

INFOSYS SPRINGBOARD INTERNSHIP



Predicting Obesity Levels using ML

Under The Mentorship of Mr. Narrendra Kumar

Group:4

Group member's Name::
Rishi Kumar Sah
Rehan Abdul Nayeem
Gurralla Nandhikar
Bejjam Charitha
Hemanth Sri Surya
Thota Manikant
Jafrash

Abstraction:

Obesity, a complex and prevalent global health issue, is influenced by various lifestyle and physical factors. This project leverages machine learning (ML) and deep learning (DL) models to predict obesity levels based on individuals' eating habits and physical conditions using a dataset from Kaggle. The models classify individuals into obesity categories and compare their performance through key evaluation metrics like accuracy, precision, and recall. Additionally, the project explores the most influential lifestyle factors driving obesity, providing actionable insights for healthcare professionals to design targeted interventions and promote healthier lifestyles. The results emphasize the potential of predictive modeling in preventive healthcare strategies.

Introduction:

we use the "Obesity based on eating habits and physical conditions" dataset from Kaggle to develop a predictive model for obesity levels. The dataset contains various attributes related to eating behaviors, physical activity, and demographic information, which serve as inputs to our models. Our approach includes both traditional machine learning techniques, such as decision trees and support vector machines (SVM), as well as more advanced deep learning models like neural networks.

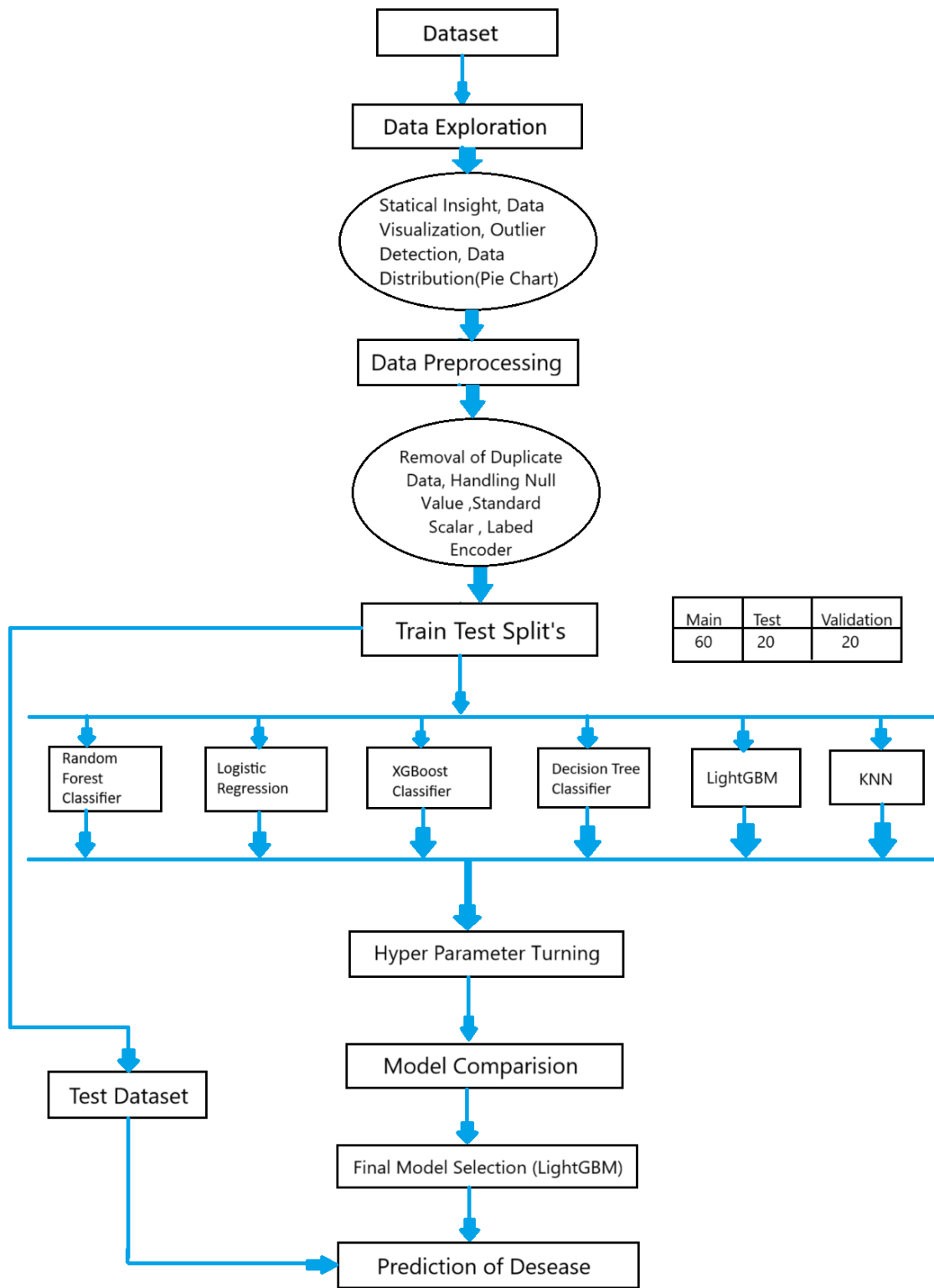
Objective:

The goal of this project is to develop predictive models that can accurately classify individuals into various obesity categories based on their eating habits and physical conditions. By leveraging both machine learning (ML) and deep learning (DL) techniques, the project aims to identify patterns in lifestyle data that correspond to different obesity levels. Additionally, the performance of these models will be compared using evaluation metrics such as accuracy, precision, recall, and others to determine which approach is more effective. Another key objective is to provide insights into the lifestyle factors that contribute the most to obesity prediction, allowing healthcare professionals to design more targeted and personalized interventions for individuals at risk of obesity. This comprehensive analysis not only enhances the predictive power of the models but also offers valuable information for promoting healthier lifestyle choices.

Methodology:

1. Data Exploration
2. Data Preprocessing
3. Splitting Data
4. Feature Selecting
5. Model Emplementation
6. Model comparison
7. Hyperparameter Tuning
8. Final Model Selection
9. WebApp

Flow chart



Data Exploration

Dataset Overview:

The dataset, titled "Obesity based on eating habits and physical conditions" which prepared by merging two different dataset and latter remaned as final_dataset.csv, contains information about individuals' physical conditions and eating habits to predict obesity levels.

1. Dataset Dimensions

After merging and renaming columns, the final dataset consists of 22,869 rows and 18 columns. These columns represent a mix of numerical and categorical variables that capture various lifestyle factors, such as eating habits and physical activity, and their relationship with obesity.

2. Renaming Columns for Better Understanding

To improve clarity, the original column names were renamed as follows:

FAVC → FCOHCF (Frequent consumption of high-calorie food)

FCVC → FCOV (Frequent consumption of vegetables)

NCP → NMM (Number of main meals)

CAEC → COFBM (Consumption of food between meals)

CH2O → CH2O (Daily water consumption)

SCC → Calorie_Consump_Monitoring (Monitoring calorie consumption)

FAF → Physical_Activity_F (Physical activity frequency)

TUE → Time_using techno_D (Time using technological devices)

CALC → Consumption_Alc (Alcohol consumption)

MTRANS → MTRANS (Mode of transportation)

NObeyesdad → NObesity (Obesity levels)

These changes make it easier to interpret and analyze the dataset by reflecting the meaning of each variable.

3. Data Types

The dataset contains a mixture of categorical and numerical variables:

Categorical Variables: Gender, Family history of overweight, FCOHCF, COFBM, Smoking habits, Calorie consumption monitoring, Alcohol consumption, MTRANS, and NObesity (target variable).

Numerical Variables: Age, Height, Weight, FCOV (vegetable consumption), NMM (number of main meals), CH2O (daily water consumption), Physical Activity Frequency, and Time using technology.

4. Statistical and individual variable exploration

4.1 Statistical Summary of Numerical Columns

Age, Height, Weight: These variables show significant variance, with potential outliers identified, particularly in the weight and height columns. These outliers will be handled during the preprocessing stage.

4.2 Statistical Summary of Categorical Columns

Categorical Variables: Each categorical feature shows a proper arrangement of data, with no missing values. They capture lifestyle habits such as smoking, calorie consumption, and transportation modes.

4.3 Gender Distribution

Male: 11404

Female: 11465 The gender distribution is almost equal, which ensures that the dataset is well-balanced for this feature.

4.4 Family History of Overweight

Yes: 18740

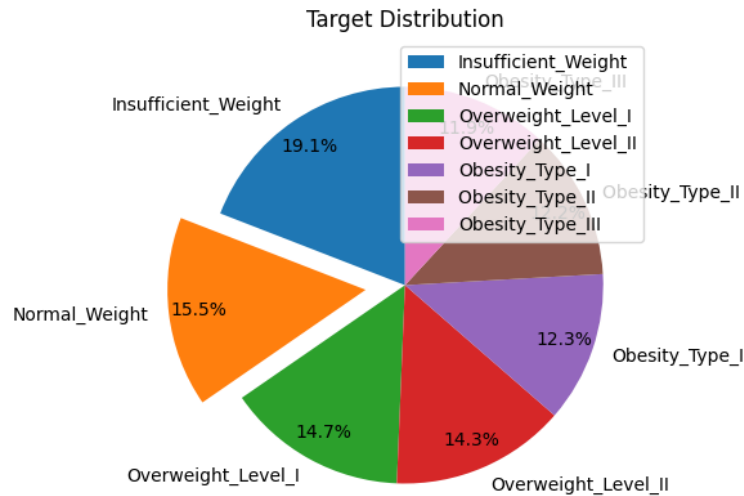
No: 4129 A significant number of individuals have a family history of being overweight, indicating a potential genetic or lifestyle link to obesity.

4.5 Obesity Levels (NObesity)

The dataset includes multiple obesity categories. The "Normal Weight" category is the most frequent, followed by various overweight and obesity types.

4.6 Target Variable Distribution

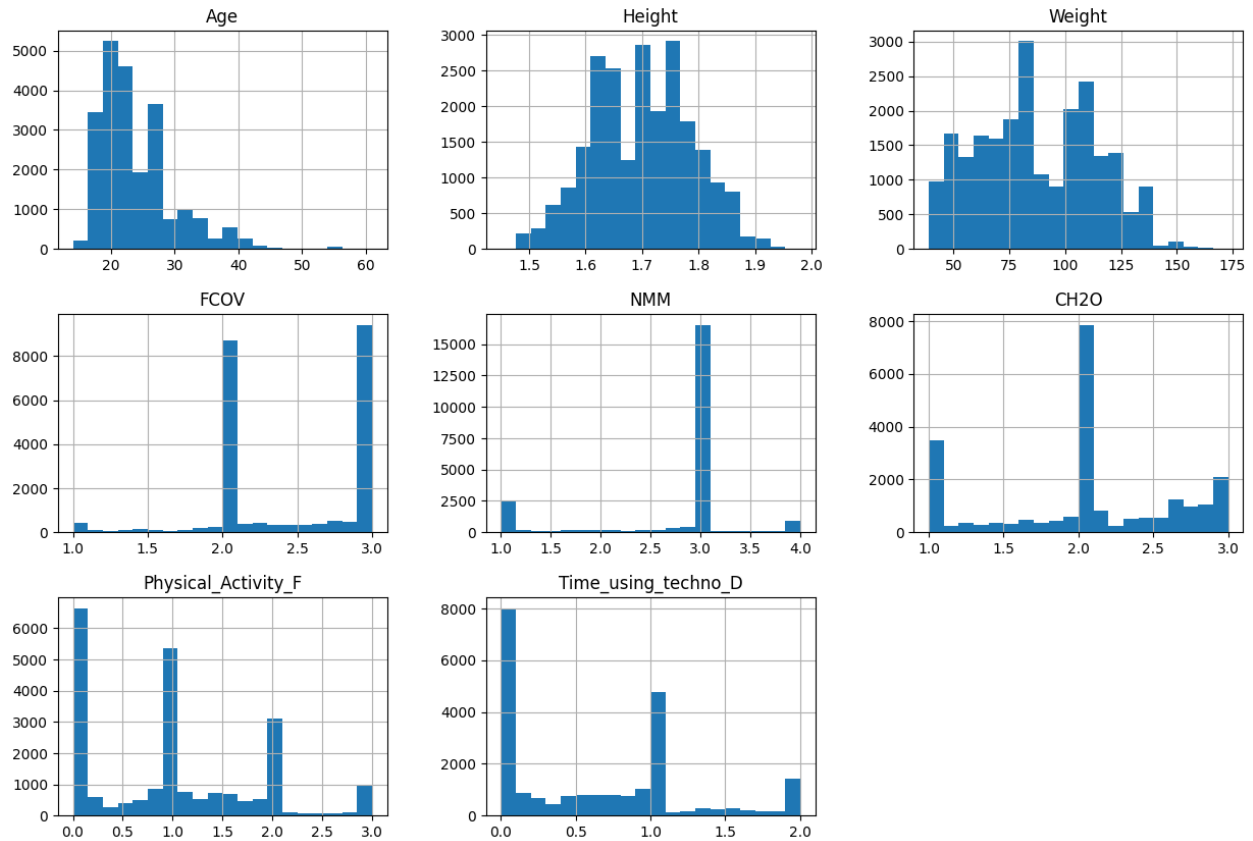
A pie chart was generated to visualize the distribution of obesity levels. The plot highlights that "Normal Weight" and "Obesity Type I" are the dominant categories, giving insight into class distribution.



5. Data Visualization

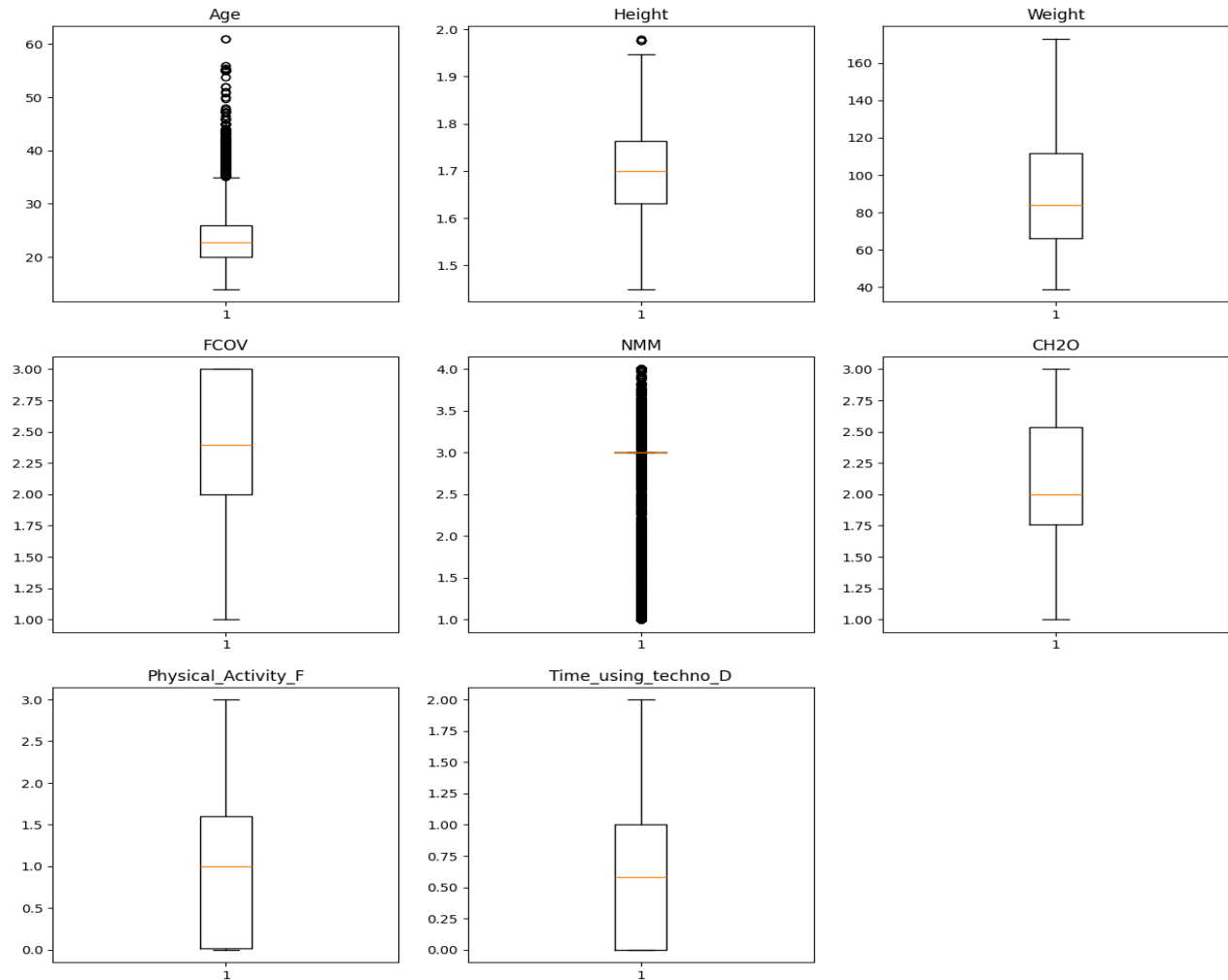
5.1 Histograms for Numerical Variables

Histograms were plotted for numerical features such as age, height, weight, and physical activity frequency to explore their distributions. The distributions reveal normal patterns in some features but skewness and outliers in others, such as weight and height.



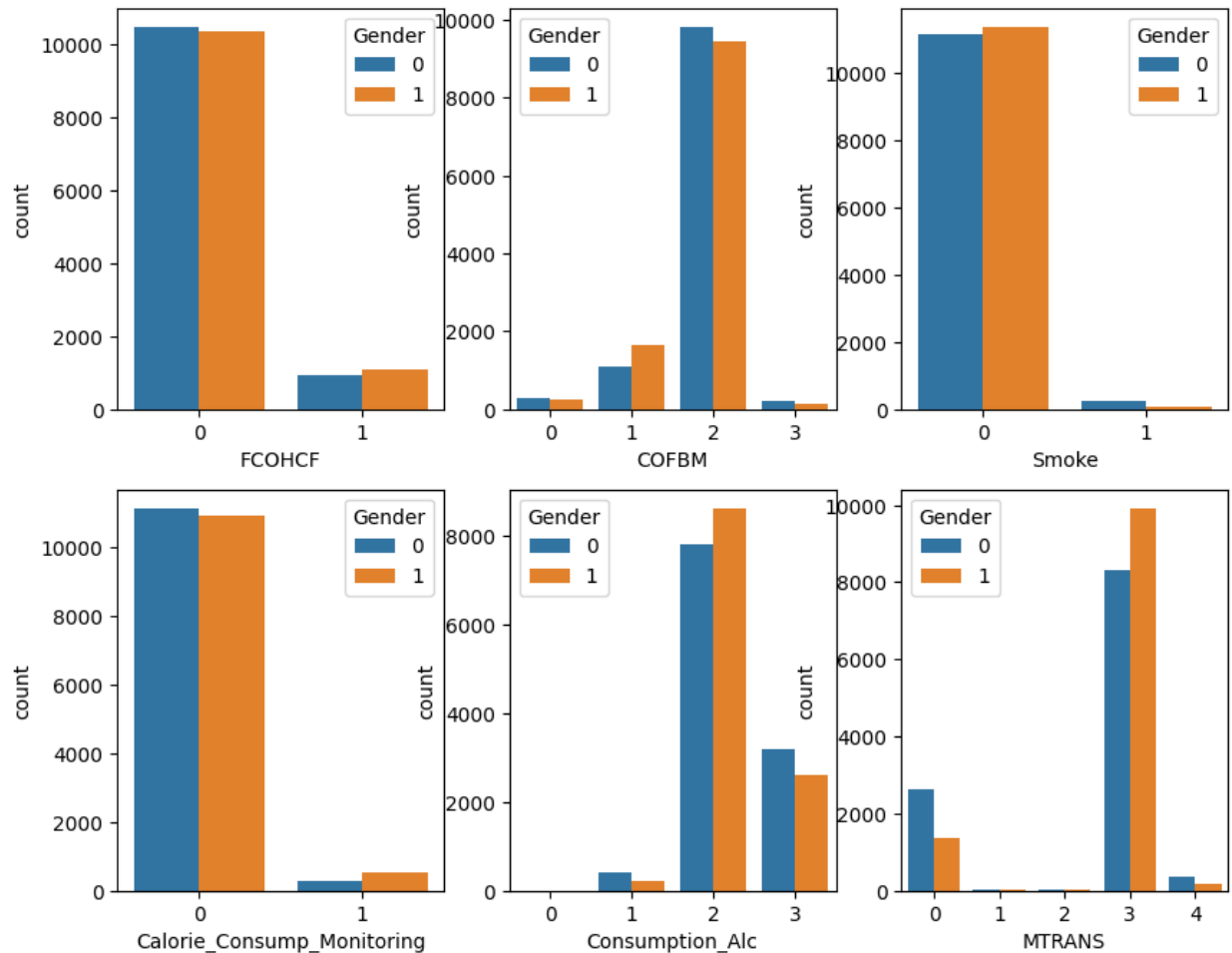
5.2 Boxplots for Outlier Detection

Boxplots were used to identify potential outliers. Features like Age, Weight, Height, and CH2O (daily water consumption) exhibited outliers, which were addressed during preprocessing.



5.3 Countplots for Categorical Variables

Countplots were generated to visualize categorical variables like Gender, FCOHCF (frequent consumption of high-calorie food), COFBM (food consumption between meals), and smoking habits. These plots give insights into the distribution and possible imbalance of categories within the dataset.



Data preprocessing

6.1 Handling Missing Values

The dataset had no missing values. This was confirmed by checking for NaN values and plotting a heatmap, which showed no missing data points.

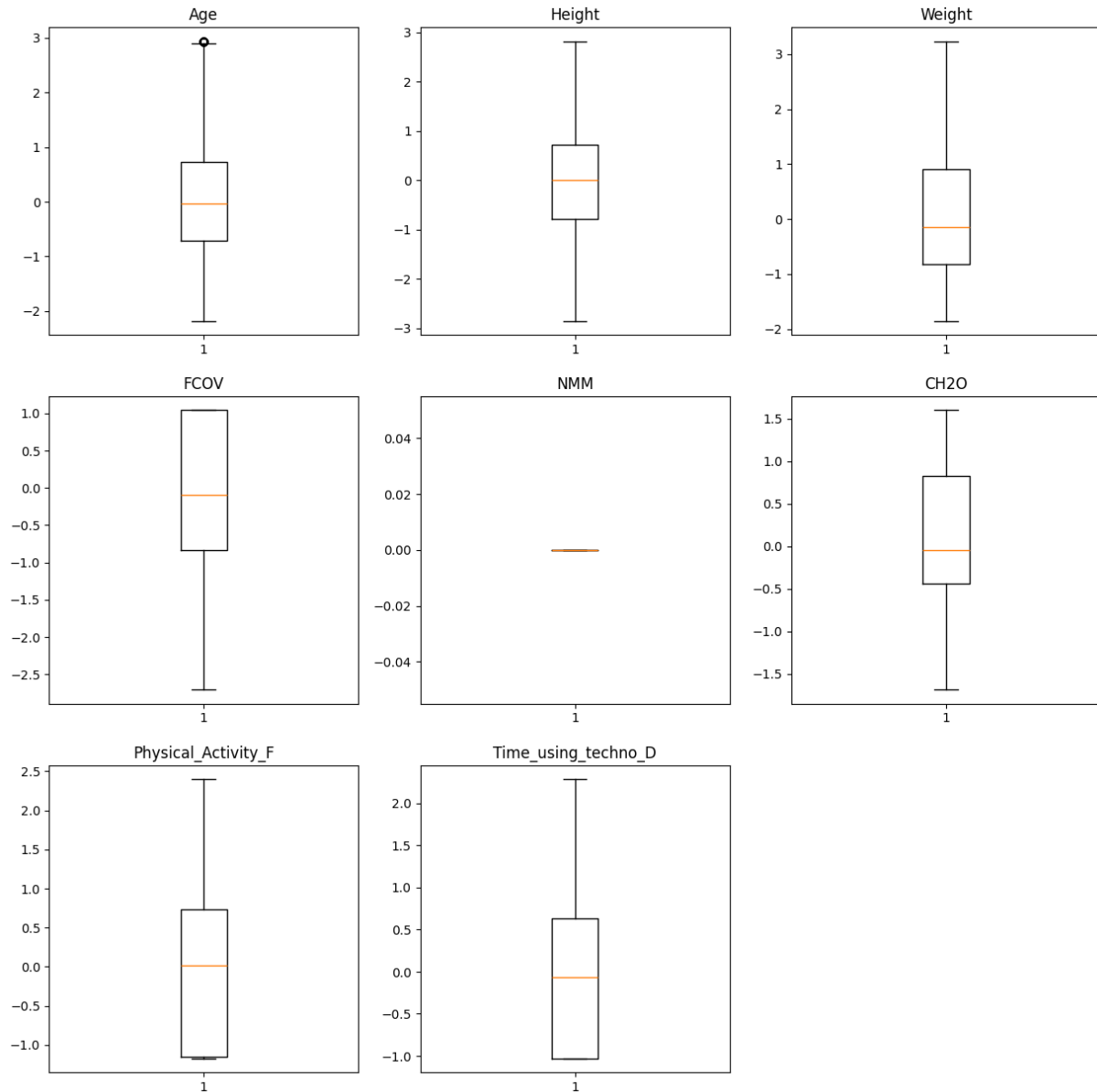
6.2 Duplicate Records

Duplicate records were detected, and they were removed to ensure the integrity of the dataset. This step reduced the dataset to 22,845 rows and 17 columns, with all unique entries ready for analysis.

6.3 Handling Outliers

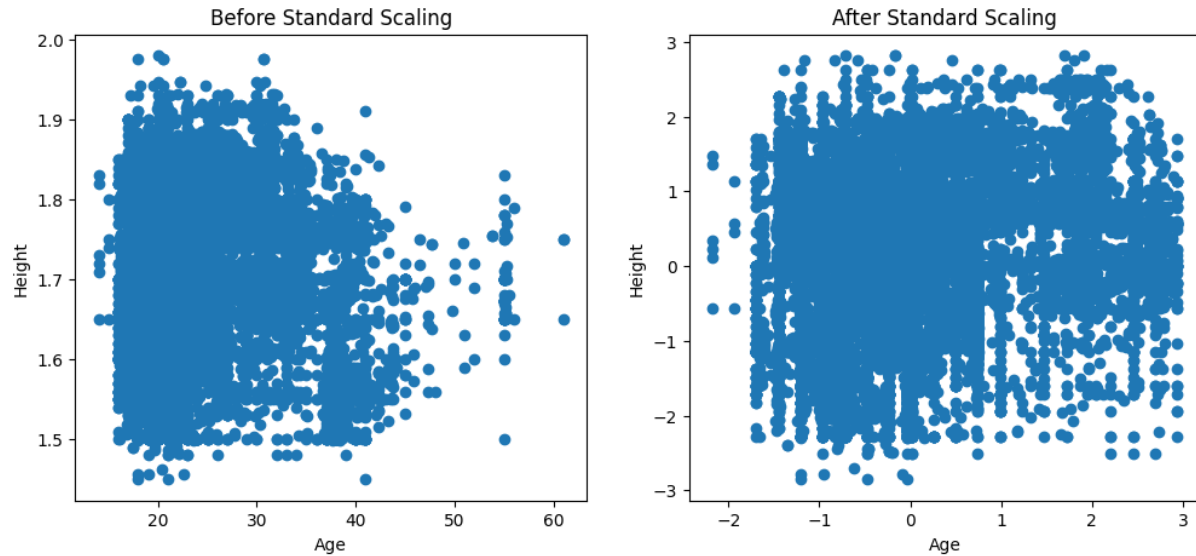
Outliers in numerical columns such as Age, Height, Weight, and NMM were addressed using the Interquartile Range (IQR) method. Outliers were replaced with the median of the respective columns to maintain the distribution without skewing the data.

After Outlier Removed:



6.4 Standard Scaling

To normalize numerical features and reduce the effect of differences in scale, StandardScaler was applied. The scaler transformed the data, and scatter plots were used to visually compare the dataset before and after scaling. For example, Age and Height showed significant transformations, confirming that scaling effectively handled these discrepancies.



7. Encoding Categorical Variables

7.1 Label Encoding

The target variable NObesity and other categorical variables such as COFBM (consumption of food between meals), MTRANS (mode of transportation), and Consumption_Alc (alcohol consumption) were label-encoded to convert them into numerical formats suitable for machine learning models.

7.2 One-Hot Encoding

For other categorical variables such as Gender, Family history with overweight, and FCOHCF, one-hot encoding was applied to preserve relationships without introducing bias through label encoding. This method ensures that the machine learning model interprets these variables correctly.

9. Final Prepared Data

After preprocessing, the dataset consists of 22,845 rows with fully cleaned and encoded numerical and categorical variables. This dataset is now ready for machine learning model development, with each feature properly scaled and all outliers addressed.

10. Data Splitting

To ensure robust model evaluation, the data is split into training, validation, and test sets. The train set is used for model training, the validation set for tuning hyperparameters and avoiding overfitting, and the test set for final performance evaluation.

1. separating independent(feature) variables and independent(target) variable
2. Train, Test, and Validation Split (60,20,20)

11. Feature Selection Using Decision Tree

A Decision Tree classifier is used to determine the importance of each feature. By examining the importance scores, we can select the top features that significantly contribute to predicting obesity levels.

We fit a Decision Tree on the training data and use its built-in feature importance scores to rank the features.

1. Weight
2. Height
3. Gender
4. Age
5. Consumption_Alc (Alcohol Consumption)
6. FCOHCF (Frequent Consumption of High Caloric Food)
7. FCOV (Frequent Consumption of Vegetables)
8. COFBM (Consumption of Food Between Meals)
9. Physical_Activity_F (Frequency of Physical Activity)
10. MTRANS (Mode of Transportation)
11. CH2O (Daily Water Consumption)
12. Time_using_techno_D (Time Using Technology Daily)

12. Model Implementation

12.1 Random Forest Classifier

The Random Forest model is an ensemble learning method that uses multiple decision trees to improve classification performance.

Results:

model	accuracy	precision	recall	f1-score
Randomforest	90	91	91	90

12.2 Logistic Regression

Logistic Regression is a linear model for binary and multiclass classification. It assumes a linear relationship between features and the log-odds of the target class.

Results:

model	accuracy	precision	recall	f1-score
Logistic Regression	87	86	86	87

12.3 XGBoost Classifier

XGBoost is a powerful gradient boosting algorithm known for its high performance on structured/tabular data

Results:

model	accuracy	precision	recall	f1-score
-------	----------	-----------	--------	----------

XGBoost Classifier	91	90	90	91
--------------------	----	----	----	----

Training: 1.0 , validation = 0.91 : overfitting seen

12.4 Decision Tree Classifier

A Decision Tree is a simple and interpretable model that splits data into branches based on feature values.

Results:

model	accuracy	precision	recall	f1-score
DecisionTree Classifier	76	77	76	76

12.5 LightGBM

LightGBM (Light Gradient Boosting Machine) is a fast, efficient gradient boosting framework based on decision tree algorithms, designed for high performance and accuracy in large datasets

model	accuracy	precision	recall	f1-score
LightGBM	90	90	91	90

13. Model Comparison

Model	Accuracy	Precision	Recall	F1-Score
Random Forest Classifier	89	89	89	90
Logistic Regression	87	86	87	87
XGBoost Classifier	90	90	90	91
Decision Tree Classifier	76	77	76	76
LightGBM	90	91	90	90
KNN	84	84	85	85

14. Hyperparameter Tuning

Hyperparameter tuning involves adjusting parameters to improve the model's performance and prevent overfitting.

Model	Hyperparameters	Training Accuracy	Validation Accuracy	Test Accuracy
XGBoost	{'objective': 'multi:softprob', 'base_score': ...}	0.926972	0.910265	0.909389
LightGBM	{'boosting_type': 'gbdt', 'class_weight': ...}	0.940104	0.912235	0.913767

	None...}			
KNN	{'algorithm': 'auto', 'leaf_size': 30, 'metric': ...}	0.840155	0.774787	0.755526
Logistic Regression	{'C': 1.0, 'class_weight': None, 'dual': False...}	0.862406	0.862114	0.866492

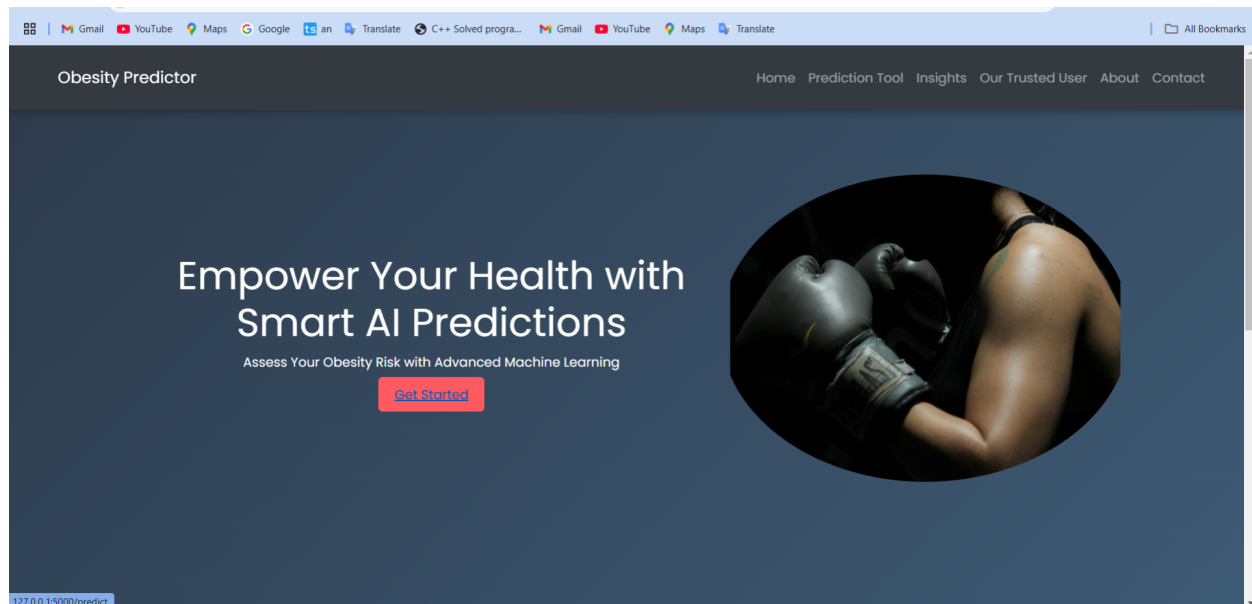
15.Final model selection

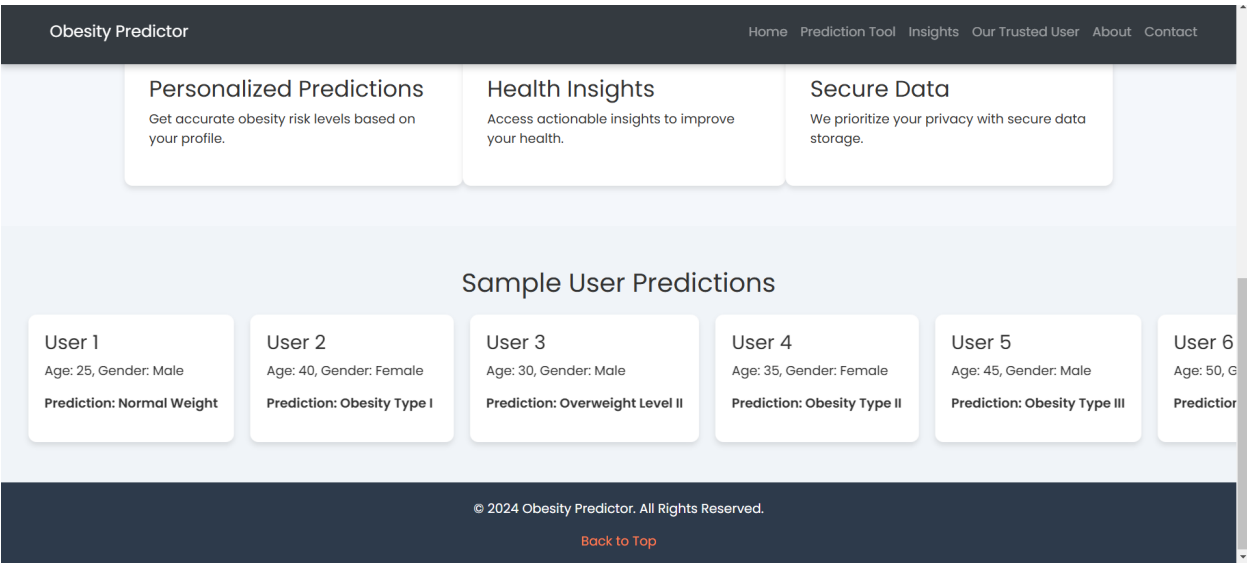
From the above hyperparameter tuning result we conclude that 'LightGBM' is the best fit model for our obesity level prediction. so we selected LightGBM model.

16.WebApp

16.1 UserInterface(Homepage):

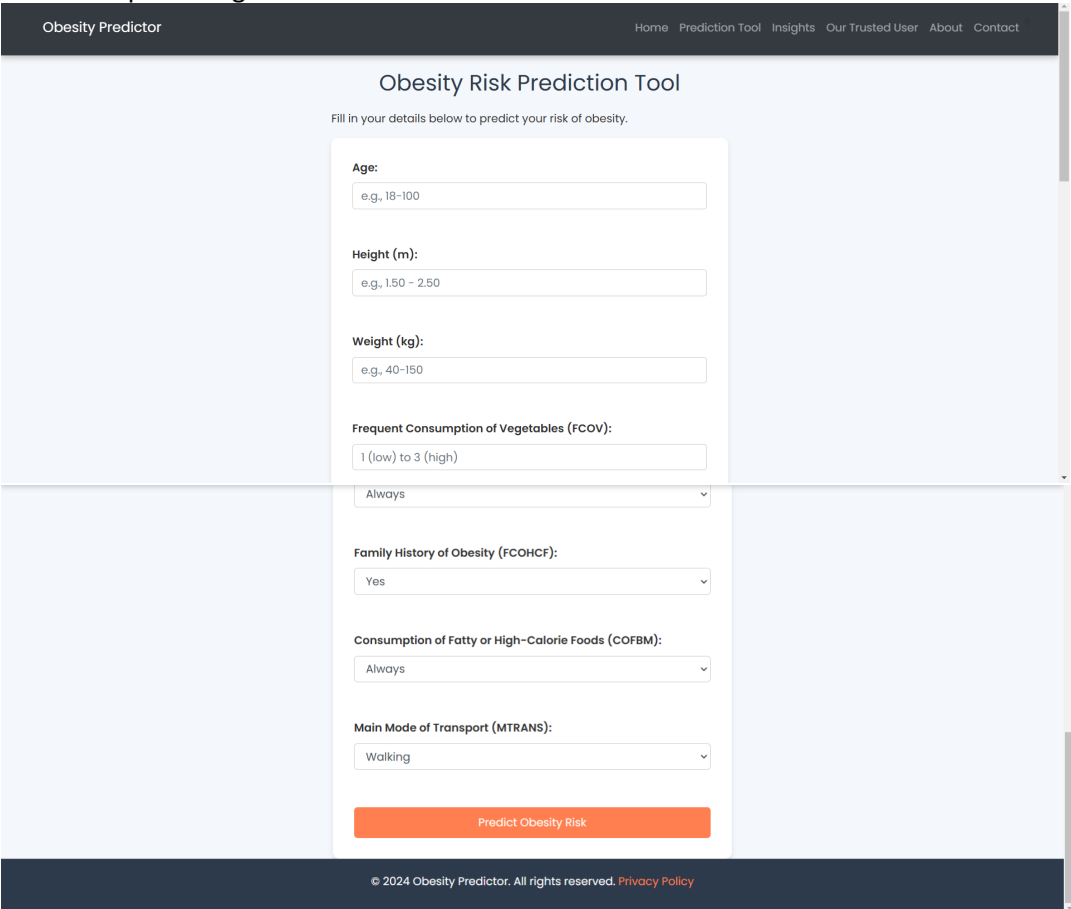
It Provides users with an intuitive and welcoming entry point to access the main features of the web app, including navigation to prediction tools and insights.





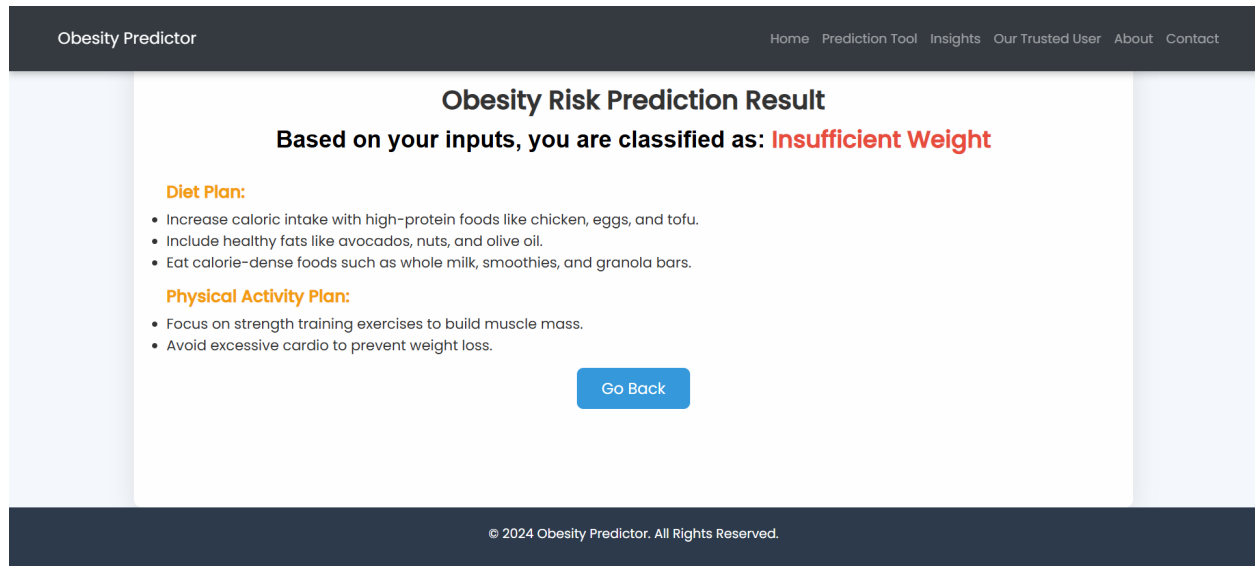
16.2 .PredictionTool

Offers an interactive module where users can input data to receive AI-driven predictions and insights based on specific algorithms or models.



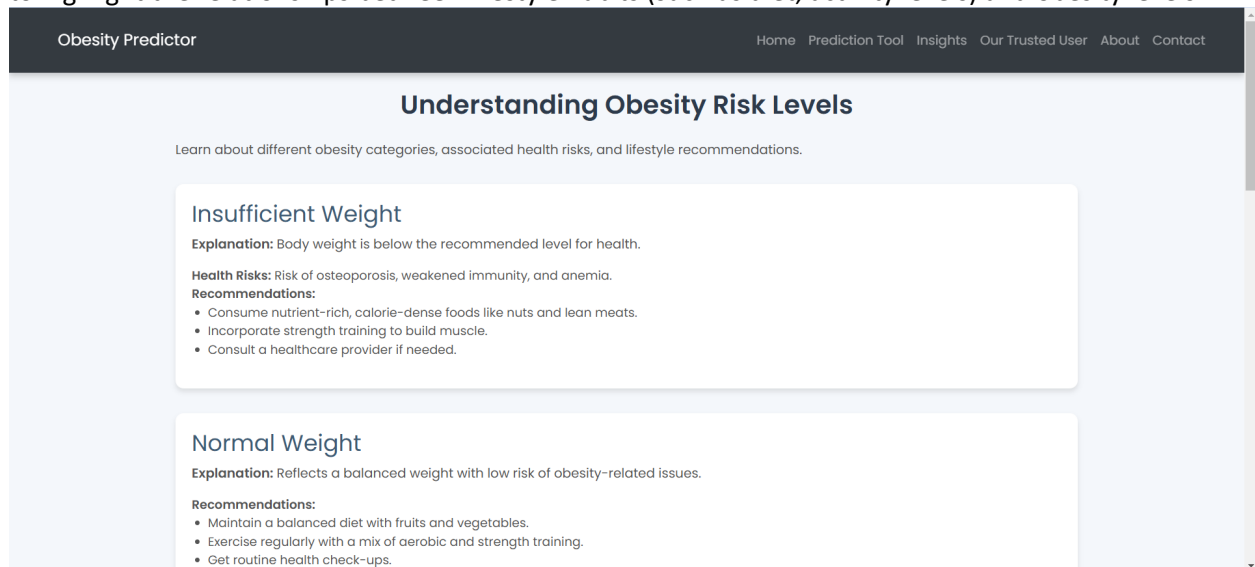
16.3 Result page

This page will show the Result and according to the result it shows some suggestion such as Diet plans and Exercise to balance you obesity level.



16.3 Insight

Presents data-driven insights into obesity trends, risk factors, and impacts on health, using visualizations to highlight the relationships between lifestyle habits (such as diet, activity levels) and obesity levels.



16.4 Our Trusted user

This page stores their data and shows the results show that they can access their result anytime anywhere
(Provide avabilality)

Obesity Predictor

HomePrediction ToolInsightsOur Trusted UserAboutContact

User Access History

User ID	Predicted Category	Timestamp
3	Normal_Weight	2024-11-12 19:27:17
2	Insufficient_Weight	2024-11-09 10:00:50
1	Obesity_Type_III	2024-11-09 09:59:08

Back to Home

© 2024 Obesity Predictor. All rights reserved.

Contact Us

16.5 About

Provides an overview of the web app’s purpose, mission, and team, offering users background information on its development and goals.

Obesity Predictor

HomePrediction ToolInsightsOur Trusted UserAboutContact

About the Obesity Prediction Model

Introduction

The Obesity Predictor is a tool designed to help individuals assess their obesity risk based on various lifestyle and demographic factors. This tool aims to raise awareness about potential health risks associated with obesity and provides insights for healthier lifestyle choices.

Explanation of the Machine Learning Model

Our prediction model combines both traditional machine learning algorithms and deep learning techniques. The model uses:

- Decision Trees: For quick and efficient classification based on clear decision rules.
- Neural Networks: To capture deeper and more complex patterns in the data.

This hybrid approach ensures a higher accuracy rate by leveraging the strengths of both techniques. It identifies clear lifestyle-related patterns while accounting for complex correlations between features like age, height, weight, and family history.

Data Sources

The model was trained using publicly available datasets, ensuring a comprehensive understanding of obesity-related factors. Key data sources include:

16.6 Contact

Includes information for reaching support or the development team, with options for feedback, queries, and customer assistance.

Obesity Predictor

HomePrediction ToolInsightsOur Trusted UserAboutContact

Contact Us

If you have any questions or feedback, feel free to reach out to us!

Email: support@obesitypredictor.com

Phone: (123) 456-7890

We appreciate your interest in our platform and look forward to assisting you.

References::

<https://www.kaggle.com/code/ddolddi/eda-with-plotly-ml-obesity-risk/input>

<https://www.kaggle.com/datasets/ankurbajaj9/obesity-levels?resource=download>