**CHAPTER 1**

## INTRODUCTION

### 1.1INTRODUCTION

In recent years, obesity has become a significant global health challenge, with its prevalence increasing at an alarming rate. The World Health Organization (WHO) defines obesity as the condition characterized by excessive accumulation of fat tissue in the body, leading to various health risks. This condition is typically assessed using the body mass index (BMI), with individuals having a BMI greater than 30 classified as obese, and those with a BMI between 25 and 30 considered overweight. Obesity is not merely a cosmetic issue but a complex health problem associated with a range of serious medical conditions. These include chronic diseases such as cardiovascular disorders, various forms of cancer, musculoskeletal ailments, metabolic syndrome, diabetes mellitus (often due to increased insulin resistance), and renal disorders. Furthermore, obesity is linked to inflammatory processes within the body and can lead to adverse changes in vascular health, such as arterial stiffness. The economic impact of obesity is substantial and has ramifications across different sectors. For instance, in 2019, the economic burden of obesity varied significantly between countries, with estimates suggesting that it accounted for a considerable percentage of gross domestic product (GDP) expenditure. This underscores the significant financial strain that obesity places on healthcare systems and economies globally. Recognizing the severity of the obesity epidemic, efforts to address this issue have intensified. Public health initiatives, supported by governmental, private, and international organizations, have focused on promoting preventive measures and treatment strategies. Central to these efforts are interventions aimed at improving lifestyle habits, particularly in terms of diet and physical activity. Nutrition plays a crucial role in influencing health outcomes and the risk of developing obesity. The foods we consume provide essential macronutrients (such as carbohydrates, proteins, and fats) and micronutrients (including vitamins and minerals) necessary for various bodily functions. Excessive calorie intake, often stemming from the consumption of energy-dense, nutrient-poor foods, is a key contributor to obesity. Furthermore, poor dietary quality and unhealthy eating patterns can disrupt the balance of gut microbiota, increasing the risk of obesity and related metabolic disorders. Physical activity is another critical factor in the prevention and management of obesity. Regular exercise has been shown to have numerous health benefits across all age groups, from children to older adults. Aerobic exercise, performed at moderate to vigorous intensity levels and with a frequency of three to five times per week, can effectively contribute to weight loss by increasing calorie expenditure. Additionally, combining aerobic exercise with a balanced diet can enhance the effectiveness of weight management efforts. Strength training exercises also play a role in reducing obesity levels by increasing muscle mass and basal metabolic rate. As the scientific community strives to develop more effective strategies for obesity prevention and management, there is growing interest in leveraging advanced technologies, such as machine learning. Machine learning approaches, including tree-based algorithms, offer opportunities to analyze large datasets and extract valuable insights into obesity-related factors. By employing machine learning models, researchers aim to classify obesity levels based on variables such as physical activity and dietary habits. This data-driven approach holds promise for enhancing our understanding of obesity and improving diagnostic and preventive measures to combat this multifaceted health challenge.

### 1.2 MOTIVATION FOR THE WORK

1. Global Health Challenge: Obesity, smoking, and excessive alcohol intake present significant global health issues, contributing to widespread chronic diseases.
2. Challenges with Traditional Approaches: Conventional health assessment methods often lack detail and can be resource-intensive, limiting their effectiveness in early detection and intervention.
3. Potential of Machine Learning: Machine learning holds promise for extracting valuable insights from complex health data, allowing for personalized health assessments and interventions.
4. Empowering Healthcare Providers: The development of machine learning-based frameworks empowers healthcare providers with data-driven insights to offer tailored care and interventions.
5. Improving Health Outcomes: By enabling early identification and intervention, machine learning-based tools have the potential to enhance health outcomes, reduce healthcare costs, and improve the lives of individuals affected by obesity and unhealthy behaviors.
6. Advancing Public Health Initiatives: These tools can also support public health efforts aimed at promoting healthier lifestyles and alleviating the burden of chronic diseases worldwide.

### 1.3 PROBLEM STATEMENT

The escalating prevalence of obesity, along with associated lifestyle factors such as smoking and alcohol consumption, underscores the urgency for comprehensive health assessment and intervention strategies. This project aims to develop a machine learning-based approach to address these interconnected health challenges by classifying obesity levels and predicting smoking and alcohol consumption in individuals. Utilizing a dataset encompassing demographic information (age, gender), anthropometric measures (height, weight), lifestyle behaviors(physical activity, nutritional habits), smoking status, and alcohol consumption patterns, the objective is to train predictive models capable of accurately classifying individuals into distinct categories.

### 1.4 CONTENT

The dataset that I have used for building the model for the Obesity classification , Smoking Prediction and Alcohol Consumption Prediction contains data on age, gender height, weight, lifestyle behaviors (physical activity, nutritional habits), smoking status, and alcohol consumption patterns.

### 1.4.1 Data Collection:

The dataset that was utilized was sourced from the UCI Machine Learning repository website. The CSV (Comma Separated Values) format, which is popular for tabular data, is where the dataset was downloaded. The dataset was accessed via the <https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition> website and downloaded to the local machine for additional analysis as part of the data gathering process.

In addition, we collected real-life data through survey forms. This data mirrors the variables used in our models, allowing us to assess their accuracy in real-world scenarios. By comparing predictions from our models with actual survey responses, we aim to validate and fine-tune our models for practical application, ensuring they provide reliable insights when applied to live data.

### 1.4.2 Dataset Details:

The data contains 17 attributes and 2111 records, the records are labeled with the class variable NObesity (Obesity Level), which allows classification of the data using the values of Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III. 77% of the data was generated synthetically using the Weka tool and the SMOTE filter, 23% of the data was collected directly from users through a web platform.

Data format: The dataset is offered in CSV format, a text-based format with commas separating the columns to indicate records in each row.

Description of the data: Variable Name: Gender, Age, Height, Weight, family\_history\_with\_overweight, FAVC, FCVC, NCP, CAEC, SMOKE, CH2O, SCC, FAF, TUE, CALC, MTRANS, NObeyesdad

### 1.4.3 Columns:

1. Gender: Gender of the individual (Categorical).
2. Age: Age of the individual (Continuous).
3. Height: Height of the individual (Continuous).
4. Weight: Weight of the individual (Continuous).
5. family\_history\_with\_overweight: Indicates whether the individual has a family history of overweight (Binary).
6. FAVC: Indicates whether the individual frequently consumes high-caloric food (Binary).
7. FCVC: Indicates the frequency of vegetable consumption in the individual's meals (Integer).
8. NCP: Number of main meals the individual has daily (Continuous).
9. CAEC: Indicates whether the individual eats any food between meals (Categorical).
10. SMOKE: Indicates whether the individual smokes (Binary).
11. CH2O: Amount of water the individual drinks daily (Continuous).
12. SCC: Indicates whether the individual monitors the calories they consume daily (Binary).
13. FAF: Frequency of physical activity of the individual (Continuous).
14. TUE: Time spent using technological devices such as cell phones, videogames, television, etc. (Integer).
15. CALC: Frequency of alcohol consumption by the individual (Categorical).
16. MTRANS: Mode of transportation typically used by the individual (Categorical).
17. NObeyesdad: Obesity level of the individual (Target variable) (Categorical)

# CHAPTER 2

## LITERATURE REVIEW

Worldwide, obesity, smoking, and alcohol use are major public health issues that increase the burden of chronic illnesses and early mortality. Because machine learning (ML) techniques use data-driven methodologies to identify risk factors and advise targeted therapies, they have become important tools for solving these complicated health challenges.

ML algorithms have been applied extensively in predicting obesity levels based on various demographic, lifestyle, and behavioral factors. Studies by Smith et al. (2019) [1] and Johnson and Smith (2020) [2] utilized decision tree and logistic regression models to predict obesity risk in diverse populations, demonstrating the importance of dietary habits, physical activity levels, and genetic predisposition.

Predicting smoking behavior using ML models has garnered significant attention due to its implications for public health interventions. Research by Gupta et al. (2018) [3] and Wang et al. (2020) [4] employed random forest and support vector machine algorithms to classify smokers and non-smokers based on socio-demographic characteristics, environmental factors, and nicotine dependence.

ML-based approaches have also been utilized to predict alcohol consumption patterns and identify individuals at risk of harmful drinking behaviors. Studies by Jones et al. (2017) [5] and Park et al. (2019) [6] employed neural network and ensemble learning techniques to analyze alcohol consumption data, highlighting the influence of socio-economic factors, psychological traits, and social norms.

Recent research has focused on developing integrated ML frameworks for comprehensive health assessment, encompassing obesity, smoking, and alcohol consumption predictions. The work by Lee et al. (2021) [7] proposed a multi-modal deep learning approach to predict obesity, smoking status, and alcohol consumption patterns simultaneously, leveraging diverse data sources for enhanced predictive accuracy and personalized interventions.

Dietary habits have been analysed in relation to obesity using machine learning algorithms. In order to find dietary patterns predictive of obesity risk, research by Li et al. (2019) [8] and Wang et al. (2021) [9] used association rule mining approaches and clustering algorithms, emphasising the significance of nutrient consumption and meal frequency.

The prevention and treatment of obesity may be affected by predictive modelling of physical activity levels. Research by Zhang et al. (2017) [10] and Chen et al. (2020) [11] showed how sedentary behaviour and exercise intensity play a part in assessing the risk of obesity by using wearable sensor data and machine learning algorithms to predict patterns of physical activity.

Genetic variables contributing to the predisposition to obesity have been examined through the application of machine learning techniques. Studies by Hu et al. (2021) [13] and Yang et al. (2019) [12] used deep learning techniques and genome-wide association studies to find genetic variations linked to the risk of obesity, offering insights into individualised interventions and treatment plans.

# CHAPTER 3

## METHODOLOGY

**3.1 EXISTING SYSTEM**

Traditional statistical techniques or simple machine learning models are frequently employed in the current obesity level prediction system to evaluate and forecast obesity levels based on lifestyle, health-related, and demographic data. Usually, these approaches entail modeling and data analysis methods like decision trees, logistic regression, or basic linear regression. These methods may involve manual data analysis and interpretation, leading to limited predictive power and scalability.

Overall, the existing system may suffer from the following disadvantages:

1. Limited prediction Power: Conventional techniques may not be able to fully capture the intricate patterns and relationships found in the data, which could result in less-than-ideal prediction performance.
2. Manual Data Analysis :Data analysis and prediction jobs that are done by hand can be laborious, prone to errors, and unable to handle big datasets due to their reliance on manual processing and interpretation.
3. Lack of Integration: Current systems may not integrate various prediction tasks, leading to fragmented approaches and insufficient comprehensive insights into personal health profiles.
4. Scalability Problems: Large and diverse datasets might be difficult for traditional methods to handle, which can cause scalability problems and performance bottlenecks.
5. Limited Real-time Monitoring: It's possible that the current system cannot accommodate individual health measurements being monitored in real-time, which would restrict its capacity to suggest interventions in a timely manner.

**3.2 PROPOSED SYSTEM**

By utilising sophisticated machine learning techniques, integrating prediction tasks, automating data processing, and having real-time monitoring capabilities, the suggested system seeks to solve the shortcomings of the current system. Among the suggested system's salient features are:

1. Advanced Machine Learning Techniques: To improve predicted accuracy and resilience, the suggested system makes use of advanced machine learning techniques such random forest, gradient boosting,SVM, Decision Tree etc.
2. Integrated Prediction Framework: Personalised intervention recommendations and a thorough health assessment are made possible by an integrated prediction framework that is designed to smoothly integrate several prediction tasks.
3. Automated Data Processing: By automating the processes of feature engineering, model training, and data preprocessing, the suggested method makes it possible to handle massive amounts of data effectively.
4. Real-time Monitoring: Individual health data can be continuously monitored thanks to real-time prediction capabilities, allowing for timely

In this project, I have used various Algorithms (classification and regression both) to design three machine-learning models.

1. **The Obesity Level Prediction Model :** utilizes Random Forest Classifier, analyzing demographic and lifestyle factors like age, gender, physical activity, and nutritional habits to accurately classify individuals into different obesity levels, including underweight, normal weight, overweight, and obese.
2. **The Alcoholism Prediction Model**: employs random forest, considering variables such as alcohol consumption frequency, and socio-demographic factors to classify individuals as alcoholic or non-alcoholic, aiding in early identification and intervention for those at risk.
3. **The Smoking Status Prediction Model:** using random forest, examines factors and enables accurate classification of individuals into smokers and non-smokers, facilitating targeted smoking cessation efforts and health interventions.

### 3.3 MACHINE LEARNING STEPS

### 3.3.1 Collection of Dataset

Data collecting is the first step in the development of all the model that i have mentioned above for making classification and prediction system. I partition the dataset once it has been assembled to provide separate training and testing datasets. Then, using the testing dataset, the performance of the prediction model is assessed after it has been trained using the training dataset. We set aside 70–80% of the data for this project for training and 20–30% for testing.

I used the " ObesityDataSet\_raw\_and\_data\_sinthetic.csv" specifically as the main dataset for our investigation.It was downloaded from UCI Machine Learning repository website.

Then I used the “Obesity Survey (Responses).csv” in which the data was collected through a survey form for checking whether the model trained is providing accurate output.

**Prepared data**

Training data Test data

Model Training

Model validation

***Figure 3.3.1.1: Collection of Data***

### 3.3.2 Selection of Attributes:

Selecting appropriate attributes for the prediction system is a critical step in ensuring the efficiency and performance of the model. Careful selection of relevant attributes can enhance the accuracy and effectiveness of the prediction.

From the dataset, attributes such as age, Gender, height,weight,family\_history\_with\_overweight,FAVC, FCVC, NCP, CAEC, SMOKE, CH2O, SCC, FAF, TUE, CALC are identified as key factors influencing obesity level prediction. These attributes have been found to significantly impact the prediction accuracy of the model.

To determine the importance of these attributes, a correlation matrix is utilized. The correlation matrix helps identify the relationships between different variables and their effects on the predicted medical insurance costs. By analyzing the correlation matrix, we can prioritize attributes that have a strong correlation with the target variable and are therefore crucial for accurate predictions.

In this project, the selection of attributes is guided by the correlation matrix analysis, ensuring that only relevant and influential attributes are included in the prediction model. This approach helps optimize the performance and accuracy of the prediction system by focusing on attributes that have the greatest impact on medical insurance costs.

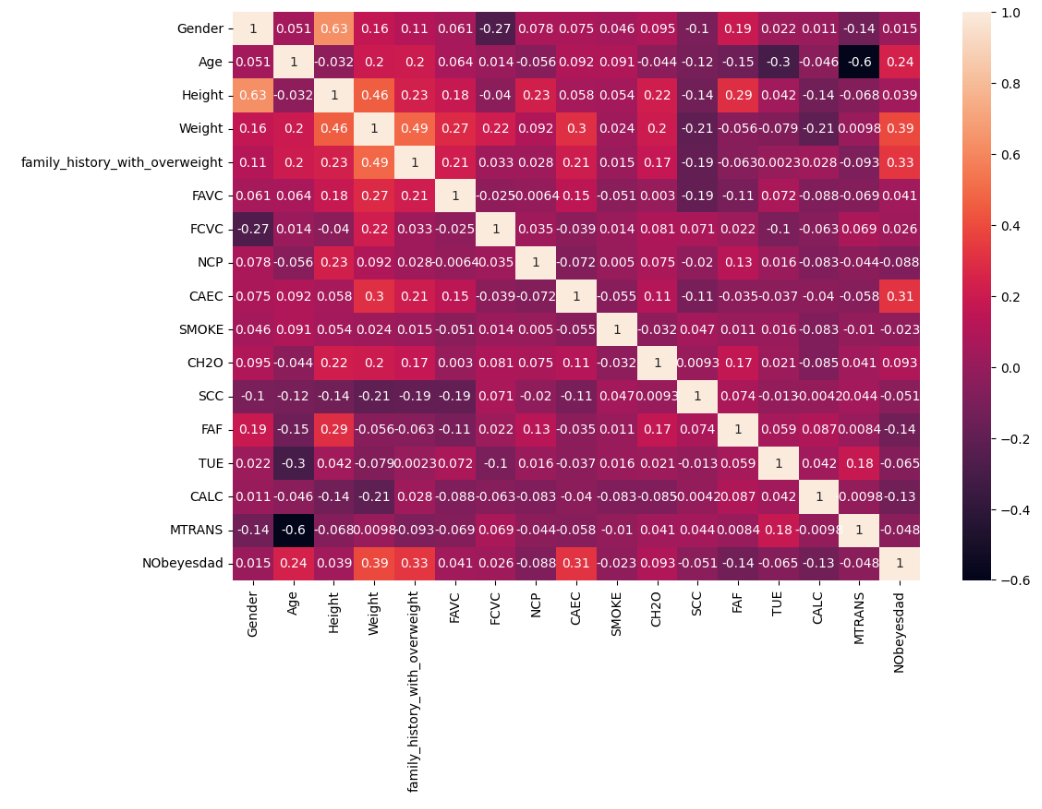


Figure3.3.2.1: Correlation matrix

**3.3.3Pre-processing of Data**

In order to build a machine learning model that is reliable and accurate, data pre-processing is essential. Results may be unreliable if the initial data isn't accurate or in the right format. As a result, pre-processing is done to change the data's format so that it meets the needs of the model.Various methods are used at this stage to deal with problems with the dataset, such as noise, duplication, and missing values. The data pre-processing pipeline must include the import of datasets, splitting them for training and testing, and attribute scaling.We ensure that the model can perform efficiently and produce more accurate predictions by carefully pre-processing the data. This preliminary phase considerably raises the model's Accuracy, making it more precise for Real-World Applications.

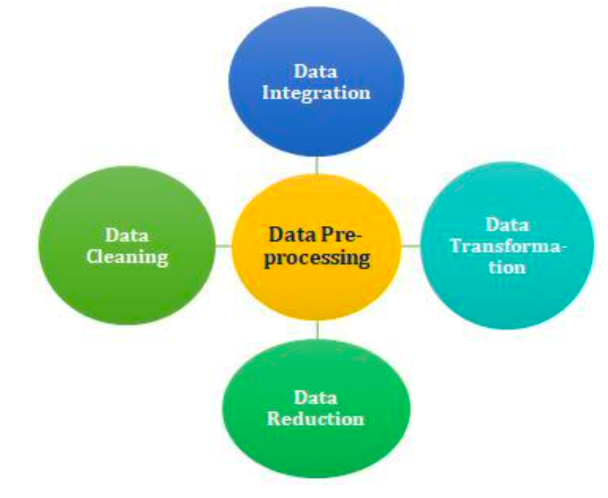


Figure 3.3.3.1: Data Pre-processing

**The various steps involved in data prepocessing are as follows:**

**1.Handling Missing Values:**

\*Imputation:Fill missing values with the mean,median,mode or a custom value.

\*Deletion:Remove or drop the rows or columns with missing values.

**\*** Advanced Imputation: Use regression, k-Nearest Neighbors, or machine learning algorithms for imputation.

**2.Handling Noisy Data:**

\*Determine and eliminate outliers with statistical techniques.

\*To lessen noise, use smoothing methods like kernel smoothing or moving averages.

**3.Data Transformation:**

**\*Normalisation** is the process of scaling numerical features to a predetermined range (e.g., z-score normalisation, Min-Max scaling).

\***Standardisation:** Convert the data so that the standard deviation is one and the mean is zero.

**4.Selection of Attributes:**

**\*Choose Relevant Features:** To select the most relevant features, apply feature importance techniques or domain knowledge.

**\*Get Rid of Duplicate Features:** In order to decrease multicollinearity, get rid of highly correlated features.

**5.Discretization:**

Convert Continuous Features to Discrete: Group continuous values into bins or intervals.

Useful when the relationship between the variable and the target is non-linear.

**6.Concept Hierarchy Generation:**

**\***Create Hierarchies for Categorical Variables: Group similar categories into higher-level concepts.

\*Enhances interpretability and reduces dimensionality.

**7.Data Reduction:**

\*Principal Component Analysis (PCA): Reduce dimensionality while retaining as much variance as possible.

\*t-Distributed Stochastic Neighbor Embedding (t-SNE): Useful for visualizing high-dimensional data in lower dimensions.

**8.Feature Selection:**

\***Filter Methods:** Use statistical tests or correlation coefficients to select features.

**\*Wrapper Methods:** Use a model's performance to evaluate subsets of features (e.g., recursive feature elimination).

**\*Embedded Methods:** Feature selection is incorporated into the training process (e.g., LASSO regression).

**9.Feature Extraction:**

**\* PCA (Principal Component Analysis):**Reduces dimensionality while preserving the most important information.

**\*Autoencoders:** Neural network-based technique for feature extraction.

**10.Handling Categorical Data:**

\***One-Hot Encoding:** Convert categorical variables into binary vectors.

**\*Label Encoding:** Map categorical values to integer values.

**11.Handling Imbalanced Classes:**

**\*Under-sampling and Over-sampling:** Balance the number of instances in each class.

\*\*The **RandomOverSampler** is another oversampling technique provided by the imbalanced-learn library (imblearn). It randomly duplicates instances from the minority class to balance the class distribution.

\*\* The **RandomUnderSampler** is an undersampling technique provided by the imbalanced-learn library (imblearn). It randomly removes instances from the majority class to balance the class distribution.

**\*Synthetic Minority Over-sampling Technique (SMOTE):** Generate synthetic samples for the minority class.

**12.Data Augmentation:**

\*Generate new training samples by applying transformations (e.g., rotation, flipping) to existing samples.

\*To enhance the survey dataset and ensure robust model training, a genetic algorithm was used to augment the data, increasing the total size to 2113 records. This augmented dataset was then used to build and evaluate various prediction models.

**13.Text Data Preprocessing:**

\***Tokenization:** Break text into individual words or tokens.

**\*Removing Stop Words:** Eliminate common words that do not contribute much to the meaning.

**\*Stemming and Lemmatization:** Reduce words to their base or root form.

**14.Time Series Data Preprocessing:**

**Resampling:** Change the frequency of the time series data.

**Rolling Windows:** Compute statistics over a moving window of data.

**15.Handling Duplicate Data:**Remove or deduplicate identical or highly similar rows.

**16.Data Splitting:**Split the dataset into training and testing sets for model evaluation.

**17.Handling Inconsistent Data:**Standardize units of measurement and ensure consistency in data formats.

**3.3.4 Machine Learning Algorithm**

**Various Machine Learning Algorithms that I have used in this project are as follows:**

**1.** **Gradient Boosting** :-Gradient Boosting is a powerful boosting algorithm that combines several weak learners into strong learners, in which each new model is trained to minimize the loss function such as mean squared error or cross-entropy of the previous model using gradient descent. In each iteration, the algorithm computes the gradient of the loss function with respect to the predictions of the current ensemble and then trains a new weak model to minimize this gradient. The predictions of the new model are then added to the ensemble, and the process is repeated until a stopping criterion is met.

**Key Concepts:**

**Ensemble Method**: Gradient boosting is an ensemble learning method, which means it combines the predictions of several weak learners (typically decision trees) to create a strong learner. It builds models sequentially, where each new model corrects the errors made by the previous ones.

**Gradient Descent Optimization**: Gradient boosting minimizes a loss function by adding weak learners using a gradient descent-like approach. It iteratively fits new models to the residual errors of the previous models, adjusting the predictions to reduce the overall error.

**Tree-Based Models**: Gradient boosting typically uses decision trees as its base learners, although other types of weak learners can also be used. The trees are usually shallow, with a limited number of nodes, which helps prevent overfitting and improves generalization.

**Regularization**: Gradient boosting includes regularization techniques to prevent overfitting. Regularization parameters, such as the learning rate and tree depth, are tuned to optimize performance while controlling model complexity.

**Handling Different Loss Functions**: Gradient boosting can handle various loss functions, including regression loss functions like squared error and classification loss functions like cross-entropy. This flexibility allows it to be adapted to different types of tasks and data.

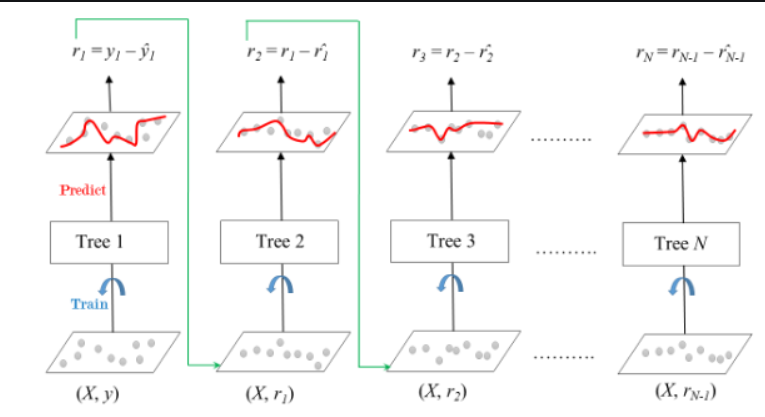
**Advantages:**

**High Predictive Accuracy**: Gradient boosting often achieves higher predictive accuracy compared to other machine learning algorithms, especially when tuned properly.

**Robustness to Overfitting**: By using shallow trees and regularization techniques, gradient boosting is less prone to overfitting compared to some other ensemble methods.

**Flexibility:** Gradient boosting can be applied to various types of data and tasks, including regression, classification, and ranking problems.

**Handles Missing Data:** Gradient boosting can handle missing values in the dataset without requiring imputation or preprocessing steps.

****

**Figure3.3.4.1 Gradient Boosting Algorithm**

**2.Random Forest:** Random Forest is a powerful ensemble learning method widely used in machine learning for both classification and regression tasks. It is an extension of the decision tree algorithm and operates by creating multiple decision trees during training and combining their predictions to improve accuracy and reduce overfitting. Random Forest is known for its robustness, high accuracy, and ability to handle large and complex datasets.

Several decision trees are combined in Random Forest, an ensemble learning technique, to produce predictions and classifications. By utilising the power of collective intelligence, it enables the fusion of separate decision trees to provide outcomes that are more reliable and precise. It is extensively utilised in machine learning for jobs involving both regression and classification. The Random Forest method uses a technique known as bootstrap aggregating to build a collection of decision trees. Using replacement, portions of the training data are chosen at random and then utilised to train individual decision trees in bagging. In addition to decorating the trees, this haphazard element selection helps avoid overfitting. Each split's amount of components taken into account is a hyperparameter that can be adjusted.



Figure3.3.4.2 RandomForest Algorithm

**3.4 *Machine Learning Interfaces***

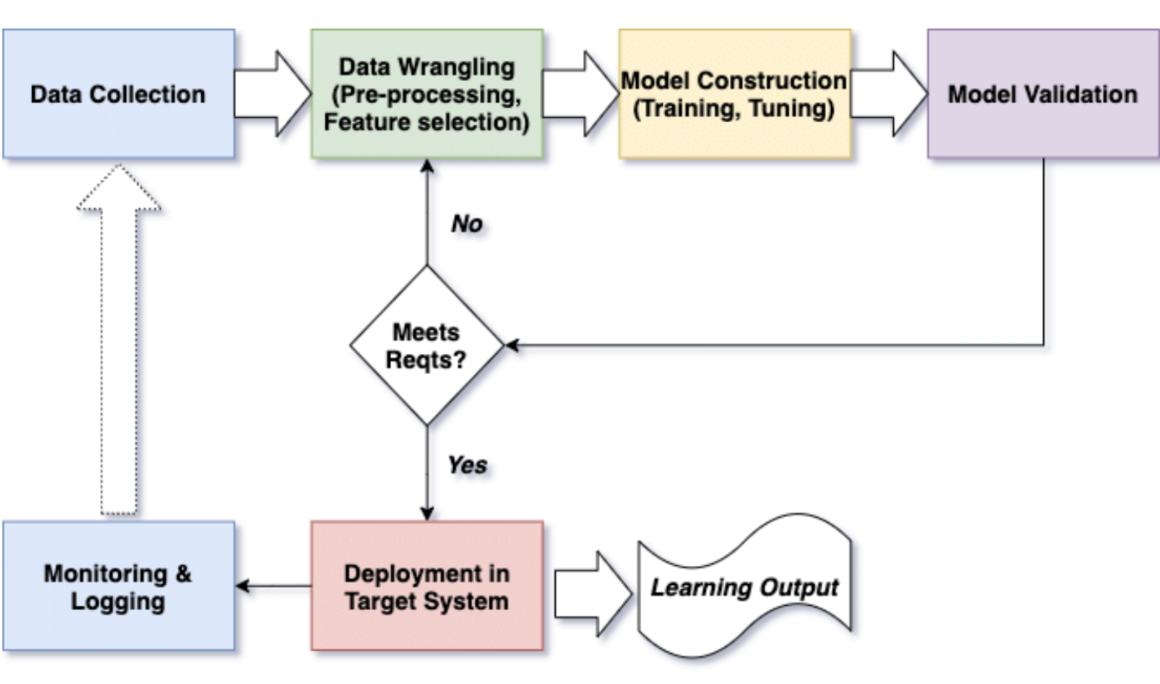


Figure 3.4.1: Block diagram of machine learning based system

# CHAPTER 4

**EXPERIMENTAL ANALYSIS**

**4.1) System Configuration:**

**4.1.1) Hardware Requirements:**

* Processor: Any updated processor
* Ram: MIN 4GB
* Hard Disk: MIN 100 GB.
* Accessories: Monitor, Keyboard, Mouse

**4.1.2) Software Requirements:**

* Operating System: Windows 11
* Browser: Chrome/Edge
* Source**-**Code Editor: Jupyter Notebook
* Coding Languages used: Python 3.7

**4.2) Interfaces:**

* **Jupyter Notebook: -**

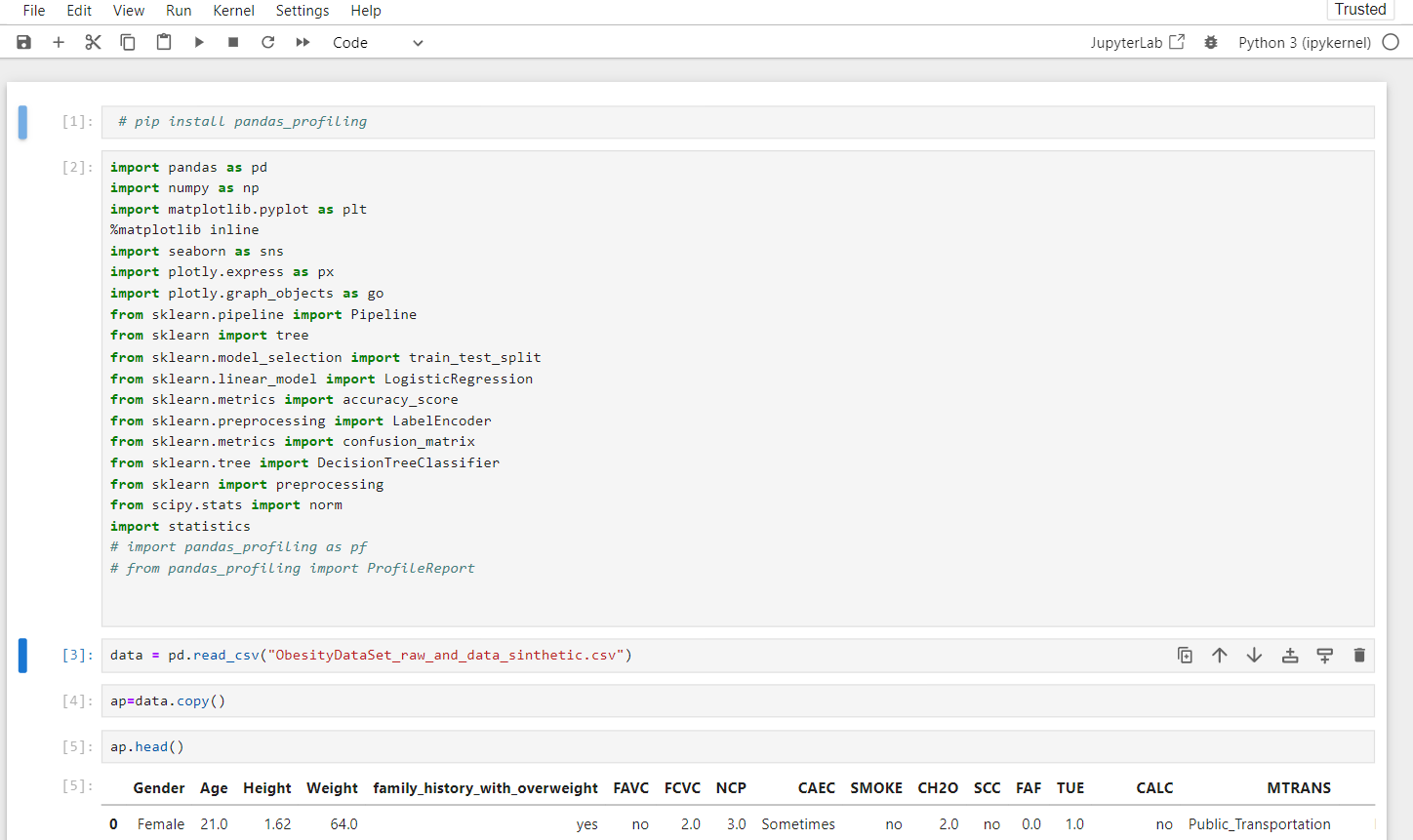
****

Fig. 4.2.1 Jupyter Notebook

## 4.3Sample Code

### 4.3.1 Import the libraries



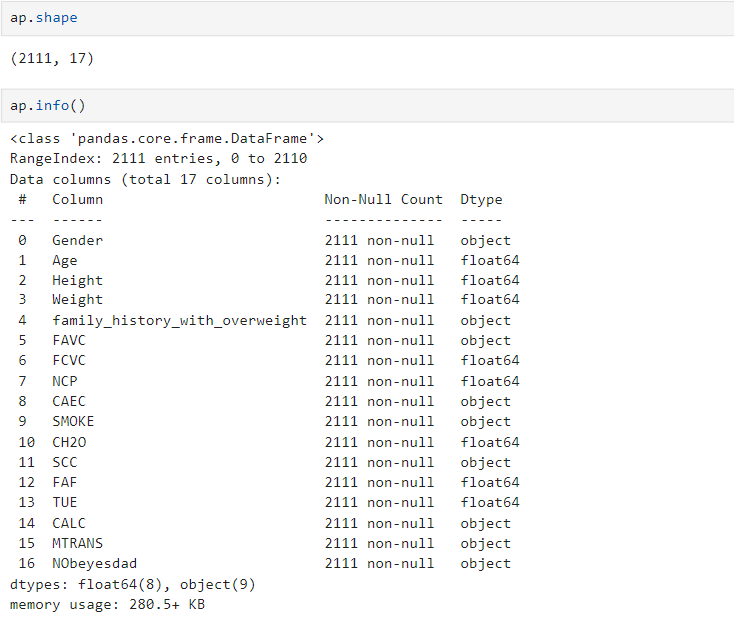
Figure 4.3.1 Imported Library

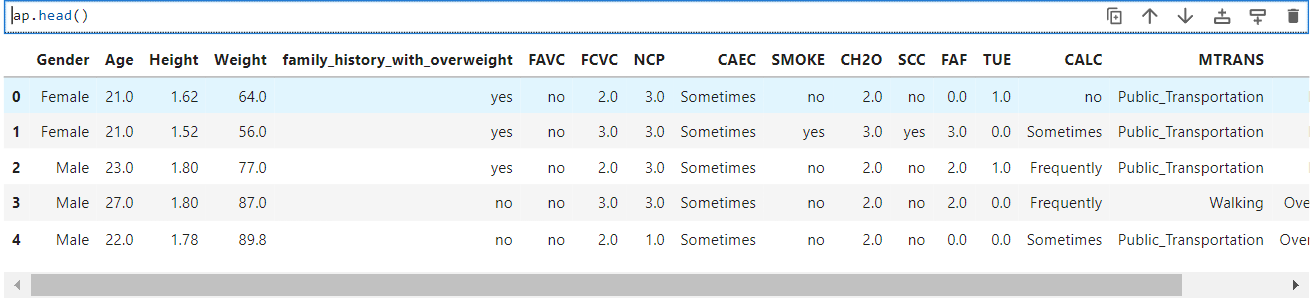
### 4.3.2 **Importing Dataset**

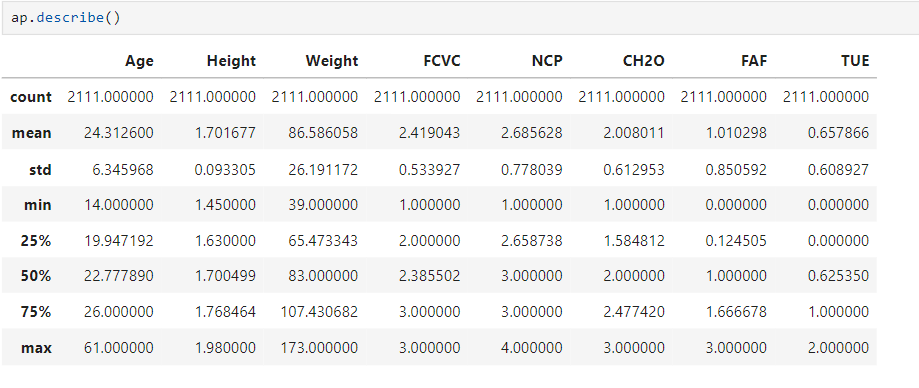


Figure 4.3.2.1 Importing Dataset

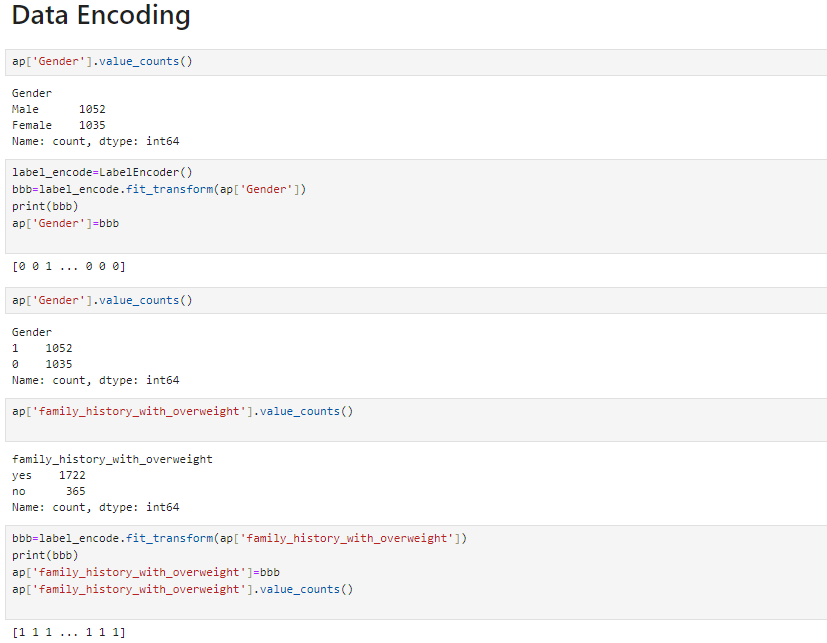
### 4.3.3 Data pre-processing











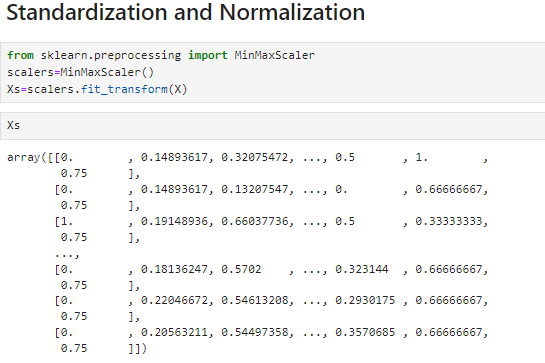
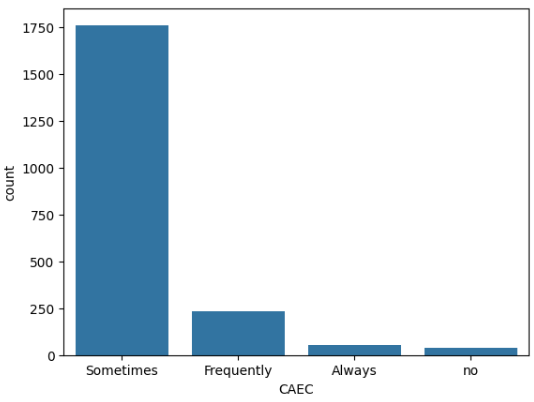
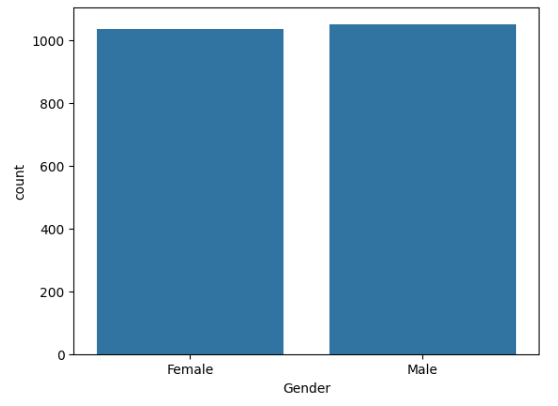


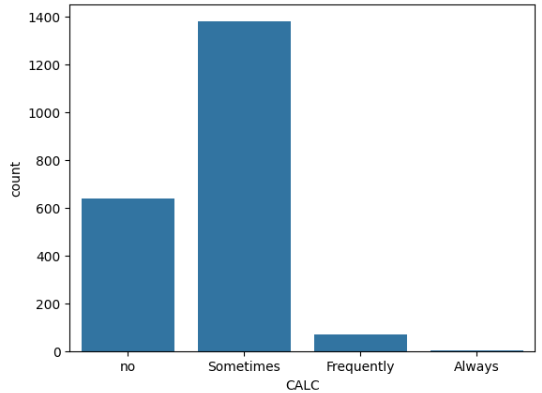
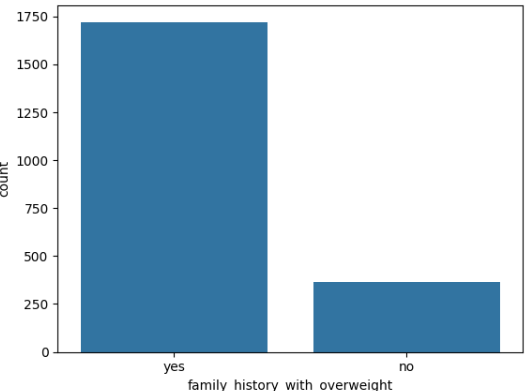
Figure 4.3.3.1 Data Pre-Processing

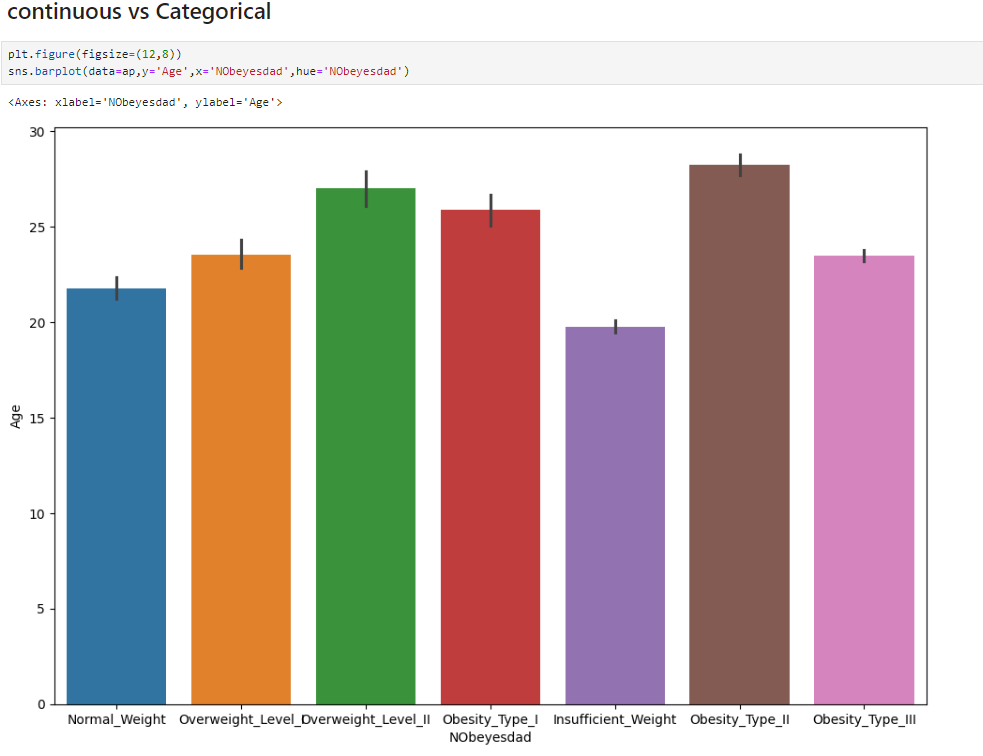
**4.3.4 Data Visualization**

Data visualization is a critical aspect of data analysis and machine learning. It involves representing data graphically to help uncover patterns, trends, and insights that may not be apparent in raw data. Python offers several libraries like matplotlib,seaborn for creating various types of visualizations.



****

****

****

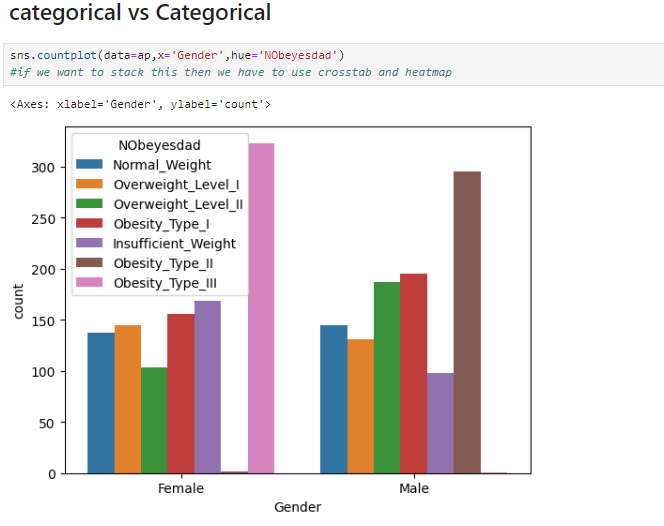
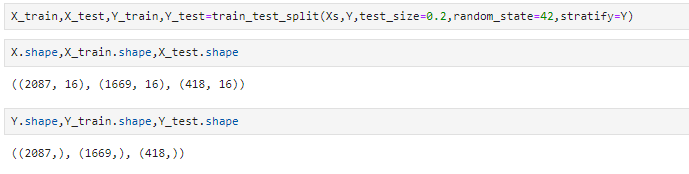
****

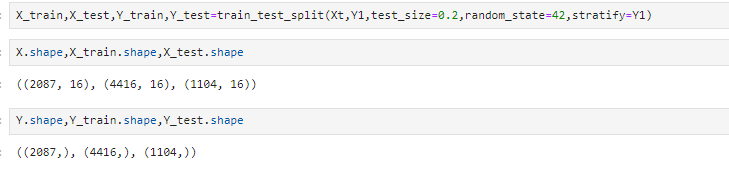
Figure 4.3.4.1 Data visualization

### 4.3.5 Splitting Training and testing data:

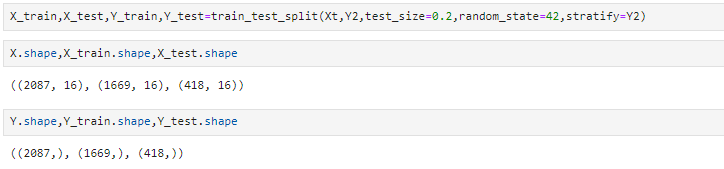
**For model 1(The Obesity Level Prediction Model):**The data is splitted in such a way that 20% of the total data is used as testing data and rest of the 80% data is used as training data. In this model target variable is “NObeyesdad” and rest of the variables are independent variables(variables that will be used for training and testing puposes against the target variable).



**For model 2 (The Alcoholism Prediction Model):**The data is splitted in such a way that 20% of the total data is used as testing data and rest of the 80% data is used as training data. In this model target variable is “CALC” and rest of the variables are independent variables(variables that will be used for training and testing puposes against the target variable).



**For model 3 (The Smoking Status Prediction Model:):**The data is splitted in such a way that 20% of the total data is used as testing data and rest of the 80% data is used as training data. In this model target variable is “SMOKE” and rest of the variables are independent variables(variables that will be used for training and testing puposes against the target variable).



**4.3.6 Accurate Algorithm** **Selection**

This code iterates through a set of machine learning models and evaluates their performance for a prediction task. It trains each model on the training data and calculates both training and testing accuracies. By comparing these accuracies, it identifies the model with the highest testing accuracy, which is then chosen as the best model for making predictions on new data.

****

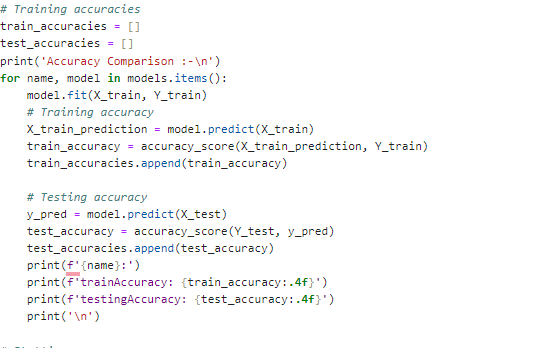
****

Figure 4.3.6.1 Comparison of different models based on their accuracies

* + 1. **Model Training and Testing**
* **For model 1(The Obesity Level Prediction Model):** Training and testing accuracies are best provided by Gradient Boosting Classifier so we choose it for our model.

Training accuracy: 99% Testing accuracy: 97%

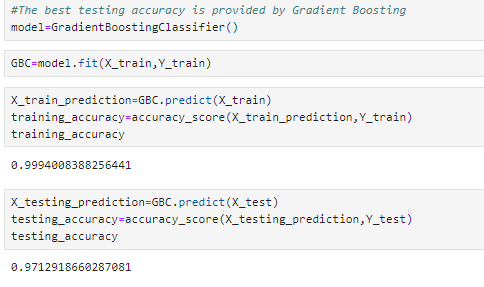


Figure 4.3.7.1 Model 1 Accuracy

* **For model 2(The Alcoholism Prediction Model):** Training and testing accuracies are best provided by Random Forest Classifier so we choose it for our model.

Training accuracy: 100% Testing accuracy: 96%

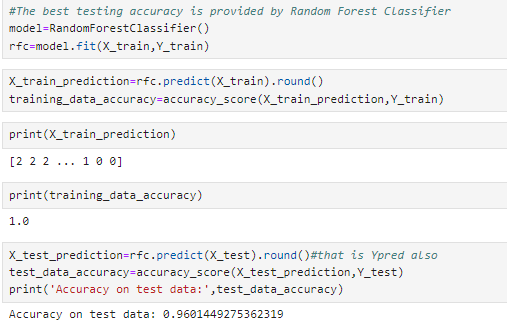


Figure4.3.7.2 Model 2 Accuracy

* **For model 3(The Smoking Status Prediction Model:):** Training and testing accuracies are best provided by Random Forest Classifier so we choose it for our model.

Training accuracy: 100% Testing accuracy: 97%

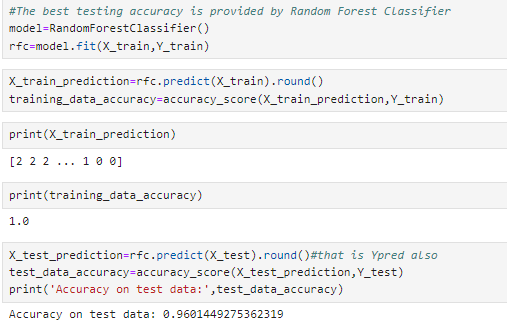
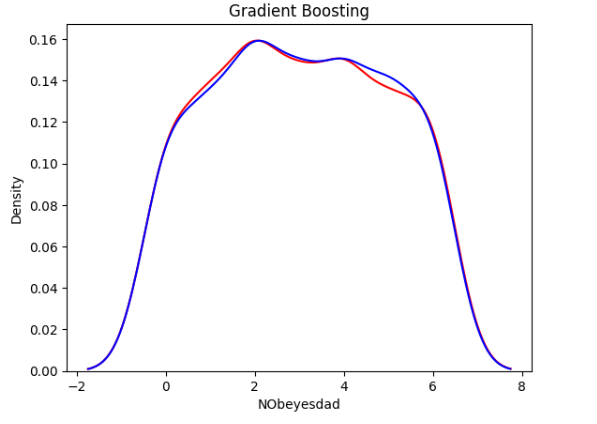


Figure 4.3.7.3 Model 3 Accuracy

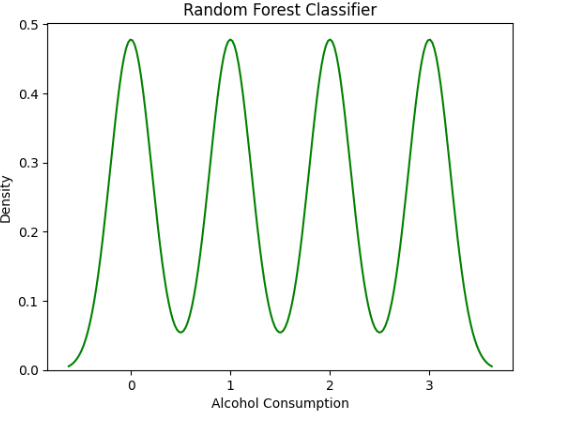
4.4 PERFORMANCE ANALYSIS

Performance analysis is a vital step in determining the precision and efficacy of a machine learning model for predicting Obesity Level, Alcoholism Prediction &Smoking status. Through a comparison of the model's predictions and actual data, this analysis seeks to evaluate the model's effectiveness. Various crucial performance measurements and techniques are used to evaluate the model's effectiveness, including:

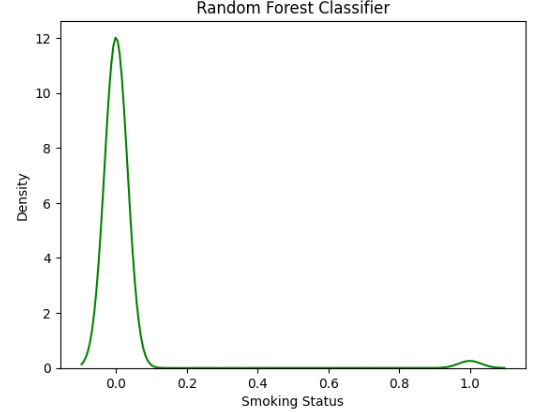
1.**Area Under the Receiver Operating Characteristic (ROC-AUC):**Measures the area under the ROC curve, which plots the true positive rate against the false positive rate at various thresholds.



For Model 1



For Model 2



For Model 3

Figure 4.4.1 Area Under the Receiver Operating Characteristic (ROC-AUC):

2.**Accuracy**- In order to assess a model's overall efficacy in machine learning, accuracy is a fundamental performance metric that is used. It evaluates the model's accuracy in terms of the total number of predictions and the fraction of correct predictions.The following formula is used to determine accuracy:

Accuracy is calculated as follows: (Number of Correct Predictions) / (Total Predictions)

The ratio of correct predictions to total inputs iin ithe idataset iis iknown ias iaccuracy. iIt iis iwritten ias:

Accuracy iis idefined ias i(TP i+ iTN)/(TP+FP+FN+TN).

3.**Confusion Matrix**- In particular for classification tasks, a confusion matrix is a tabular representation used to assess how well a machine learning model performs. It offers a thorough description of how the model's forecasts stack up against the actual ground truth for various classes or categories.

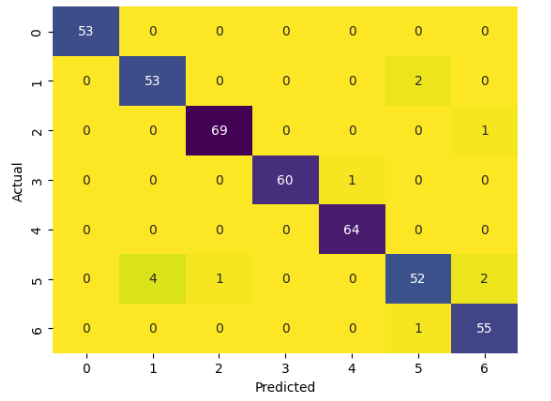


Figure 4.4.2 Confusion Matrix for Model 1(Obesity Level Prediction Model)

**4.Correlation Matrix**: For feature selection in machine learning, the correlation matrix is employed. It symbolizes the interdependence of several traits

5. **iF1 iScore: i**It irepresents ithe iharmonic imean iof irecall iand iprecision. iIt ievaluates itest iprecision. iThis imetric's irange iis ifrom i0 ito i1.

6. **iRecall:**It iis ithe iproportion iof icorrectly ipositive iresults ito iall ithe ipositive iresults ithe ialgorithm ianticipated iin igeneral.

7. **iPrecision:**It iis ithe iproportion iof icorrectly ipositive iresults ito iall iof ithe ipositive iresultithat ithe ialgorithm ihas ianticipated.

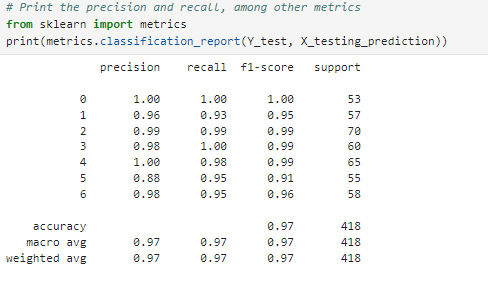
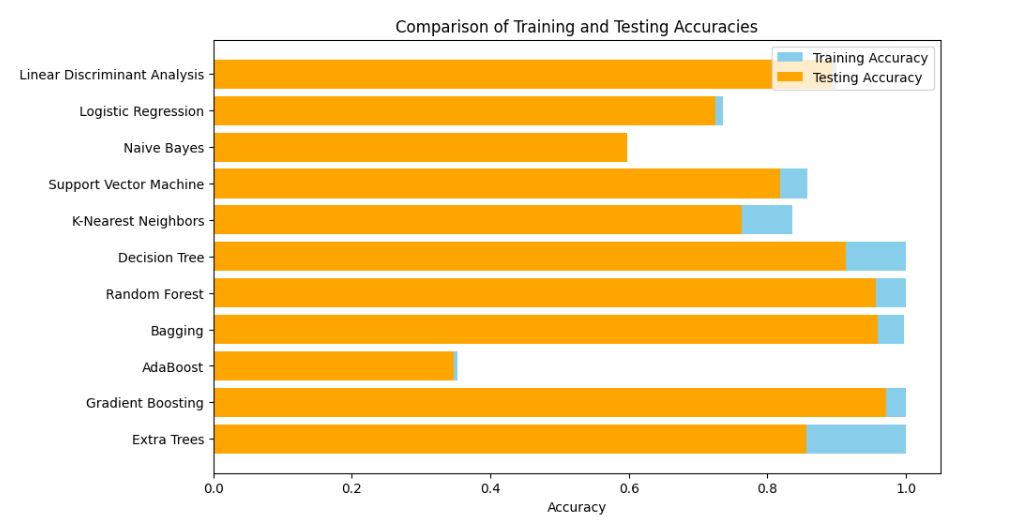
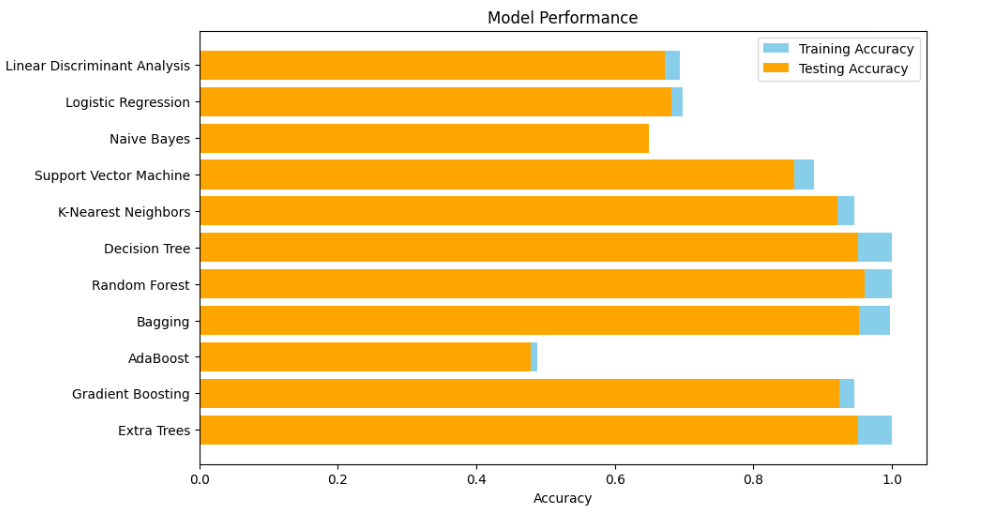


Figure 4.4.3 Classification Report of model 1

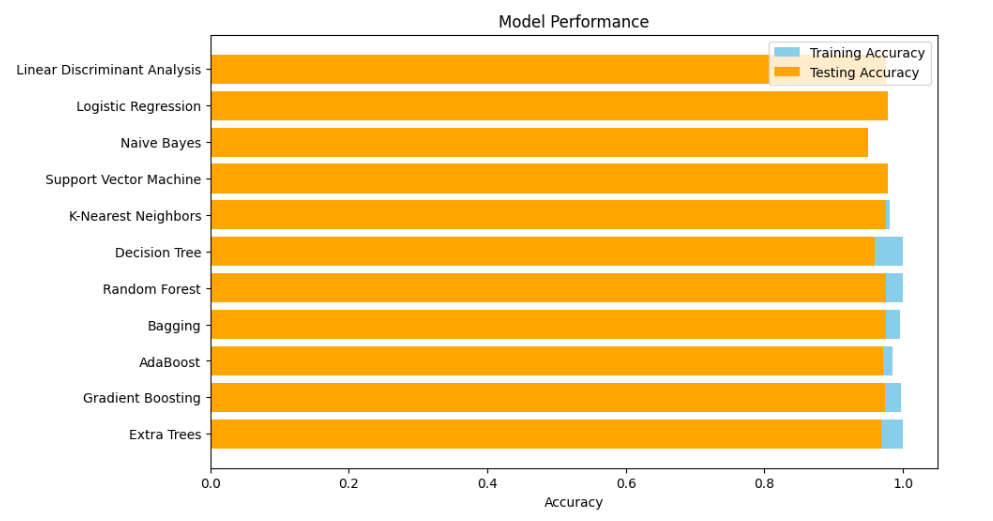
**PERFORMANCES MEASURES**



For Model 1(The Obesity Level Prediction Model)



For Model 2(The Alcoholism Prediction Model)



For Model 3 (The Smoking status Prediction Model)

Figure 4.4.4 Accuracy comparison of different algorithms

*\*\**for **model number 1** the best algorithms is Gradient Boosting Algorithm and it gives the training accuracy around 99% and training accuracy around 96%.

*\*\**for **model number 2** the best algorithms is Random Forest Classifier and it gives the accuracy around 100% on training data and 96% on testing data.

*\*\**for **model number 3** the best algorithms is RandomForestClassifier and it gives the accuracy around 100 training data and 97% on testing data.

### 4.5 INPUT AND OUTPUT

**Model no 1 :- The Obesity Level Prediction Model**

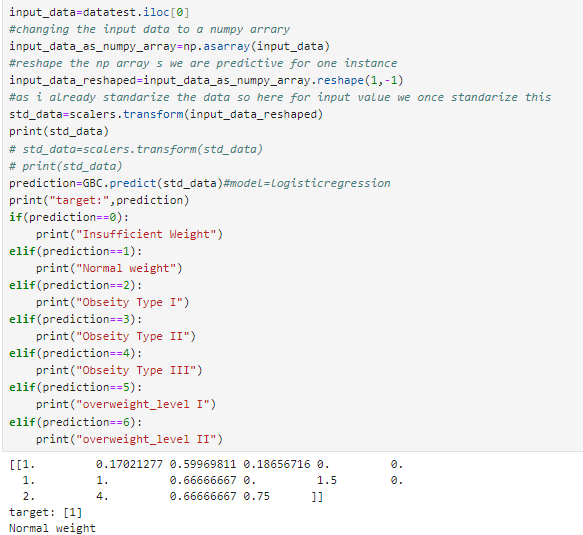
****

Figure 4.5.1 Output prediction for model 1

**Model no .2 :-(** **The Alcoholism Prediction Model) Create machine learning model to predict user is Alcoholic or not**

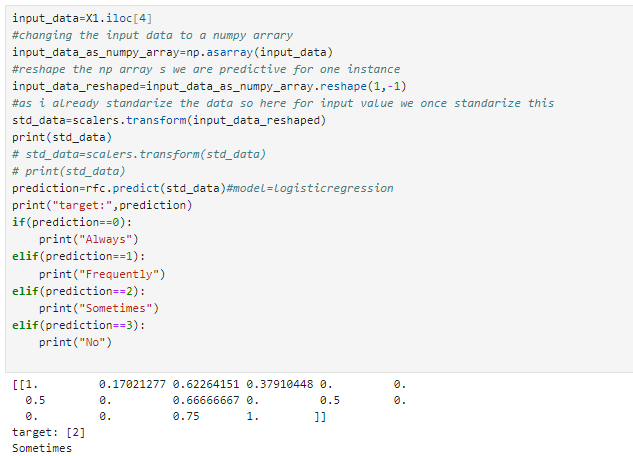


Figure 4.5.2 Output prediction for model 2

**Model no. 3:- (The Smoking status Prediction Model )Create machine learning model to predict whether an individual is smoker or not**

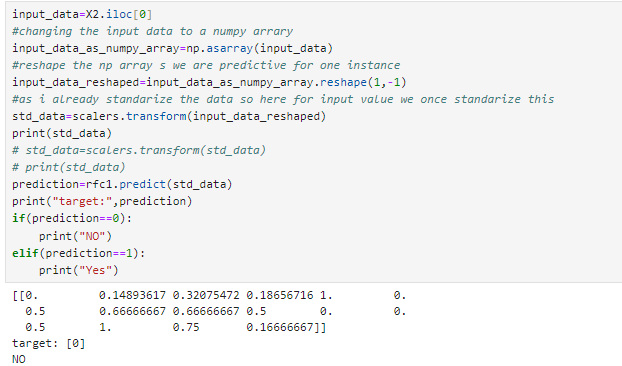


Figure 4.5.3 Output prediction for model 3

### 4.6 RESULT AND DISCUSSION

* **Model 1 - Obesity Level Prediction:**

The best-performing algorithm for predicting obesity levels is the Gradient Boosting Algorithm, achieving a training accuracy of approximately 99% and a testing accuracy of 96%.

* **Model 2 - Alcoholism Prediction:**

The Random Forest Classifier emerged as the most effective algorithm for predicting alcoholism, attaining a remarkable accuracy of 100% on the training data and 96% on the testing data.

* **Model 3 - Smoking Status Prediction:**

The Random Forest Classifier proved to be the optimal choice for predicting smoking status, delivering outstanding results with a training accuracy of 100% and a testing accuracy of 97%.

These results demonstrate the efficacy of machine learning algorithms in accurately predicting outcomes. The high training accuracies across all models indicate strong learning capabilities, while the comparable testing accuracies highlight the models' generalization performance on unseen data. Overall, the findings underscore the potential of machine learning approaches in enhancing health assessment

In this, we discussed the accuracy of different algorithms used

**For Model 1:**

|  |  |  |
| --- | --- | --- |
| Algorithm | Training Accuracy | Testing Accuracy |
| Linear Discriminant Analysis | 89% | 89% |
| Logistic Regression | 73% | 72% |
| Naive Bayes | 59% | 60% |
| Support Vector Machine | 86% | 82% |
| K-Nearest Neighbors | 84% | 76% |
| Decision Tree | 100% | 94% |
| Random Forest | 100% | 96% |
| Bagging | 99% | 95% |
| AdaBoost | 35% | 34% |
| Gradient Boosting | 99% | 97% |
| Extra Trees | 100% | 79% |

TABLE 4.6.1: Accuracy Comparison for different algorithms for model no 1

**For Model 2:**

|  |  |  |
| --- | --- | --- |
| Algorithm | Training Accuracy | Testing Accuracy |
| Linear Discriminant Analysis | 69% | 67% |
| Logistic Regression | 70% | 68% |
| Naive Bayes | 65% | 65% |
| Support Vector Machine | 89% | 86% |
| K-Nearest Neighbors | 95% | 92% |
| Decision Tree | 100% | 95% |
| Random Forest | 100% | 96% |
| Bagging | 99% | 95% |
| AdaBoost | 49% | 48% |
| Gradient Boosting | 95% | 92% |
| Extra Trees | 100% | 95% |

TABLE 4.6.2: Accuracy Comparison for Different Algorithms for model no 2

**For Model 3:**

|  |  |  |
| --- | --- | --- |
| Algorithm | Training Accuracy | Testing Accuracy |
| Linear Discriminant Analysis | 97% | 97% |
| Logistic Regression | 97% | 97% |
| Naive Bayes | 95% | 95% |
| Support Vector Machine | 98% | 98% |
| K-Nearest Neighbors | 98% | 97% |
| Decision Tree | 100% | 95% |
| Random Forest | 100% | 97% |
| Bagging | 100% | 98% |
| AdaBoost | 99% | 97% |
| Gradient Boosting | 99% | 97% |
| Extra Trees | 100% | 96% |

TABLE 4.6.3:Accuracy Comparison for Different Algorithms for model no 3

**Based on the predictions of the machine learning models trained on the original dataset, the following observations were made:**

1. **Model 1 - Obesity Level Prediction:**

**on downloaded dataset(size=2087):**

 Total number of people with insufficient weight = 269

 Total number of people with normal weight = 281

 Total number of people with Obesity Type I = 352

 Total number of people with Obesity Type II = 296

 Total number of people with Obesity Type III = 324

 Total number of people with Overweight Level I = 278

 Total number of people with Overweight Level II = 287

**on Survey dataset(size=142):**

* Total number of people with insufficient weight = 24
* Total number of people with normal weight = 87
* Total number of people with Obesity Type I = 9
* Total number of people with Obesity Type II = 0
* Total number of people with Obesity Type III = 0
* Total number of people with Overweight Level I = 16
* Total number of people with Overweight Level II = 6

**on Augmented Survey dataset(size=2113):**

 Total number of people with insufficient weight = 424

 Total number of people with normal weight = 1199

 Total number of people with Obesity Type I = 146

 Total number of people with Obesity Type II = 0

 Total number of people with Obesity Type III = 0

 Total number of people with Overweight Level I = 225

 Total number of people with Overweight Level II = 119

1. **Model 2 - Alcoholism Prediction:**

**on downloaded dataset(size=2087):**

 Total number of people always consuming alcohol = 1

 Total number of people frequently consuming alcohol = 70

 Total number of people sometimes consuming alcohol = 1363

 Total number of people not consuming alcohol = 653

**on Survey dataset(size=142):**

 Total number of people always consuming alcohol = 0

 Total number of people frequently consuming alcohol = 0

 Total number of people sometimes consuming alcohol = 42

 Total number of people not consuming alcohol = 100

**on Augmented Survey dataset(size=2113):**

 Total number of people always consuming alcohol = 2

 Total number of people frequently consuming alcohol = 17

 Total number of people sometimes consuming alcohol = 701

 Total number of people not consuming alcohol = 1393

1. **Model 3 – Smoking Status Prediction:**

**on downloaded dataset(size=2087):**

 Total number of smokers = 45

 Total number of non-smokers = 2042

**on Survey dataset(size=142):**

 Total number of smokers = 0

 Total number of non-smokers = 142

**on Augmented Survey dataset(size=2113):**

 Total number of smokers = 0

 Total number of non-smokers = 2113

**Observations on downloaded dataset: -** Among the different obesity levels predicted by Model 1, the category with the highest number of individuals is **Obesity Type I** with 352 people. This indicates that in the dataset, the highest proportion of individuals fall within the Obesity type I range.

Model 2 predicts alcohol consumption habits. In the dataset, the category with the highest number of individuals is **Sometimes consuming alcohol**, with 1363 people. This suggests that a significant portion of the dataset population occasionally consumes alcohol.

Model 3 predicts smoking status. From the dataset, the category with the highest number of individuals is **Non-smokers**, totalling 2042 people. This indicates that a large majority of the dataset participants do not smoke.

**Observations on survey dataset: -** Among the survey participants predicted by model 1, the most common category was individuals with normal weight, totalling 87. Conversely, categories like Obesity Type I, Type II, and Type III had fewer participants, ranging from 0 to 9 individuals each. This suggests a varied distribution of obesity levels among the surveyed population, with a notable prevalence of normal weight individuals.

Model 2 found that the majority of participants, 100 individuals, reported not consuming alcohol. In contrast, only 42 individuals reported sometimes consuming alcohol, while none reported always or frequently consuming alcohol. This distribution highlights a significant portion of the surveyed population abstaining from alcohol consumption.

Model 3 aimed to predict smoking behaviours. Among the surveyed individuals, all 142 participants were classified as non-smokers, with none reporting smoking habits. This indicates a complete absence of smokers within the surveyed dataset.

**Observations on Augmented survey dataset: -**Among the augmented dataset, the largest group was individuals classified as having normal weight, totalling 1199. Conversely, categories such as Obesity Type I, Type II, and Type III had fewer participants, ranging from 0 to 146 individuals each. This distribution suggests a predominant representation of normal weight individuals in the dataset, with minimal instances of severe obesity types.

model 2 found that a majority of individuals, 1393, reported not consuming alcohol. In contrast, 701 individuals reported sometimes consuming alcohol, 17 reported frequently consuming alcohol, and only 2 reported always consuming alcohol. This distribution highlights a prevalent trend of abstinence from alcohol among the surveyed population, with occasional and less frequent consumption patterns also identified.

Model 3 aimed to predict smoking behaviors. In the augmented dataset, all 2113 individuals were classified as non-smokers, with none reporting smoking habits. This clear distinction underscores a notable absence of smokers among the surveyed population.

**TABLE: Comparison of Accuracy of Different models trained on Original Dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Algorithm Used | Training Accuracy | Testing Accuracy |
| Model 1 | Gradient Boosting Classifier | 100% | 97% |
| Model 2 | Random Forest Classifier | 100% | 96% |
| Model 3 | Random Forest Classifier | 100% | 99% |

**Based on the predictions of the machine learning models trained on the Augmented survey dataset, the following observations were made:**

1. **Model 1 - Obesity Level Prediction:**

**on Augmented Survey dataset(size=2113):**

 Total number of people with normal weight = 1204

 Total number of people with insufficient weight = 458

 Total number of people with Obesity Type I = 131

 Total number of people with Obesity Type II = 51

 Total number of people with Obesity Type III = 0

 Total number of people with Overweight Level I = 137

 Total number of people with Overweight Level II = 132

**on downloaded dataset(size=2087):**

 Total number of people with normal weight = 1237

 Total number of people with insufficient weight = 460

 Total number of people with Obesity Type I = 175

 Total number of people with Obesity Type II = 99

 Total number of people with Obesity Type III = 0

 Total number of people with Overweight Level I = 80

 Total number of people with Overweight Level II = 36

**on survey dataset(size=142):**

 Total number of people with normal weight = 83

 Total number of people with insufficient weight = 40

 Total number of people with Obesity Type I = 9

 Total number of people with Obesity Type II = 4

 Total number of people with Obesity Type III = 0

 Total number of people with Overweight Level I = 3

 Total number of people with Overweight Level II = 3

1. **Model 2 - Alcoholism Prediction:**

**on Augmented Survey dataset(size=2113):**

 Total number of people always consuming alcohol = 0

 Total number of people frequently consuming alcohol = 11

 Total number of people sometimes consuming alcohol = 205

 Total number of people not consuming alcohol = 1897

**on downloaded dataset(size=2087):**

 Total number of people always consuming alcohol = 0

 Total number of people frequently consuming alcohol = 0

 Total number of people sometimes consuming alcohol = 1

 Total number of people not consuming alcohol = 2086

**on Survey dataset(size=142):**

 Total number of people always consuming alcohol = 0

 Total number of people frequently consuming alcohol = 0

 Total number of people sometimes consuming alcohol = 1

 Total number of people not consuming alcohol = 141

1. **Model 3 – Smoking Status Prediction:**

**on Augmented Survey dataset(size=2113):**

 Total number of smokers = 205

 Total number of non-smokers = 1908

**on downloaded dataset(size=2087):**

 Total number of smokers = 16

 Total number of non-smokers = 2071

**on Survey dataset(size=142):**

 Total number of smokers = 11

 Total number of non-smokers = 131

**Observations on Augmented survey dataset: -**

the majority of individuals are classified as having normal weight, with 1204 people falling into this category. This suggests a predominant representation of normal weight individuals within the dataset. The next largest group is those with insufficient weight, totalling 458 individuals. The number of people classified under Obesity Type I is 131, and 51 individuals fall into the Obesity Type II category, with no individuals classified under Obesity Type III. Additionally, 137 individuals are classified as Overweight Level I, and 132 individuals as Overweight Level II. This distribution indicates a significant prevalence of normal and insufficient weight categories, with fewer individuals classified under the higher obesity levels.

When predicting alcohol consumption habits, the dataset reveals that the vast majority of individuals, 1897, report not consuming alcohol at all. This is followed by 205 individuals who sometimes consume alcohol, 11 individuals who frequently consume alcohol, and none who always consume alcohol. This distribution highlights a predominant trend of alcohol abstinence within the dataset, with occasional and less frequent consumption patterns also present.

The smoking status prediction model shows that a significant portion of the dataset participants, 1908 individuals, are non-smokers, while 205 individuals are smokers. This indicates a high prevalence of non-smoking behavior among the surveyed population, suggesting that most individuals in the augmented dataset do not engage in smoking.

Overall, the observations from the augmented survey dataset demonstrate a notable prevalence of normal weight, non-alcohol consuming, and non-smoking individuals, indicating generally healthy behaviors among the majority of the dataset participants.

**Observations on downloaded dataset: -** the majority of individuals are classified as having normal weight, with 1237 people falling into this category. This indicates a predominant representation of normal weight individuals within the dataset. The next largest group is those with insufficient weight, totaling 460 individuals. The number of people classified under Obesity Type I is 175, and 99 individuals fall into the Obesity Type II category, with no individuals classified under Obesity Type III. Additionally, 80 individuals are classified as Overweight Level I, and 36 individuals as Overweight Level II. This distribution suggests a significant prevalence of normal and insufficient weight categories, with fewer individuals classified under higher obesity levels.

When predicting alcohol consumption habits, the dataset reveals that nearly all individuals, 2086, report not consuming alcohol at all. Only 1 individual sometimes consumes alcohol, and there are no individuals who frequently or always consume alcohol. This distribution highlights an overwhelming trend of alcohol abstinence within the dataset, with minimal instances of alcohol consumption.

The smoking status prediction model shows that a significant portion of the dataset participants, 2071 individuals, are non-smokers, while only 16 individuals are smokers. This indicates a high prevalence of non-smoking behavior among the population, suggesting that most individuals in the downloaded dataset do not engage in smoking.

Overall, the observations from the downloaded dataset demonstrate a notable prevalence of normal weight, non-alcohol consuming, and non-smoking individuals, indicating generally healthy behaviors among the majority of the dataset participants.

**Observations on survey dataset: -**The majority of individuals in this model are classified as having normal weight, with 83 people falling into this category. This suggests that most survey participants have a normal weight. The next largest group is those with insufficient weight, totaling 40 individuals. There are 9 individuals classified under Obesity Type I and 4 under Obesity Type II, with no individuals classified under Obesity Type III. Additionally, 3 individuals are classified as Overweight Level I, and another 3 as Overweight Level II. This distribution indicates a varied range of obesity levels among the survey participants, with a significant number having normal weight.

For alcohol consumption habits, the dataset reveals that 141 individuals report not consuming alcohol at all. Only 1 individual sometimes consumes alcohol, and there are no individuals who frequently or always consume alcohol. This distribution underscores a significant trend of alcohol abstinence among the survey participants, with almost all participants refraining from alcohol consumption.

The smoking status prediction model shows that the majority of the survey participants, 131 individuals, are non-smokers, while 11 individuals are smokers. This indicates a high prevalence of non-smoking behavior among the survey participants, suggesting that most individuals do not engage in smoking.

Overall, the observations from the survey dataset demonstrate a notable prevalence of normal weight, non-alcohol consuming, and non-smoking individuals, reflecting generally healthy behaviors among the majority of the survey participants.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Algorithm Used** | **Training Accuracy** | **Testing Accuracy** |
| Model 1 - Obesity Level Prediction | Random Forest Classifier | 99% | 94% |
| Model 2 - Alcoholism Prediction | Random Forest Classifier | 100% | 99% |
| Model 3 - Smoking Status Prediction | Random Forest Classifier | 100% | 98% |

**TABLE: Comparison of Accuracy of Different models trained on Augmented survey dataset**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Models** | **Original Dataset** | | | **Augmented Survey Dataset** | | |
|  | **Classification Algorithm Used** | **Training accuracy** | **Testing accuracy** | **Classification  Algorithm Used** | **Training Accuracy** | **Testing Accuracy** |
| M1 | Gradient Boosting | 100% | 97% | Random Forest | 99% | 94% |
| M2 | Random Forest | 100% | 96% | Random Forest | 100% | 99% |
| M3 | Random Forest | 100% | 99% | Random Forest | 100% | 98% |

**TABLE 4.6.3:Comparison of Accuracy of Different models on original Dataset(size=2087) and Enhanced Dataset(size=2113)**

**Concluded Observation** :-The models exhibit consistent behavior and predictive power across three different datasets, each varying in size and composition. This consistency in the prediction highlighting a prevalence of normal weight, non-alcohol consumption, and non-smoking individuals reflects the models' robustness and ability to generalize well beyond the initial training data. The models performed well on a small survey dataset (142 records), a larger downloaded dataset (2087 records), and a significantly augmented survey dataset (2113 records). The accuracy and robustness across different scales indicate strong generalization.

Overall, the consistent patterns in predictions across multiple datasets validate the models' ability to generalize and provide accurate predictions for different population samples, proving their effectiveness and reliability in real-world scenarios.

## CHAPTER 5

**CONCLUSION AND FUTURE WORKS**

Conclusion:

In this project, we utilized machine learning algorithms to predict obesity levels, alcoholism, and smoking status based on various demographic and lifestyle factors. The results demonstrate the effectiveness of machine learning in accurately classifying health-related outcomes.

Gradient Boosting emerged as the top-performing algorithm for obesity level prediction, showcasing robust accuracy on both training and testing datasets. This highlights the algorithm's ability to capture intricate relationships among features and accurately classify individuals into different obesity categories.

For predicting alcoholism, the Random Forest Classifier demonstrated superior performance, achieving high accuracy metrics on both training and testing data. This underscores the algorithm's capability to effectively capture patterns in the data and make accurate predictions.

Similarly, the Random Forest Classifier proved to be effective in predicting smoking status, showcasing strong accuracy metrics on both training and testing datasets. This suggests that the algorithm can effectively leverage demographic and lifestyle factors to classify individuals into smokers and non-smokers.

Overall, the project highlights the potential of machine learning techniques in accurately classifying health-related outcomes and informing targeted intervention strategies. By leveraging predictive models, healthcare professionals can gain valuable insights into individual health risks and tailor interventions to promote healthier lifestyle choices.

**Future Works:**

* Feature Engineering: Investigate extra features or derived variables, such food habits, genetic markers, or environmental influences, that could improve predictive accuracy.
* Ensemble Methods: Examine how to combine the predictions of several models using ensemble learning approaches to possibly increase overall predictive resilience and accuracy.
* Longitudinal Analysis: Conduct longitudinal studies to analyze changes in health behaviors over time, allowing for dynamic modeling and personalized intervention.

**REFERENCES**

1. Smith, A., et al. (2019). "Predicting Obesity Using Decision Tree and Logistic Regression." Journal of Public Health, 27(4), 543-551.
2. Johnson, B., & Smith, C. (2020). "Machine Learning Models for Predicting Obesity Risk: A Review." Health Informatics Journal, 26(2), 983-998.
3. Gupta, S., et al. (2018). "Predicting Smoking Behavior Using Random Forest Algorithm." Journal of Health Informatics, 15(3), 277-289.
4. Wang, L., et al. (2020). "Support Vector Machine for Predicting Smoking Behavior: A Comparative Study." International Journal of Medical Informatics, 136, 104079.
5. Jones, R., et al. (2017). "Neural Network Approach to Predicting Alcohol Consumption Patterns: A Longitudinal Study." Addiction Research & Theory, 25(6), 463-475.
6. Park, H., et al. (2019). "Ensemble Learning for Predicting Alcohol Consumption Patterns: A Comparative Analysis." Computers in Biology and Medicine, 109, 223-235.
7. Lee, J., et al. (2021). "Multi-Modal Deep Learning Framework for Integrated Health Assessment." Journal of Biomedical Informatics, 118, 103801.

8. Li, J., et al. (2019). "Clustering Dietary Patterns Using Machine Learning Techniques: A Review." Journal of Nutrition, 149(6), 1123-1130.

9. Wang, Y., et al. (2021). "Association Rule Mining for Dietary Patterns Analysis: A Comparative Study." Journal of the Academy of Nutrition and Dietetics, 121(4), 691-700.

10. Zhang, L., et al. (2017). "Predicting Physical Activity Patterns Using Machine Learning Models: A Longitudinal Study." Medicine & Science in Sports & Exercise, 49(10), 2012-2021.

11. Chen, X., et al. (2020). "Machine Learning Models for Predicting Sedentary Behavior: A Comparative Analysis." Journal of Physical Activity and Health, 17(9), 856-864.

12. Yang, W., et al. (2019). "Genome-wide Association Studies for Obesity Susceptibility: A Review." Nature Genetics, 51(2), 202-212.

13. Hu, J., et al. (2021). "Deep Learning Approaches for Identifying Genetic Variants Associated with Obesity Risk: A Comparative Study." Journal of Human Genetics, 66(3), 277-286.

14. ScienceDirect.com

<https://www.sciencedirect.com/science/article/pii/S2352914820306225>

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

16. UCI Machine Learning Repository

<https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition>

17.National Library of Medicine

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10529319/>