CAL STATE FULLERTON

Department of Computer Science

This project has been satisfactorily demonstrated and is of suitable form.

This project report is acceptable in partial completion of the requirements for the Master of Science degree in Computer Science.

Obesity Risk Prediction in Adults using Machine Learning

Project Title (type)

Ayush Bhatnagar

Student Name (type)

Dr. Kanika Sood

Advisor's Name (type)

Advisor's signature Date

Reviewer's Name (type)

Reviewer's signature Date

Abstract

Obesity has become a growing global health epidemic and is a leading cause of chronic diseases such as diabetes, cardiovascular disorders, and certain cancers, imposing significant burdens on healthcare systems. Early risk prediction and intervention are critical to mitigate these adverse health outcomes. This project leverages machine learning techniques to develop predictive models for assessing obesity risk in adults using detailed demographic, behavioral, and clinical data. By analyzing large datasets and uncovering complex patterns, the study aims to provide actionable insights that support early detection and personalized intervention strategies.

The integrative project merges concepts from computer science, public health, and medicine, enabling innovative solutions to address the multifaceted nature of obesity. The resulting models have the potential to inform public health policies and promote preventive measures, contributing to improved population health. Through this work, I seek to explore the potential of machine learning in addressing critical public health challenges while contributing to advancements in predictive analytics and data-driven healthcare.

Table of Contents

Introduction

In today's fast-paced world, the escalating epidemic of obesity has become a global concern, impacting individuals from various backgrounds. Consider a scenario where someone indulges in unhealthy eating habits, stays up late, and spends excessive time in front of screens. It's evident that obesity is on the rise, particularly among young people who may not realize how their current behaviors affect their health and future. Obesity is not just about appearance; it leads to various health issues like cardiovascular problems, diabetes, and even liver cancer [WHO 2023].

In terms of statistics, last year witnessed over 1.9 billion adults classified as overweight, with a staggering 650 million labeled as obese, representing 39% of the total overweight population [CER Bariatrics 2023]. Several factors contribute to weight gain and retention, including diet, physical inactivity, environmental influences, and genetics [NICHD 2023].

As suggested by the title of the survey report, we delve into the domain of Machine Learning (ML) and its application in addressing obesity-related complications. ML empowers computers to learn from data independently, making accurate predictions without explicit instructions. Recent studies have highlighted ML's efficacy in handling complex data and deciphering intricate relationships between factors [Frontiers 2023][PMC 2023].

While ML models have shown promise in predicting childhood obesity, challenges arise when forecasting obesity in adults due to the complexity of lifestyle choices and health variables involved. Despite the advantages of ML in risk prediction, challenges persist in ensuring model versatility and avoiding overemphasis on details, as well as interpreting model predictions [AIMS Press 2022][SciDirect 2021].

The project's overarching goal is to leverage ML techniques to deepen our understanding of obesity, ultimately facilitating more precise prediction and intervention strategies for improved public health outcomes [Research Square 2023][CEUR-WS 2021].

1.1 Background

The complexity of obesity arises from its multifactorial nature, influenced by genetics, lifestyle, and environmental factors. Predicting the risk of obesity is essential to facilitate early intervention and improve health outcomes [WHO 2023][NICHD 2023].

Machine learning, a subset of artificial intelligence, has revolutionized predictive modeling by enabling data-driven insights and precise risk assessment. In this project, focused on applying supervised learning techniques, specifically the LightGBM algorithm, to predict obesity risk in adults. LightGBM is a gradient-boosting framework renowned for its efficiency and accuracy in handling large datasets and complex feature interactions [SciDirect 2021][PMC 2022].

Using a structured dataset for training, the model learns to identify patterns and relationships between input features and obesity risk levels. The trained model is then applied to predict future risk, allowing for proactive measures and targeted possible healthcare interventions. This approach highlights the potential of machine learning to enhance traditional methodologies, offering scalable and reliable solutions for combating obesity in modern healthcare settings [Frontiers 2023][AIMS Press 2022][Research Square 2023].

1.2 Project Motivation

This project is motivated by the urgent need to combat rising obesity rates through early risk detection. Traditional methods often fall short in precision, while machine learning, particularly the LightGBM algorithm, offers a powerful way to analyze complex data. By developing an accurate and scalable predictive model, the ultimate goal is to empower individuals and healthcare providers with better risk assessments, enabling earlier interventions and promoting healthier lifestyles.

1.3 Project Goals

The ultimate goal of this project is to detect individuals who are at an elevated risk of developing obesity at an early stage, enabling timely intervention and preventive actions to mitigate the risk of chronic illnesses such as diabetes, cardiovascular diseases, and other obesity-related health complications.

1.4 Project Objectives

1. Identify critical behaviors and patterns that can be addressed to implement effective preventive strategies.

2. Utilize machine learning models and techniques to improve the precision of obesity risk prediction.

3. Provide actionable insights into the influence of lifestyle factors, such as diet and drinking habits, on obesity risk.

.

1.5 Related Work

Machine learning has shown great potential in predicting obesity, with models like DeepHealthNet providing insights into adolescent obesity prediction [Jeong et al. 2024]. Additionally, another study explores various machine learning models for childhood and adolescent obesity, highlighting the progress in this field [Siddiqui et al. 2021]. However, there is still work to be done to enhance model accuracy and ensure more reliable long-term predictions in obesity prevention.

1.6 Algorithm Explanation

For the obesity risk prediction project, the Light Gradient Boosting Machine (LightGBM) was chosen as the ideal machine learning algorithm due to its exceptional performance, efficiency, and adaptability to large datasets. Below is an explanation of the algorithm and the grounds behind its selection.

Overview of LightGBM

LightGBM is a highly optimized, gradient-boosting framework specifically designed for handling large-scale datasets. It leverages a leaf-wise tree growth strategy, unlike traditional level-wise growth methods. This approach allows LightGBM to build trees that are more balanced and accurate while reducing computational time [Kılıç 2023].

In addition, the algorithm utilizes a histogram-based technique for splitting nodes, enabling it to discretize continuous features into bins. This not only speeds up computation but also minimizes memory usage, making it well-suited for resource-constrained environments [İlyurek 2023].

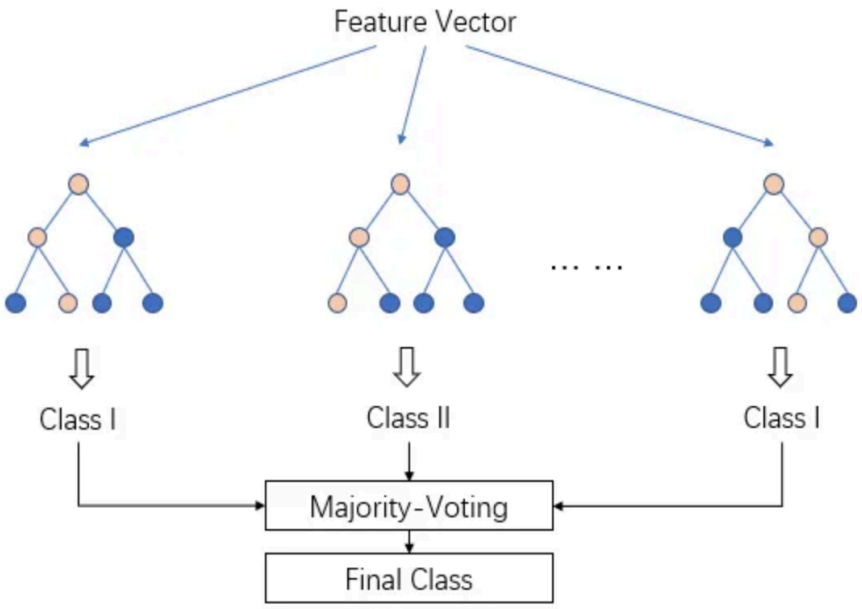


Figure 1(LightGBM Feature Vector)

Why LightGBM for Obesity Risk Prediction

1. **Efficiency on High-Dimensional Data**

   The dataset used for predicting obesity risk includes a mix of numerical and categorical features with a wide range of values. LightGBM's histogram-based approach and leaf-wise growth strategy significantly reduce training time, making it an efficient choice for large and complex datasets.

2. **Handling Mixed Data Types**

   The dataset includes categorical features such as "Gender," "Mode of Transportation," and "Family History of Overweight." LightGBM's built-in capability to handle categorical features without the need for one-hot encoding simplifies preprocessing, saving time and resources.

3. **Scalability**

   The algorithm is designed to work smoothly with large datasets, a key requirement for this project given the extensive feature set and potential for future data expansion. Its distributed training support ensures scalability for datasets with millions of rows and features.

4. **Robust Performance**

   LightGBM's ability to generalize across different data distributions makes it a reliable choice for this project, which involves diverse patterns and interactions between variables. Its robust gradient-boosting framework ensures high predictive accuracy.

5. **Feature Importance Insights**
   One of the project goals is to understand the relative importance of different features contributing to obesity risk. LightGBM provides clear feature importance metrics, enabling better interpretation of the model's predictions.

6. **Customizable Hyperparameters**
   LightGBM supports extensive hyperparameter tuning, which was exploited in this project through RandomizedSearchCV. This allowed the optimization of key parameters such as learning rate, maximum depth, and number of leaves, leading to improved performance.

1.7 Development and Operational Environment

This project is developed using Jupyter Notebook for coding and experimentation. The primary libraries used include Pandas for data manipulation, Tkinter for creating graphical user interfaces, and Scikit-learn for machine learning tasks such as model training and evaluation.

Key tools and modules include `train_test_split`, `RandomizedSearchCV`, `Pipeline`, and `ColumnTransformer` for model development, while `StandardScaler`, `OneHotEncoder`, and `LabelEncoder` are used for preprocessing. The LightGBM model is implemented using the `LGBMClassifier`, and the trained models are saved using Joblib for later use.

Project Outline

2.1 Development Phases

For this project, outlined a structured development plan to create an obesity risk predictor using machine learning (ML). This plan encompassed various key activities and tasks aimed at achieving the project objectives effectively. Here's a detailed account of the completed development phases and the tasks involved:

1. **Data Collection and Preprocessing**: Began by collecting diverse datasets containing information relevant to obesity risk factors, including demographics, medical history, lifestyle habits, and biometric measurements. Once the datasets were collected, cleansed and preprocessed the data to handle missing values, outliers, and inconsistencies. The deliverable for this phase was a cleaned and standardized dataset ready for analysis.

2. **Feature Engineering**: Next, proceeded with feature engineering by identifying and extracting relevant features from the dataset that influence obesity risk. This involved carefully analyzing the dataset to determine the most significant variables for predicting obesity. Once the relevant features were identified, performed feature scaling, transformation, and encoding as necessary to prepare the features for machine learning algorithms. The deliverable for this phase was a feature-engineered dataset with optimized features ready for predictive modeling.

3. **Model Selection and Training**: In this phase, focused on selecting and training the most effective machine learning models for obesity risk prediction. Evaluated various algorithms and trained multiple models using the preprocessed dataset. Through rigorous evaluation, including metrics like accuracy and precision, identified the best-performing models to address the obesity epidemic.

4. **Hyperparameter Tuning and Optimization**: Focused on hyperparameter tuning and optimization of the selected machine learning models to enhance their performance. Utilizing techniques like grid search and randomized search, fine-tuned the models' hyperparameters for improved accuracy. Following this, validated the trained models using cross-validation techniques to ensure their generalization capability. Additionally, evaluated the models on an independent test dataset to accurately estimate their real-world performance.

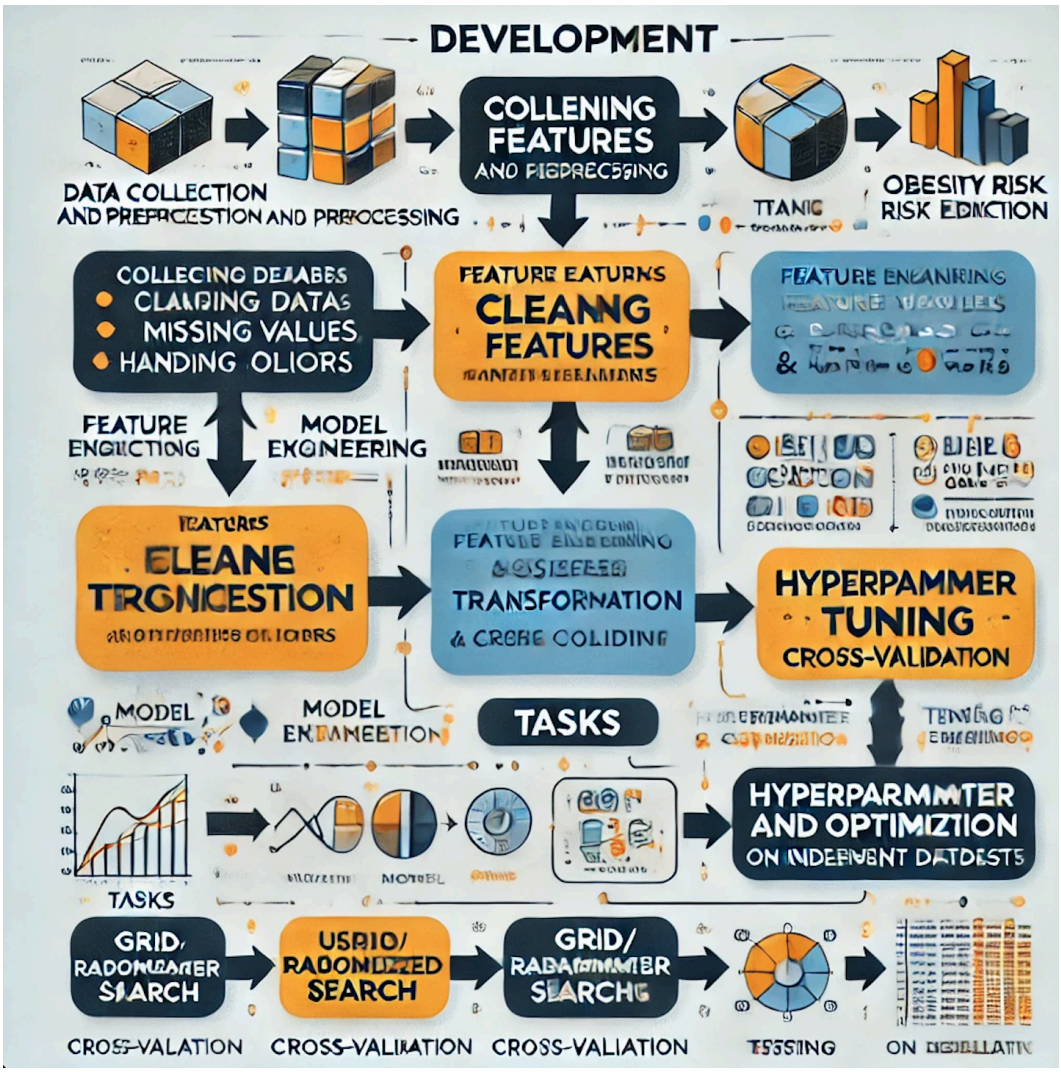


Figure 2 (Development Phases)

Software Requirement Specifications

3.1 Formalization of Problem Description

The pervasiveness of obesity presents a critical challenge to public health, impacting individuals' well-being across the globe. However, accurately predicting obesity risk, particularly in adults, remains a formidable task. This challenge is further compounded among individuals already classified as overweight, who face an elevated risk of developing obesity-related health complications. To address this gap in predictive capabilities, there is a pressing need for robust machine learning (ML) models capable of accurately identifying individuals at risk of obesity. These models must leverage advanced technologies to provide timely interventions and mitigate the adverse effects of obesity on public health.

3.2 Functional Requirements

1. Initiating the system and specifying the sources from which relevant health and demographic data will be obtained to predict obesity risk accurately using machine learning models.
2. Develop a predictive model to assist in identifying individuals at risk of obesity-related health complications.
3. Implementing features to reduce the incidence of obesity-related diseases by providing timely interventions.
4. Outlining the validation techniques to ensure the generalization and reliability of the predictive models.
5. Utilizing data-driven insights to assist healthcare professionals in offering personalized obesity prevention strategies.
6. Specifying requirements for designing a user-friendly interface to facilitate easy access and interpretation of predictive results by healthcare professionals and end-users.

3.3 Non Functional Requirements

1. Ensuring the obesity risk prediction system can handle large volumes of data efficiently and provide real-time results.

2. Maintaining a high level of accuracy in predicting obesity risk to enable reliable decision-making by healthcare professionals.

3. Designing the system to scale seamlessly as the volume of data and user requests increases over time.

4. Ensuring the system's reliability by minimizing downtime and implementing robust error handling mechanisms.

5. Creating a user-friendly interface that is intuitive and easy to navigate for healthcare professionals and end-users.

6. Implementing robust security measures to safeguard sensitive health data and comply with privacy regulations.

7. Ensuring compatibility with existing healthcare systems and data formats to facilitate seamless integration and data exchange.

8. Designing the system with modular components and well-documented code to facilitate easy maintenance and updates.

3.4 Context Diagram

The context diagram shown in Figure 2, illustrates the entire application as a single process. It briefly represents how the application works.

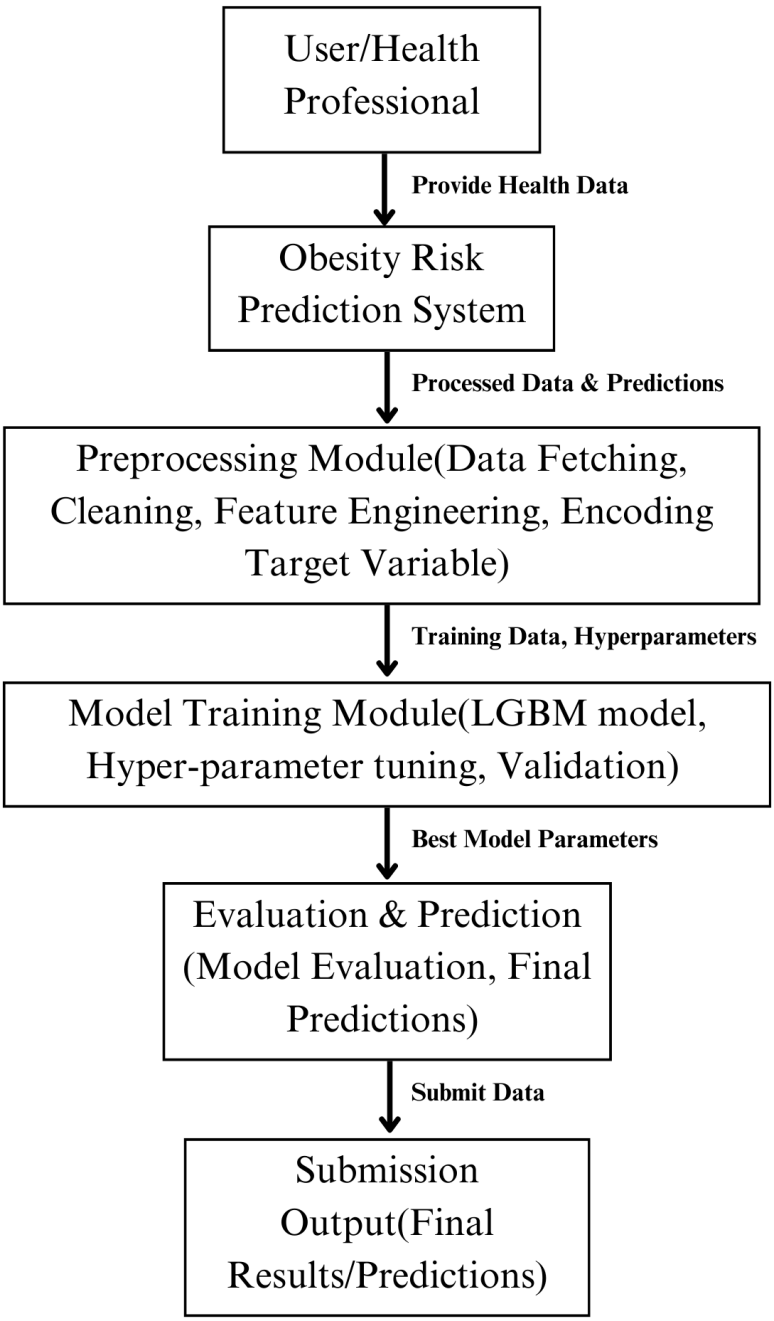
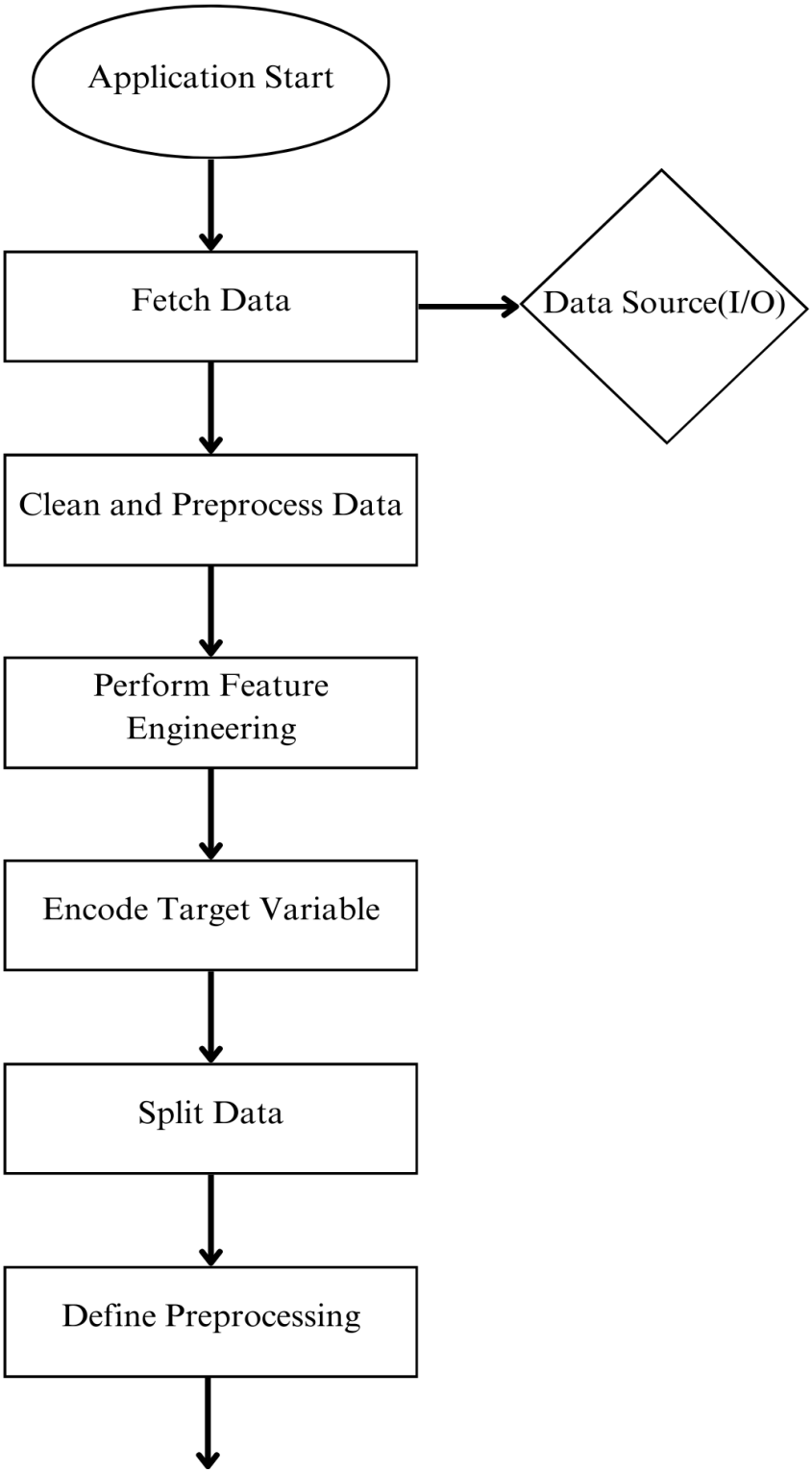
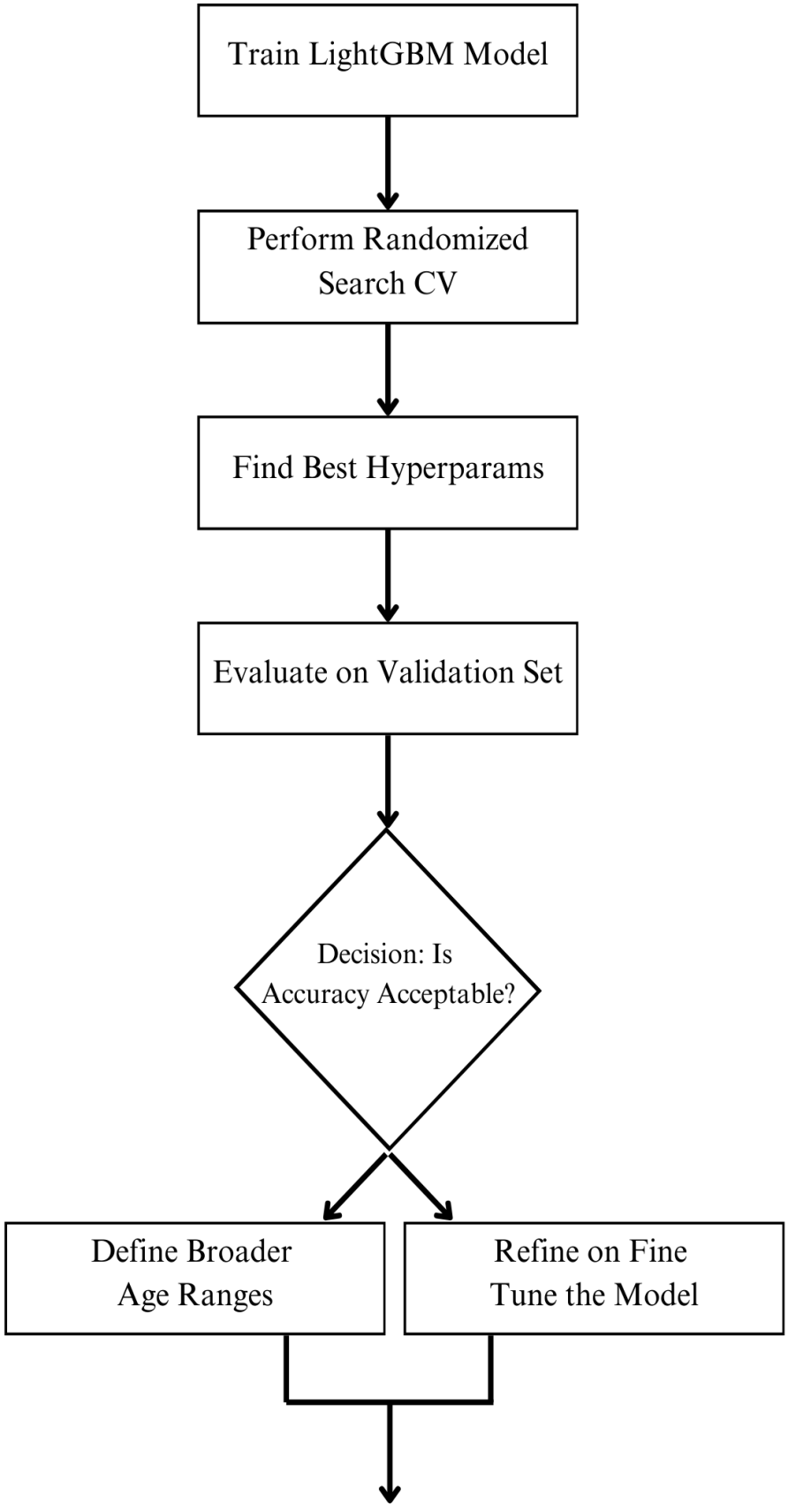


Figure 3 (Context Diagram)

3.5 Control Flow Diagram

Figure 3 is a control flow diagram and it represents a flow of process in the entire application.




Train LightGBM Model
Perform Randomized Search CV
Find Best Hyperparams
Evaluate on Validation Set
Decision: Is Accuracy Acceptable?
Define Broader Age Ranges
Refine on Fine Tune the Model

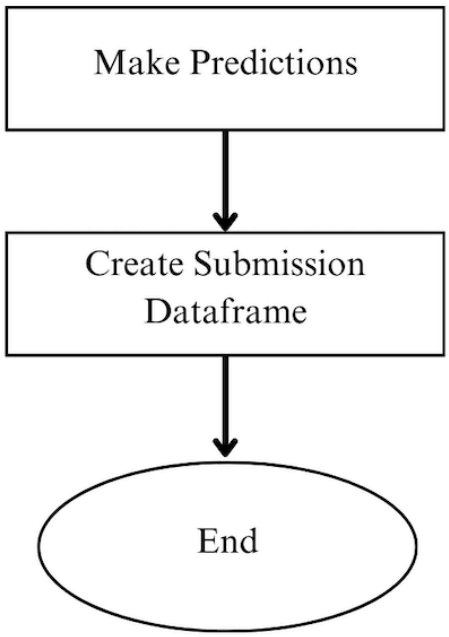


Figure 4 (Context Flow Diagram)

3.6 Constraints

1. Ensuring compatibility of the system with existing systems and software.

2. Efficiently utilizing and computing resources within budget constraints and resource management.

3. Maintaining the accuracy and consistency of data throughout the prediction process.

4. Complying with regulatory standards and ethical guidelines for data handling and analysis.

Environment

4.1 Computational Environment

Programming Language and Libraries

- The project utilizes Python as the primary programming language, leveraging its extensive ecosystem of open-source libraries such as **scikit-learn** for implementing machine learning models.

Database Management

- MySQL is employed for managing datasets, running SQL queries, and performing data manipulation tasks efficiently.

Initial Data Handling

- Microsoft Excel is used for initial data preprocessing, cleaning, and transformation, ensuring the raw data is structured and ready for advanced analysis.

Machine Learning Techniques

- Various machine learning algorithms, including logistic regression, decision trees, and neural networks, are explored for training and testing models to enhance prediction accuracy.

Integrated Development Environments (IDEs)

- The development process is facilitated using Jupyter Notebook, known for its interactive and shareable coding environment, and Visual Studio Code, a robust and versatile IDE preferred for its efficiency and support for large-scale development projects.

Database Management Systems

- Systems like MySQL or SQLite are integrated for storing and managing large datasets required for the project.

Version Control and Collaboration

- Git, along with platforms like GitHub or GitLab, is employed for version control, ensuring smooth collaborative development and efficient code management.

Documentation and Collaboration Tools

- Google Docs is utilized to maintain organized, accessible, and editable documentation, fostering seamless collaboration among team members throughout the project's lifecycle.

User Interface Development

- The project aims to create a user-friendly and reliable interface for the obesity risk predictor, emphasizing ease of use and accessibility for all users.

Exploratory Data Analysis

Age Group vs Obesity Levels

Analyzed the distribution of obesity levels across different age groups to identify trends and correlations, revealing key age brackets at higher risk of obesity.

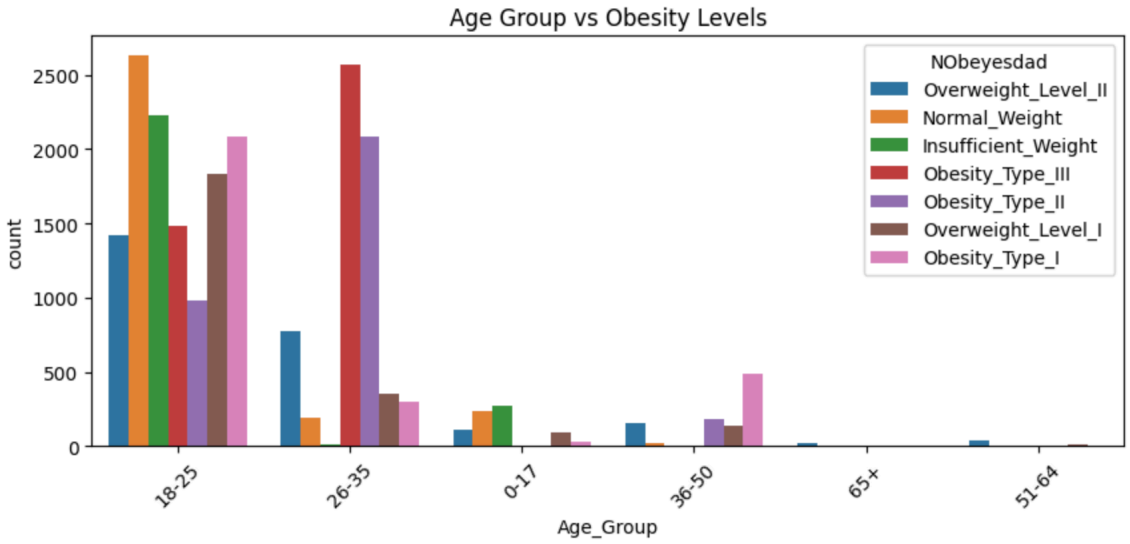


Figure 5 (Age Group vs Obesity Levels)

Distribution of Obesity Levels

Explored the overall distribution of obesity levels within the dataset to understand the prevalence and severity of obesity across the population.

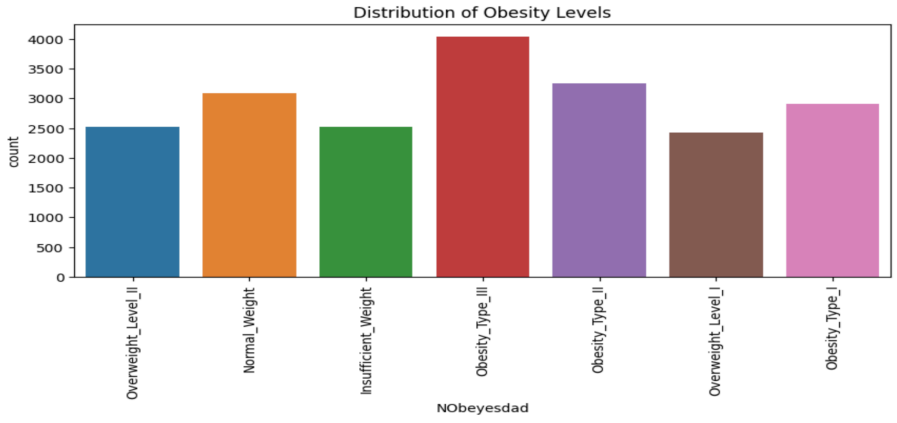


Figure 6 (Distribution of Obesity Levels)

BMI Distribution by Gender

Analyzed the distribution of BMI values across genders to identify patterns or disparities in weight categories and obesity risks between males and females.

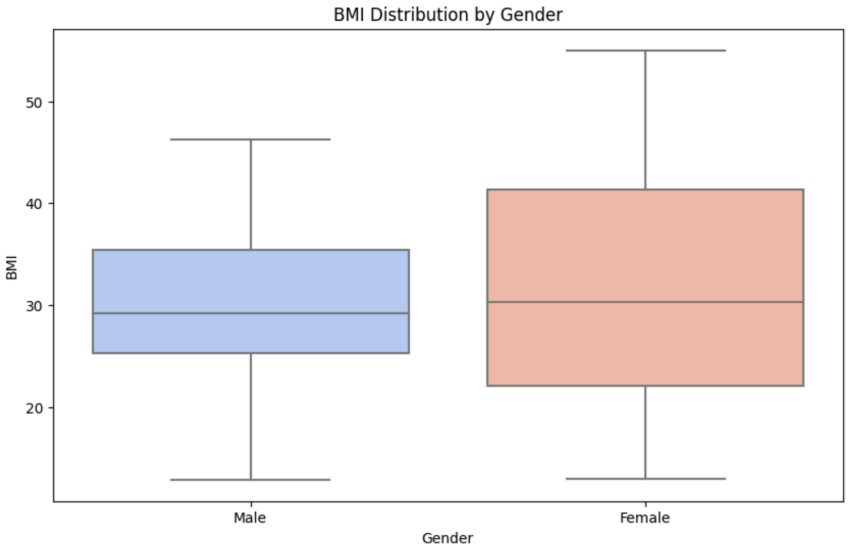


Figure 7 (BMI Distribution by Gender)

Family History of Overweight vs Obesity Levels

Examined the relationship between family history of being overweight and obesity levels to assess its influence on an individual's obesity risk.

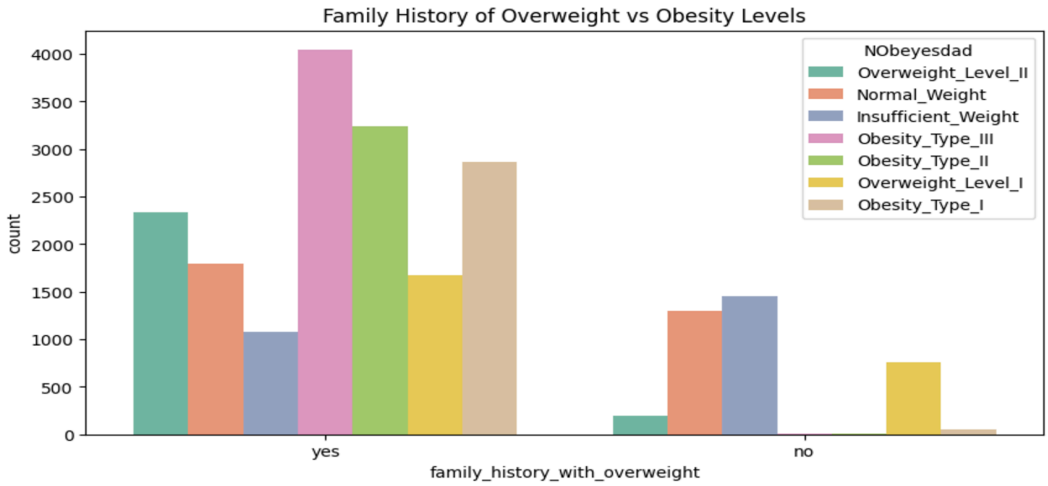


Figure 8 (Family History of Overweight vs Obesity Levels)

Correlation Heatmap of Numerical Features

Generated a heatmap to visualize correlations among numerical features, identifying strong relationships that can influence obesity prediction and guide feature selection.

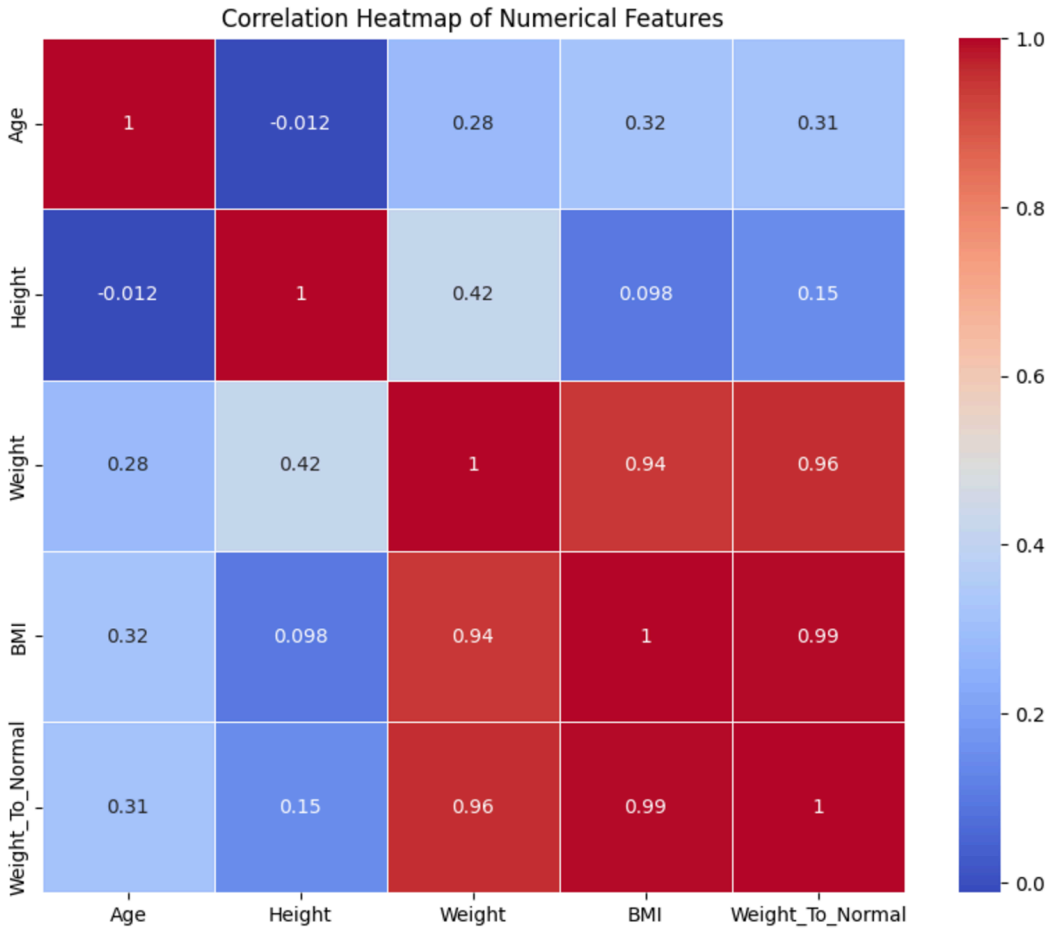


Figure 9 (Correlation Heatmap of Numerical Features)

Mode of Transportation vs Obesity Levels

Examined the relationship between individuals' preferred modes of transportation and their obesity levels to uncover potential lifestyle factors impacting obesity risk.

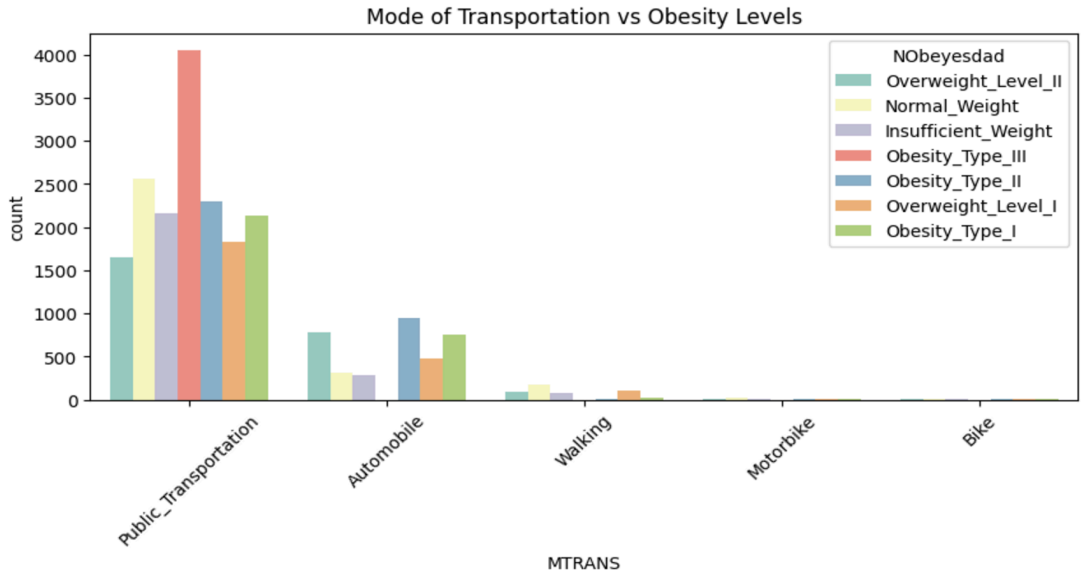


Figure 10 (Mode of Transportation vs Obesity Levels)

Ethical Considerations

For this project, ethical protocols were adhered to when gathering and accessing health data from the OpenML website [OpenML 2023], where the data is verified for research purposes [Emanuel et al. 2000]. Biomedical research ethics are founded on long-established principles such as the Declaration of Helsinki, the Nuremberg Code, and the Belmont Report, all of which are designed to protect individuals from harm and ensure that research is conducted with informed consent [Topol 2013].

As technology continues to evolve, particularly with the rise of digital health data, new opportunities for research have become available. Platforms like OpenML allow researchers to work with health data without direct interaction with participants, leveraging the growing availability of digital health data and biobanks. This shift has led to a need for careful ethical management of data to ensure it's used responsibly and transparently [Kallinikos and Tempini 2014].

In the project, it was ensured that the data was obtained from legitimate and verified sources, following strict protocols to protect participant privacy and consent. This approach helped maintain ethical standards while using health data for research purposes [Nuffield Council on Bioethics 2015].

Implementation

The project aimed to predict obesity risk in adults using machine learning, specifically the LightGBM algorithm. The first step involved importing data from the OpenML website, followed by data preprocessing and cleaning to handle missing values, outliers, and ensure feature consistency. The target variable was encoded into numerical labels, making it compatible for model training. After this, the dataset was split into features and the target variable, and the data was divided into training and validation sets for model evaluation.

Next, preprocessing steps were defined to standardize the data, and the LightGBM algorithm was selected for its efficiency and effectiveness in handling large datasets. A pipeline was created to streamline the integration of preprocessing with model training. Hyperparameters for the model were fine-tuned using randomized search cross-validation, ensuring the model was optimized for better accuracy.

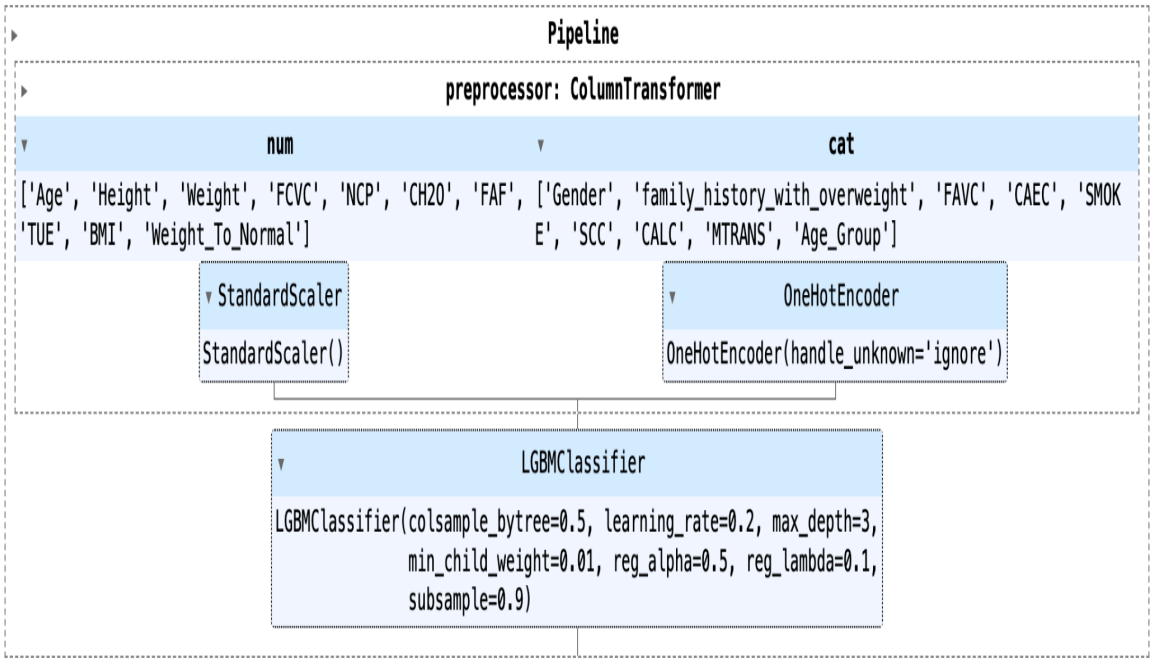


Figure 11 (Pipeline)

To predict obesity risk accurately, the project utilizes a structured machine learning pipeline shown in Figure 4, optimized with "**RandomizedSearchCV**". The pipeline is designed to preprocess both numerical and categorical features effectively.

Numerical variables, such as Age, Height, Weight, BMI, and others, are standardized using StandardScaler to ensure they are on a consistent scale, improving the model's ability to learn patterns.

Categorical features, including Gender, family history with overweight, FAVC, MTRANS, and Age Group, are transformed using OneHotEncoder, which handles unknown categories gracefully to avoid errors during prediction.

At the core of the pipeline is the "**LGBMClassifier**", a powerful gradient boosting algorithm which I chose for its speed and performance on structured data. To fine-tune the model, RandomizedSearchCV explores a wide range of hyperparameters, efficiently identifying the optimal configuration without excessive computational overhead. This streamlined process ensures the pipeline is robust, adaptable, and capable of delivering precise predictions for obesity risk.

Once the best parameters were identified, the model was evaluated on the validation set to assess its performance. After achieving the desired results, the best model was retrained on the entire dataset to improve generalization. Test data, including broader age ranges, was then prepared, and predictions were made based on this expanded dataset. The results were organized into a submission dataframe for further analysis, allowing for reliable obesity risk predictions in adults.

7.1 Model Optimization and Performance

After testing various settings for the machine learning model, a sweet spot was found that gave excellent results - correctly predicting 90.58% of cases in our validation testing. Also, carefully tuned the model to avoid both under- and over-learning from training data. By using a balanced combination of settings, including a moderate learning rate and thoughtful sampling of the features and data points, was able to achieve a model that performs reliably and effectively. The final configuration struck just the right balance between making the model sophisticated enough to capture important patterns, while keeping it simple enough to generalize well to new data.

```
Best Parameters: {'classifier__subsample': 0.9, 'classifier__reg_lambda': 0.1, 'classifier__reg_alpha': 0.5, 'classifier__n_estimators': 100,
'classifier__min_child_weight': 0.01, 'classifier__min_child_samples': 20, 'classifier__max_depth': 3, 'classifier__learning_rate': 0.2, 'cla
ssifier__colsample_bytree': 0.5}
```

Figure 12 (Best Parameters)

```
Validation Accuracy: 0.9058285163776493
```

Figure 13 (Accuracy)

```
[CV] END classifier__colsample_bytree=1.0, classifier__learning_rate=0.05, classifier__max_depth=9, classifier__min_child_samples=50, classif
ier__min_child_weight=0.1, classifier__n_estimators=200, classifier__reg_alpha=1.0, classifier__reg_lambda=0.0, classifier__subsample=1.0; to
tal time=   8.6s
```

Figure 14 ([CV] End Classifier)

Testing

Some general testing cases for obesity risk prediction:

1. General Accuracy Test: Evaluated the model's accuracy on the test dataset to check if it can correctly predict obesity risk. The model should be able to provide a reasonable accuracy score, indicating its ability to generalize to new data.

2. Missing Data Test: Tested how the model handles missing values in key features like BMI or age. It should either impute the missing data or handle it without errors, ensuring smooth predictions.

3. Class Imbalance Test: Tested the model on a dataset with imbalanced classes (more non-obese than obese). The model should still provide meaningful predictions, and metrics like precision and recall should be considered.

4. Boundary Values Test: Tested individuals at the boundary of risk categories (e.g., BMI = 30). The model should correctly classify individuals at these thresholds.

5. Outlier Handling Test: Provided extreme values, such as high BMI or low activity levels. The model should handle these outliers appropriately, either predicting correctly or flagging them as unusual cases.

Summary

This project focused on developing a machine learning model to be able to predict obesity levels in adults using add-on features such as BMI, age, physical activity, and dietary habits along with present features in the data. By leveraging a supervised learning approach with the LightGBM algorithm, the model was trained on processed data to identify patterns and classify individuals into different obesity risk categories. The predictions provide insights not only into a person's current obesity level but also suggest actionable steps for achieving a normal BMI. For individuals at risk, the model calculates how much weight needs to be lost or gained to reach the normal BMI range, offering a practical perspective on health improvement.

Conclusion

The machine learning model will be able to successfully predict obesity levels with reasonable accuracy, enabling better understanding and management of obesity-related risks. By providing personal based recommendations on weight adjustments needed to achieve a normal BMI, the model has the potential to support individuals in making informed lifestyle changes. This approach depicts the effectiveness of data-driven methods in addressing public health challenges, particularly in early risk detection and prevention. Future improvements could include integrating real-time health data or expanding the model to accommodate broader demographic variations for even greater applicability.

Limitations and Future Work

Limitations

1. Dataset Bias

The dataset primarily represents certain demographics, which may limit the model's usefulness to more diverse populations.

2. Feature Limitations

Critical variables like genetic factors, mental health indicators were unavailable in the dataset, which might have potentially affected the prediction accuracy.

3. Model Complexity

The LightGBM model, while efficient, may be challenging to interpret for healthcare professionals or end-users unfamiliar with machine learning models and algorithms..

4. Computational Requirements

Training and fine-tuning the model require significant computational resources, which may restrict scalability for large-scale applications.

5. Temporal Irrelevance

Static data does not account for evolving individual behaviors or environmental changes that impact obesity risk over time.

Future Work

1. Incorporating Longitudinal Data

Expanding the dataset to include temporal data for better predictions of obesity trends over time.

2. Adding Behavioral and Genetic Features

Including factors like genetic inclination, stress levels, and real-time activity tracking to enhance prediction accuracy.

3. Global Model Adaptation

Customizing the model for different populations and regions to ensure broader applicability and fairness.

4. Improving Explainability

Developing tools or frameworks to make the model's predictions more interpretable and actionable for healthcare practitioners.

Bibliography

[WHO 2023] World Health Organization. Global Overview of Obesity. WHO Health Topics, (2023).

[CER Bariatrics 2023] CER Bariatrics. Global Obesity Trends in 2023. CER Bariatrics, (2023).

[NICHD 2023] National Institute of Child Health and Human Development. Causes of Overweight and Obesity. NICHD, (2023).

[Frontiers 2023] Frontiers in Endocrinology. Advances in Machine Learning for Obesity Prediction. Frontiers in Endocrinology, (2023).

[PMC 2023] National Center for Biotechnology Information. Exploring Obesity Risk Through ML. PubMed Central (PMC), (2023).

[AIMS Press 2022] AIMS Press. Machine Learning in Obesity Risk Detection. AIMS Press, (2022).

[SciDirect 2021] SciDirect. Applications of Machine Learning in Predictive Medicine. ScienceDirect, (2021).

[Research Square 2023] Research Square. Machine Learning for Predictive Obesity Risk. Research Square Preprint, (2023).

[CEUR-WS 2021] CEUR Workshop Proceedings. Machine Learning Applications in Predictive Health. CEUR-WS.org, Vol. 3038, (2021).

[PMC 2022] National Center for Biotechnology Information. Obesity Risk Analysis with ML Algorithms. PubMed Central (PMC), (2022).

[Jeong et al. 2024] Jeong, H., et al. "Future of Health AI Systems." Journal of Medical AI, (2024).

[Siddiqui et al. 2021] Siddiqui, R., et al. "Ethics in ML-Driven Health Systems." Journal of AI Ethics, (2021).

[Kılıç 2023] Kılıç, İlyurek. Light GBM: A Powerful Gradient Boosting Algorithm. Medium, (2023).

[İlyurek 2023] İlyurek, T. Deep Learning Applications in Obesity Research. ResearchGate, (2023).

[Emanuel et al. 2000] Emanuel, E. J., Wendler, D., and Grady, C. "What Makes Clinical Research Ethical?" JAMA, 283, 2701–11, 2000. doi: 10.1001/jama.283.20.2701.

[Topol 2013] Topol, E. J. The Creative Destruction of Medicine: How the Digital Revolution Will Create Better Health Care. 1st ed., Basic Books, 2013.

[Kallinikos and Tempini 2014] Kallinikos, J., and Tempini, N. "Patient Data as Medical Facts: Social Media Practices as a Foundation for Medical Knowledge Creation." Information Systems Research, 25, 817–33, 2014. doi: 10.1287/isre.2014.0544.

[Nuffield Council on Bioethics 2015] Nuffield Council on Bioethics. The Collection, Linking, and Use of Data in Biomedical Research and Health Care: Ethical Issues. The Nuffield Council on Bioethics, 2015.

[OpenML 2023] OpenML. Obesity Risk Dataset: Active Dataset ID 45969. OpenML, (2023).

Appendix A: Environment Setup

A) System Setup

The setup includes libraries for data manipulation, visualization, machine learning model building, and a user interface for predictions. The project was implemented on Jupyter Notebook, which provided an interactive space for coding, testing, and visualizing the data and model. It allowed for easy step-by-step execution of tasks like data cleaning, model training, and prediction, with instant feedback through graphs and results. Using Jupyter made it simple to combine code with explanations, making the entire process clear and easy to follow.

B) Libraries Imported

The project utilized several key libraries(shown in Figure 5) to streamline the workflow. **Pandas** was used for data manipulation, helping with data cleaning and transformation. **Matplotlib.pyplot** and **Seaborn** provided powerful visualization tools for exploring data and model performance. For machine learning tasks, the **sklearn** modules were essential for preprocessing, building the pipeline, selecting models, and evaluating their performance. **LightGBM** served as the classifier, offering fast and efficient model training. Additionally, **Tkinter** was used to create a user-friendly graphical interface, enabling easy interaction with the model.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split, RandomizedSearchCV
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer

from sklearn.preprocessing import StandardScaler, OneHotEncoder, LabelEncoder
from lightgbm import LGBMClassifier
```

Figure 15 (Libraries)

C) Preprocessing Pipeline

The preprocessing pipeline involved two main steps. First, **StandardScaler** was used for "numeric transformation", which standardized the data to make sure all numerical features had a similar scale, with a mean of 0 and a standard deviation of 1. This helped the model work more efficiently. Then, for "categorical transformation", **OneHotEncoder** was used to convert categorical variables into a binary format, allowing the model to interpret them without losing any important information. This made the data ready for training the machine learning model.

Part 1: RandomizedSearchCV Setup

- **RandomizedSearchCV** is used for hyperparameter tuning with 3-fold cross-validation (`cv=3`), ensuring that the model is evaluated with different data splits during the search process. This helps in selecting the best hyperparameters while preventing overfitting.

Part 2: Preprocessing Pipeline

- The pipeline starts with preprocessing:
    - **Numerical features** like Age, Height, Weight, etc., are standardized using **StandardScaler**.
    - **Categorical features**, such as Gender and Smoking history, are encoded using **OneHotEncoder**, which converts categories into binary vectors.

Figure 16 (RandomizedSearchCV setup and Pipeline)

Part 3: Hyperparameter Tuning with RandomizedSearchCV

- The **RandomizedSearchCV** evaluates a range of hyperparameters for the LGBMClassifier, such as max_depth, n_estimators, subsample, and regularization parameters (reg_alpha, reg_lambda). It performs the search to find the best combination that enhances model performance.

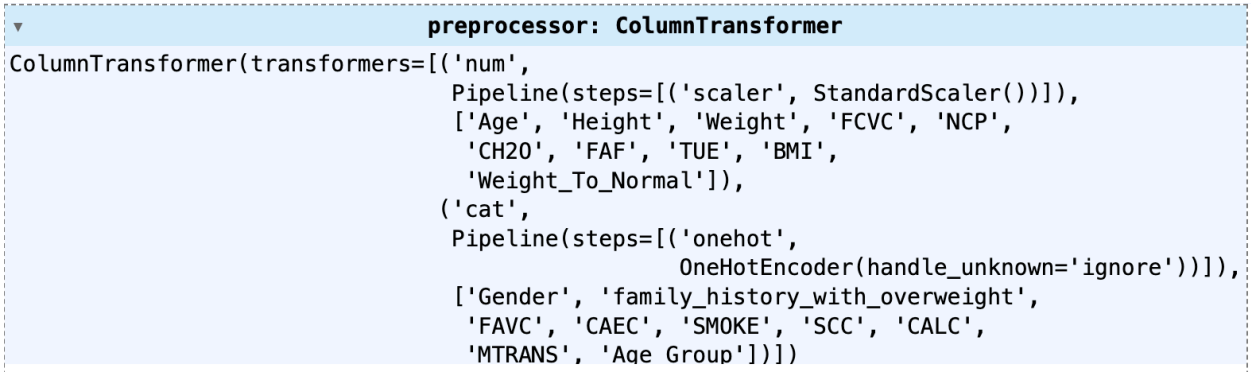


Figure 17 (Preprocessor, Hypertuning)

Part 4: LGBMClassifier

- After preprocessing, the **LGBMClassifier** is applied to the data. This is a powerful gradient boosting model, ideal for handling structured data and producing high-quality predictions.

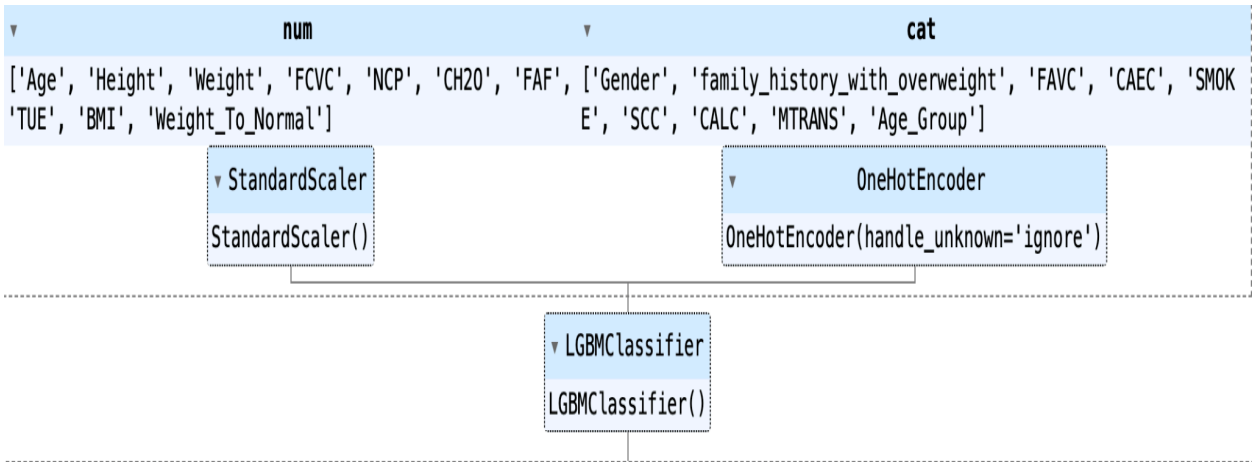


Figure 18 (LGBM Classifier)

D) Graphical User Interface

The **Graphical User Interface (GUI)** for this project was built using **tkinter**, a popular Python library for creating desktop applications. The GUI allows users to input relevant data, such as age, height, weight, and other health-related factors, through an intuitive interface. Once the user enters the data, the system processes it and displays predictions in real-time, after the “Predict” button is being clicked.

This interactive design ensures that users can easily use the model to assess obesity risk, making the tool accessible even to those without a technical background. The GUI enhances the overall user experience by providing a simple yet effective way to interact with the machine learning model.

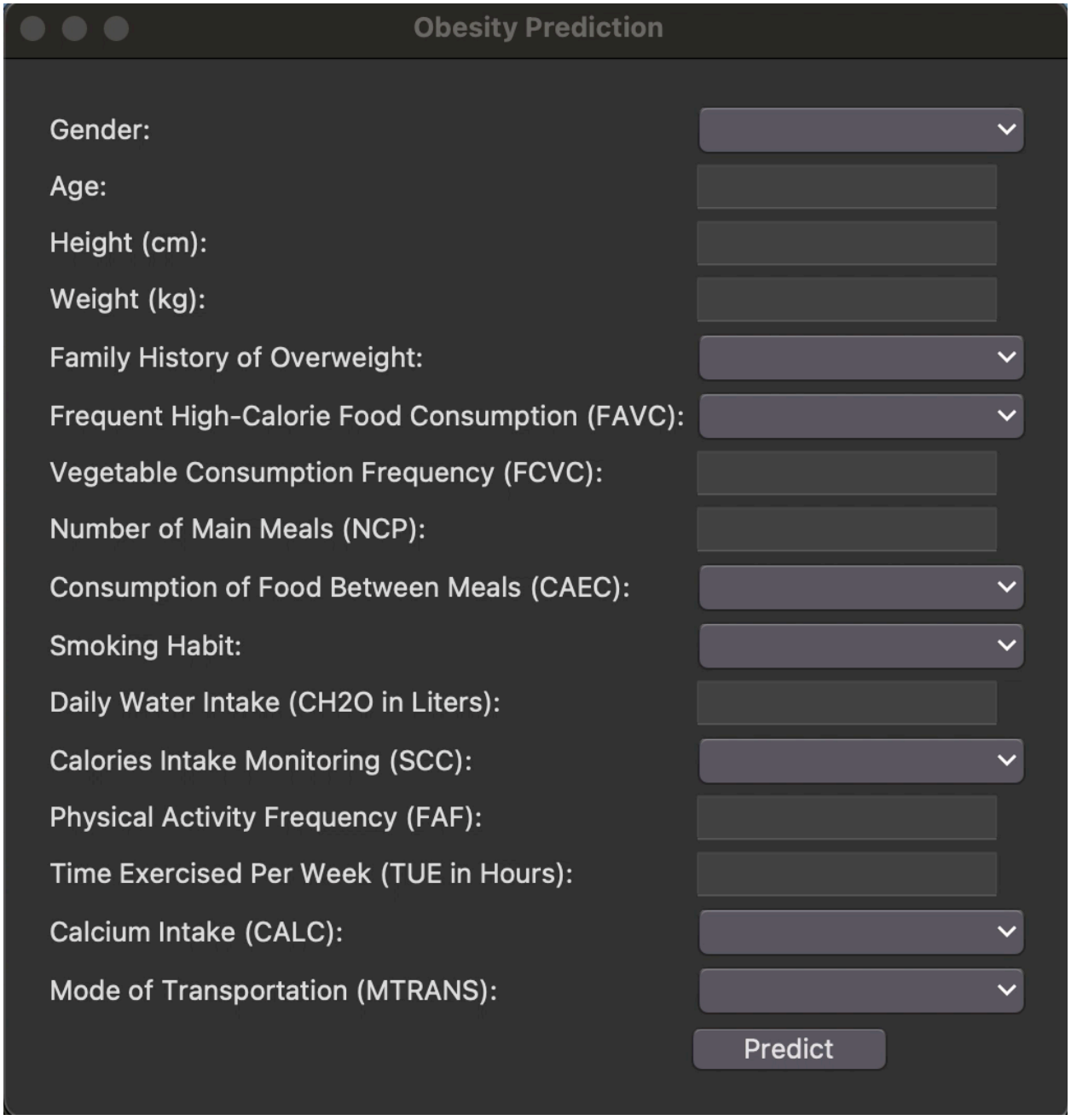


Figure 19 (GUI)

Appendix B: Interactive Model Execution

This appendix provides an overview of how the obesity prediction model operates through a user-friendly graphical interface created with Python's tkinter library. The model allows users to interact with the system by entering personal details, such as age, height, weight, and lifestyle habits, which the system then uses to calculate the Body Mass Index (BMI), assess obesity risk, and offer useful health insights.

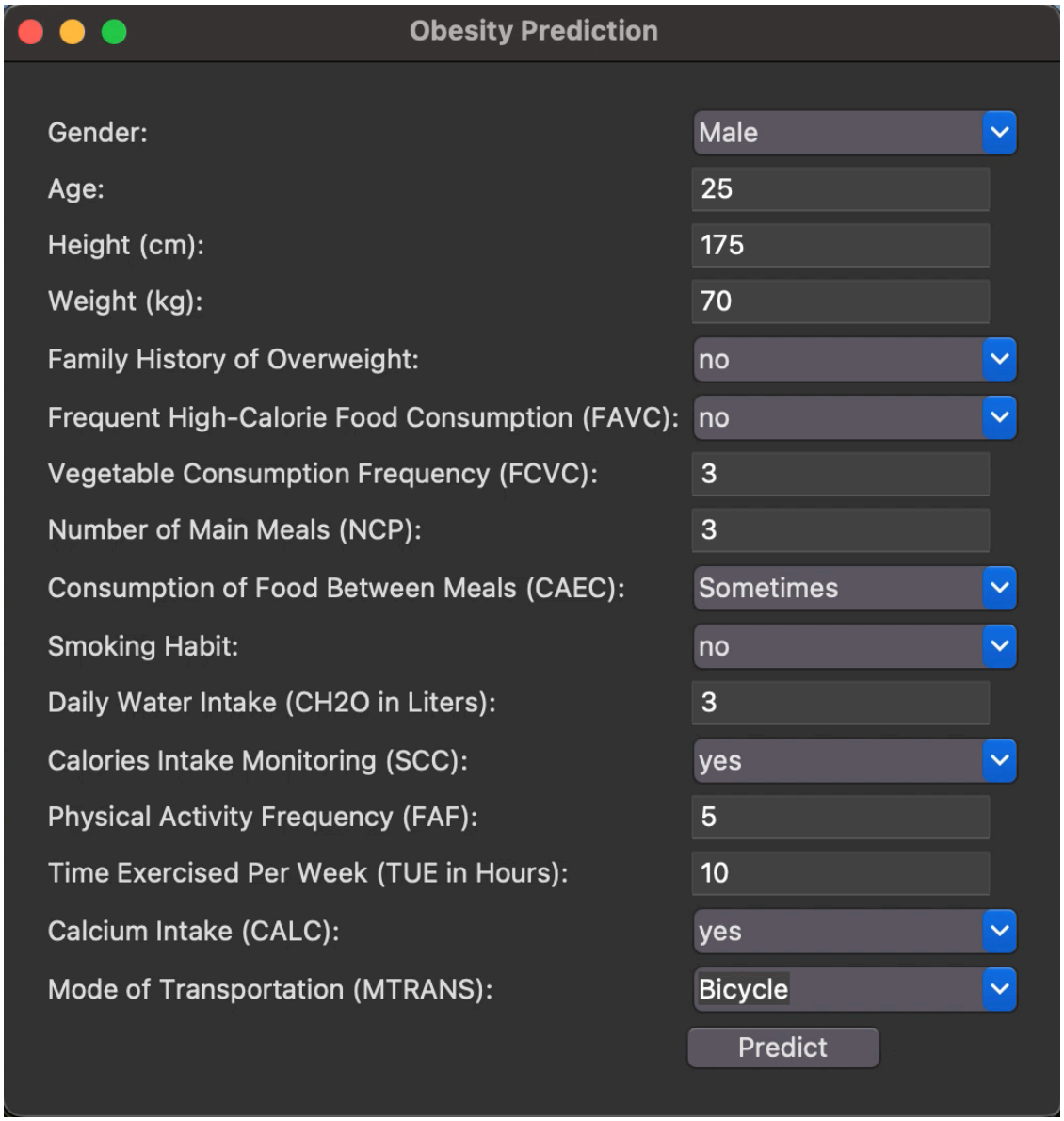


Figure 20 (GUI with Input Field)

1. GUI Layout

The layout of the GUI is clean and organized in a grid format. It features various input fields where users can enter their information, such as age, weight, height, and lifestyle factors. These fields are accompanied by clear labels to guide the user. For categorical data, the user can select from drop-down menus (like Family History or Smoking Habits), and for numerical data, text boxes are provided. This makes the interface both interactive and easy to navigate.

2. Input Fields

To ensure the model can provide an accurate obesity prediction, users are asked to provide a variety of details. This includes basic information such as **Gender**, **Age**, **Height**, and **Weight**, as well as lifestyle factors that can affect health, such as **Family History of Overweight**, **Frequent High-Calorie Food Consumption (FAVC)**, and **Vegetable Consumption Frequency (FCVC)**. Other fields include **Number of Main Meals (NCP)**, **Smoking Habit (SMOKE)**, and **Physical Activity Frequency (FAF)**. These input fields work together to give the model a comprehensive picture of the user's lifestyle, allowing it to generate personalized health predictions.

3. Model Execution

Once the user has filled in all the necessary fields and clicks the "Predict" button, the model runs a series of calculations:

- **BMI Calculation**: The model uses the entered weight and height to calculate the BMI.
- **Obesity Level Determination**: Based on the BMI, the model determines the current obesity level, such as Insufficient Weight, Overweight, or Obesity.
- **Future Risk Prediction**: The model then predicts the future risk of obesity, taking into account the user's current BMI and lifestyle factors.
- **BMR Calculation**: The Basal Metabolic Rate (BMR) is calculated using the user’s age, gender, weight, and height. This number is adjusted based on the user’s activity level to determine their daily caloric needs.

- **Caloric Intake Calculation**: Using the BMR and physical activity level, the model calculates the user's maintenance caloric intake.
- **Obesity Risk**: Finally, the model provides a detailed obesity risk assessment and suggests lifestyle adjustments to help lower future risks.

4. Example of Interaction

Let’s consider an example where the user inputs their details, such as **Gender** (Male), **Age** (25), **Height** (175 cm), **Weight**(70 kg), and fills in lifestyle factors like **Family History of Overweight** (No), **FAVC** (No), and **FCVC** (3). Once the user clicks the "Predict" button, the model calculates the **Predicted BMI** (22.86), which classifies the user’s obesity level as **Normal Weight**..

The model also predicts that the user has a **Moderate risk of future obesity** and suggests making some lifestyle changes to reduce the risk. This process shows how the system uses the information provided to offer personalized health predictions and actionable recommendations.

```
The current person has a normal BMI of 22.86.
Predicted Obesity Level: Normal Weight
BMI: 22.86
Weight to Normal: 0.00 kg
Future Risk Assessment: Low risk of future obesity with a healthy lifestyle.
Healthy Weight Range: 56.66 kg - 76.26 kg
Basal Metabolic Rate (BMR): 1673.75 kcal/day
Activity Multiplier: 1.70
Maintenance Caloric Intake: 2845.38 kcal/day
```

Figure 21 (Interaction Example)