Abstract and Introduction

This report will adopt a wide-ranging dataset (Kaggle, 2024) that includes demographic, clinical, and lifestyle attributes to uncover the association of several behaviors with obesity. It shall utilize neural network analysis, decision trees, ensemble learning, and random forests in the hope of discovering trends, locating high-risk groups, and determining possible outcomes highly accurately. These results will further contribute to the literature in the field of obesity studies, hence aiding evidence-based decision-making in public health and preventive medicine.

Obesity has been counted as one of the most global health problems (Organization, World Health, 2021), affecting millions of people across different populations. The condition is intricately linked to a host of chronic diseases, including cardiovascular diseases (CVDs), diabetes, and cancer (Hu A. H., 2015) making it a pressing concern for healthcare systems worldwide. The consequences of obesity go beyond the clinical field, into societal and economic structures (al, 2011). From increasing healthcare costs to the reduced workforce, people became less productive. Obesity prevalence has risen over the last decades due to complex interplay among genetic, behavioural, environmental, and socio-economic factors (El-Serag, 2010) (MD, 2010). To avoid the consequences of obesity, it needs a variety of approaches, involving public health policies, enhancing lifestyle and community involvement. Solutions to the obesity pandemic are needed, ranging from changes in public policy on health and nutrition to innovative technology solutions. Among data science interventions, machine learning techniques can present the ability to derive actionable insights into complex relations between obesity and related health conditions, particularly CVDs.

Leveraging Data Mining Techniques to Understand Drivers of Obesity
Authors: R. Salehnejad, R. Allmendinger, Y.-W. Chen, M. Ali, A. Shahgholian, P. Yiapanis, and M. Mansur
Publication Date: 2017

The machine learning algorithms used by us including neural networks, decision trees, ensemble learning and random forests were able to predict the prevalence of obesity with an accuracy ranging from 91% to 96%. Neural networks are highly capable of capturing intricate non-linear relationships and decision trees offer less level of accuracy with the added benefit of interpretable data. Random forests are stronger in prediction using ensemble approaches and they also provide us with valuable information on the importance of different features. In this case we emphasize the key determinants of obesity as weight, height and age.

In contrast to the journals that used regression trees and LASSO for modeling obesity and interactions of different determinants, our models aim at having maximum predictive accuracy. The journal emphasizes the importance of describing multivariable relationships such as those between income, physical activity and mental health since it is in these relationships that the causes of obesity must be found then not in the variables themselves. The regression trees in the journal present complex patterns illustrating a declining rate of obesity for individuals with higher income and physical activity levels which emphasizes our own results about demographic and lifestyle factors. The journal uses interaction and the LASSO technique for feature selection thus providing us with a better cause-and-effect insight. Our models achieve this indirectly through feature importance quantification mediated by feature importance.

Overall, our models have increased ability to predict obesity prediction accuracy, but the techniques used in the journal provide added value by revealing complex cause and effect mechanisms. The integration of techniques like unbalanced regression trees in our modeling can increase transparency and clarify interactions between different determinants which can be the cause for more efficient interventions to promote obesity

Exploratory Data Analysis (EDA) Report for Obesity Dataset

Our dataset includes information about people's demographics, health, and lifestyle. It has 2,111 rows and 17 columns, featuring details like age, height, weight, physical activity levels, and eating habits, among others. The dataset does not have any missing values, making it complete and reliable for analysis. The features are the characteristics such as numbers like age and weight, and categories like gender and history of overweight in the family. All these features combine to assist in categorizing individuals into various obesity levels. The subsequent analysis provides information regarding the data and how it was pre-processed for machine learning algorithms.

To help with our machine learning study, we show important statistics for numbers that help classify people into obesity levels. These include Age, Height, Weight, Frequency of Physical Activity (FAF), and Technology Use (TUE). The statistics are shown in **[Table 1**]. These statistics provide helpful information on the shape of the dataset and how it differs. The high range and spread of meaningful features assist the machine learning algorithm in performing efficiently with varying levels of obesity. This information also assists in preprocessing activities, such as scaling and encoding, to enable the model to perform better.

Feature	Mean	Median	Min	Max	Std Dev
Age	24.3	22.8	14	61	6.35
Height (m)	1.70	1.70	1.45	1.98	0.09
Weight (kg)	86.6	83.0	39	173	26.19
FAF (times/week)	1.01	1.00	0	3	0.85
TUE (hours/day)	0.66	0.63	0	2	0.61

Table 1: Descriptive statistics

The target column is NObeyesdad, which has multiple levels of obesity as classes, normal weight, overweight, and other levels of obesity. The distribution, graphed in **Figure 1**, is pretty well-balanced, with no high skewness in classes. "Normal Weight" and "Obesity Type I" classes happen more than others, being a general trend within this dataset.

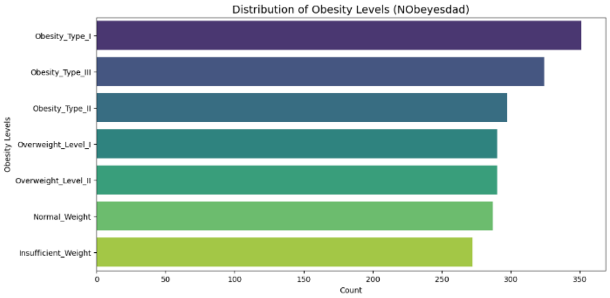


Figure 1 - Distribution of Obesity Levels (NObeyesdad)

Numerical feature analysis provides information about various features of the dataset concerning age, height, weight, and physical activity and the level of using technology among the respondents. The distribution of height appears to be normal as depicted in **Figure 2**, while the distribution of weight (**Figure 3**) is highly dispersed; the higher the weights, the higher the levels of obesity. Physical activity, according to FAF Frequency of Physical Activity, and as illustrated in **Figure 4**, follows a pattern whereby lower activity is associated with a rising proportion of obesity. Similarly, **Figure 5** shows the scatter of Technology Use - TUE - with the feature that higher usage is associated with a higher prevalence of obesity. These findings hint at the effects of lifestyle factors like technology use and exercise, coupled with inherent physical attributes like weight and height, on obesity trends within the set.

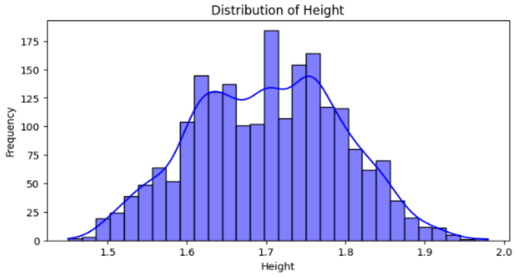


Figure 2 - Distribution of Height

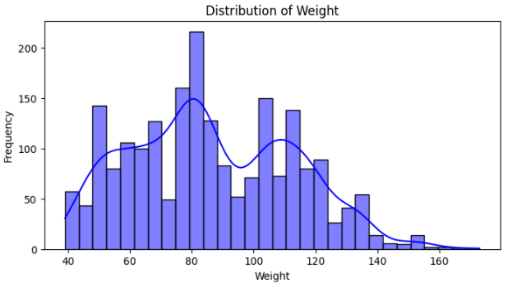


Figure 3 - Distribution of Weight

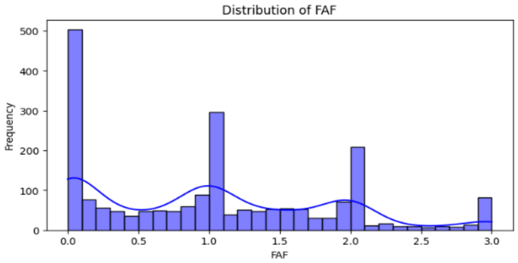


Figure 4 - Distribution of FAF

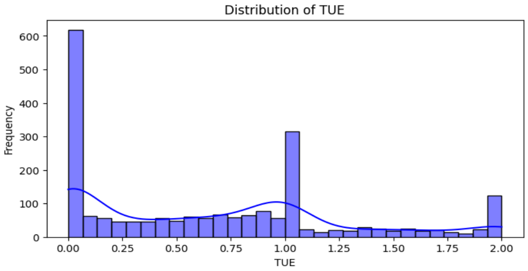


Figure 5 - Distribution of TUE

Figure 6 presents the correlation analysis, which provides a summary of the relationship among some of the significant numeric variables within the data set. Weight is positively correlated with the levels of obesity, indicating its direct contribution towards obesity trends. Physical activity, as quantified by FAF, is negatively correlated with obesity, indicating that with increased physical activity, obesity levels are lower. On the contrary, TUE is weakly but positively associated with obesity, explaining that the higher the technology use the higher the obesity. These findings bring out the interaction between activities of life style and physical attributes to impact the outcome on obesity.

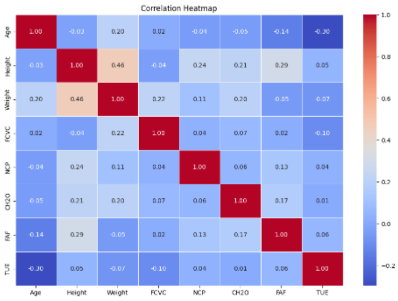


Figure 6 - Correlation Heatmap

As it can be seen from **Figure 7**, the weight distribution and degree of obesity definitely have distinct differences between obesity classes. The normal-weight class has a distribution that is tightly dispersed, while higher classes such as Obesity Type II and Type III have more spread in weight, and the median weight seems to increase. Insufficient weight has a very low spread compared to other classes. This chart highlights weight as a major determinant of obesity levels and emphasizes how important it is in identifying and classifying trends relating to obesity.

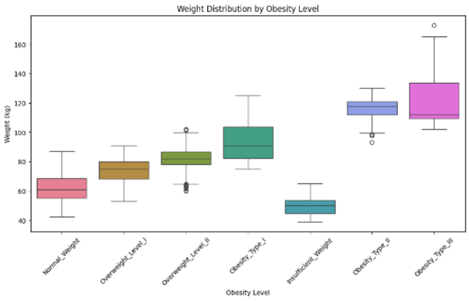
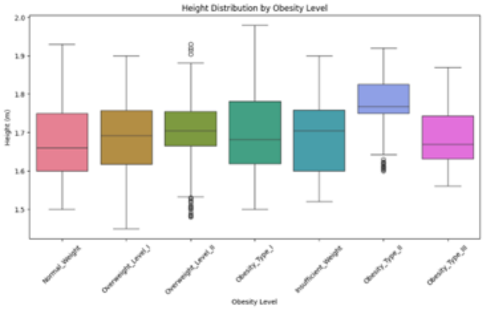


Figure 7 - Weight Distribution by Obesity Level

The following analysis provides key findings to obesity and its determinants through several visualizations. **Figure 8** shows the distribution of height for the different levels of obesity. One can observe from this that people of normal weight, or with a declining level of obesity, have a relatively consistent height distribution, while increasing levels of obesity have greater dispersion. **Figure 9**, Boxplot of Height The boxplot of height presents the median height and some outliers on the higher side that need to be examined for accuracy of data verification. **Figure 10**: Boxplot of Weight The boxplot of weight shows the centrality and variability in weight with some outliers on the higher side, which could be an extreme case and could influence the analysis. **Figure 11**: Pairplot of numerical features. A summary of pairwise relations for different numerical variables, which includes age, height, weight, among others, to illustrate patterns, clusters, and possible correlations helpful in further exploratory analysis and modelling exercises. These figures collectively impart valuable information concerning this dataset and bring into focus the need for outlier management and also identifying important predictive variables which could go to further advance the models for classifying and



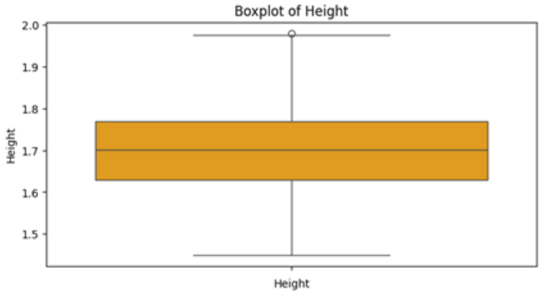
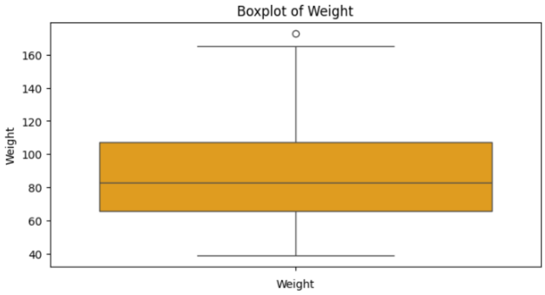


Figure 8 - Height Distribution by Obesity Level

Figure 9 - Boxplot of Height



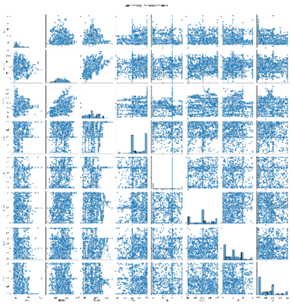


Figure 10 - Boxplot of Weight

Figure 11 - Pairplot of Numerical Features

Preprocessing

The objective is to manage both numerical and categorical features appropriately, normalize numerical features for uniformity in terms of a single measurement, encode target variable for enhanced model compatibility, and transform both training and testing datasets through a uniform pipeline.

The preprocessing processes began with encoding categorical variables like Gender, MTRANS, and CAEC, which could not directly be used in machine learning algorithms. One-Hot Encoding was applied for encoding such variables, creating binary columns for each group (e.g., Gender_Male, MTRANS_Public_Transport), and deleting the first category for each feature to avoid dummy variable trap. As a result, categorical variables were replaced with binary columns, for example, partitioning Gender into Gender_Male (1 for Male, 0 for Female). Target variable (NObeyesdad), with categories for obesity, was transformed into numerical labels using Label Encoding, such that, for example, Insufficient_Weight was 0 and Obesity_Type_I was 1, compatible with machine learning algorithms. Numerical features like Age, Weight, and Height, with variable scales, were normalized using StandardScaler to have a mean of 0 and a standard deviation of 1, improving model performance and interpretability. Outliers in numerical features were checked via visualization with boxplots, but no processing was done for them since the dataset appeared uniformly distributed. For uniformity, a pipeline was constructed to combine all preprocessing processes, including One-Hot Encoding for categorical values and scaling for numerical values, such that the same transformation is performed for training and testing sets alike. Finally, the dataset was split between training (80%) and testing (20%) sets via stratified sampling in order to maintain balanced class distribution, and data leakage was prevented via preprocessing post-split.

Feature Type	Transformation Applied
Categorical Features	One-Hot Encoding (e.g., Gender, MTRANS, etc.)
Target Variable	Label Encoding (NObeyesdad → 0 to 6)
Numerical Features	Standard Scaling (Age, Height, Weight, etc.)
Train-Test Split	80% Training, 20% Testing (Stratified by NObeyesdad)
Preprocessing Pipeline	Combined transformations into a reusable pipeline for efficiency

Table 2: ***Summary of Pre-processed Data***

The preprocessing involved an increase in feature numbers using One-Hot Encoding, with the initial 17 features expanding as categorical in a manner that, for instance, feature MTRANS, one column, increased into numerous binary columns. The datasets were split into a training partition for 80% and a testing partition for 20%, with both sets having balanced classes. All numerical features were normalized in an attempt to have a uniformity in model scales, allowing them to work and compare effectively together.

The preprocessing stage bestows numerous benefits, including uniformity through a pipeline that performs a uniform transformation for all sets of the data. It converts categorical and numerical features to a state of compatibility with most machine algorithms, getting them analysis ready. With additional scaling and encoding, preprocessing maximizes algorithm performance and acceleration in convergence. To a similar extent, stratified sampling reduces bias through balanced distribution of classes in training and test sets, maximising reliable and proper model output.

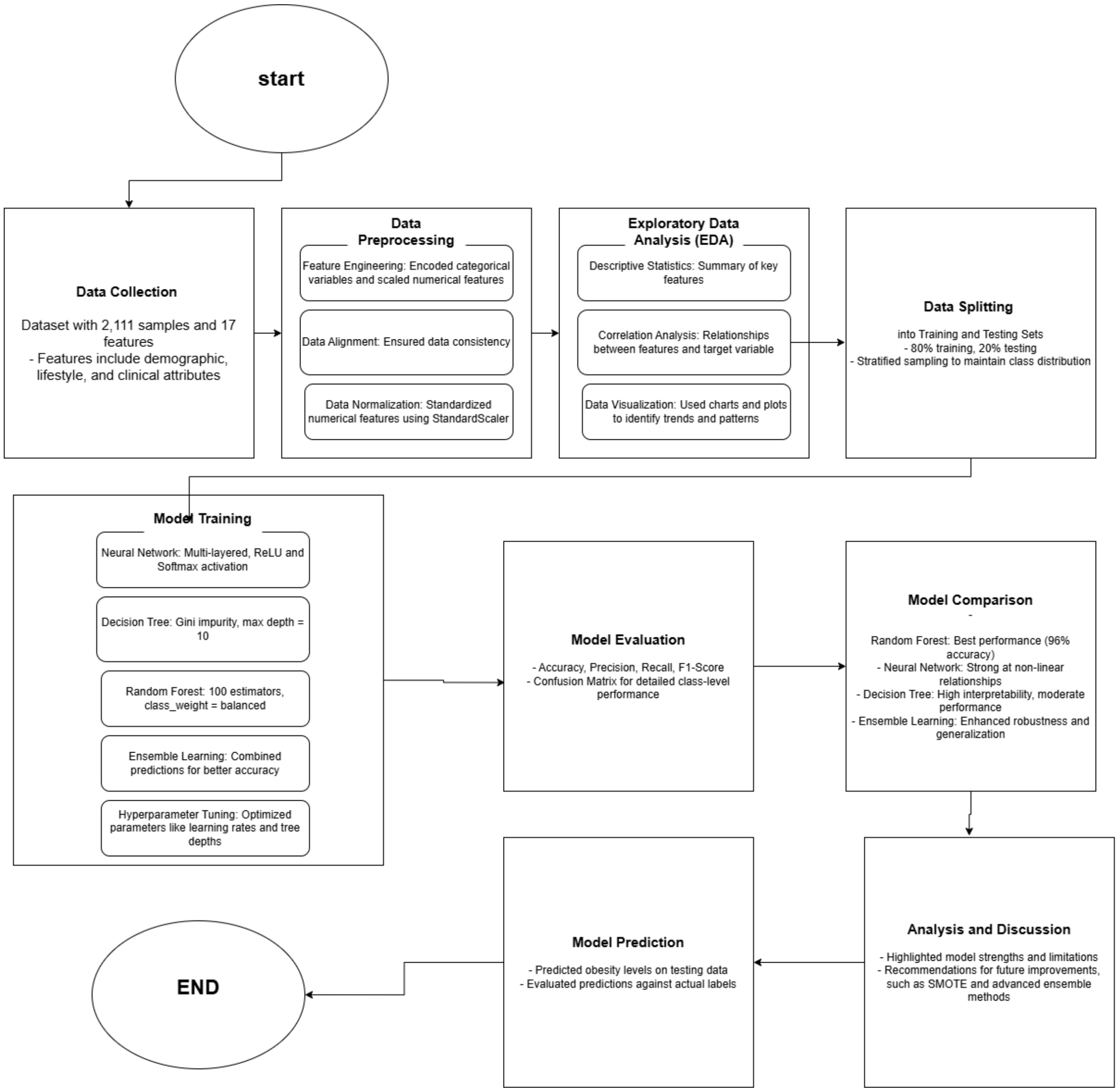

start
Data Collection
Dataset with 2,111 samples and 17 features
- Features include demographic, lifestyle, and clinical attributes
Data Preprocessing
Feature Engineering: Encoded categorical variables and scaled numerical features
Data Alignment: Ensured data consistency
Data Normalization: Standardized numerical features using StandardScaler
Exploratory Data Analysis (EDA)
Descriptive Statistics: Summary of key features
Correlation Analysis: Relationships between features and target variable
Data Visualization: Used charts and plots to identify trends and patterns
Data Splitting
into Training and Testing Sets
- 80% training, 20% testing
- Stratified sampling to maintain class distribution
Model Training
Neural Network: Multi-layered, ReLU and Softmax activation
Decision Tree: Gini impurity, max depth = 10
Random Forest: 100 estimators, class_weight = balanced
Ensemble Learning: Combined predictions for better accuracy
Hyperparameter Tuning: Optimized parameters like learning rates and tree depths
Model Evaluation
- Accuracy, Precision, Recall, F1-Score
- Confusion Matrix for detailed class-level performance
Model Comparison
-
Random Forest: Best performance (96% accuracy)
- Neural Network: Strong at non-linear relationships
- Decision Tree: High interpretability, moderate performance
- Ensemble Learning: Enhanced robustness and generalization
Analysis and Discussion
- Highlighted model strengths and limitations
- Recommendations for future improvements, such as SMOTE and advanced ensemble methods
Model Prediction
- Predicted obesity levels on testing data
- Evaluated predictions against actual labels
END

Figure 12. Methodology flowchart

1. Random Forest

The Random Forest model is developed for multi-class classification of individuals into obesity categories such as Normal_Weight, Obesity_Type_I, etc., with the use of physiological data (e.g., age, height, and weight), lifestyle, and demographics in "ObesityDataSet." An aggregation of decision trees is employed in the model to obtain accuracy and robustness, and it is best suited for multi-class classification.

The training process involved a series of significant steps. First, preprocessing of data involved label encoding and standarization of categorical feature such as Gender and encoding through MTRANS and standarization of continuous feature for uniformity in scales to enhance model performance. Second, the dataset was split into training (80%) and testing (20%) sets with a fixed constant for reproduction (random_state=42). The model employed hyperparameters such as n_estimators=100, with 100 trees for an efficient strong ensemble, balancing computation efficiency, and max_depth=None for growing trees to its full for effective capture of complex relationships between data, with random_state=42 for reliable reproduction of its performance. Analysis of feature importance revealed that most contributing factors towards prediction included Weight (33.62%), Height (10.48%), and Age (9.69%) and at 5.17%, FCVC (frequency of consumption of vegetables). **Figure 1**3 presents these values for feature importance, with a considerable contribution towards prediction for obesity level.

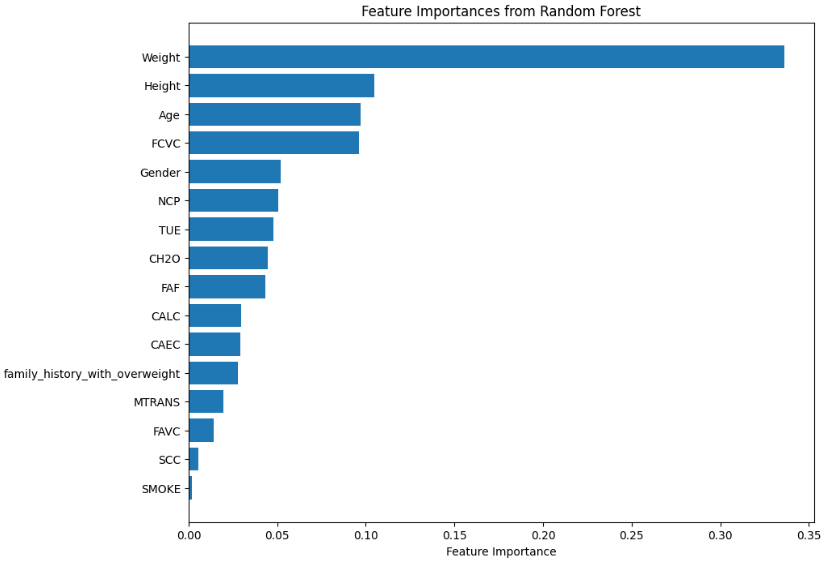


Figure 13. Feature Importance from Random Forest

The evaluation revealed high performance of the model with 96% accuracy in the test set. In the confusion matrix **(Figure 14)**, strong performance in all obesity classes with minor misclassifications, particularly between neighboring classes like Normal_Weight and Overweight_Level_I, was observed. Classification report **(Figure 15)** contained precision, recall, and F1-value for every class, and weighted averages of 96% for all three values, representing balanced performance in that model handles all obesity classes with no bias towards any sets of classes.

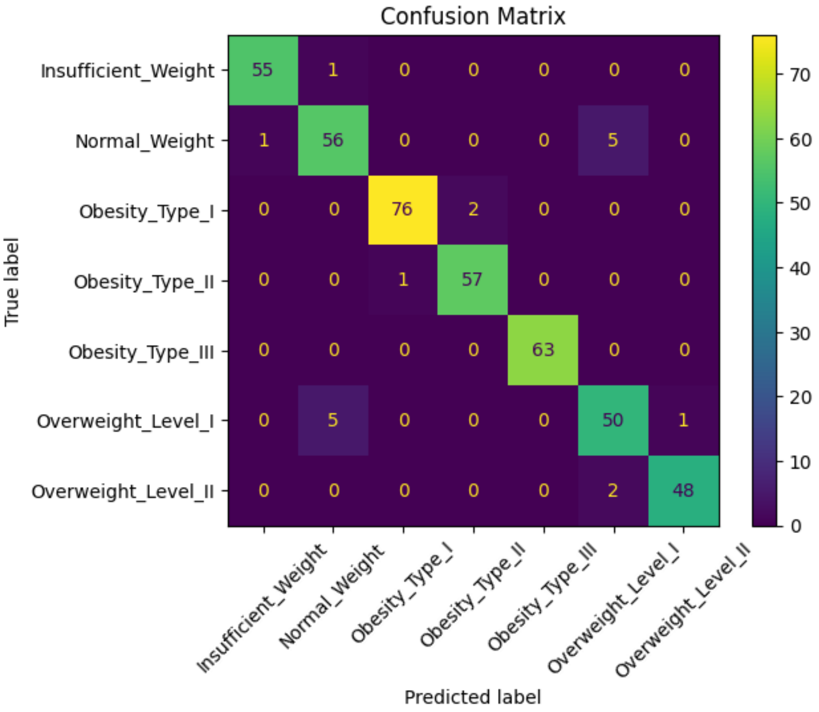
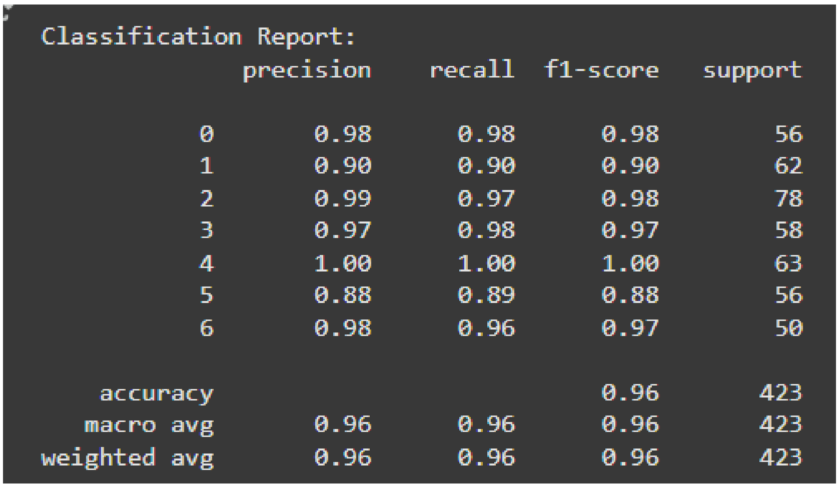


Figure 14. Confusion Matrix



```
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.98      0.98        56
           1       0.90      0.90      0.90        62
           2       0.99      0.97      0.98        78
           3       0.97      0.98      0.97        58
           4       1.00      1.00      1.00        63
           5       0.88      0.89      0.88        56
           6       0.98      0.96      0.97        50

    accuracy                           0.96       423
   macro avg       0.96      0.96      0.96       423
weighted avg       0.96      0.96      0.96       423
```

Figure 15 - Classification Report - 96%

The model holds high accuracy and balanced prediction for all classes, and its salient values such as Weight, Height, and Age validate domain expert insights in terms of obesity factors. There are, however, a small number of misclassifications, for example, Prediction of Normal_Weight as Overweight_Level_I, and perhaps due to feature boundary overlaps, and perhaps a class imbalance in the training data, possibly impacting prediction for under-class categories.

To enhance the model, underrepresented categories can be supplemented with extra samples in the dataset in order to mitigate overbalance and extra information can be included using a feature such as an individual's activity level, level of metabolization, or nutritional information for added information. Model optimization can involve hyperparameters such as max_depth and n_estimators being optimized through grid and random search and testing with alternative models such as XGBoost or Gradient Boosting for a performance evaluation. Performance can be measured with added effectiveness through cross-validation for assurance of performance for a range of datasets and with sophisticated evaluation using a metric such as an ROC curve or an AUC for a deeper analysis.

The Random Forest model is seen to have a high level of performance for "ObesityDataSet" with high accuracy and interpretability. With slight improvements in the model and model fine-tuning in the dataset, it can become even efficient and powerful for obesity categories classification. Figures 13 (Confusion Matrix), 14 (Classification Report), and 15 (Feature Importance) included confirm full analysis and proper visualization of output.

2. Decision tree

A study is conducted of a decision tree classifier on a dataset specialized in obesity level classification. The dataset contains a set of attributes based on which the obesity classes of individuals are predicted. The aim is to build a decision tree model with the ability to classify individuals into different obesity levels based on their respective features. The data has categorical and numerical features and the target feature **NObeyesdad** has different categories of obesity. Data preprocessing involved one-hot encoding the categorical features using **Label Encoder** and scaling numerical features using **Standard Scaler**.

Categorical features were encoded using **Label Encoder** to convert them into numerical variables. Numerical features were scaled using **Standard Scaler** to normalize the range of independent variables. The Dataset was split into training and testing sets in a ratio of 80-20, making sure that it was stratified so that the distribution of the target variable is maintained. A decision tree classifier was initialized with the following parameters. **criterion= gini** which employed the Gini impurity criterion to estimate the quality of the splits. **max_depth=10** which showed the depth of the tree was set to 10 to avoid overfitting. **random_state=42** which sets the random seed to a specific value to make the process reproducible. The model was trained on the training dataset.

The model achieved 91.02% accuracy on the test data. The classification report provided precision, recall and F1-score for each class and overall accuracy. The ratio of correctly predicted positive observations to total predicted positives. The model performed with high precision for all classes, from 0.78 to 1.00. The proportion of positively predicted observations to all observations in the true class. The recall values were between 0.84 and 0.98. The harmonic means of precision and recall. The F1-scores were high throughout, reflecting a proper balance between precision and recall.

The confusion matrix was plotted to illustrate the performance of the model in classifying every category of obesity.

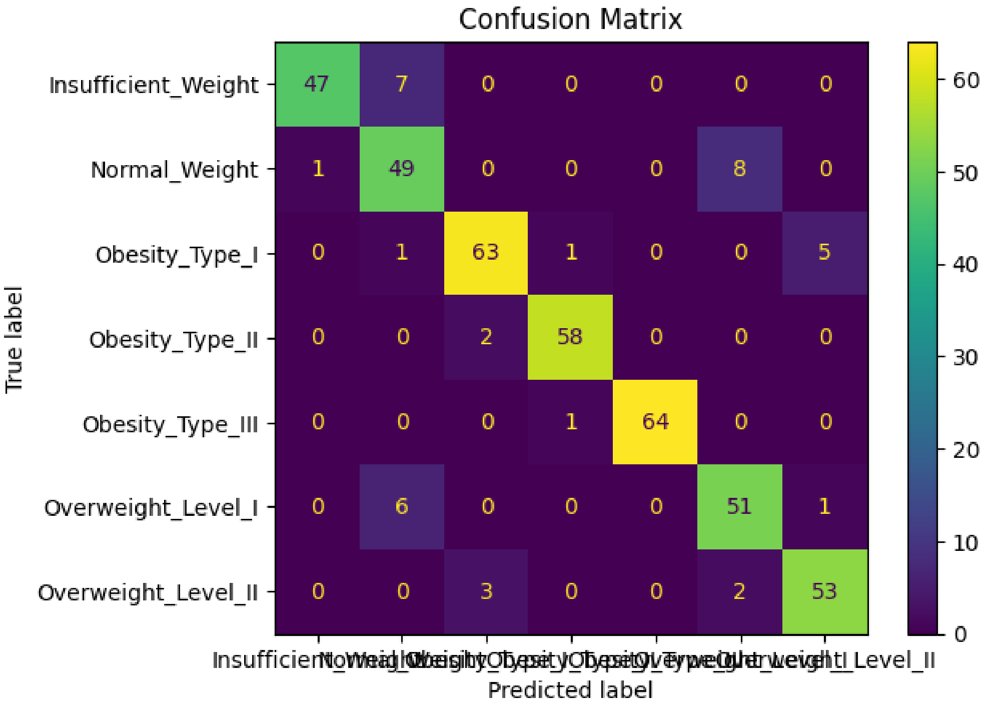


Figure 16. Confusion Matrix

A simple visualization of the Decision Tree was plotted, showing the initial three levels of the tree for the sake of readability. The model's decision-making process visualization becomes easier to interpret.

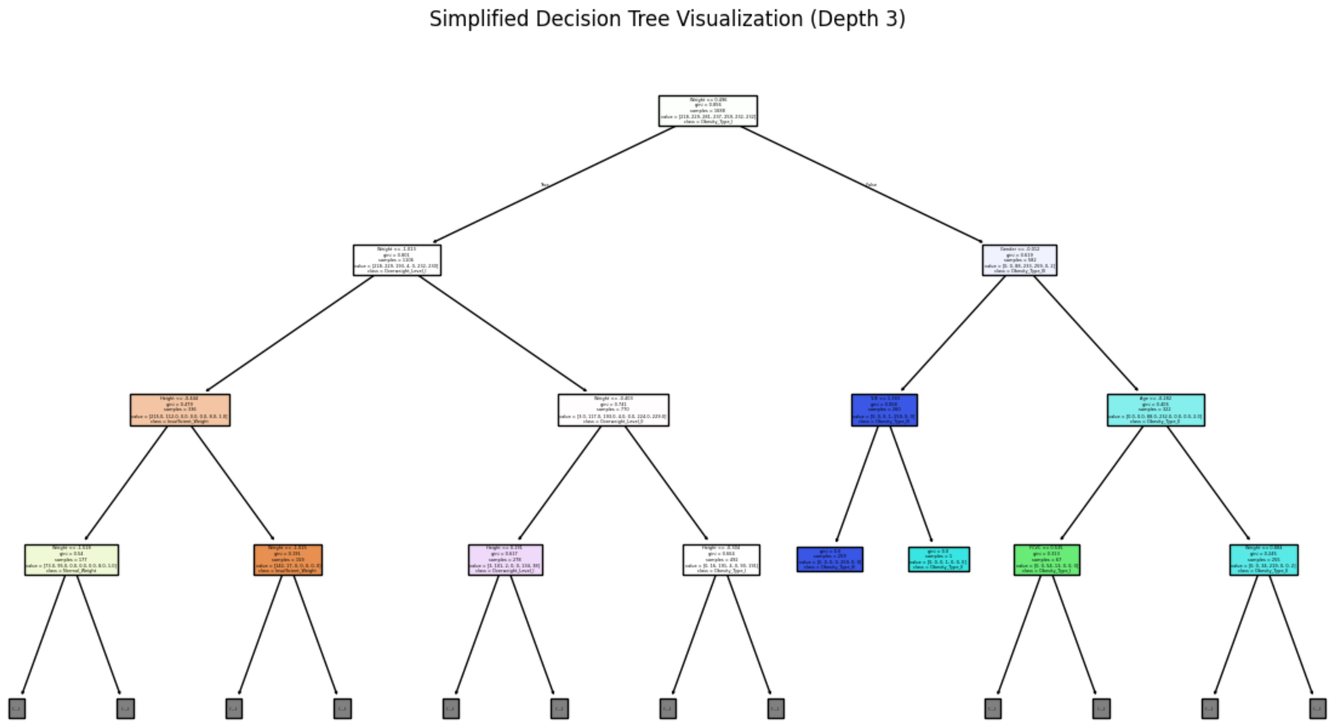


Figure 17. Decision tree

The decision boundary was visualized based on 2 most important features, weight and height. This plot shows how this model separates these classes using weight and height features. There is a lot of influence from these features and their importance in the model is very high as shown by their high values.

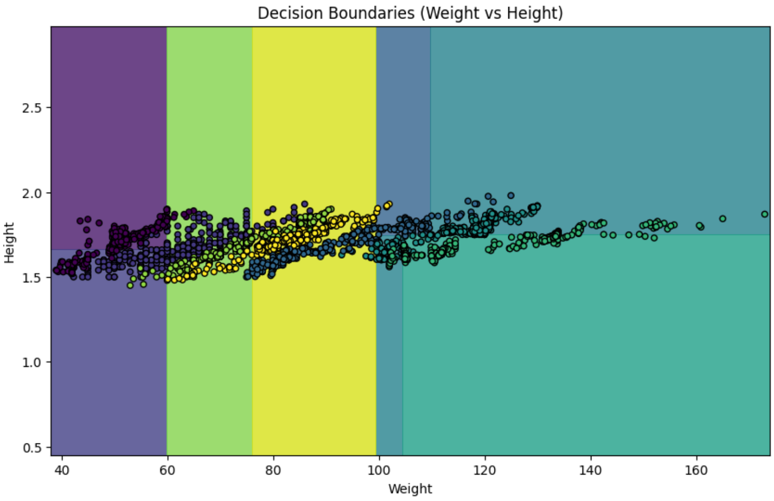


Figure 18. Decision Boundary

In the given analysis, evidence that weight and height constitute the main support for the decisions returned by this model. The decision boundary depends on a feature with low scores regarding importance. Using this insight simplifies the model for only the important features, aiming to enhance its interpretability as well as performances.

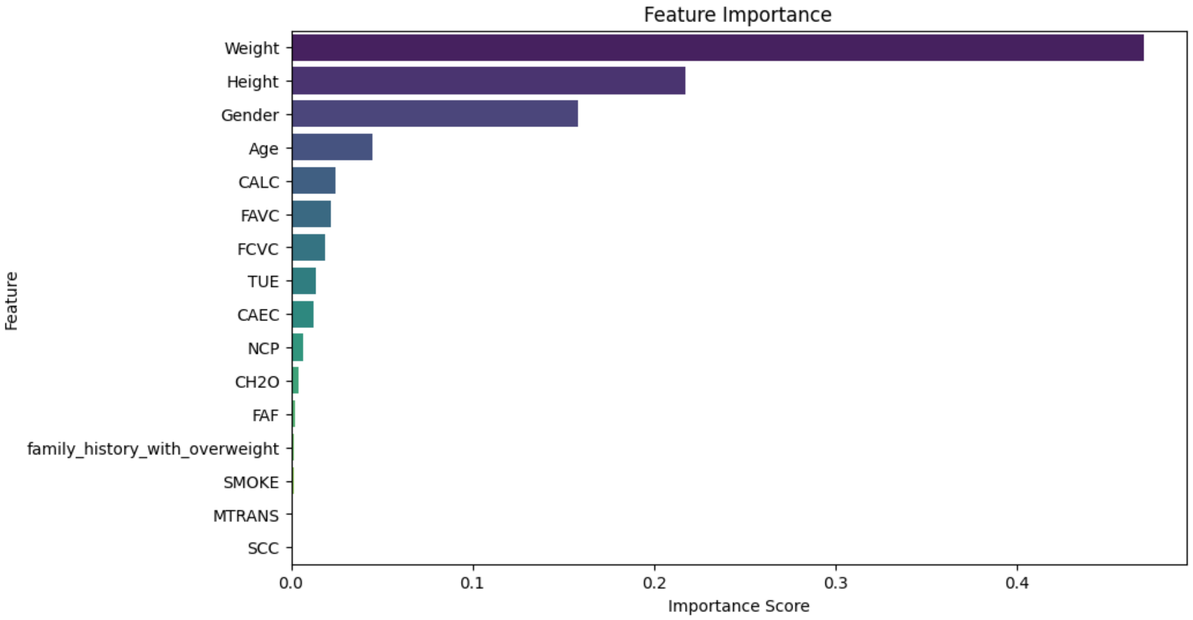


Figure 19. Feature Importance

The feature importance plot, represented in the graph in ***figure 19,*** illustrates the relative importance of every feature towards the model's ability to make predictions. Weight is the most influential feature with an importance score of more than 0.4 followed very closely by height with a score of a little over 0.2. The two features are the key factors in the model's decision-making process, reflecting their high influence on the resulting predictions. Gender, Age and family history with overweight are moderately significant attributes, all with a value of less than 0.1. While they are beneficial to the model's functionality, their impact is far less significant than Weight and Height. CALC, FAVC, FCVC, TUE, CAEC, NCP, CH2O, and FAF are less significant attributes with lower values showing that they play little role in the model's predictive power.

The model achieved an overall accuracy of 91.02%. The model performed well for all the obesity categories with high precision, recall and F1-scores. The lowest precision was 0.78 for the **Normal Weight** class and the lowest recall was 0.84 for the **Insufficient Weight** class. The confusion matrix showed that the model misclassified very few cases, with most predictions being equal to the actual labels. The Decision Tree classifier worked very well on the obesity dataset, with high accuracy and balanced precision and recall for all classes. The model's capability to classify people into various obesity classes correctly makes it a useful tool for obesity research and interventions.

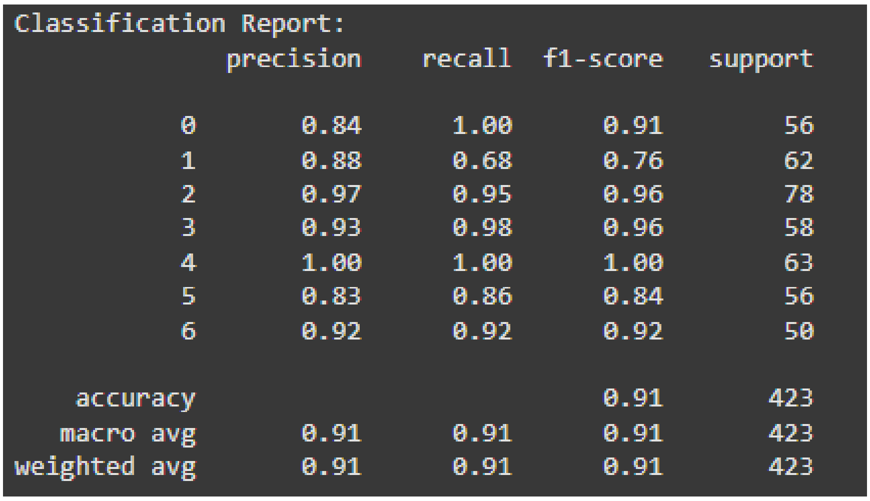
Although the model is performing well, additional fine-tuning of hyperparameters such as modifying **max_depth**, **min_samples_split** can lead to performance gains. Feature importance analysis can provide a clue about which features contribute most to the classification and further enhance the model.

3.Neural Network

The neural network model has 3 layers which are input layer, hidden layers and output layer. The input layer focuses on the behavior of each person that leads to a result in **the NObeyesdad column, the age, height, weight, family history with overweigh and smoking habits e**tc, all of this are categorized in the input layer as neurons, each neuron ask a specific question, for instance, one neuron might ask, "What is the age of the person being classified?" In total, the input layer contains 16 neurons, one for each feature. For the hidden layers, first hidden layer included 64 neurons using ReLU activation function, this layer takes the input layer data and applies a linear transformation using weights and biases, and passes it through the ReLU activation function, in the second hidden layer taking 32 neurons, takes the output which is **NObeyesdad** from the first hidden layer and do another activation function to introduce non-linearity. The output layer contains 6 neurons as the optional results are Normal_Weight, Overweight_Level_I, Overweight_Level_II, Obesity_Type_I, Obesity_Type_II, Insufficient_Weight. Using soft-max activation function to calculate the probability for each class, ensuring the outputs add up to 100%, the class with the highest probability is chosen as the model's prediction. Neural networks are highly effective for modelling complex relationships and patterns in datasets like this, where features like age, eating habits, and physical activity have non-linear dependencies. The model can automatically learn feature importance and interactions without requiring extensive manual feature engineering.

The neural network model trained to classify people into obesity levels by minimizing prediction errors, by using loss function, adam optimizer and evaluation metric. Loss function are used because it is ideal for multi-class classification, it calculates the difference between the predicted probabilities and the actual class labels. (Medium, n.d.), in the other hand adam optimizer was used due to its efficiency in updating weights, combining the benefits of adaptive learning rates and momentum. The training data was processed in batches of 43 samples at a time, balancing computational efficiency and learning speed, and for the epochs training iterated through the entire dataset for 50 times. Early stopping was applied to prevent overfitting by monitoring the validation loss. The learning phase for the neural network model divided to four steps, firstly the model predicts class probabilities for the input data with its current weights, secondly using the loss function that calculates and measures the errors between predictions and actual labels, thirdly the backpropagation algorithm adjusts the weights to minimize the loss, using gradients computed from the error, this process repeats for multiple epochs and in each epochs its gradually improving its ability to classify the obesity levels.

Careful tuning of (Hyperparameter tuning, n.d.) was carried out for the neural network, including learning rate, batch size and number of epochs. Convergence was looked out with an effective learning rate of 0.001, and the batch size was set as 43 to balance between computational efficiency and model performance. It used 50 epochs, with early stoppage to prevent overfitting while monitoring validation loss that converges at approximately epoch 20. This architecture was chosen by iterative tests because with 64 neurons in the first in the first layer and 32 in the second, one achieves a good trade-off between computational cost and accuracy that leads the model to be clear enough for classifying obesity levels.



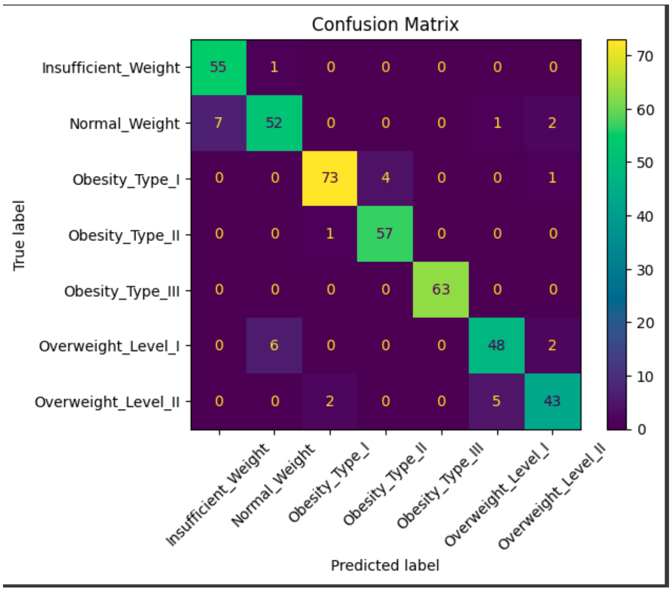
```
Classification Report:
              precision    recall  f1-score   support

           0       0.84      1.00      0.91        56
           1       0.88      0.68      0.76        62
           2       0.97      0.95      0.96        78
           3       0.93      0.98      0.96        58
           4       1.00      1.00      1.00        63
           5       0.83      0.86      0.84        56
           6       0.92      0.92      0.92        50

    accuracy                           0.91       423
   macro avg       0.91      0.91      0.91       423
weighted avg       0.91      0.91      0.91       423
```

Figure 20. Classification Report

The trained neural network achieved a test accuracy of **91%** showing its ability to classify individuals into different obesity levels based on their features. Based on the classification report in ***figure 20*****,** precision indicates the proportion of correct predictions for each class, the model performed exceptionally well in predicting classes like *obesity_Type_I(0.97)* and *Obesity_Type_II(1.00).* Where the recall represents the proportion of actual instances that are correctly identified, the recall for *Normal_Weight* (1.00) and *obesity_Type_II*(1.00) indicates perfect identification, while *Overweight_Level_I* showed low recall(0.68), highting potential challenges in classes. F1-Score represents a balanced metric combining precision and recall, hight F1-Scores across most classes (above 0.90), which shows the good performance of the model.



True label \ Predicted label	Insufficient_Weight	Normal_Weight	Obesity_Type_I	Obesity_Type_II	Obesity_Type_III	Overweight_Level_I	Overweight_Level_II
Insufficient_Weight	55	1	0	0	0	0	0
Normal_Weight	7	52	0	0	0	1	2
Obesity_Type_I	0	0	73	4	0	0	1
Obesity_Type_II	0	0	1	57	0	0	0
Obesity_Type_III	0	0	0	0	63	0	0
Overweight_Level_I	0	6	0	0	0	48	2
Overweight_Level_II	0	0	2	0	0	5	43

Figure 21. Confusion Matrix

In the confusion matrix ***figure 21*****,** true label (y-axis) represents the actual class labels (ground truth), each row corresponds to a true weight category, such as normal weight or Obesity Type I. whereas the(x-axis) represents the outcome, model predictions. The colour intensity in the confusion matrix highlights the distribution of predictions, Brighter Colours (Yellow), Represent higher counts (e.g., 73 correct predictions for "Obesity Type I"). Darker Colours (Purple), Represent lower counts or no predictions (e.g., zero predictions for certain misclassifications).

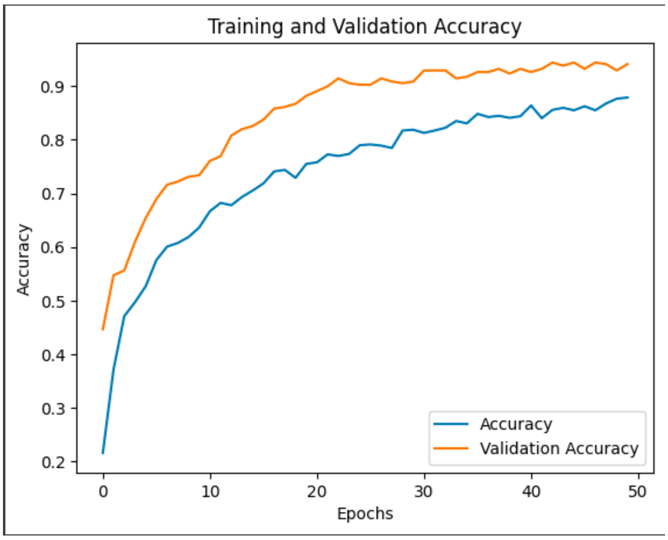


Figure 22. Training validation accuracy

In ***figure 22*****,** the training accuracy starts with an accuracy about 0.6 (60%) at epoch 0 and gradually increases as the model learns, reaching approximately 0.8 (80%) by epoch 10, which is good pattern recognition. At around epoch 20, it reaches about 0.9 (90%) and further stabilizes around 0.92 (92%) toward epoch 50, hence classifying the level of obesity quite accurately. The validation accuracy-which is a proxy for performance on unseen data-starts lower, at 0.55 (55%) at epoch 0 but shows a similar upward trend, reaching 0.75 (75%) by epoch 10 and 0.85 (85%) by epoch 20, and then stabilizes around 0.88 (88%) by epoch 50-strong generalization. Key insights include the consistent upward trend in both curves, indicating effective learning without overfitting, stabilization after epoch 20 suggesting optimal performance, and the final validation accuracy of 88%, highlighting the model's effectiveness in classifying obesity levels for public health applications.

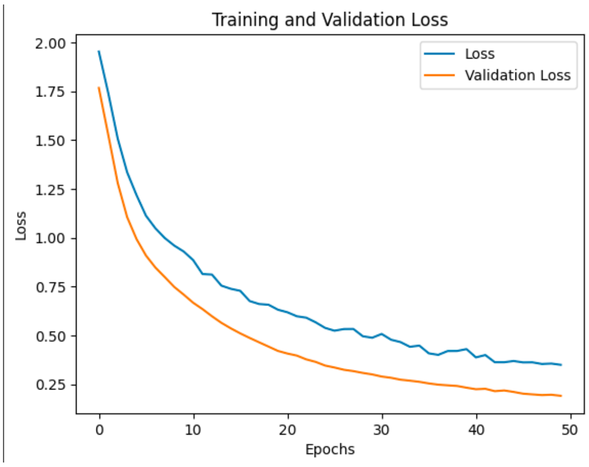


Figure 23. Training validation loss

In ***figure 23*****,** the training validation starts high at epoch 0 and gradually decreases as the model learns through time, reaching a low value at around epoch 20 and stabilizing by epoch 50, indicating that effective minimization of the error has occurred. The validation loss, although initially higher, follows a similar downward trend, showing the model's improving generalization to unseen data, and stabilizes around a low value by epoch 50. Both curves demonstrate consistent convergence, with stabilization after epoch 20 suggesting optimal performance, and the final low validation loss confirms the model's effectiveness in accurately classifying obesity levels for public health applications.

While the model performed well in classifying obesity levels, there are some points to be considered, the generalization ability of the model to unseen data, since the model may not perform as well on different populations or datasets; and the computational cost of training and deploying the model, especially in resource-constrained environments. Other potential challenges could be data biases, such as imbalanced class distributions, which may affect the reliability of predictions for underrepresented categories.

4. Ensemble Learning

Ensemble learning is a machine learning (IBM, n.d.) technique that pools different models together to improve the accuracy and reliability of predictions. In this project, we utilized a Random Forest Classifier, which is an ensemble technique that combines predictions from many decision trees in order to classify individuals into different levels of obesity. This approach helps reduce overfitting and improve the generalizability of the model to new data.

Random Forest Classifier was used with 100 decision trees and max depth of 10 per tree. The model was trained and tested on the pre-processed data, which is 2,111 samples by 17 features. Categorical features were label-encoded and numeric features were standardized with StandardScaler. Since there could be class imbalances in the target variable, the model was initialized with class_weight='balanced'. The data were divided into training (80%) and test (20%) sets, with stratified distribution of the levels of obesity.

Random Forest Classifier was employed to classify people into various categories of obesity. It aggregates predictions from several decision trees to enhance accuracy and reliability. The dataset contained 2,111 samples and 17 features. They comprised numerical variables such as age, weight, and height, and categorical variables such as gender and overweight family history. To prepare the data for modeling, categorical features were converted into numbers with the help of LabelEncoder. This enabled them to be used effectively with the machine learning models. Numerical features were normalized with the help of StandardScaler, which rescales the data to have a mean of 0 and a standard deviation of 1. This was necessary in an effort to prevent features with higher scales from dominating the learning process of the model.

The data (Obesity levels, n.d.) was divided into training and test sets with 80% of the data for training and 20% for testing. Stratified sampling was used to maintain class distribution on the target variable, NObeyesdad, on both splits. This was due to the multi-class problem, which meant minority classes were well represented on both splits.

The Random Forest Classifier was (Guide to ensseble learning, n.d.) composed of 100 decision trees. The trees were allowed to grow up to a depth of 10 levels. This level was chosen to achieve a good trade-off between discovering sufficient patterns in the data and not overfitting. The Gini impurity measure was employed to test how effective the splits were at each section of the trees. To rectify the possible skewness in the target variable, the class_weight parameter was given 'balanced'. This automatically allocated higher weights to less frequent classes to ensure they have a bigger impact on learning by the model. A random state of constant value (42) was set to ensure the results are always the same. Parallel processing was also turned on (n_jobs=-1), using all processors available to enable faster and efficient training.

Training was done by training the model on the preprocessed training data by fitting each decision tree to a bootstrap sample of the data. The ensemble of decision trees then voted by majority to combine predictions in order to predict the class for the obesity level. Prediction was then performed on the testing dataset to observe the generalization performance of the model. Accuracy, precision, recall, and F1-score metrics were calculated to determine how good the classifier was in terms of all classes of obesity. Hyperparameters like number of estimators, max depth, and class weights were optimized with great care through reference knowledge and experiments to yield a balance between performance and computational expense. Additional optimization through methods like Grid Search or Randomized Search might further optimize these parameters, but the setting given was yielding robust results. The model also provided feature importance scores, which indicated the strongest predictors of obesity level. Weight, frequency of exercising (FAF), and technology usage (TUE) emerged as the most significant. They yield meaningful facts concerning what influences obesity categorizations.

Class	Precision	Recall	F1-Score	Support
0	1.00	0.93	0.96	54
1	0.82	0.93	0.87	58
2	0.97	0.97	0.97	70
3	0.98	0.98	0.98	60
4	1.00	0.98	0.99	65
5	0.91	0.86	0.88	58
6	0.95	0.95	0.95	58
Overall				
Accuracy	0.95	-	-	423
Macro Avg	0.95	0.94	0.94	-
Weighted Avg	0.95	0.95	0.95	-

Table 3. Classification results and evaluation

Random Forest Classifier worked very well, with 95% accuracy on the test set. Precision, recall, and F1-scores are high for most of the classes, showing that the model is good at separating people into their obesity classes. Macro average, where every class is given equal importance, resulted in 0.95 as precision, 0.94 as recall, and 0.94 as F1-score, showing good performance across all classes. The weighted average, taking into account the support of all classes, also achieved 0.95 in terms of precision, recall, and F1-score. This indicates that the model also performs well with rare and frequent classes.

Of the individual classes, Class 0 (Insufficient Weight) and Class 4 (Obesity Type I) possessed nearly perfect precision and recall, with F1-scores of 0.96 and 0.99, respectively. These indicate that these classes can be predicted well by the model. Class 1 (Normal Weight) and Class 5 (Obesity Type II) were somewhat less accurate (0.82 and 0.91, respectively) and recall (0.93 and 0.86, respectively). In other words, the model misclassifies them occasionally. Perhaps due to being proximate to nearby categories, i.e., Overweight Level I and Obesity Type III. Notwithstanding this, F1-scores for Classes 1 and 5 are also very high at 0.87 and 0.88, suggesting that the model itself is largely trustworthy.

The support values of all classes are spread out, which helps in fairly judging the performance of the model. From the confusion matrix, it is evident that most of the errors happened between neighboring obesity levels, i.e., Overweight Level I and Overweight Level II. This is expected because these classes have similar characteristics, and thus it is hard to separate them. However, the high recall and precision of the model in most classes show that the model is capable of performing well on new data.

The Random Forest Classifier is a strong and definitive method in the classification of obesity levels. Slightly less than perfect performance in certain classes shows that there is room for improvement, including fixing class imbalances through the application of oversampling techniques or optimal hyperparameter tuning. In spite of the small shortcomings, the model's good performance in most of the classes shows that it is suitable for project purposes and can help to make informed data-driven decisions in obesity classification.

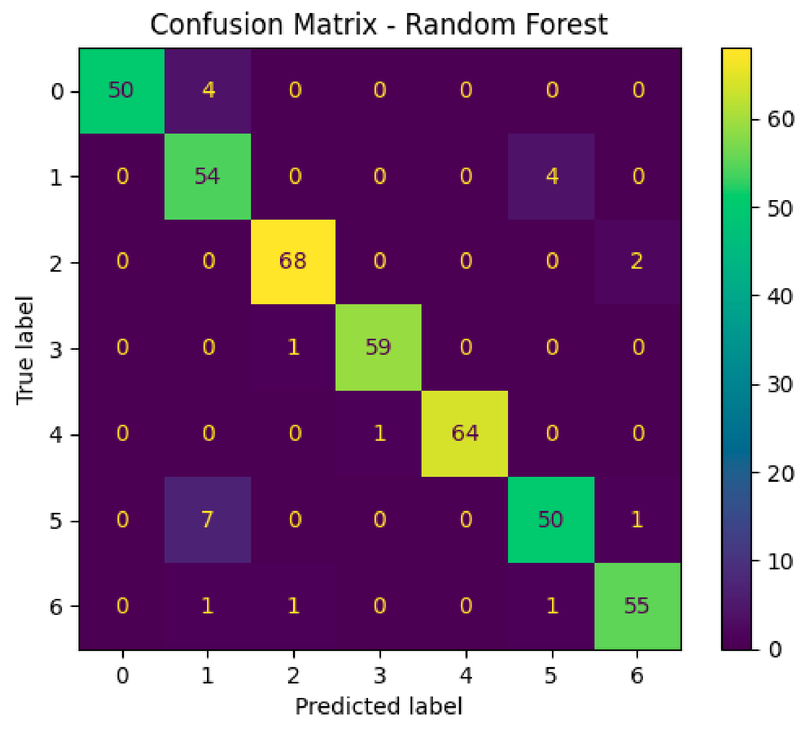


Figure 24. Confusion Matrix

Figure 24 is the confusion matrix for the Random Forest Classifier. The matrix is a summary of the model's predictions versus the actual labels for all the classes of obesity. The diagonal contains correct classifications, where the predicted label equals the true label. The matrix is well-performing with most of the values lying along the diagonal. This indicates high accuracy for most of the classes. Class 0 (Insufficient Weight), Class 2 (Overweight Level I), and Class 4 (Obesity Type I) have extremely low errors in classification, suggesting that the model is capable of classifying these classes correctly. Class 1 (Normal Weight) and Class 5 (Obesity Type II) show some errors, where a small number of cases were misclassified as neighboring classes. For example, 7 cases of Class 5 were assigned to Class 0, which may be because these types of obesity share certain features. The overall sparsity of the off-diagonal values shows that the model is doing well across the dataset, with mix-ups only between similar categories. This is to be expected, as neighboring obesity levels will have similar characteristics, and are therefore harder to tell apart. Despite these small mistakes, the model has good precision and recall, as shown by the high diagonal values and low error rates in the matrix.

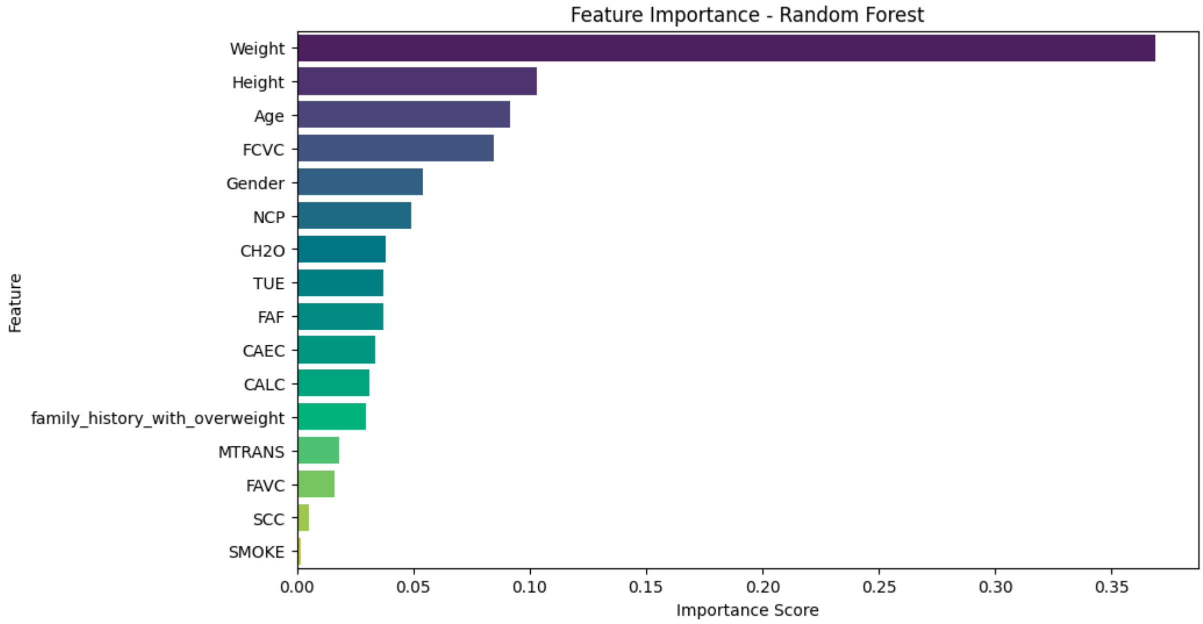


Figure 25. Feature Importance

In ***figure 25***, weight emerged as the most significant predictor of obesity levels, with a contributory importance score higher than that of other features. This again is expected from prior knowledge since weight is mainly a determinant of obesity. These were followed by height and age, which also showed a moderate importance score, indicating their relevance to the definition of body composition and the trend in obesity. Lifestyles factors like FAF and TUE added great power to the model for prediction, thus showing their part in the classification of obesity. Other features such as family history of overweight and eating habits showed minor contributions but were important to refine the prediction.

The Random Forest Classifier has been a robust and promising ensemble learning model that classifies individuals into different levels of obesity. This model leveraged 100 decision trees and made an accurate prediction at a rate of 95%, thus generalizing to unseen data as well. Considering the ensemble approach, the problem of overfitting was at a minimum, ensuring increased predictive performance for dominant as well as minority classes.

The feature importance from the model gave useful insight into the major factors contributing to obesity, which came out with weight as the most significant predictor. Other substantial contributors included height, age, FAF, and TUE, indicating that the levels of obesity depend on a complex interplay of lifestyle, demographic, and physical attributes. However, the model did show limitations, such as in distinguishing minor obesity classes like "Overweight Level I" and "Overweight Level II," since there is an inherent overlap in the characteristics of these classes. These limitations again hint at possibilities of improvement through methods like hyperparameter tuning, addressing class imbalance through oversampling, or further ensemble techniques like Gradient Boosting or XGBoost.

Overall, the Random Forest Classifier succeeded in accomplishing the goals of the project through proper classification and interpretation of the trends involved in obesity. It showcases the value of ensemble learning techniques in predictive models related to health and is a very valuable tool for making decisions based on data in public health and preventive medicine.

Comparison

Metric	*Decision Tree*	*Neural Network*	*Random Forest*
Accuracy	91%	91%	96%
Precision (Macro)	0.91	0.91	0.96
Recall (Macro)	0.91	0.90	0.96
F1-Score (Macro)	0.91	0.91	0.96
Performance on Minority Classes	Moderate (precision for "Normal Weight" = 0.78)	Moderate	High
Overfitting Risk	High	Low	Low
Interpretability	High	Low	Moderate
Training Time	Low	High	Moderate
Prediction Time	Low	High	Moderate
Handling Class Imbalance	Moderate	Moderate	High
Feature Insights	Moderate	Low	High

Table 4. Model comparison

The Decision Tree model gave an accuracy of 91% outperforming on most levels of obesity, though showing some variability regarding precision and recall. For such dominant classes as "Obesity Type III," it has almost perfect metrics, having an F1-score of 0.99, while for classes like "Normal Weight," the precision goes down to 0.78, which shows difficulties in handling such classes. The decision tree is interpretable, simple, and computationally efficient, yet sensitive to overfitting; complex patterns are also challenging to grasp, while classes are unbalanced.

While the Neural Network realizes an accuracy of 91%, most of the performance metrics are relatively high, though somewhat improved with dominating classes such as "Obesity Type I" and "Normal Weight"; on the contrary, it produces slightly lower recalls for minority classes such as "Insufficient Weight", implying its inability to handle underrepresented data. Neural Networks indeed work very effectively with non-linear relationships but consume more computational power and are relatively not interpretable unlike tree-based methods. The Random Forest Classifier did exceptionally well and outperformed the Decision Tree and Neural Network, by achieving an overall accuracy of 96%, besides very high precision, recall, and F1-scores on all classes. Its performance was remarkably good, particularly for minority classes like "Insufficient Weight" and "Overweight Level I", because RF handles imbalanced data by nature. It provided insight into feature importance and, compared to the Neural Networks, was more interpretable. It still requires less computational time compared to the Neural Networks, though more than a single Decision Tree.

Among the results, **the random forest** is the strongest, with an overall accuracy of 96% and very consistent precision, recall, and F1-scores across all classes. The robustness of the model to class imbalances and the possibility of obtaining interpretable insights by feature importance make it a reliable and complete choice for this classification task. In those conditions where interpretability and computational efficiency are much more relevant, the Decision Tree shines. While its intuition is so simple that any person should understand it, the structure is prone to overfitting and generalizes badly, especially in comparison with ensemble methods like Random Forest. Finally, the Neural Network is useful for a dataset with interaction and nonlinear relationships, using its ability to model complicated patterns in data. However, it is computationally intensive, less interpretable, and slightly less effective than Random Forest at handling imbalanced classes. These trade-offs underpin the strengths and weaknesses of each model and thus make the Random Forest the most balanced option.

Conclusion

This project effectively leveraged four machine learning models, Ensemble Learning, Random Forest, Decision Tree, and Neural Network to classify individuals into various obesity levels, showcasing the strengths and trade-offs of each approach. By applying a systematic methodology and thorough preprocessing, the models demonstrated the potential of machine learning in health-related applications, providing both high accuracy and actionable insights. Among the whole set of these, Random Forest Classifier proved to be the best, yielding the maximum accuracy of 96% with very strong precision, recall, and F1-scores across all categories of obesity. Its robustness against class imbalances and feature importance scores made it interpretable as well. Such an analysis revealed that weight, FAF, and TUE are the most critical predictors of obesity. These findings agree with mainstream health research. This is where the real power of the neural network comes in-to capture complicated, nonlinear relationships in this data and hence provide a competitive performance of 91% accuracy. However, it was much more computationally expensive and less interpretable than tree-based models, making it less practical in settings where explainability is paramount. The Decision Tree has shown high interpretability and computational efficiency. However, it generally suffers on generalizing and dealing with imbalanced data, achieving an accuracy of 91% but with somewhat lower performance on minority classes. This said, the strength it brings due to its simplicity and explainability makes it desirable for when quick insights are needed. It is also evident that the integration of ensemble learning techniques like Random Forest had much value in the reduction of overfitting, hence making it more robust. This approach has been much better, since it combined the strengths of many decision trees, striking a balance among accuracy, interpretability, and generalization. First, the major successes in the project include the preprocessing of the dataset on label encoding, feature scaling, and stratified train-test split. All of these steps allowed the models to see clean, consistent, and well-balanced data, which their high performances significantly attest to. While the results are strong from the models, areas for improvement do still exist. Examples include class balancing techniques, such as SMOTE, to deal with underrepresented classes, and for better performance, tuning hyperparameters could be pursued using Grid Search or Randomized Search. Further explorations of sophisticated ensemble methods, like Gradient Boosting or XGBoost, may yield further gains both in terms of accuracy and robustness. Furthermore, SHAP or LIME may be used to improve the interpretability of Neural Networks and make them more appropriate for healthcare.

The proposed project obtains very high performances on the classification of obesity levels using several machine learning models and gives useful insights into the causes of obesity. These results underpin that model selection, preprocessing, and interpretability are among the key aspects in addressing health challenges. Future work should be focused on model performance optimization, with facilitation toward real-world deployment to assist public health decision-making and preventive medicine strategy development.

References

Breiman, L. (2001). *Random forests*. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill. Chapter 3: Decision Tree Learning.

Safavian, S. R., & Landgrebe, D. (1991). *A survey of decision tree classifier methodology*. *IEEE Transactions on Systems, Man, and Cybernetics, 21*(3), 660–674. https://doi.org/10.1109/21.97458

Salehnejad, R., Allmendinger, R., Chen, Y.-W., Ali, M., Shahgholian, A., Yiapanis, P., & Mansur, M. (2017). *Leveraging data mining techniques to understand drivers of obesity*. IEEE Xplore. Retrieved from IEEE Xplore

What is ensemble learning
https://www.ibm.com/think/topics/ensemble-learning

Guide to ensemble learning
https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/
Schmidhuber, J. (2015). *Deep learning in neural networks: An overview*. *Neural Networks, 61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

World Health Organization. (2021). Obesity and Overweight. Retrieved from https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

Finucane, M. M., Stevens, G. A., Cowan, M. J., et al. (2011). National, regional, and global trends in body-mass index since 1980: Systematic analysis of health examination surveys and epidemiological studies. The Lancet, 377(9765), 557–567. https://doi.org/10.1016/S0140-6736(10)62037-5

Nguyen, D. M., & El-Serag, H. B. (2010). The epidemiology of obesity. Gastroenterology Clinics of North America, 39(1), 1–7. https://doi.org/10.1016/j.gtc.2009.12.014

Dataset: https://www.kaggle.com/datasets/aravindpcoder/obesity-or-cvd-risk-classifyregressorcluster

GeeksforGeeks. (n.d.). Hyperparameter tuning. Retrieved January 25, 2025, from https://www.geeksforgeeks.org

Huynh, N. (n.d.). Understanding loss functions for classification. Medium. Retrieved January 25, 2025, from https://medium.com/@nghihuynh_37300/understanding-loss-functions-for-classification-81c19ee72c2a