Semester Project (Final Deliverable)

Course: Introduction to Data Science

Semester: Fall 2025

Intelligent Obesity Level Detection from Health Indicators

Submitted by:

Ali Tayyab

01-134232-037

Zameer Hussain

01-134231-096

Submitted To:

Dr. Arif ur Rahman

Submission Date:

2 December 2025

Bahria University Islamabad Campus

Department of Computer Science

Table of Contents

Intelligent Obesity Detection Project

Abstract

Obesity is one of the most critical public health challenges worldwide, leading to severe medical conditions such as **diabetes**, **cardiovascular diseases**, and reduced life expectancy. Early detection and classification of obesity levels can play a vital role in prevention and intervention. This project, titled **"Intelligent Obesity Detection Project"**, focuses on analysing lifestyle, physical, and behavioural factors using data science techniques to **classify obesity levels** accurately. An openly available obesity dataset was utilised, followed by data preprocessing, exploratory data analysis (EDA), and the application of machine learning classification models. **Random Forest** and **Logistic Regression** models were implemented and evaluated. In addition, an interactive Graphical User Interface (GUI) was developed to provide real-time obesity level prediction. The results demonstrate that data-driven approaches can effectively support obesity risk analysis and decision-making in health-related applications.

1. Introduction

Obesity has emerged as a global epidemic affecting individuals across all age groups. Rapid urbanisation, sedentary lifestyles, **unhealthy dietary habits**, and **genetic factors** have significantly contributed to the rise in obesity rates. Traditional methods of obesity assessment often rely on limited indicators such as **Body Mass Index (BMI)**, which may not capture the complete lifestyle context of an individual.

With the advancement of data science and machine learning, it is now possible to analyse multiple contributing factors simultaneously and build intelligent systems for **obesity detection**. This project aims to develop a data-driven obesity classification system that not only achieves high predictive accuracy but also provides an interactive interface for user-friendly analysis.

2. Dataset Selection and Justification

An openly available obesity dataset obtained from a public data repository (Kaggle/UCI) was used in this project. The dataset contains demographic, physical, and lifestyle-related attributes such as age, gender, eating habits, physical activity levels, screen time, and transportation methods.

Justification for Dataset Selection:

- The dataset is health-related and directly relevant to obesity analysis.
- It includes both numerical and categorical features, making it suitable for applying various preprocessing and modelling techniques.
- The target variable represents multiple obesity levels, enabling multi-class classification.

3. Data Preprocessing

Data preprocessing was a critical step to ensure model **reliability** and **validity**. The following steps were performed:

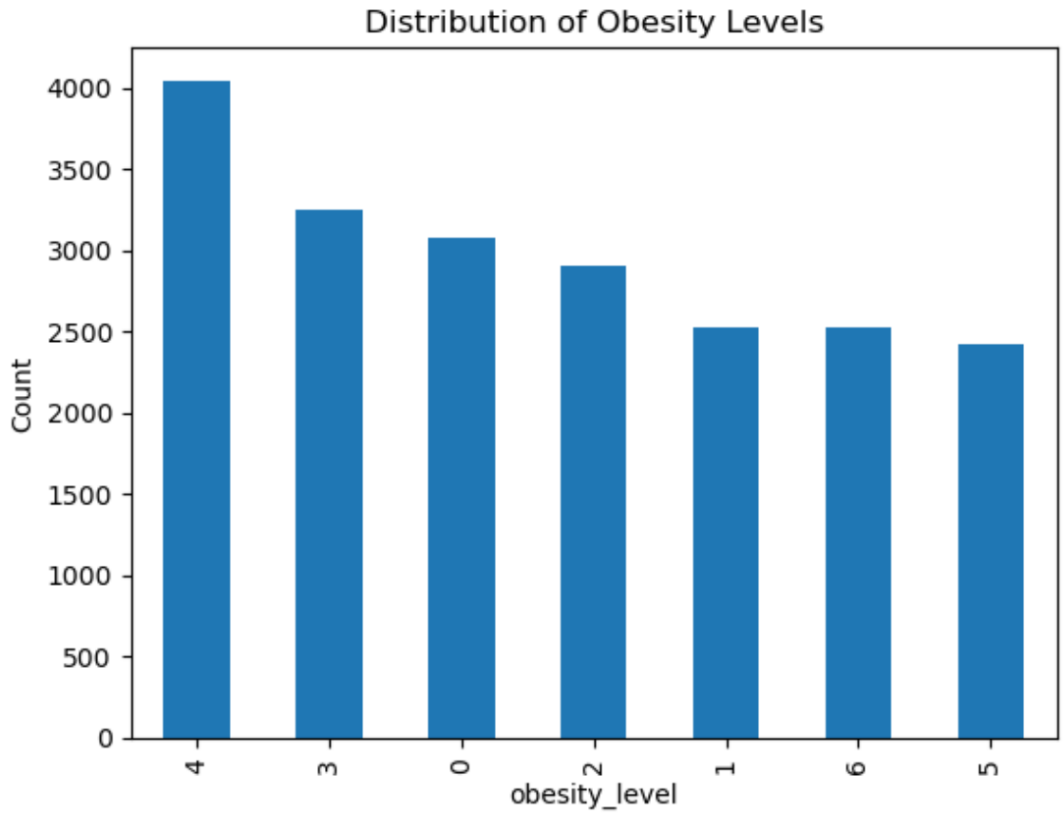
- Removal of irrelevant and redundant features.
- Handling of missing and inconsistent values.
- Encoding of categorical variables using appropriate techniques:
    - Ordinal Encoding for ordered features (e.g., eating frequency).
    - One-Hot Encoding for nominal features (e.g., transportation mode).
- Feature engineering by calculating **BMI** from height and weight and removing the original height and weight columns.
- Feature scaling using standardisation to ensure fair contribution of numerical features.

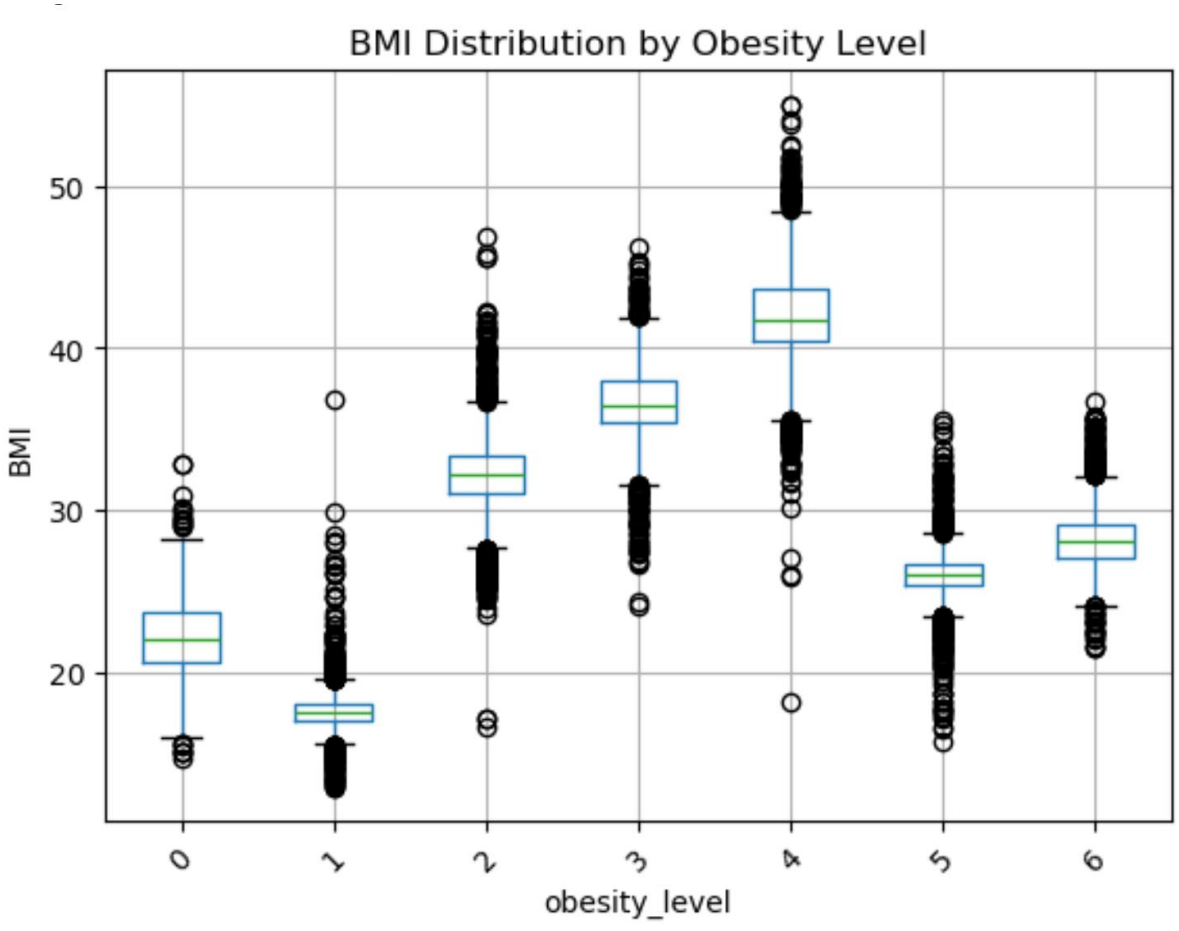
These steps helped reduce noise, prevent data leakage, and improve model performance.

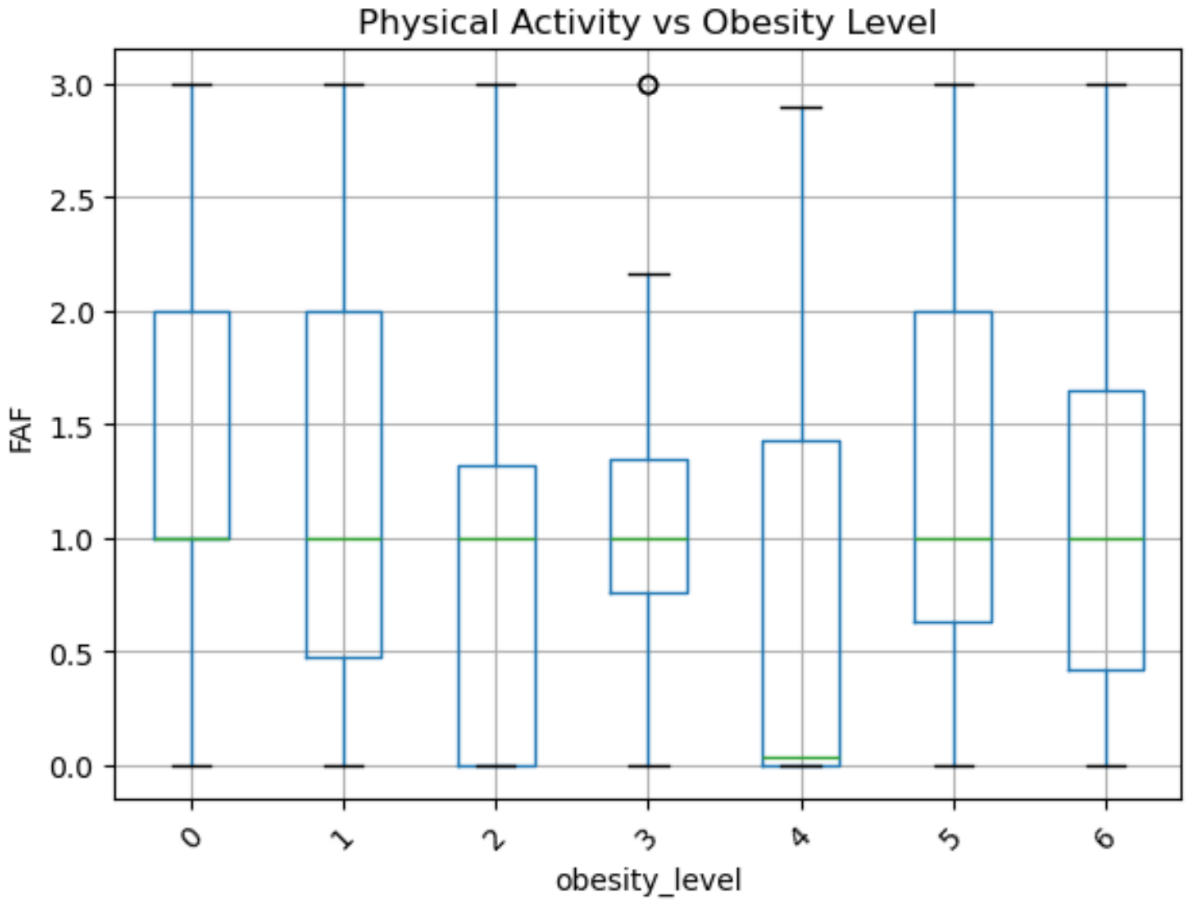
4. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand data distribution and relationships among features. The following techniques were applied:

- Descriptive statistics (mean, median, standard deviation).
- Visualisations including histograms, box plots, and correlation heatmaps.


Distribution of Obesity Levels
Count
4000
3500
3000
2500
2000
1500
1000
500
0
4
3
0
2
1
6
5
obesity_level


BMI Distribution by Obesity Level
BMI
50
40
30
20
0
1
2
3
4
5
6
obesity_level



Key Insights from EDA:

1. **BMI** showed a **strong correlation** with **obesity levels**, confirming it as a primary predictor.
2. Low physical activity (FAF) and high screen time (TUE) were associated with higher obesity categories.
3. Individuals with a family history of obesity were more likely to fall into overweight or obese classes.

These insights validated the importance of lifestyle and behavioural features in obesity prediction.

5. Modelling and Evaluation

5. 1 Machine Learning Models Used

Two classification models were implemented and compared in this project:

Random Forest Classifier

Random Forest is an ensemble learning technique that builds multiple decision trees and combines their predictions. It reduces overfitting and performs well on complex, non-linear datasets.

- Supports multi-class classification
- Robust to noise and feature interactions
- Achieved approximately **86–87% accuracy** using cross-validation

Logistic Regression

Logistic Regression is a linear classification algorithm that predicts class probabilities using a logistic function.

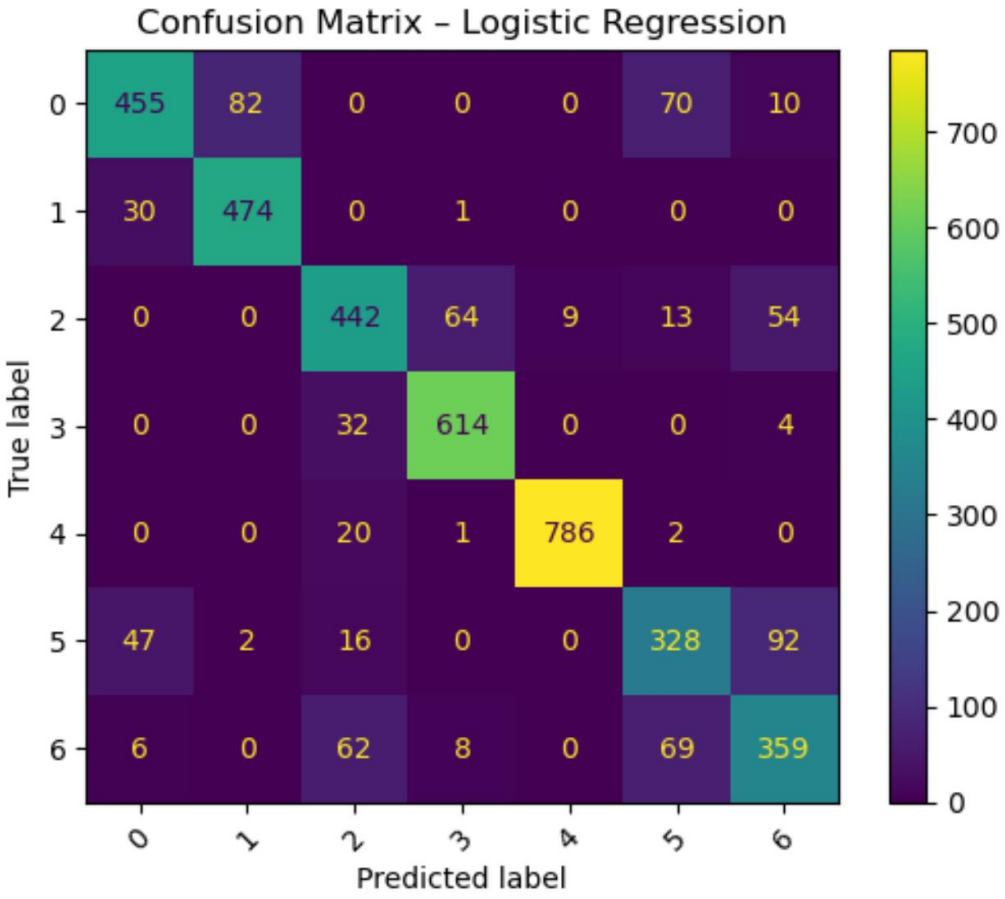
- Simple and interpretable
- Computationally efficient
- Selected for GUI deployment due to its stability and ease of integration
- And achieved approximately **83% accuracy.**

5.2 Evaluation Metrics

After training the models and generating predictions on the test dataset, performance was evaluated using multiple metrics.

Confusion Matrix:

A confusion matrix was generated **after model training and prediction** to analyse class-wise performance. **Diagonal** values represent **correct predictions**, while off-diagonal values indicate misclassifications.



Evaluation Metrics Table

Obesity Class	Precision	Recall (Sensitivity)	F1-Score
Class 0	0.846	0.737	0.788
Class 1	0.849	0.939	0.892
Class 2	0.773	0.759	0.766
Class 3	0.892	0.945	0.918
Class 4 (Severe Obesity)	**0.989**	**0.972**	**0.980**
Class 5	0.680	0.676	0.678
Class 6	0.692	0.712	0.702

Overall Model Performance:

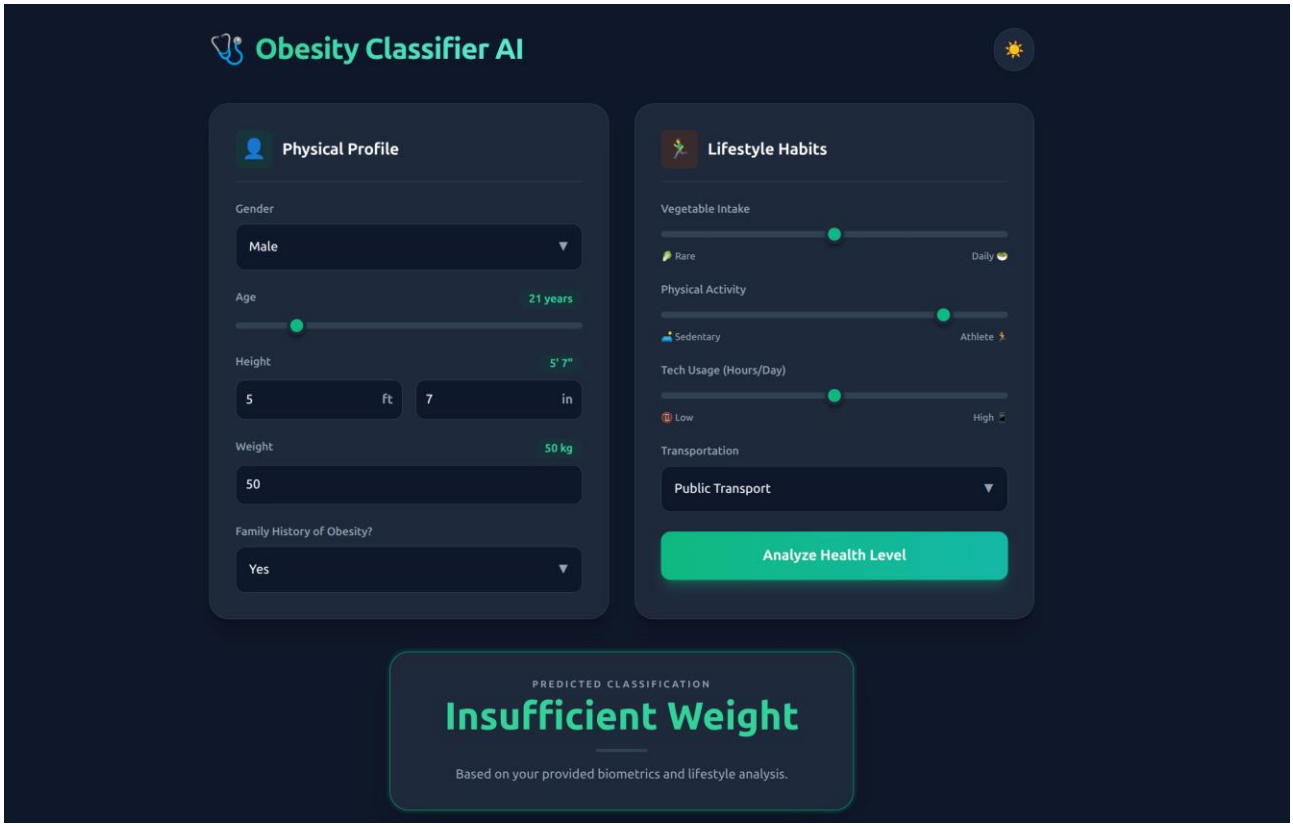
Metric	Value
Overall Accuracy	**83.29%**

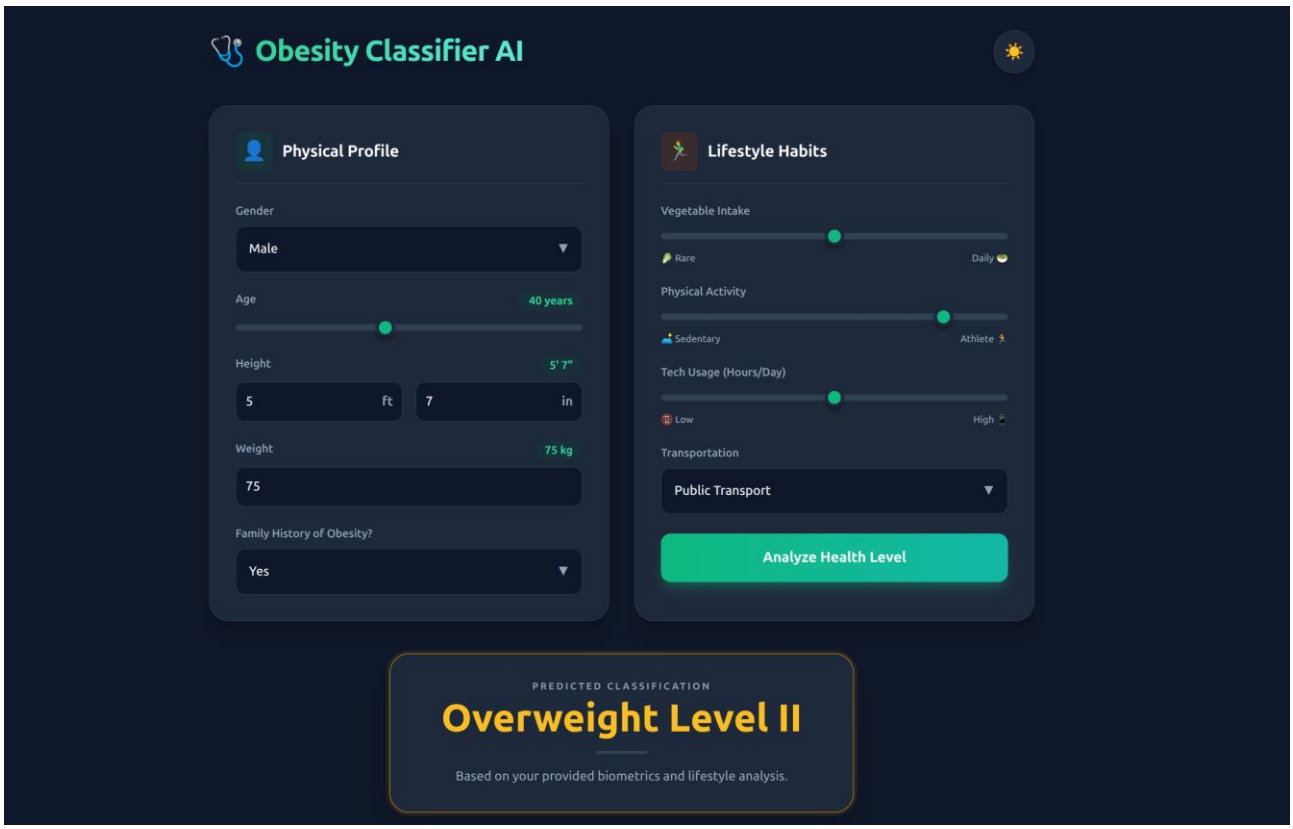
Key Insights:

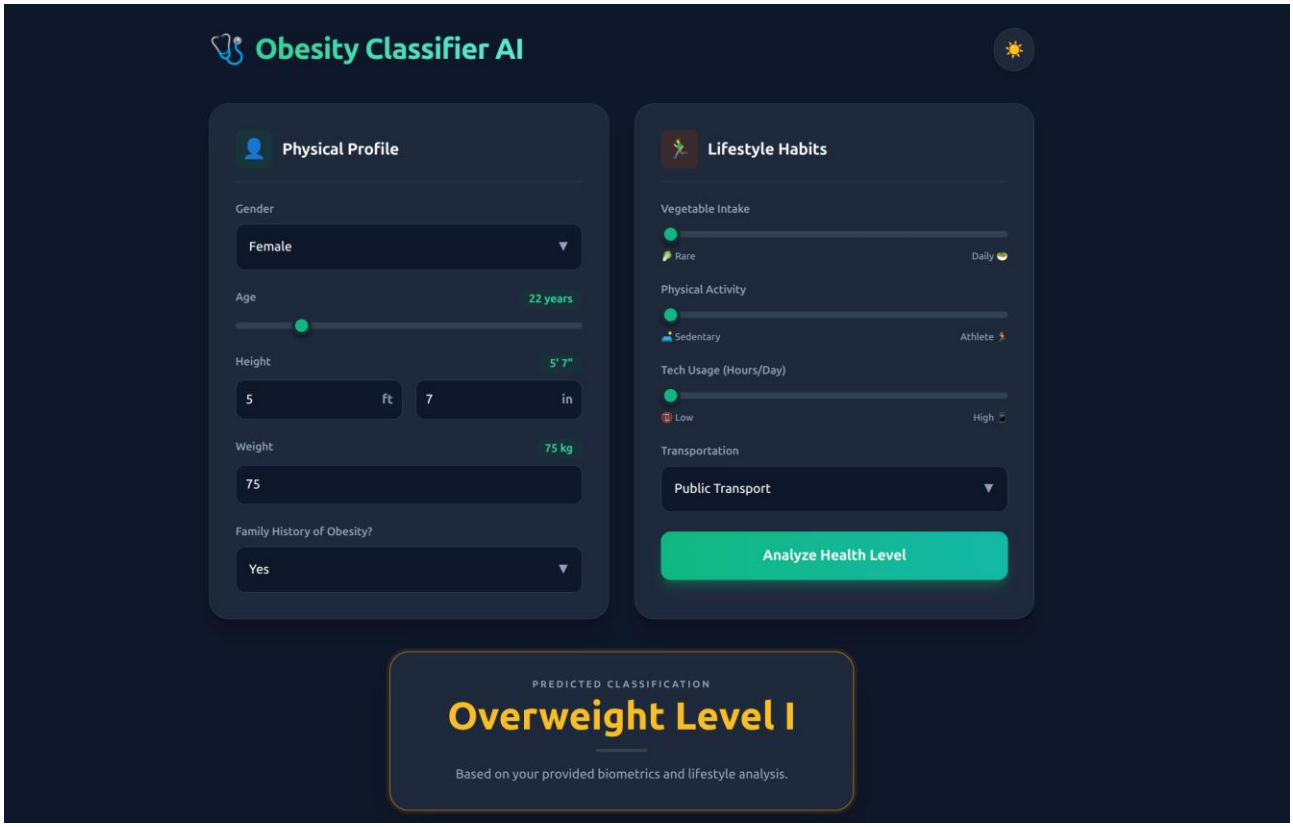
- **Most predictions** lie along the **diagonal**, indicating **strong** classification performance.
- **Misclassifications** mainly occur between **neighbouring obesity categories,** which is realistic in health-related multi-class problems.
- Severe obesity classes show higher recall, demonstrating effective high-risk detection.
- **Class 4 (Severe Obesity)** achieves the **highest precision, recall, and F1-score**, confirming excellent detection of high-risk individuals.
- Classes with **adjacent obesity levels** show slightly lower precision and recall, which is expected due to overlapping physiological features.
- The **balanced F1-scores** indicate that the Logistic Regression model performs consistently across most obesity categories.

6. Graphical User Interface (GUI)

A professional and interactive Graphical User Interface (GUI) was developed using **Streamlit** to demonstrate real-time obesity level prediction. The GUI allows users to input lifestyle and health-related attributes through dropdown menus and numeric inputs.







Key features of the GUI include:

- User-friendly and intuitive interface

- Real-time obesity level prediction
- Colour-coded output indicating obesity risk level

Logistic Regression was used for GUI deployment due to its simplicity, interpretability, and smooth integration.

7. Project Demonstration & Online Resources

To ensure transparency, reproducibility, and professional presentation, the project has been shared through online platforms.

7.1 LinkedIn Video Demonstration:

A demonstration video explaining the project workflow and GUI functionality has been uploaded on LinkedIn.

LinkedIn Video Link:
https://bit.ly/3KZAxHy

GitHub Repository:

To ensure transparency and allow others to replicate or explore our work, we uploaded the entire project to a public GitHub repository. This repository includes all essential components of the project, such as:

- Data cleaning and preprocessing scripts
- Exploratory Data Analysis (EDA) notebook
- Model training and evaluation code
- The complete GUI application
- Documentation and supporting files

The repository is available at the following link:
https://github.com/ALITAYYAB2K1/NutritionModel

8. Conclusion

The **Intelligent Obesity Detection Project** successfully applies data science and machine learning techniques to a real-world health problem. Through careful preprocessing, insightful exploratory analysis, and robust classification models, the system accurately predicts obesity levels. The integration of an interactive GUI further enhances practical usability. Overall, the project demonstrates the effectiveness of intelligent systems in supporting early obesity detection and promoting health awareness.

9. Future Work

Future improvements may include:

- Incorporation of medical and clinical data for enhanced accuracy
- Use of advanced ensemble models such as CatBoost or XGBoost
- Deployment as a mobile application

Continuous learning using real-time dat

10. References

- World Health Organization (WHO) – Obesity and Overweight Reports
- Kaggle Obesity Dataset
- Scikit-learn Documentation
- Streamlit Documentation