TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**FINAL REPORT**

*Người hướng dẫn*: **GV LÊ ANH CƯỜNG**

*Người thực hiện*: **NGUYỄN KHẮC HUY – 521H0502**

**GIANG HOẰNG ĐẠT – 521H0498**

**PHẠM PHÚ BÌNH – 521H0495**

Lớp **: 21H50301**

Khoá  **: 25**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**FINAL REPORT**

*Người hướng dẫn*: **GV LÊ ANH CƯỜNG**

*Người thực hiện*: **NGUYỄN KHẮC HUY**

**GIANG HOẰNG ĐẠT**

**PHẠM PHÚ BÌNH**

Lớp **: 21H50301**

Khoá  **: 25**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023**

LỜI CẢM ƠN

To complete this mid-term project, we would like to express our gratitude to Ton Duc Thang University for providing the necessary facilities and to the teachers who have supported us throughout our studies at the university. Moreover, we love to give thanks to our parents who gave us chances to approach knowledge at TDTU and motivated, supported us a lot.

Especially, we would like to thank Mr. Le Anh Cuong for teaching us with great dedication and detail, so that we have enough knowledge to use for this essay. Due to our limited experience and knowledge, we are sure that there are some mistakes in our work. we sincerely hope to receive feedbacks and constructive criticism from our teacher so that we can complete this essay more effectively.

We would like to express my heartfelt thanks and wish teacher good health.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Chúng tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của TS Lê Anh Cường;. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 23 tháng 12 năm 2023*

*Tác giả*

Nguyễn Khắc Huy

Phạm Phú Bình

Giang Hoằng Đạt

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

TÓM TẮT

This essay will envolves four chapters about The goal of creating a machine learning model, the methods or algorithms for machine learning models and the criteria of learning, the problems and pros and cons of each models.

Firstly, the chapter one demonstrate about the pusposes of creating a machine learning model.

In the second chapter, it illustrates about several methods or algorithms to implement or “teach” the computer how to learn from already dataset in order to do it function. Moreover, in this chapter also has the criteria about the learning. As we know learning something new that we have to have some rules to ensure that the computer will learn correctly. There is the disscussion about four models: kNN, Linear Regression, Naïve Bayes classifiers and the last is Decision Tree. In detail, in this chapter we analyzed each model, we compare them in the cirteria that we have mentioned in chapter one.

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc154257175)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii](#_Toc154257176)

[TÓM TẮT v](#_Toc154257177)

[MỤC LỤC 1](#_Toc154257178)

[CHƯƠNG 1 – EXPLORATORY DATA ANALYSIS AND FEATURE ROLE ASSESSMENT 4](#_Toc154257179)

[1.1 Statistical analysis on data: 4](#_Toc154257180)

[1.2 Evaluate specific roles of this problem: 9](#_Toc154257181)

[CHƯƠNG 2 – APPLICATION OF BASIC AND ENSEMBLE MACHINE LEARNING MODELS 12](#_Toc154257182)

[2.1 Machine Learning Models: 12](#_Toc154257183)

[2.1.1 Application of Machine Learning Models for this problem: 13](#_Toc154257184)

[2.1.2 Ensemble Machine Learning Models: 18](#_Toc154257185)

[CHƯƠNG 3 – UTILIZING FEED FORWARD AND RECURRENT NEURAL NETWORKS 19](#_Toc154257186)

[3.1 Feedforward Neural Network (FNN): 20](#_Toc154257187)

[**3.2 Recurrent Neural Network (RNN):** 20](#_Toc154257188)

[CHƯƠNG 4 – OVERFITTING AND METHODS TO ADDRESS THIS PROBLEM 21](#_Toc154257189)

[4.1 What is the Overfitting ? 21](#_Toc154257190)

[4.2 How to fix the Overfitting in machine learning ? 22](#_Toc154257191)

[4.2.1 Feature selection: 22](#_Toc154257192)

[4.2.2 Use Regularization ( for Neural Network ) : 22](#_Toc154257193)

[4.2.3 Drop out ( for Neural Network ) : 23](#_Toc154257194)

[CHƯƠNG 5 –ENHANCING MACHINE LEARNING ACCURACY 23](#_Toc154257195)

[5.1 Methods for enhancing accuracy : 24](#_Toc154257196)

[**5.1.1 Hyperparameter Tuning:** 24](#_Toc154257197)

[5.1.2 Data Augmentation: 24](#_Toc154257198)

[5.1.3 Deep Learning Architectures: 24](#_Toc154257199)

[5.1.4 Regularization: 25](#_Toc154257200)

[5.1.5 Ensemble Models: 25](#_Toc154257201)

[5.1.6 Check and Handle Noisy Data: 25](#_Toc154257202)

[5.1.7 Model Architecture Changes: 25](#_Toc154257203)

[5.1.8 Optimize Model Size: 26](#_Toc154257204)

[**5.2** **Analysis and evaluation the problem:** 26](#_Toc154257205)

[5.2.1 Model Selection and Tuning: 26](#_Toc154257206)

[5.2.3 Data Preprocessing: 27](#_Toc154257207)

[5.2.4 Cross-Validation: 28](#_Toc154257208)

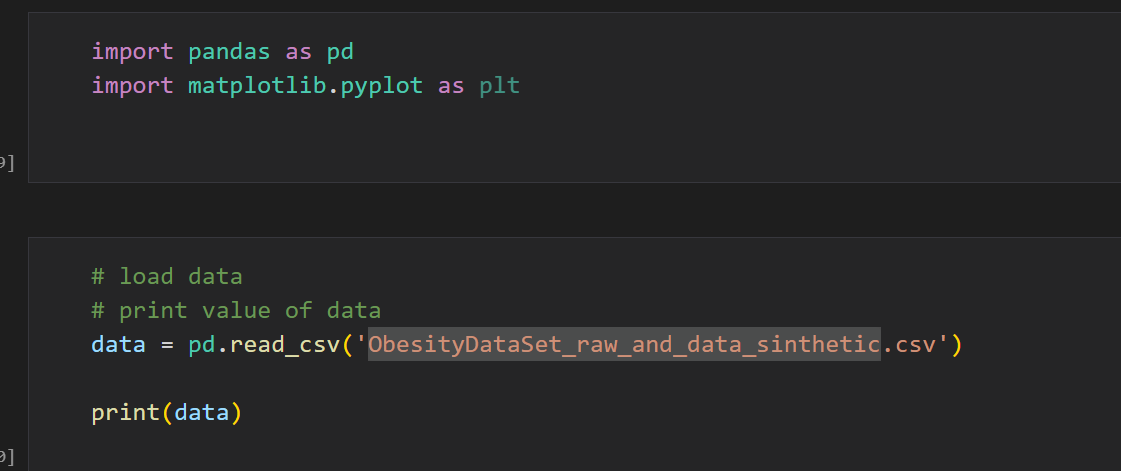
[**5.2.5 Evaluation Metrics:** 28](#_Toc154257209)

[TÀI LIỆU THAM KHẢO 29](#_Toc154257210)

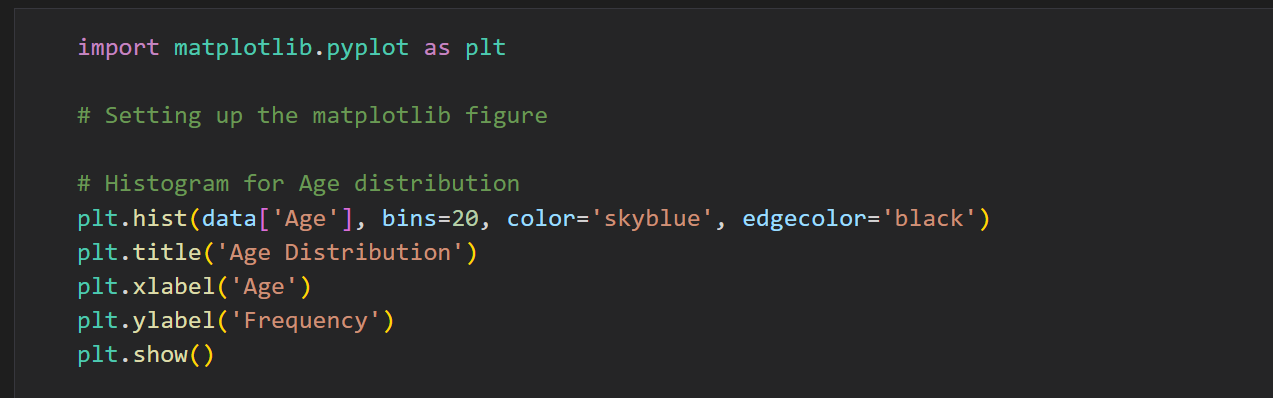
[PHỤ LỤC 30](#_Toc154257211)

CHƯƠNG 1 – EXPLORATORY DATA ANALYSIS AND FEATURE ROLE ASSESSMENT

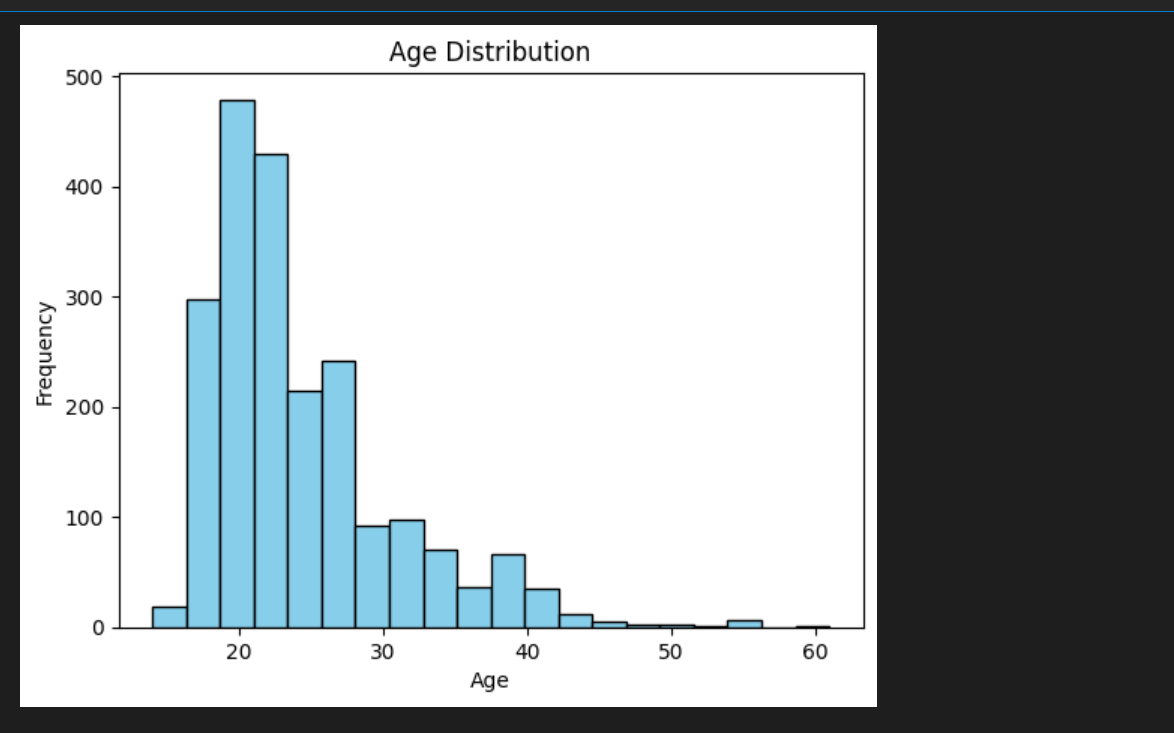
* 1. Statistical analysis on data:
* Importing dataset :
  + Features include various types: categorical and numerical.
  + The data should remain untouched, not utilized for learning, classroom exercises, or homework in machine learning.



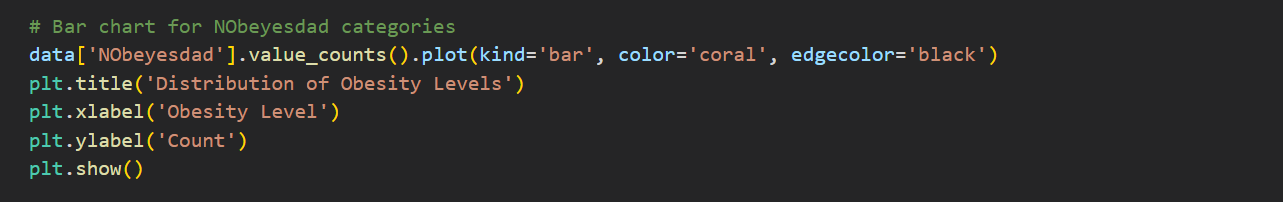
* **Visualizes the age distribution in the dataset using a histogram.**



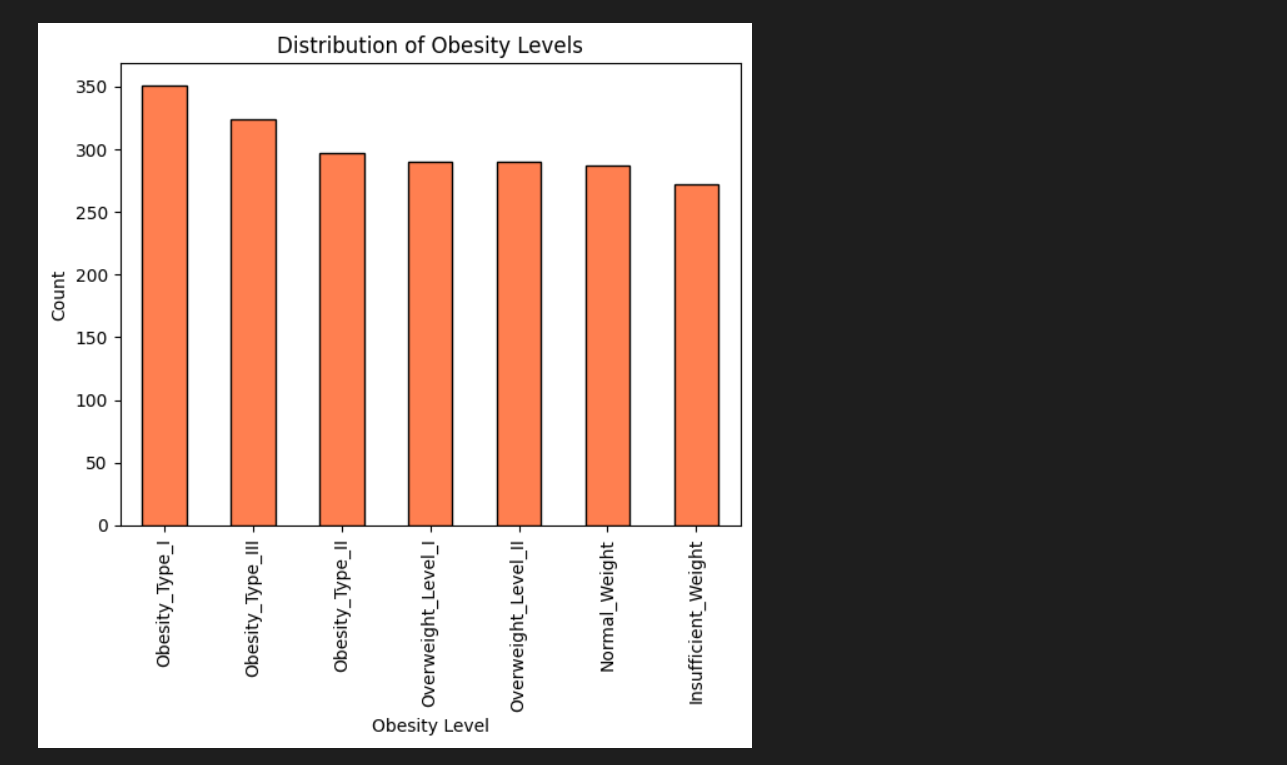
* **Histogram for Age Distribution.**



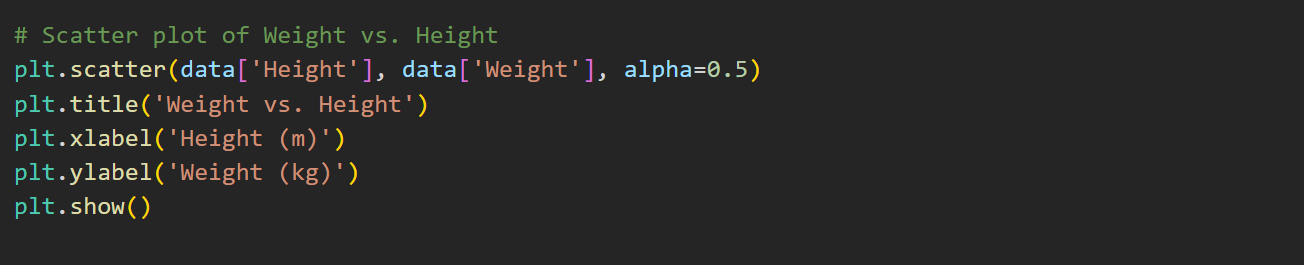
* **Represents the distribution of obesity levels in the dataset using a bar chart.**



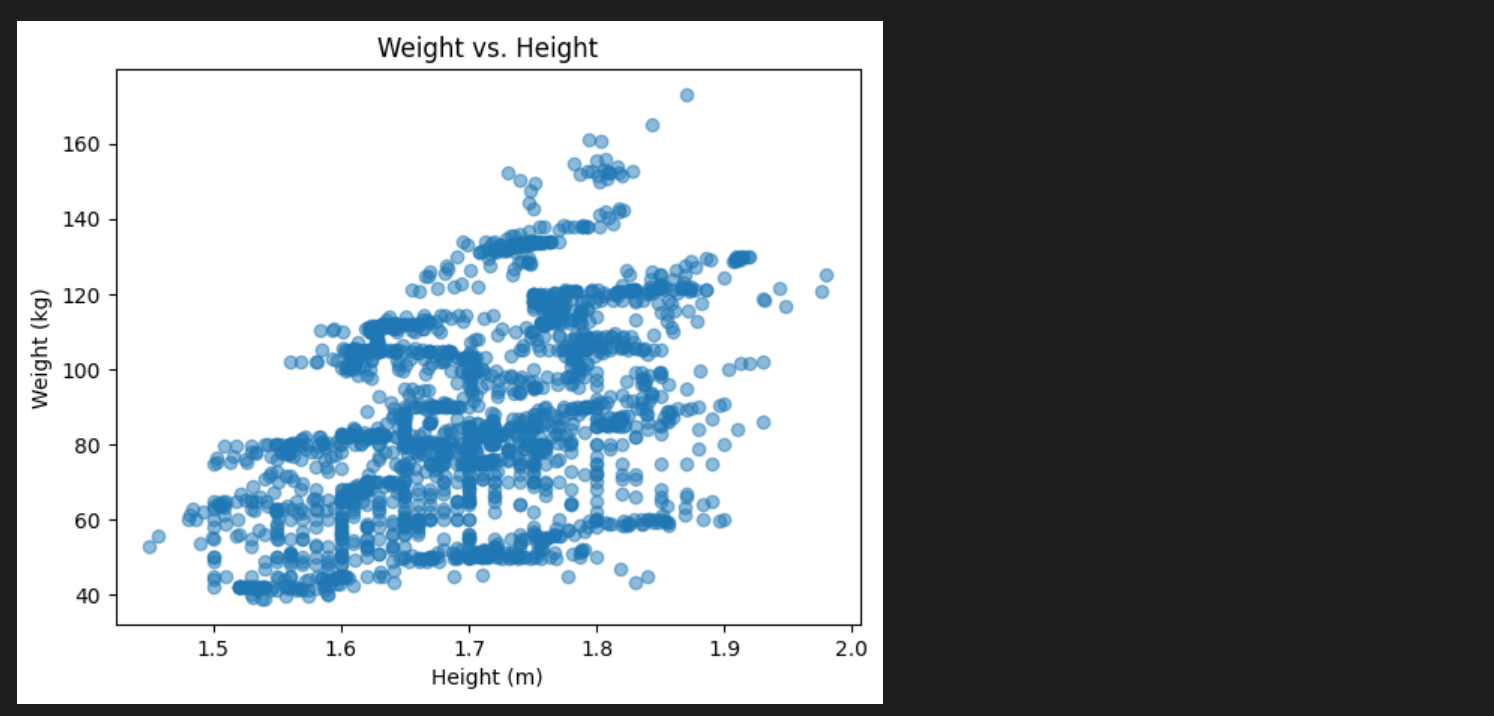
* **Histogram for Nobeyesdad categories.**



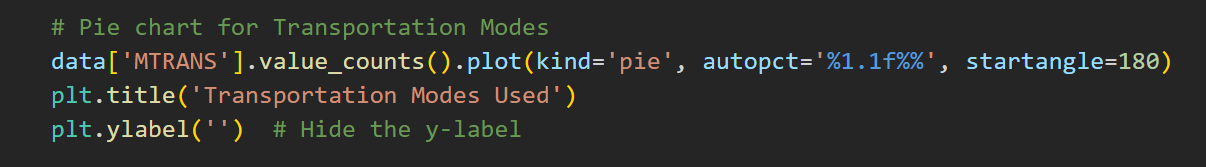
* **Represents the relationship between 'Height' and 'Weight' using a scatter plot.**



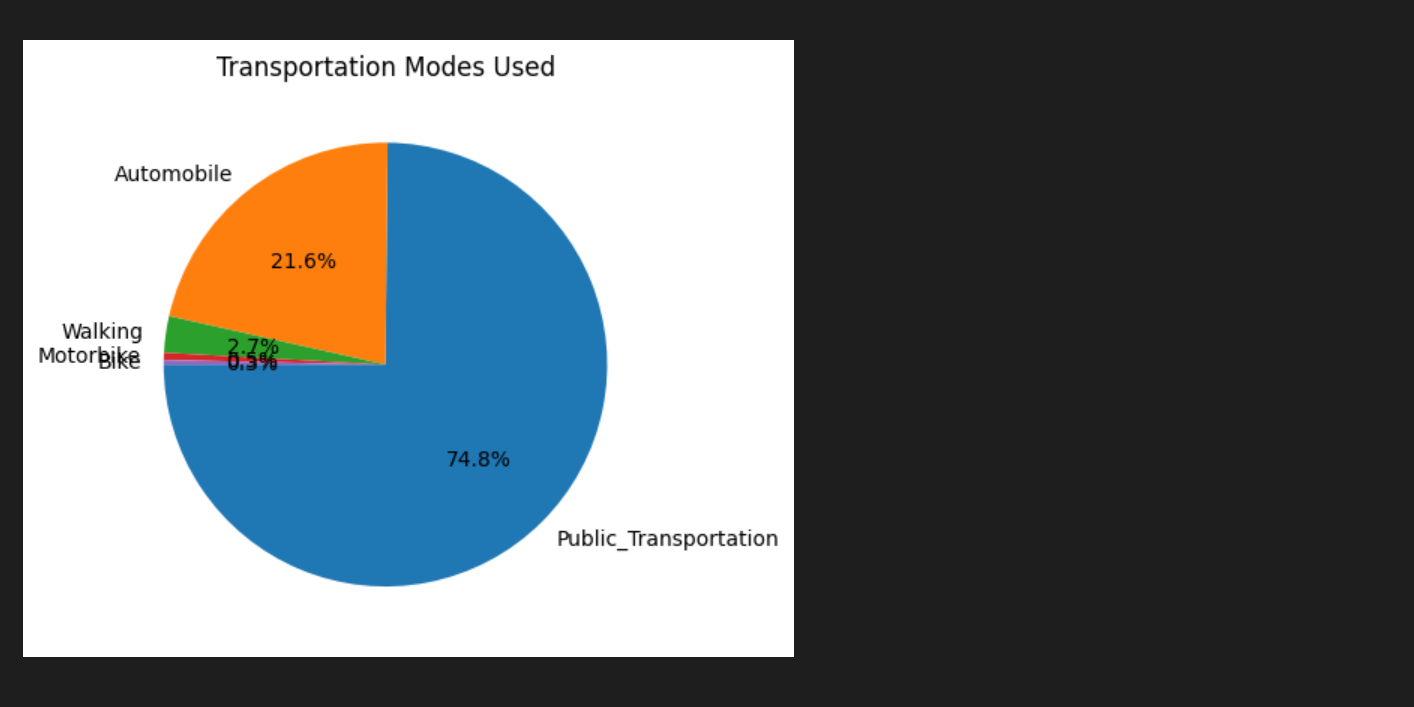
* **Scatter plot of Weight and Height.**



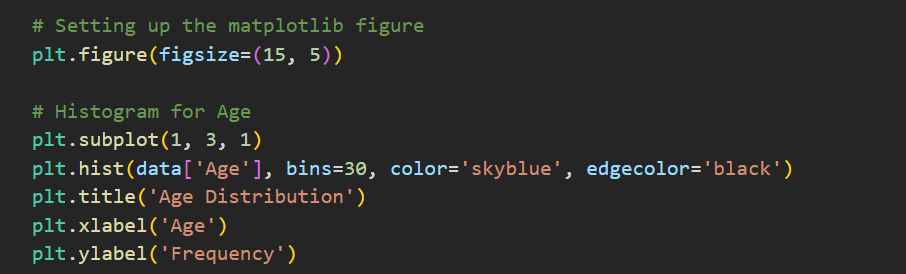
* **Represents the distribution of transportation modes in the dataset using a pie chart.**

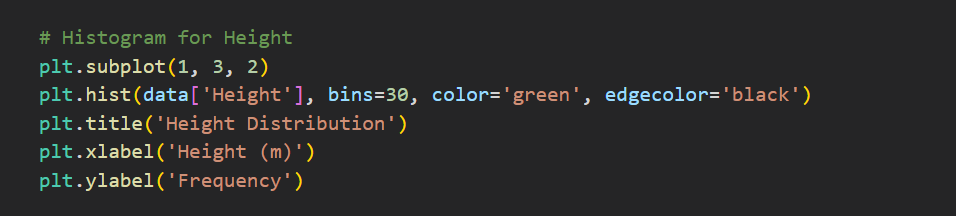


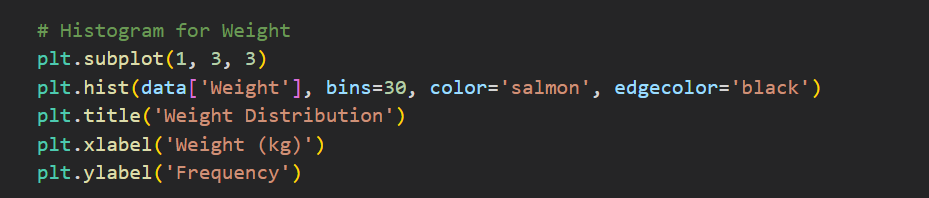
* **Pie Chart for Transpotations Modes.**



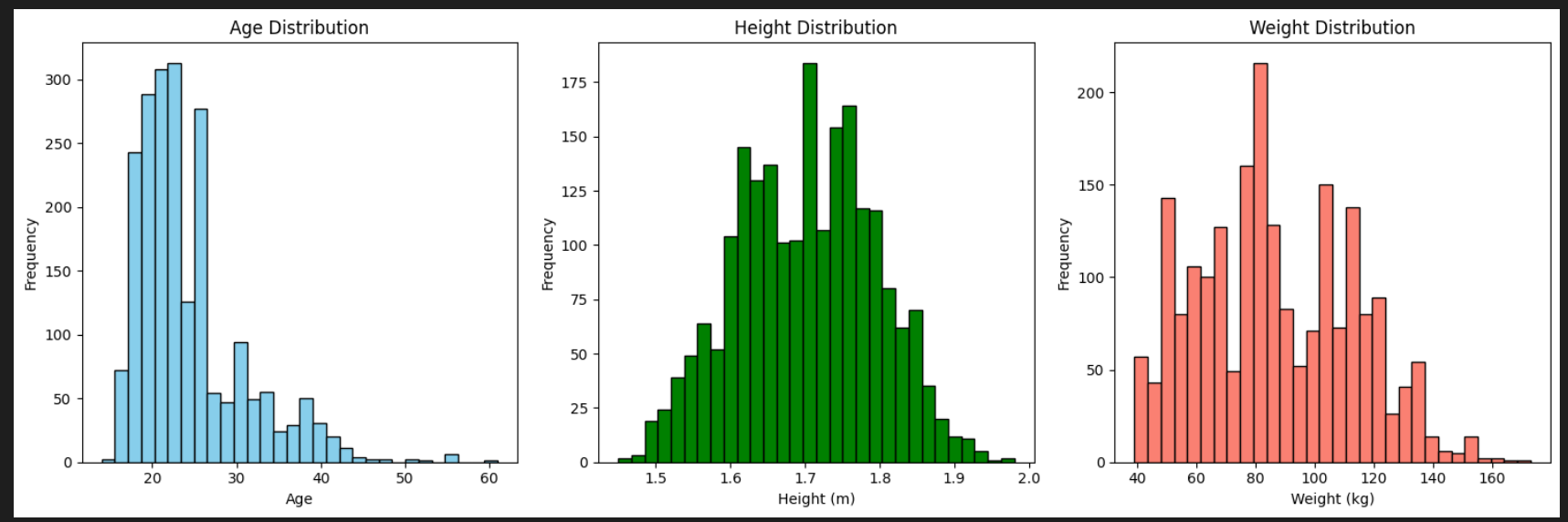
* **configures a matplotlib and larger visualization setup, and additional subplots to the figure.**



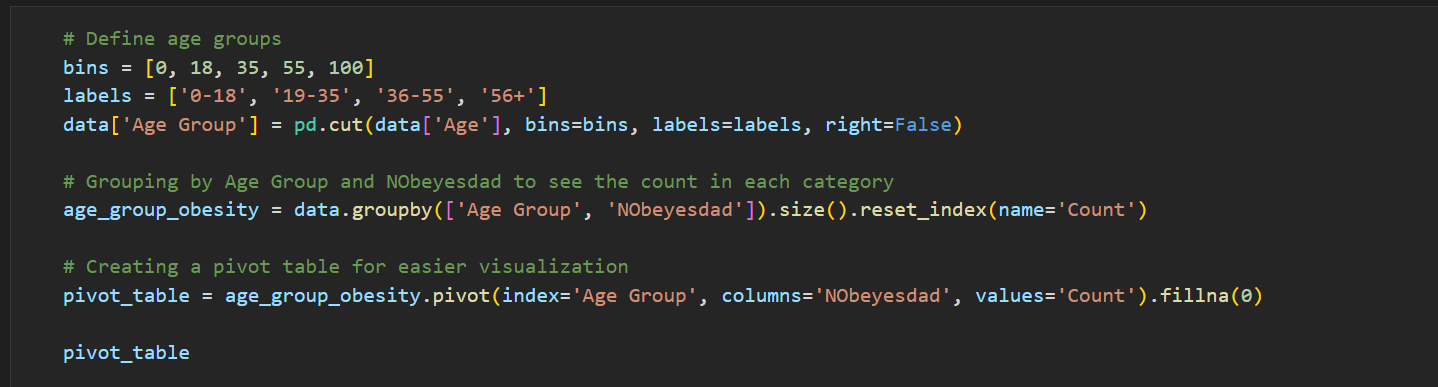




* **Histogram for each attributes.**



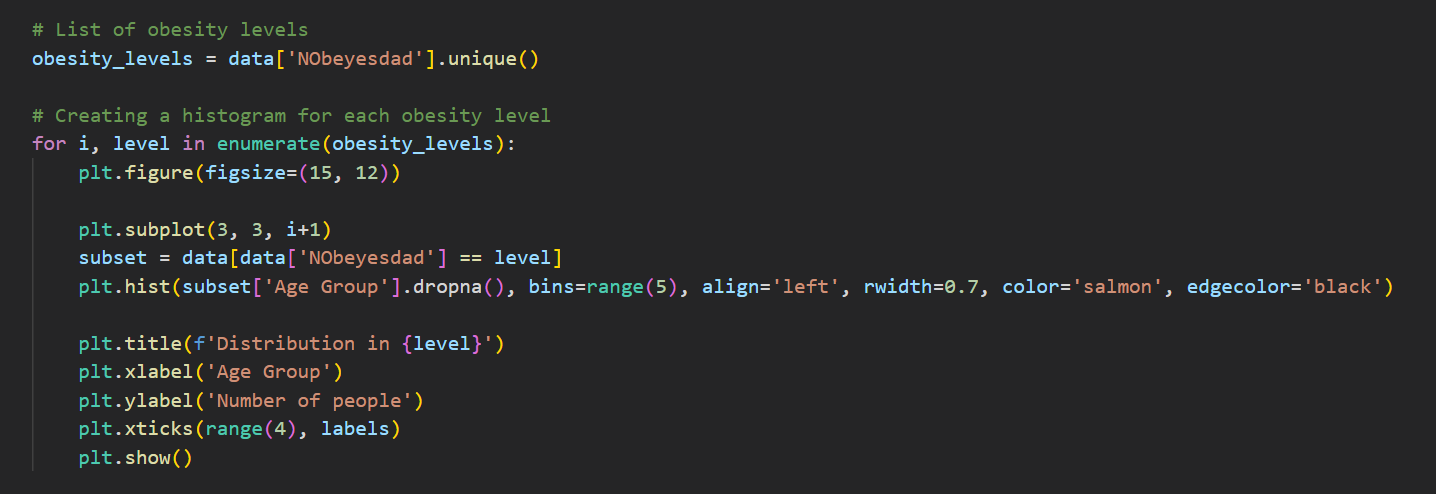
* 1. Evaluate specific roles of this problem:
* Examining obesity levels across various age groups by organizing the data and generating a concise pivot table for visualization and interpretation



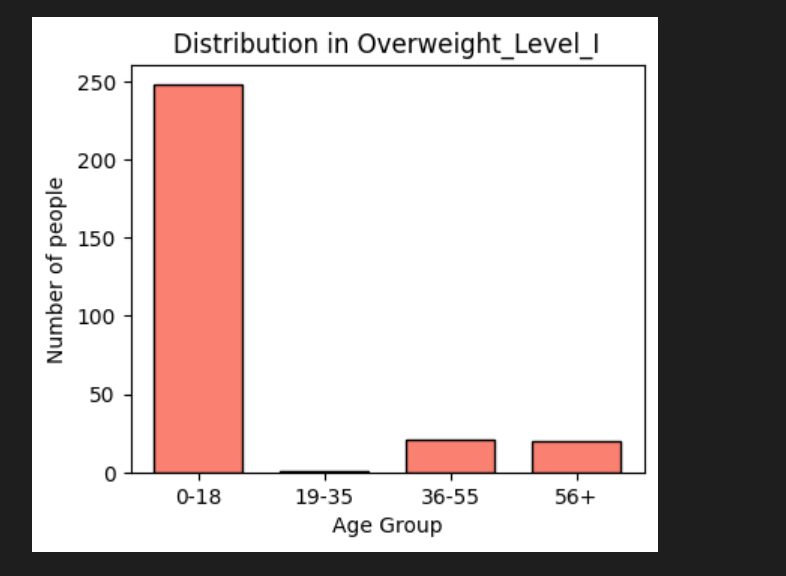
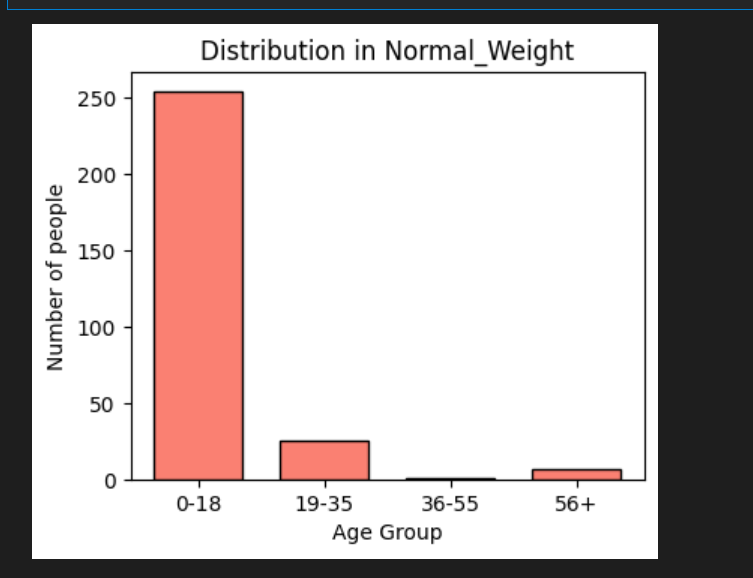
* Pivot table

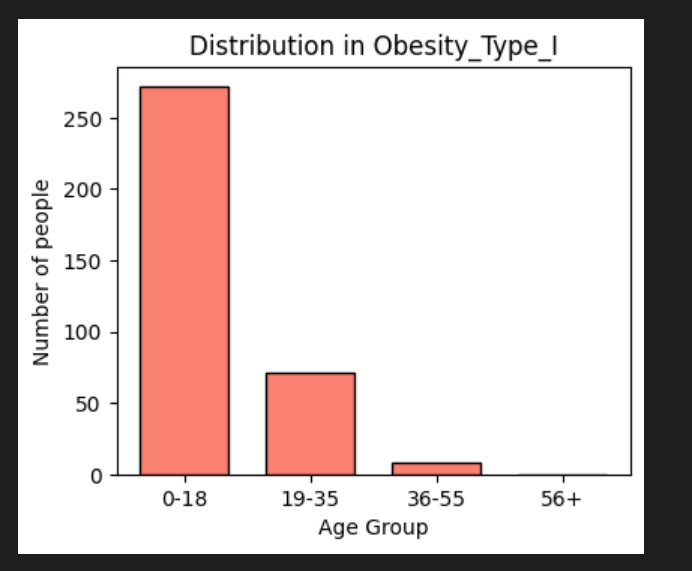
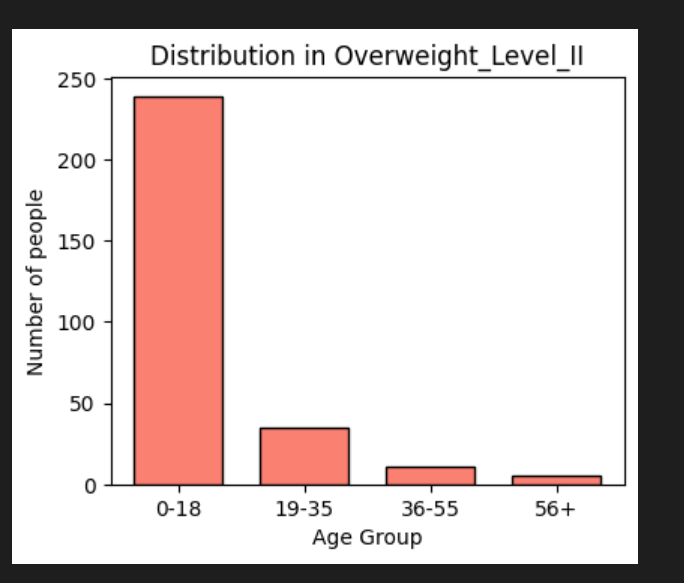


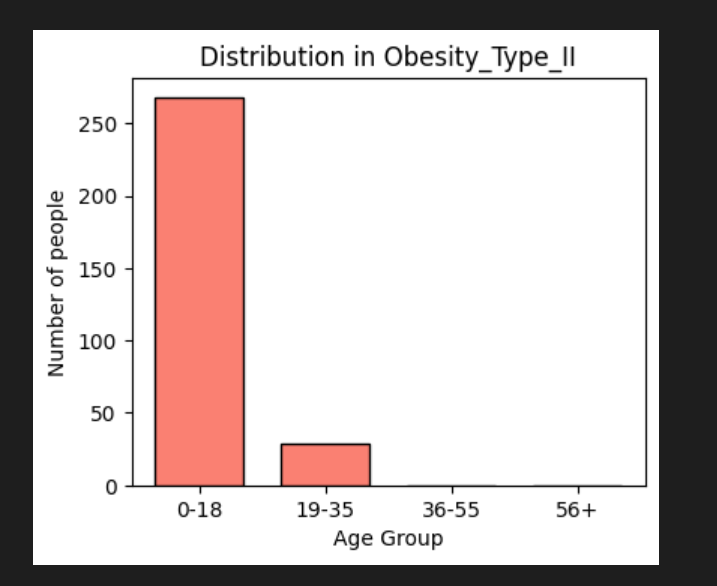
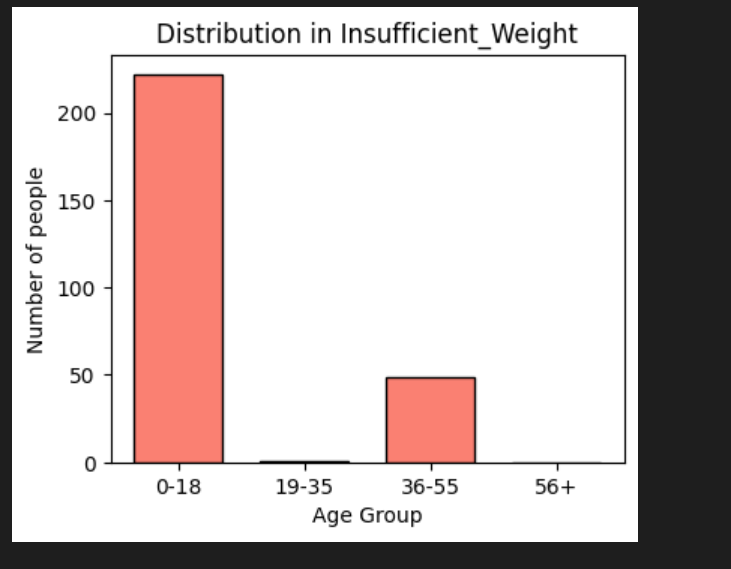
* provide a visual exploration of the distribution of age groups within each obesity level in each histograms.

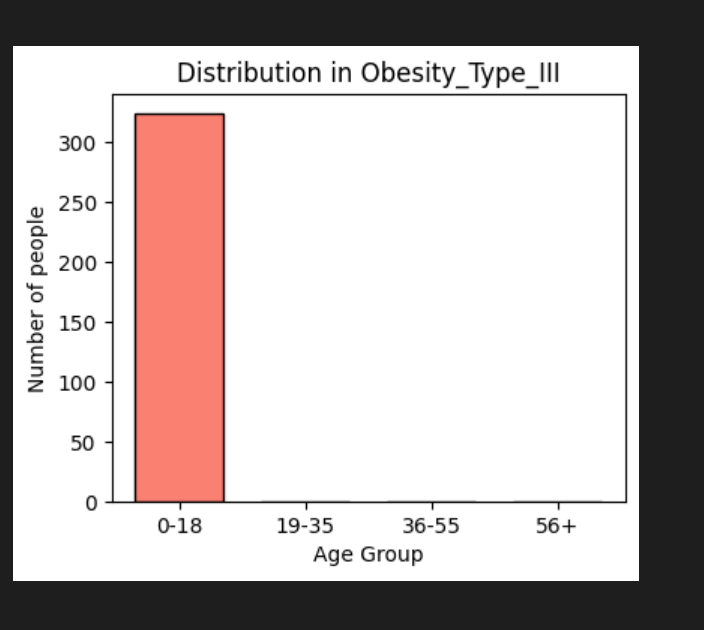


* Histograms of each obesity level









CHƯƠNG 2 – APPLICATION OF BASIC AND ENSEMBLE MACHINE LEARNING MODELS

Machine learning, a dynamic field at the intersection of computer science and statistics, leverages algorithms and models to enable systems to learn and make predictions or decisions without explicit programming. In this situation, two prominent categories of models, Basic Machine Learning Models and Ensemble Machine Learning Models, play pivotal roles in solving a myriad of problems.

2.1 Machine Learning Models:

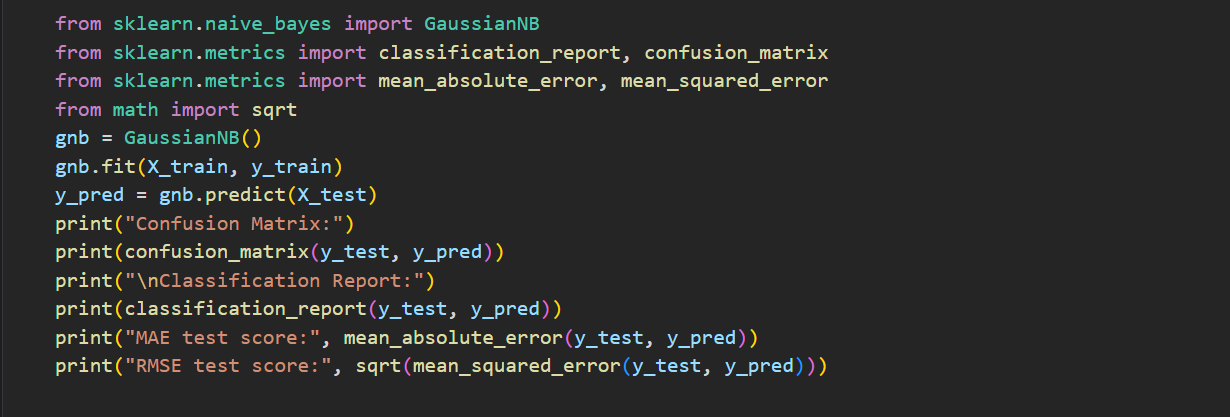
2.1.1 Application of Machine Learning Models for this problem:

Basic machine learning models are individual models that are relatively simple in structure and can be used independently for making predictions. For example, decision trees, logistic regression, K-nearest neighbors (KNN), Support Vector Machines (SVM), linear regression, and so on. Base models are often the building blocks of ensemble models.

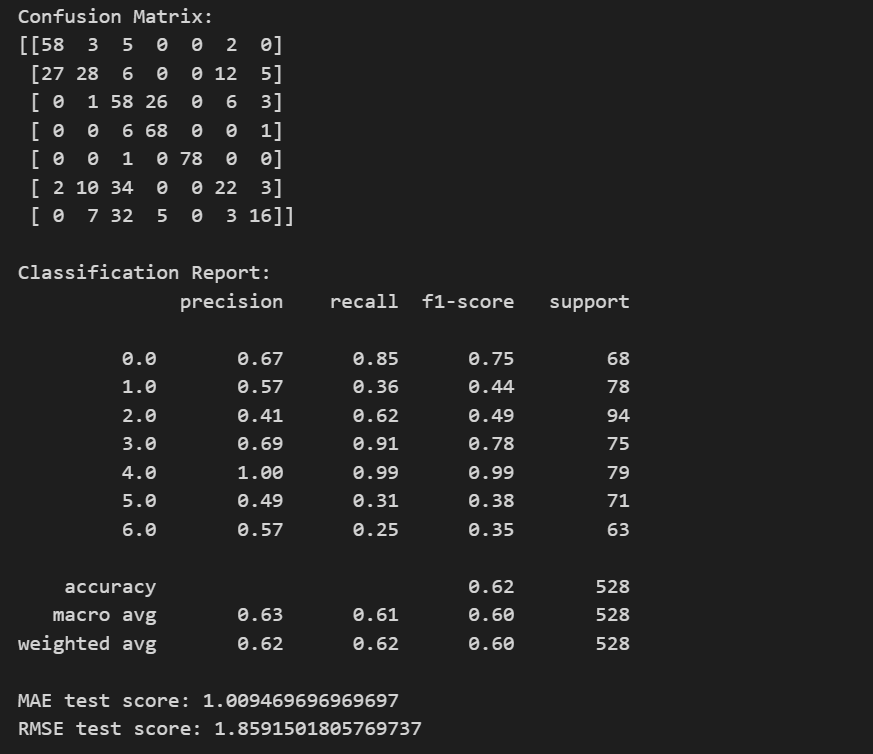
2.1.1.1 Gaussian Naive Bayes (GaussianNB):

Gaussian Naive Bayes is a model belonging to the Naive Bayes family, primarily used for classification problems. It assumes that features are independent and follow a Gaussian distribution (or a uniform distribution).

* Advantages:
* It's simple , requiring minimal training data, and is particularly effective for text classification.
* Disadvantages:
* It assumes feature independence, potentially limiting its performance in real-world scenarios, struggles with complex relationships due to its simplicity, and can be sensitive to outliers, impacting its effectiveness.
* **Solve this problem using GaussianNB.**



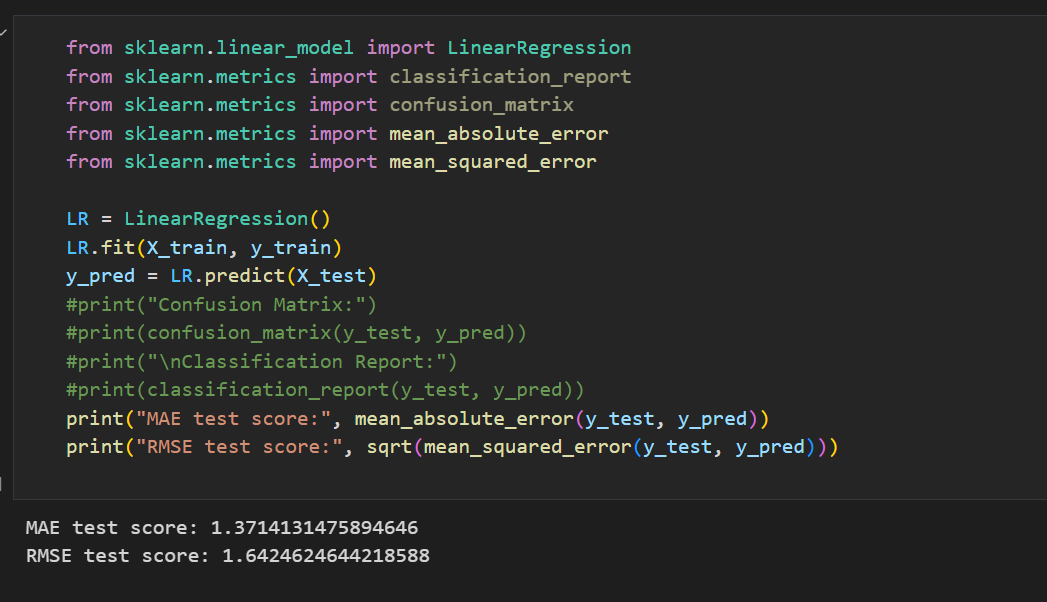
* The result of GaussianNB:



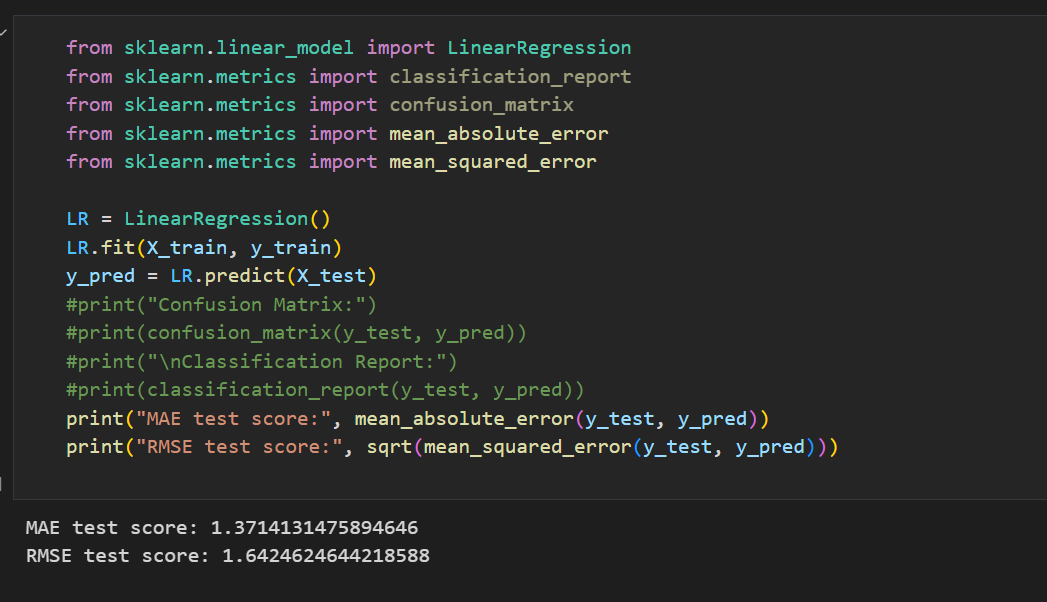
2.1.1.2 Linear Regression:

Linear regression is used when the dependent variable is continuous and the independent variables are either continuous or discrete. It assumes a linear relationship between dependent and independent variables.

* Advantages:
* It's simple, fast, and efficient for small datasets or large datasets with fewer features.
* Disadvantages:
* It assumes a linear relationship between dependent and independent variables, which may not always hold true. It's sensitive to outliers
* **Solve this problem using Linear Regression.**



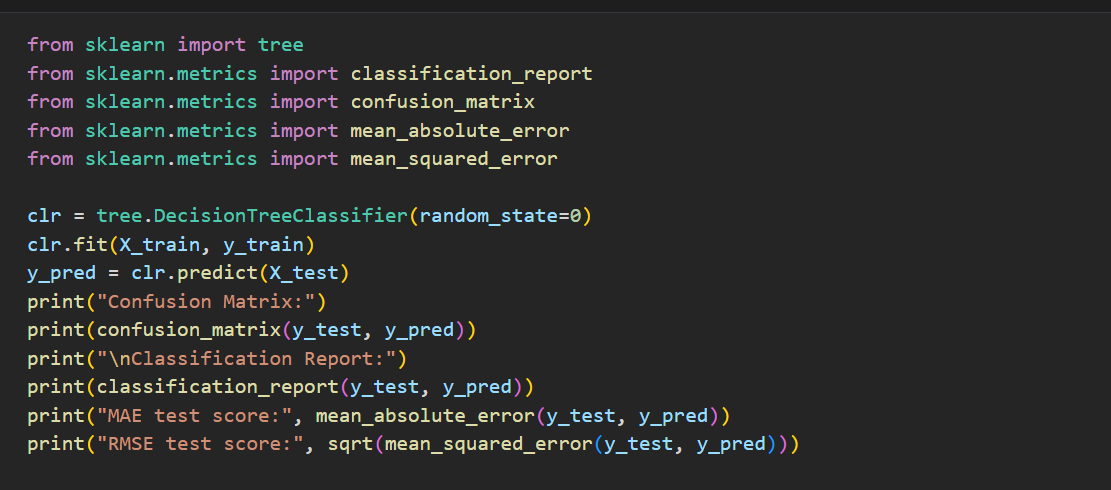
* The result of Linear Regression:



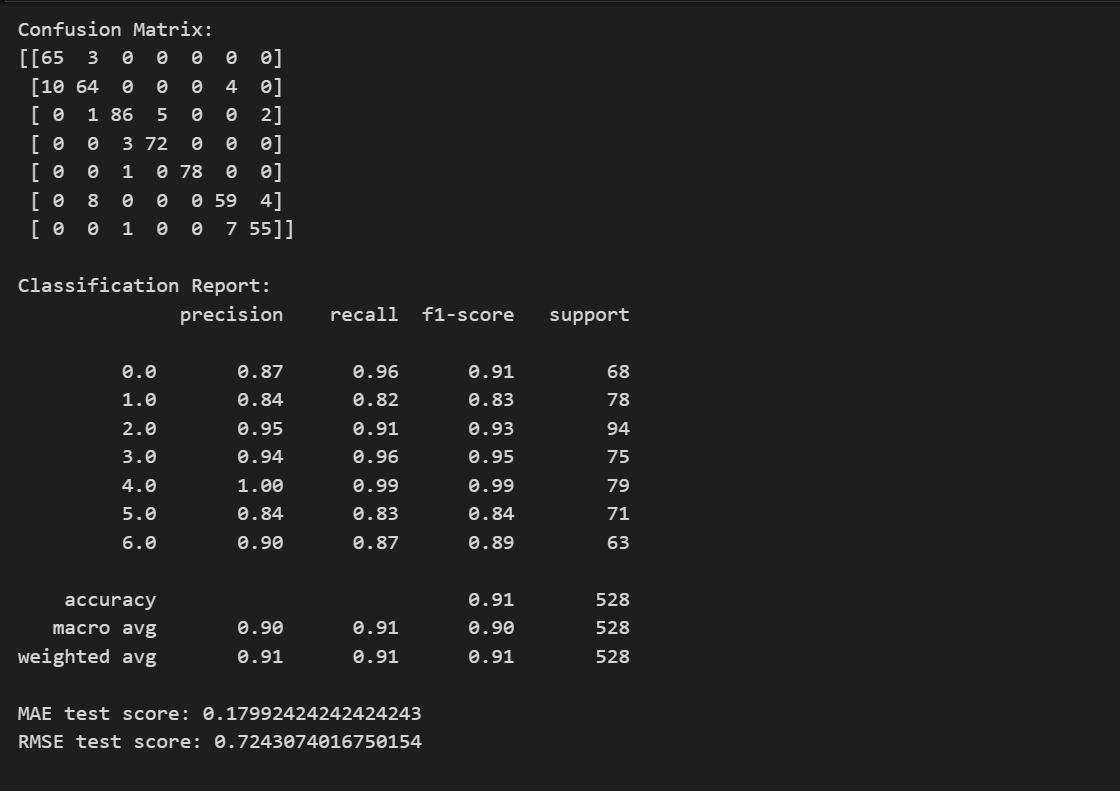
2.1.1.3 DecisionTreeClassifier:

The DecisionTreeClassifier is a model used for both classification and regression tasks. This model creates a decision tree by splitting the data based on features to achieve the best classification or prediction. Each leaf node of the tree represents a class (in classification) or a value (in regression).

* Advantages:
* It's Interpretable and versatile with numerical and categorical data,
* Disadvantages:
* Prone to overfitting, instability, and less effective for complex relationships.
* **Solve this problem using Linear Regression.**

****

* **The result of DecisionTreeClassifier:**

****

2.1.2 Ensemble Machine Learning Models:

Ensemble machine learning models involve the combination of multiple individual models to create a more robust and accurate predictive model.

An ensemble learning model is a machine learning model combinates the predictions of multiple basic machine learning models simultaneously to improve predictive performance compared to using a single model. This approach leverages the diversity and strength of multiple models to create an aggregate model with enhanced capabilities.

* Popular Ensemble Learning techniques:
* Bagging (Bootstrap Aggregating): Utilizes multiple independently trained models on different subsets of the training data. The results are combined by taking the average (regression problems) or using majority voting (classification problems).
* Boosting: Builds multiple models sequentially, with each model attempting to correct the errors of the previous one. Predictions from the models are combined with different weights.
* Stacking (Meta-Ensemble): Uses predictions from base models as input for a second-stage "meta-model," often a linear model, to generate the final predictions.
* Advantages:
* Diversity of Models: Ensemble learning combines the strengths of diverse models like GaussianNB, DecisionTreeClassifier, and Linear Regression, leveraging their unique characteristics.
* Improved Predictive Performance: Ensemble methods often enhance overall accuracy by mitigating the weaknesses of individual models through combination.
* Versatility: Can handle various types of data and capture different types of patterns due to the inclusion of models with different assumptions and structures.
* Disadvantages:
* Increased Complexity: Ensemble models can be more complex, making them harder to interpret and explain compared to individual models.
* Computational Intensity: Training and combining multiple models can be computationally expensive, especially with large datasets.
* Overfitting Risk: If not properly tuned, ensemble models can still be susceptible to overfitting, particularly if the base models are overfitting.

CHƯƠNG 3 – UTILIZING FEED FORWARD AND RECURRENT NEURAL NETWORKS

3.1 Feedforward Neural Network (FNN):

FNN is a type of artificial neural network where information flows in one direction—from input to output—without cycles or loops. It excels at tasks with structured data but struggles with sequences and temporal dependencies.

* Advantages:
* Universal Approximator: Capable of representing any continuous function.
* Parallel Processing: Allows for faster execution with parallelized training and inference.
* Structured Data: Effective for tasks involving structured data.
* Disadvantages:
* Limited Sequential Handling: Struggles with capturing information from sequential or temporal data.
* Fixed Input Size: Requires a fixed-size input, limiting suitability for variable-length sequences.
* Sensitivity to Noise: Can be sensitive to noisy input data, potentially leading to overfitting.

**3.2 Recurrent Neural Network (RNN):**

RNN is a type of artificial neural network designed to handle sequential data by maintaining hidden states that capture information from past inputs. It is well-suited for tasks involving variable-length sequences and temporal dependencies but faces challenges with vanishing gradients and computational complexity.

* **Advantages:**
* Sequential Handling: Excels at sequential data by maintaining a hidden state.
* Variable Input Length: Suitable for variable-length sequences in tasks like natural language processing.
* Memory Mechanism: Can learn and utilize memory, considering context over sequences.
* **Disadvantages:**
* Gradient Issues: Faces challenges with vanishing or exploding gradients, impacting long-range dependencies.
* Computational Intensity: Training can be computationally intensive, requiring specialized architectures.
* Limited Parallelization: Sequential nature hinders efficient parallelization, limiting speed on certain hardware.

CHƯƠNG 4 – OVERFITTING AND METHODS TO ADDRESS THIS PROBLEM

4.1 What is the Overfitting ?

When a model learns the training data too thoroughly, overfitting is a common issue in machine learning has occured, capturing noise and random fluctuations rather than the underlying pattern. An overfit model therefore performs well on training data but poorly on fresh, untried data. Because the main objective of machine learning is to create models that generalize to new data, overfitting is a serious problem.

## 4.2 How to fix the Overfitting in machine learning ?

Addressing overfitting is a crucial aspect of machine learning model development, as it ensures that the model can make accurate predictions on new, unseen data. And we have the sample code to fix this problem which is attached with this report. There are some way to fix that issue:

### 4.2.1 Feature selection:

Feature selection is crucial for improving the performance of machine learning models and enhancing their generalization ability. Here are some key applications:

* **Speeding up Training:** Reducing the number of features accelerates the training process.
* **Preventing Overfitting:** Selecting important features helps prevent overfitting.
* **Enhancing Generalization:** Eliminating unimportant features improves the model's generalization ability.
* **Reducing Resource Consumption:** Models with fewer features require fewer computational resources.

### 4.2.2 Use Regularization ( for Neural Network ) :

Regularization adds a penalty to the model's loss function based on the magnitudes of its weights. It helps prevent overfitting by discouraging overly large weights, promoting a simpler and more generalizable model.

* **Overfitting Prevention**: Regularization prevents overfitting by penalizing complex models, promoting simplicity.
* **Improved Generalization**: Enhances the model's ability to generalize to unseen data.
* **Stability**: Contributes to stable learning, reducing sensitivity to small variations in training data.

### 4.2.3 Drop out ( for Neural Network ) :

Dropout is a regularization method specifically designed for neural networks. During training, it randomly "drops out" a percentage of neurons or connections by setting their outputs to zero. This introduces redundancy, prevents co-adaptation of neurons, and encourages the network to learn more robust features, improving generalization.

* **Overfitting Reduction:** Dropout reduces overfitting by introducing randomness during training.
* **Ensemble Learning:** Acts as an ensemble method, training multiple subnetworks for improved generalization.
* **Adaptability:** Enhances the model's adaptability to diverse features by preventing neuron co-adaptation.

CHƯƠNG 5 –ENHANCING MACHINE LEARNING ACCURACY

"Enhancing Machine Learning accuracy" is the process of improving the performance of machine learning models by optimizing parameters, algorithms, and datasets. The goal is to minimize errors and improve the model's ability to make accurate predictions on new, unseen data in real-world applications.

* 1. Methods for enhancing accuracy :

**5.1.1 Hyperparameter Tuning:**

* Adjust hyperparameters such as learning rate, layer sizes, and units per layer to optimize model performance.
* Benefits: Optimizes model performance.
* Drawbacks: Requires significant time and effort for experimentation.

### 5.1.2 Data Augmentation:

* Generate additional data by transforming, rotating, flipping images, or applying other data augmentation techniques.
  + Benefits: Enhances data, reduces overfitting risk.
  + Drawbacks: May introduce loss of generality if applied excessively.

### 5.1.3 Deep Learning Architectures:

* Utilize deep neural network models and complex network architectures to learn intricate data representations.
  + Benefits: Capable of learning complex data representations.
  + Drawbacks: Requires abundant data and may increase computational complexity.

### 5.1.4 Regularization:

* Apply techniques like L1 or L2 regularization to control overfitting and prevent excessive complexity.
  + Benefits: Controls overfitting, reduces optimization risk.
  + Drawbacks: May result in loss of critical information.

### 5.1.5 Ensemble Models:

* Combine predictions from multiple different models to create a final prediction with higher accuracy.
  + Benefits: Boosts accuracy by combining strengths of multiple models.
  + Drawbacks: Increases computational cost and resource requirements.

### 5.1.6 Check and Handle Noisy Data:

* Inspect data to identify and address noise that could impact accuracy.
  + Benefits: Improves data quality and accuracy.
  + Drawbacks: Requires time-consuming inspection and handling.

### 5.1.7 Model Architecture Changes:

* Experiment with and adjust the model architecture, adding or removing layers to assess their impact on performance.
  + Benefits: Optimizes model architecture for specific tasks.
  + Drawbacks: Requires in-depth knowledge of model architecture.

### 5.1.8 Optimize Model Size:

* Evaluate and reduce model size to decrease complexity and increase prediction speed.
  + Benefits: Reduces complexity, speeds up predictions.
  + Drawbacks: May result in loss of learning capability.
  1. **Analysis and evaluation the problem:**

Analyzing and evaluating problems to enhance accuracy in a given context typically involves a combination of techniques such as feature selection, model tuning, and data preprocessing. Here's a brief overview of some common approaches:

### 5.2.1 Model Selection and Tuning:

The goal is to identify the most suitable machine learning model for a given task and optimize its hyperparameters to achieve optimal performance.

**Some methods that use for Model Selection and Tuning:**

**Grid Search and Random Search:**

* Grid Search: Systematically explores hyperparameter combinations in a predefined grid. Computationally expensive but exhaustive.
* Random Search: Randomly samples hyperparameter combinations, providing an efficient alternative to Grid Search in high-dimensional spaces.

**Bayesian Optimization:**

* Uses probabilistic models to predict hyperparameter performance. Adapts the search based on past evaluations, reducing the number of model evaluations compared to exhaustive methods.

**Ensemble Methods:**

* Bagging (e.g., Random Forest): Constructs parallel models, trains on random data subsets, and aggregates predictions to reduce overfitting.
* Boosting (e.g., AdaBoost, Gradient Boosting): Builds models iteratively, focusing on correcting errors from previous models. Weights models based on performance, improving overall accuracy.

### 5.2.3 Data Preprocessing:

Clean and preprocess the data to improve its quality and ensure that it is suitable for the chosen model.

**Methods:**

* **Data Cleaning**: Handle missing values, outliers, and inconsistencies in the dataset.
* **Normalization and Standardization**: Scale numerical features to a standard range to prevent dominance by certain features. (MinmaxScaler, StandardScaler)
* **Encoding Categorical Variables**: Represent categorical variables in a format suitable for machine learning models (e.g., one-hot encoding, label encoding).
* **Data Augmentation**: Generate additional training samples by applying transformations to existing data (common in image and text data).

### 5.2.4 Cross-Validation:

Assess the model's performance in a robust manner to ensure that it generalizes well to new, unseen data.

**Methods:**

* **k-Fold Cross-Validation**: Divide the dataset into k folds and train the model k times, using a different fold for validation each time. Average the results to obtain a more reliable performance estimate.
* **Stratified Cross-Validation**: Ensure that each fold maintains the same distribution of target classes as the original dataset.

**5.2.5 Evaluation Metrics:**

Choose appropriate metrics to measure the performance of the model based on the specific characteristics of the problem (e.g., accuracy, precision, recall, F1 score, ROC-AUC).

**Methods:**

* **Selecting Relevant Metrics**: Choose metrics that align with the goals of the model and provide meaningful insights into its performance.

**Threshold Adjustment**: Fine-tune classification thresholds to balance precision and recall based on the specific requirements of the problem.

# TÀI LIỆU THAM KHẢO

[1]Course “Các phương pháp tránh Overfitting”. Sep 12, 2023. <https://viblo.asia/p/cac-phuong-phap-tranh-overfitting-gDVK24AmlLj> . Address Dec 20,2023.

[2] Văn Thanh. Course “TẠI SAO LẠI SỬ DỤNG ACTIVATION FUNCTION TRONG NEURAL NETWORK”. " <https://viblo.asia/p/tai-sao-lai-su-dung-activation-function-trong-neural-network-MG24BwweJz3> ”Address Dec20,2023.

[3] Quốc Hi. Course “Optimizer- Hiểu sâu về các thuật toán tối ưu ( GD,SGD,Adam,..)

”. Dec19, 2017 . <https://viblo.asia/p/optimizer-hieu-sau-ve-cac-thuat-toan-toi-uu-gdsgdadam-Qbq5QQ9E5D8> .Address Oct 20,2023.

[4] Course “Label Encoding in Python – 2024

s” https://www.mygreatlearning.com/blog/label-encoding-in-python/#:~:text=Label%20encoding%20is%20a%20technique,only%20operate%20on%20numerical%20data. .Address Dec 20,2023.

# PHỤ LỤC