poderosa para visualizar e analisar conjuntos de dados. Através de uma caixa, bigodes e linha mediana, ele apresenta a forma da distribuição, tendência central, dispersão e presença de outliers de forma compacta e informativa.

Os bigodes, também chamados de *whiskers*, são linhas que complementam a caixa mostrando a amplitude e a presença de outliers (pontos fora dos bigodes, representados por círculos ou asteriscos e indicam valores atípicos que podem influenciar a análise) na distribuição dos dados complementando a análise realizada através da caixa e da mediana (indica simetria na distribuição, se estiver no centro, está simétrica, caso contrário, assimétrica) no Box Plot (SEABORN, 2012).

A figura 9 representa o Box Plot dos dados obridos em questão e nos mostra a distribuição, amplitude e simetria dos dados em diversos pontos e variações do m2.

Figura 9 - Box Plot da aceleração em função do m2



Fonte: Bruno Arthur Basso Silva

Observando a Figura 9, percebe-se que a grande dispersão dos valores de aceleração para as massas m2 entre os valores de 0,010 kg até 0,022 kg. Essa variabilidade dificultou o correto processo de agrupamento. Como consequência, isso pode diminuir a acurácia na resposta da IA por se tratar de valores muito próximos.

Assim, será necessário escolher somente um valor de m2 nesse intervalo a fim de concentrar as informações sobre a aceleração.

5 CONCLUSÃO PARCIAL E PRÓXIMAS ETAPAS

A aquisição de dados experimentais mostrou-se desafiadora, mas de extrema importância para um bom treinamento da árvore de decisão. Assim, nosso próximo passo é otimizar os valores de massa m2 que serão utilizados para a aquisição dos dados. Além disso, novos dados serão obtidos com o trilho de ar inclinado para termos mais um parâmetro a ser considerado na análise.

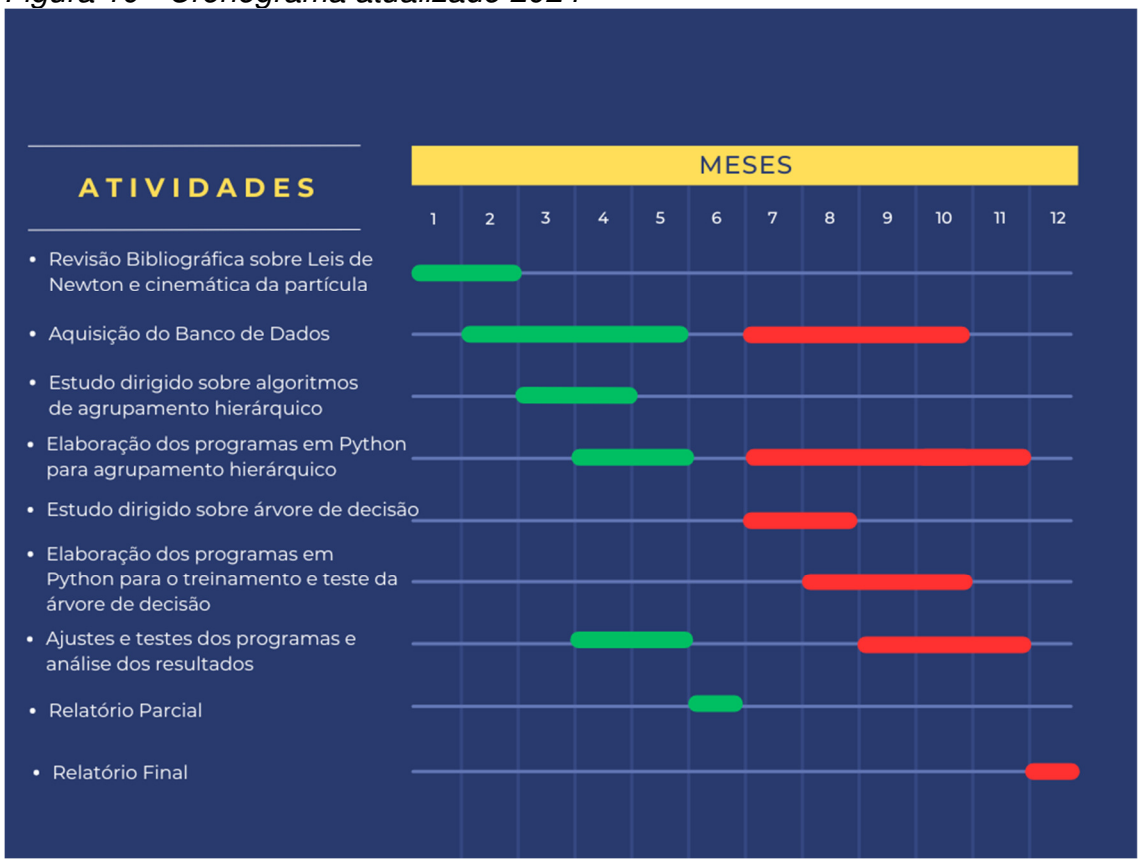
Espera-se assim melhorar a qualidade do banco de dados de tal forma que variabilidade dos dados seja refletida nos processos de agrupamento e consequentemente no treinamento da árvore de decisão. Sabe-se que a obtenção de dados reais é desafiadora, porém essencial para o treinamento de uma IA.

Assim, até a finalização desse projeto espera-se obter uma IA treinada para que em uma aula do laboratório de física 1 os alunos entrem com os valores de massa m1, inclinação do plano e aceleração experimental e a IA retorne o valor da massa m2, com uma certa acurácia.

6 CRONOGRAMA

A figura 10 apresenta o cronograma das próximas etapas.

Figura 10 - Cronograma atualizado 2024



Fonte: Bruno Arthur Basso Silva

E o foco maior será nos seguintes tópicos no período de maio - novembro:

- Aquisição do Banco de Dados;
- Elaboração dos programas em Python para o treinamento e teste da árvore de decisão;
- Ajustes e testes dos programas e análise de resultados.

REFERÊNCIAS

DANIYA, T; GEETHA, M; SURESH KUMAR, K. **CLASSIFICATION AND REGRESSION TREES WITH GINI INDEX,** 2020. disponível em: https://www.researchgate.net/profile/suresh-kumar-k-dr/publication/344385674_classification_and_regression_trees_with_gini_index/links/5f780b4d92851c14bca9e8a5/classification-and-regression-trees-with-gini-index.pdf. Acesso em: 20/06/2024.

FÁVERO. **MANUAL DE ANÁLISE DE DADOS ESTATÍSTICA E MODELAGEM MULTIVARIADA COM EXCEL, SPSS E STATA**, v.9.1 – v.9.2, p. 300 – 327, 2022. Acesso em: 08/06/2024.

HALLIDAY, David; RESNICK, Robert; WALKER, Jearl. **FUNDAMENTOS DE FÍSICA,**
VOLUME 1: MECÂNICA. 10. ed. Rio de Janeiro: LTC - Livros Técnicos e Científicos Editora LTDA, 2016. v. 1. Acesso em: 25/05/2024.

IBM**, AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the difference?,** 2023. disponível em: https://www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks/. Acesso em: 20/05/2024

KLOSTERMAN**,** Stephen. **PROJETOS DE CIÊNCIA DE DADOS COM PYTHON.** v.5. p – 207, 2020. Acesso em: 08/06/2024.

MATPLOTLIB. **MATPLOTLIB.PYPLOT**. 1991. Disponível em: https://matplotlib.org/3.5.3/api/_as_gen/matplotlib.pyplot.html. Acesso em: 07/06/2024.

MEI/CNI, Abenge. **DIRETRIZES PARA O CURSO DE ENGENHARIA**. 2018. Disponível em:

http://www.abenge.org.br/documentos/propostadcnabengemei_cni.pdf. Acesso em: 07/03/2024.

MITCHELL, Tom M. **MACHINE LEARNING.** v.3 - Decision Tree Learning. 1997. Disponível: https://www.cin.ufpe.br/~cavmj/Machine%20-%20Learning%20-%20Tom%20Mitchell.pdf. Acesso em: 01/02/2024.

PANDAS. PANDAS DOCUMENTATION. 2008. Disponível em: https://pandas.pydata.org/docs/. Acesso em: 07/06/2024.

PUIGBÒ, Pere. **DENDROUPGMA: A DENDROGRAM CONSTRUCTION UTILITY.** 2009. Disponível em: https://usuaris.tinet.cat/debb/UPGMA/DendroUPGMA_Tut.pdf. Acesso em: 20/06/2024.

PYTHON. **PYTHON 3.12.4 DOCUMENTATION**. 2001. Disponível em: https://docs.python.org/3/. Acesso em: 14/05/2024.

SCIKIT-LEARN. **SCIKIT-LEARN.ORG**. 2007. Disponível em: https://scikit-learn.org/stable/about.html#history. Acesso em: 07/06/2024.

SCIPY, api. **SCIPY.CLUSTER.HIERARCHY.LINKAGE.** docs.scipy, 2008. Disponível em: https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html. Acesso em: 06/06/2024.

SCIPY, api. **SCIPY DOCUMENTATION**. 2008. Disponível em: https://docs.scipy.org/doc/scipy/. Acesso em: 07/06/2024.

SEABORN. **SEABORN.BOXPLOT.** 2012. Disponível em: https://seaborn.pydata.org/generated/seaborn.boxplot.html. Acesso em: 08/06/2024.

W3SCHOOL. SEABORN. 2012. Disponível em: https://www.w3schools.com/python/numpy/numpy_random_seaborn.asp. Acesso em: 07/06/2024.

YELLOWBRICK. **ELBOW METHOD.** 2016. Disponível em: https://www.scikit-yb.org/en/latest/api/cluster/elbow.html. Acesso em: 08/06/2024.