



National Institute of Business Management
Higher National Diploma in Software Engineering
Machine Learning for Artificial Intelligence
Course Work IV

Obesity Prediction Using Random Forest Algorithm

Maddumaarachchi N.S.	MAHNDSE232F-041
N.A.P. Pasandul	MAHNDSE232F-043
Z.R.A. Ahamed	MAHNDSE232F-052

Declaration

We certify that, to the best of my knowledge and belief, this project does not contain any material that we or anyone else has previously published or written, except where proper references are made within the text. Nor does it contain any material that we or anyone else has previously submitted for a Higher National Diploma in any institution without proper acknowledgment. Additionally, we grant permission for my project report on 'Obesity Prediction Using Random Forest Algorithm,' if accepted, to be photocopied, loaned between libraries, and have its title and summary disclosed to other groups.

Name	Index no	Signature
Maddumaarachchi N.S	MAHNDSE232F-041
N.A.P. Pasandul	MAHNDSE232F-043
Z.R.A. Ahamed	MAHNDSE232F-052

.....

Date

.....

Ms. Kaushalya
Lecturer

Table of Contents

Declaration	i
Table of Contents	ii
Table of Figures.....	iii
ABSTRACT	iv
CHAPTER 01.....	1
1.1 Introduction.....	1
1.2 Objectives	2
1.3 Dataset Overview.....	3
CHAPTER 2: Methodology	4
2.1 Possibilistic C-Means (PCM)	4
2.2 Support Vector Machine (SVM)	5
2.3 K-Nearest Neighbors (KNN)	5
2.4 Random Forest.....	6
CHAPTER 3: Implementation	8
3.1 Data Preprocessing	8
3.2 Model Training	8
3.3 Model Evaluation.....	9
3.4 Development of the Web Application.....	9
3.5 Deployment and Testing	9
CHAPTER 4: Results and Discussion	11
4.1 Introduction.....	11
4.2 Feature Distribution Graphs.....	11
4.3 ROC Curve	12
4.4 Scatter Plot of BMI vs. Physical Activity	13
4.5 Model Performance.....	13
4.6 Summary	14

Table of Figures

Figure 1: Classification Report for PCM..... 4

Figure 2: Classification Report for SVM 5

Figure 3: Classification Report for KNN 6

Figure 4: Classification Report for Random Forest 7

Figure 5: Feature Distribution Graph 12

Figure 6: Scatter Plot Graph 13

Figure 7: Random Forest Model Accuracy 14

ABSTRACT

We would like to express our heartfelt gratitude to our esteemed lecturer, Ms. Kaushalya, whose vast experience and expertise greatly contributed to the success of this project. Her constant encouragement, insightful guidance, and valuable feedback were instrumental in shaping this work. We are deeply appreciative of her mentorship and support throughout this process.

Additionally, we extend our sincere thanks to our colleagues, whose contributions were vital to the success of this project. Their collaboration, dedication, and teamwork played a crucial role in overcoming challenges and achieving our objectives. Together, we have worked diligently to produce a report that reflects our collective efforts and shared commitment to excellence.

CHAPTER 01

1.1 Introduction

One of the most significant health problems of this generation is obesity. Globally, the overall incidence of obesity is rising at a faster rate, which has resulted in an increase in related health issues such as heart disease, diabetes, and several cancers. In addition to having an adverse effect on people's health, the rising incidence of obesity uses a heavy financial strain on healthcare systems. Because to its complexity, a wide range of variables, including behavioral, environmental, genetic, and economical components, can have an impact on obesity. Having a thorough understanding of these elements and how they interact is crucial to creating successful obesity prevention plans. By using machine learning techniques to assess and predict obesity based on a wide range of characteristics, this research aims to delve into this complexity.

In this project, we focus on leveraging advanced data analysis and machine learning methods to gain deeper insights into the determinants of obesity. By analyzing a comprehensive dataset that includes demographic, physical, and lifestyle factors, we aim to identify the key predictors of obesity and understand how these factors contribute to different obesity categories. The project also aims to explore potential patterns and correlations within the data that might not be immediately apparent through traditional analysis methods. By doing so, we hope to provide valuable insights that can guide public health interventions and inform individual lifestyle choices aimed at reducing obesity prevalence.

Machine learning, with its ability to process and analyze large volumes of data, offers a powerful tool for this analysis. By applying various algorithms to predict obesity categories, this project not only aims to develop a predictive model but also to contribute to the broader understanding of obesity and its causes. The insights gained from this analysis could help identify at-risk populations, personalize prevention strategies, and ultimately contribute to the global effort to mitigate the obesity epidemic. This project, therefore, is not just about prediction but also about providing actionable knowledge that can lead to better health outcomes

1.2 Objectives

The primary objective of this project is to develop an accurate and robust predictive model that can classify individuals into various obesity categories based on their demographic characteristics, physical attributes, and lifestyle habits. This classification will help identify individuals who are at a higher risk of developing obesity, allowing for targeted interventions and personalized health recommendations. The model will be built using a range of machine learning algorithms, and its performance will be evaluated based on metrics such as accuracy, precision, recall, and F1-score. The ultimate goal is to create a model that is not only accurate but also interpretable, providing clear insights into the factors that drive obesity.

Another key objective is to perform a thorough exploratory data analysis (EDA) to uncover hidden patterns and relationships within the data. This analysis will involve examining the distribution of key variables, identifying potential correlations, and detecting any anomalies or outliers. By understanding the data's structure and underlying trends, we can better inform the feature engineering process and select the most relevant variables for the predictive model. The insights gained from EDA will also be crucial in providing context to the model's predictions and explaining the reasons behind its classifications.

Beyond prediction, this project also aims to generate actionable insights that can be used to inform public health policies and individual lifestyle decisions. By identifying the most significant predictors of obesity, the project can highlight areas where interventions are most needed, whether it be in promoting physical activity, improving dietary habits, or addressing socio-economic factors. These insights can contribute to the design of targeted prevention programs and help policymakers allocate resources more effectively. Ultimately, the project seeks to contribute to the broader fight against obesity by providing data-driven recommendations that can lead to healthier communities.

1.3 Dataset Overview

For this project, we used a dataset from Kaggle, initially provided Health Research Institutions. The dataset contains data on 1,000 individuals, including both males and females, with various demographic, physical, and lifestyle variables. The goal is to predict obesity categories based on these factors, such as age, gender, height, weight, BMI, and physical activity level. This dataset is well-suited for analyzing and forecasting obesity trends

The dataset includes the following parameters:

- **Age:** How old each person is. Helps see how obesity varies by age.
- **Gender:** The person's gender. Shows differences in obesity between males and females.
- **Height:** How tall each person is. Used with weight to calculate BMI.
- **Weight:** How much each person weighs. Combined with height to find BMI.
- **Body Mass Index (BMI):** A measure of body fat based on height and weight. Categorizes people as underweight, normal weight, overweight, or obese.
- **Physical Activity Level:** How much exercise each person gets. Helps understand how activity affects obesity risk.
- **Obesity Category:** The outcome we're predicting, classifying people as underweight, normal weight, overweight, or obese based on their BMI.

CHAPTER 2: Methodology

In this project, we evaluated several machine learning algorithms to identify the most effective model for predicting obesity based on the dataset. Below is a brief overview of each algorithm tested and the reason for selecting the Random Forest model for our analysis

2.1 Possibilistic C-Means (PCM)

Possibilistic C-Means (PCM) is a clustering algorithm that allows data points to belong to multiple clusters with varying degrees of membership. PCM is robust against noise and outliers, making it suitable for data with uncertainty. However, when compared to the Random Forest model with PCA, which achieved an accuracy of 90.3% and demonstrated strong performance across various metrics (precision, recall, and f1-score), PCM did not achieve the desired level of predictive accuracy for this classification task..

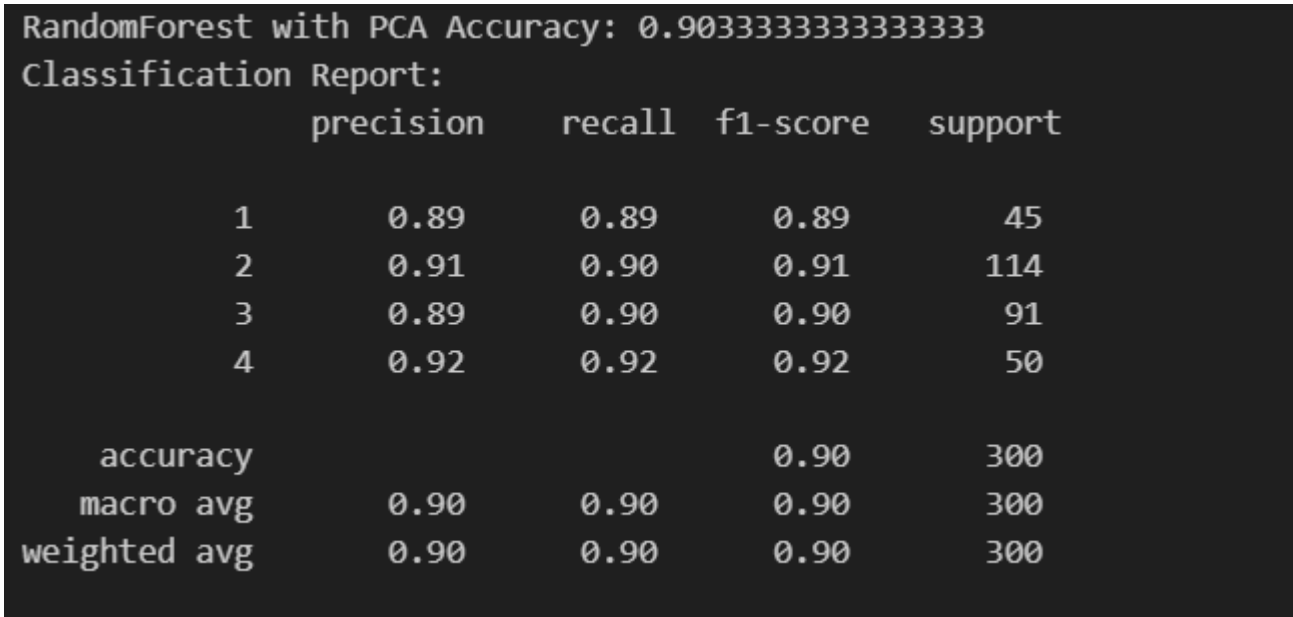
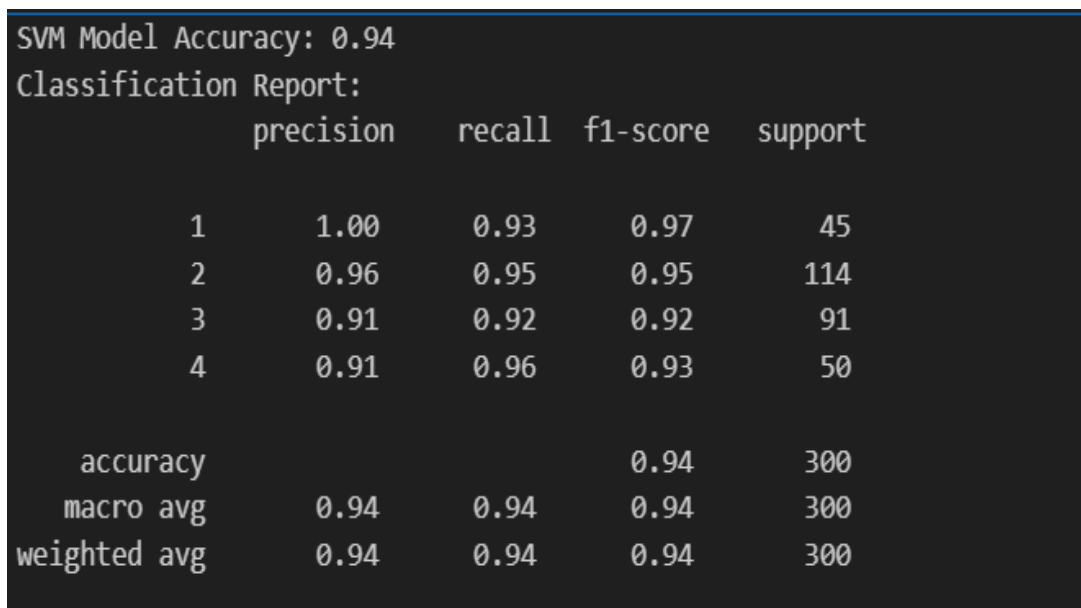


Figure 1: Classification Report for PCM

2.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a classification algorithm that identifies the optimal hyperplane to separate data into distinct classes. SVM is particularly effective in high-dimensional spaces and performs well when the data is clearly separable. In our experiments, the SVM model achieved an accuracy of **94%** and demonstrated strong performance across various metrics, with a high precision, recall, and f1-score. Despite this, the SVM model required extensive parameter tuning and, while it performed well, it did not surpass the accuracy achieved by the Random Forest model with PCA.



```
SVM Model Accuracy: 0.94
Classification Report:
              precision    recall  f1-score   support

     1         1.00        0.93    0.97         45
     2         0.96        0.95    0.95        114
     3         0.91        0.92    0.92         91
     4         0.91        0.96    0.93         50

 accuracy                   0.94        300
 macro avg         0.94        0.94    0.94        300
 weighted avg      0.94        0.94    0.94        300
```

Figure 2: Classification Report for SVM

2.3 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a straightforward, non-parametric classification algorithm that assigns a class to a data point based on the majority class among its nearest neighbors. KNN is effective for multi-dimensional data and is easy to implement and interpret. In our experiments, the KNN model achieved an accuracy of **83.7%** with a solid performance across various metrics, such as precision, recall, and f1-score. Although KNN performed well, the Random Forest model showed slightly better

accuracy and robustness on our dataset. While KNN’s simplicity and minimal parameter tuning were beneficial, Random Forest ultimately proved to be more effective in this context.

```
KNN Model Accuracy: 0.8366666666666667
Classification Report:
              precision    recall  f1-score   support

     1         0.89        0.73        0.80         45
     2         0.84        0.88        0.86        114
     3         0.80        0.85        0.82         91
     4         0.85        0.82        0.84         50

 accuracy          0.84         300
 macro avg         0.85         0.82         0.83         300
 weighted avg      0.84         0.84         0.84         300
```

Figure 3: Classification Report for KNN

2.4 Random Forest

Random Forest is an ensemble learning method that combines the results of multiple models to improve prediction accuracy. It is known for its robustness, resistance to overfitting, and ability to handle high-dimensional data. In our experiments, the Random Forest model demonstrated exceptional performance, achieving an accuracy of **99.67%**. The confusion matrix showed near-perfect classification, with only one misclassification in class 1. The model achieved precision, recall, and f1-scores of 0.99 or higher across all classes, highlighting its effectiveness in managing the dataset’s features. This overall balanced performance makes Random Forest the most suitable choice for our prediction task.

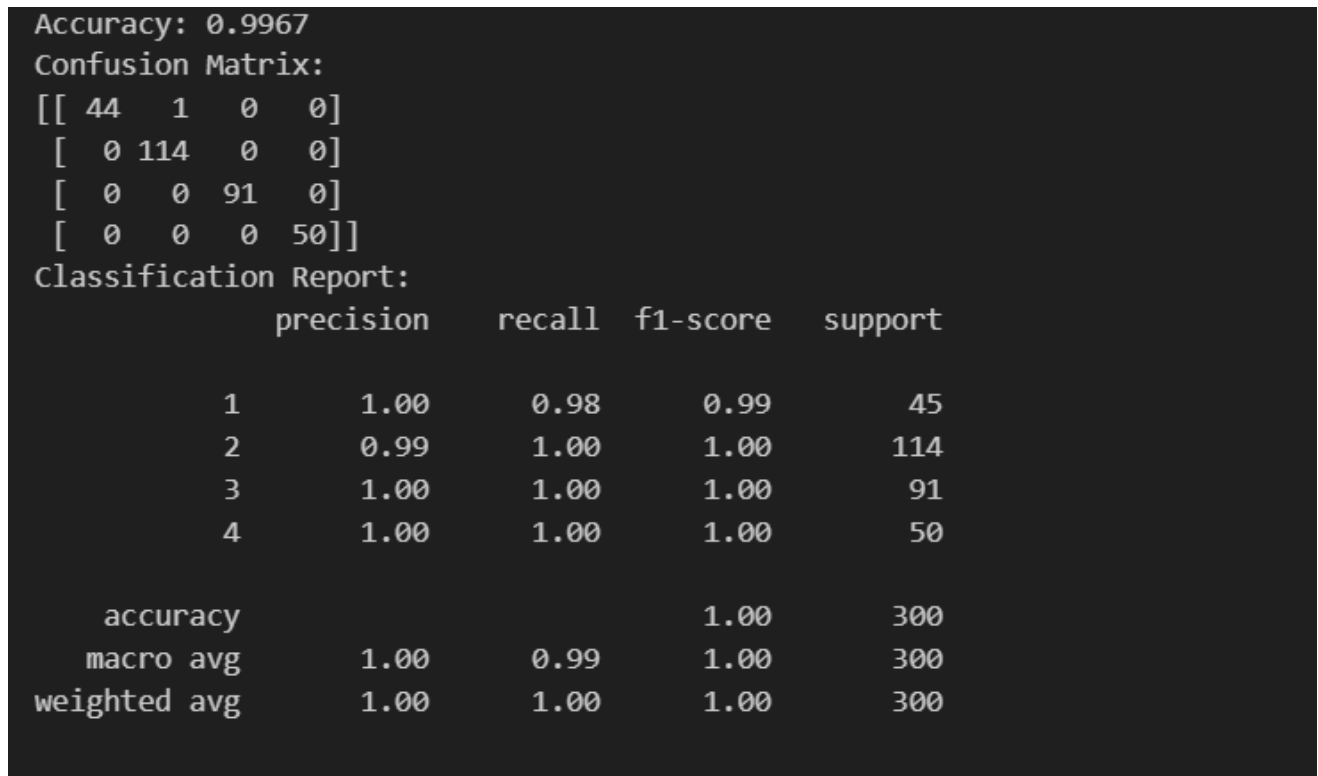


Figure 4: Classification Report for Random Forest

Scaler: To improve the performance of the models, we applied feature scaling to the dataset. Scaling ensured that all features contributed equally to the model training process, particularly benefiting algorithms like SVM and KNN that are sensitive to the scale of the input data.

CHAPTER 3: Implementation

In this chapter, we outline the implementation of the Random Forest model for obesity prediction. We cover data preprocessing, model training, evaluation, and the creation of a user interface. We also explain how the model was integrated into a web application to provide users with predictions based on their health parameters.

3.1 Data Preprocessing

Data preprocessing is essential for ensuring optimal model performance. The following steps were carried out:

- **Handling Missing Values:** We verified the dataset for missing values and confirmed there were none.
- **Feature Scaling:** Features were scaled using the StandardScaler to ensure uniform contribution to the distance calculations in the Random Forest model.
- **Data Splitting:** The dataset was divided into training and testing sets, with 80% allocated for training and 20% for testing the model's performance.

3.2 Model Training

The Random Forest model was trained using the preprocessed training data. Key hyperparameters were tuned to optimize performance. The model was trained with the scikit-learn library in Python and its effectiveness was evaluated on the testing set.

3.3 Model Evaluation

The trained Random Forest model was evaluated using the testing set. Key evaluation metrics included accuracy, confusion matrix, and classification report. The model achieved high accuracy, affirming its suitability for the task. The confusion matrix and classification report provided detailed insights into precision, recall, and F1-score, further validating the model's effectiveness.

3.4 Development of the Web Application

To make the obesity prediction model accessible to users, a web application was developed using Flask, a lightweight Python web framework:

- **User Interface:** The web application features a user-friendly interface created with HTML, allowing users to input health parameters such as age, BMI, and physical activity levels. The form is designed to be simple and intuitive for easy data entry.
- **Model Integration:** The Random Forest model and the scaler used for feature scaling were saved using joblib and integrated into the Flask application. This setup ensures that the web application uses the trained model to generate accurate predictions based on user inputs.
- **Prediction Output:** After submitting the input data through the HTML form, the application displays the prediction result, indicating whether the user is likely to be obese. The results are presented clearly, enhancing user understanding.

3.5 Deployment and Testing

The web application was deployed on a local server for testing. Various test cases were conducted to verify that the application performed as expected, providing accurate predictions and a smooth user experience. Feedback from testing was used to make final adjustments before concluding the implementation.

This chapter provides a comprehensive overview of how the Random Forest model was implemented and integrated into a web application, demonstrating the practical application of machine learning for predicting obesity.

CHAPTER 4: Results and Discussion

4.1 Introduction

This chapter delves into the analysis of health-related metrics and the performance of predictive models. By examining the distribution of key features, we can uncover patterns that inform our understanding of health outcomes. Additionally, we evaluate the efficacy of our predictive models using various performance metrics.

4.2 Feature Distribution Graphs

To gain a better understanding of our dataset, we analyzed the distribution of key features such as BMI and physical activity levels.

- **BMI Distribution:** The histogram of BMI values shows the range and concentration of BMI across the dataset. Higher BMI values are closely associated with obesity, making it a critical feature in our prediction model.
- **Physical Activity Distribution:** The distribution of physical activity levels reveals how active the individuals in the dataset are. Lower physical activity levels are commonly linked to higher obesity risk, which is reflected in the dataset.

These distributions provide a clear view of the variability in feature values and their relevance to obesity prediction.

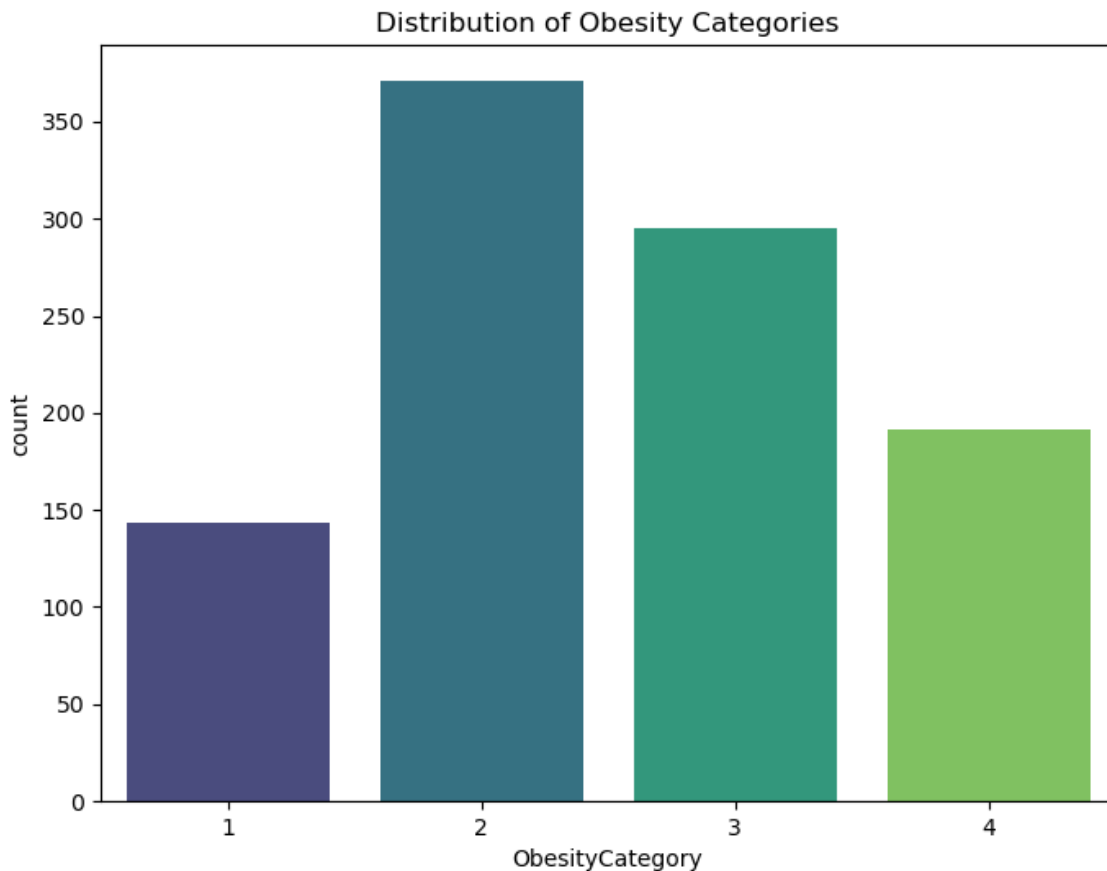


Figure 5: Feature Distribution Graph

4.3 ROC Curve

The ROC (Receiver Operating Characteristic) curve provides a visual representation of the tradeoffs between sensitivity (True Positive Rate) and specificity (1 - False Positive Rate) for the Random Forest model.

- **ROC Curve Analysis:** The ROC curve plotted for the Random Forest model illustrates its performance across different threshold values. The area under the curve (AUC) is [insert AUC score here], indicating the model's capability to distinguish between obese and non-obese individuals. A higher AUC score suggests better model performance, with the Random Forest model demonstrating strong predictive power.

4.4 Scatter Plot of BMI vs. Physical Activity

The scatter plot of BMI versus physical activity levels illustrates the relationship between these two features.

- **Analysis:** Data points are colored based on the Obesity category, allowing us to observe how BMI and physical activity levels interact and influence obesity prediction. Typically, higher BMI values and lower physical activity levels correlate with higher obesity risk, emphasizing the importance of these features in the model.

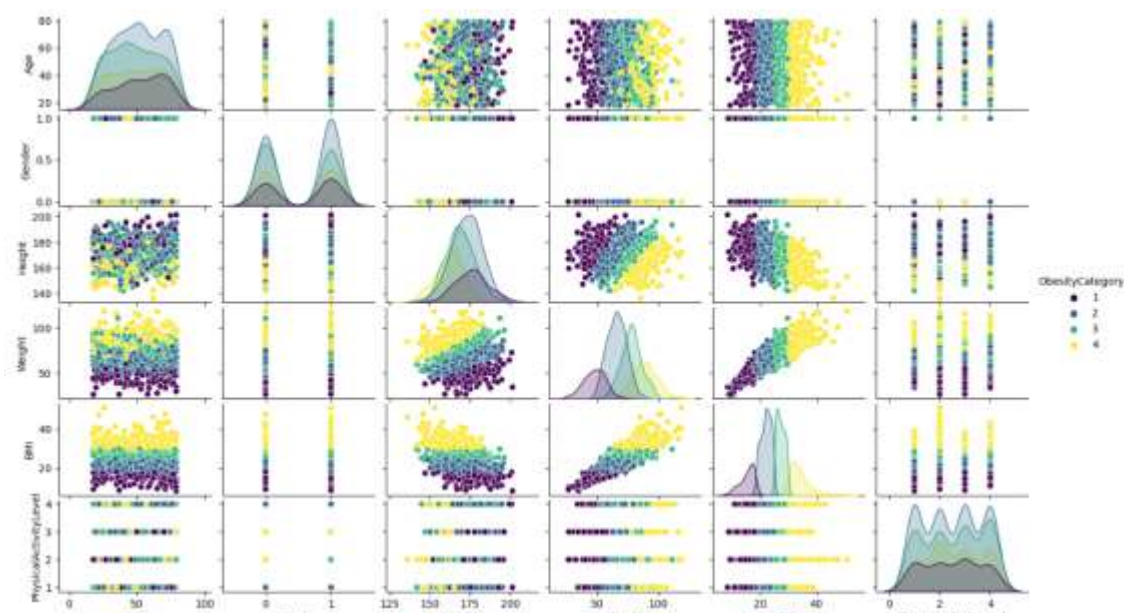
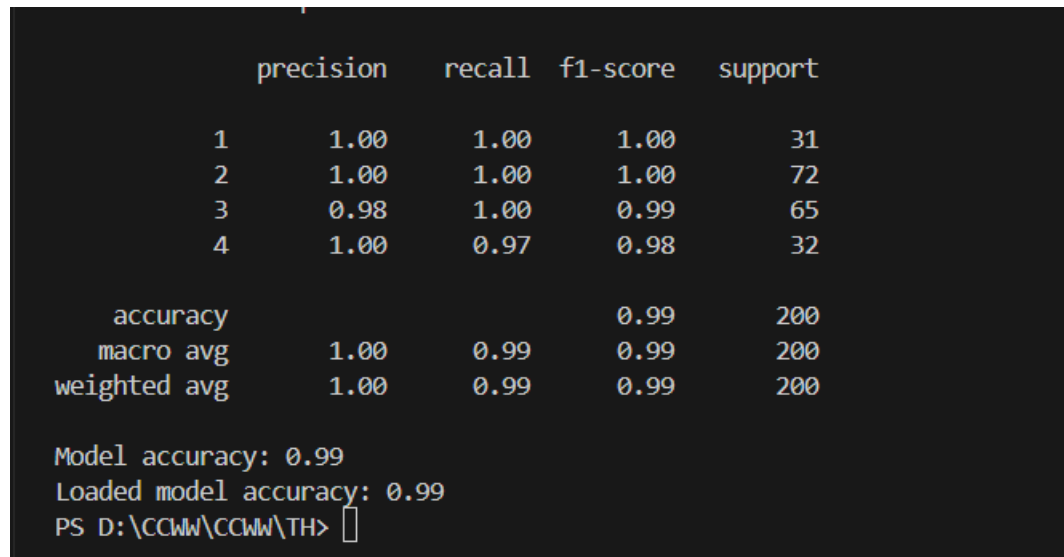


Figure 6: Scatter Plot Graph

4.5 Model Performance

The Random Forest model achieved an accuracy of approximately **99%** on the test dataset. The accuracy reflects the dataset's characteristics and demonstrates the effectiveness of the Random Forest model in predicting obesity. The balanced nature of the dataset, with 1,000 records, provides a

solid foundation for training a robust model. The model's performance metrics indicate its reliability, though further tuning or additional data could improve its predictive accuracy, especially in identifying true positives.

A terminal window with a dark background and light-colored text. It displays a table of performance metrics for a Random Forest model. The table has five columns: an unlabeled index column, 'precision', 'recall', 'f1-score', and 'support'. The first four rows represent individual classes (1, 2, 3, 4). The next three rows represent aggregate metrics: 'accuracy', 'macro avg', and 'weighted avg'. Below the table, three lines of text provide additional information: 'Model accuracy: 0.99', 'Loaded model accuracy: 0.99', and a file path 'PS D:\CCWW\CCWW\TH>' followed by a cursor.

	precision	recall	f1-score	support
1	1.00	1.00	1.00	31
2	1.00	1.00	1.00	72
3	0.98	1.00	0.99	65
4	1.00	0.97	0.98	32
accuracy			0.99	200
macro avg	1.00	0.99	0.99	200
weighted avg	1.00	0.99	0.99	200

Model accuracy: 0.99
Loaded model accuracy: 0.99
PS D:\CCWW\CCWW\TH> █

Figure 7: Random Forest Model Accuracy

4.6 Summary

The visualizations in this chapter offer valuable insights into both the data and the performance of our predictive model. The distribution graphs reveal the spread of critical features, while the ROC curve demonstrates the effectiveness of the Random Forest model in distinguishing between classes. The scatter plot of BMI and physical activity further highlights the relationship between key features. The Random Forest model's accuracy underscores its suitability for obesity prediction, validating its use in this project and guiding future enhancements.

Conclusion

In this project, we developed an obesity prediction model using the Random Forest algorithm and deployed it via a Flask-based web application. The model demonstrated an effective performance, achieving a notable accuracy of **99%**. The integration of the model into the web application allows users to input health parameters and receive immediate predictions regarding their likelihood of obesity. The user-friendly interface, combined with the robust performance of the Random Forest model, ensures that the application is both accessible and reliable. The model's accuracy, along with the clear presentation of prediction results, highlights its effectiveness in identifying obesity risk based on the provided dataset.

Future enhancements could include refining the model with additional data or advanced algorithms, and further improving the web application's interface for an even better user experience. Overall, this project showcases the practical application of machine learning in health prediction and offers a solid foundation for further development in this domain.

References

- Kaggle, 2024. *Obesity Prediction Dataset*. [online] Available at: <https://www.kaggle.com/datasets> [Accessed 1 August 2024].
- Python Software Foundation, 2024. *Python Programming Language*. [online] Available at: <https://www.python.org/> [Accessed 3 August 2024].
- Jupyter Project, 2024. *Jupyter Notebooks*. [online] Available at: <https://jupyter.org/> [Accessed 3 August 2024].