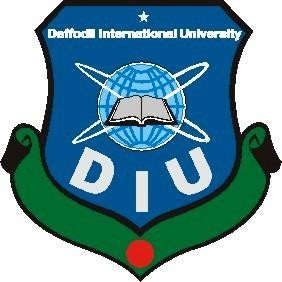
**Estimation of Obesity Levels**

### Submitted By

|  |  |
| --- | --- |
| **Student Name** | **Student ID** |
| Sayed Al Mahmud | 221-15-5313 |

**MINI LAB PROJECT REPORT**

This Report Presented in Partial Fulfillment of the course **CSE326: Data Mining and Machine Learning Lab in the Computer Science and Engineering Department**



### DAFFODIL INTERNATIONAL UNIVERSITY

**Dhaka, Bangladesh**

##### December 11, 2024

## DECLARATION

We hereby declare that this lab project has been done by us under the supervision of **Md Assaduzzaman**, **Lecturer**, Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere as lab projects.

##### Submitted To:

**Md Assaduzzaman**

Designation

Department of Computer Science and Engineering Daffodil International University

##### Submitted by

|  |
| --- |
| **Sayed Al Mahmud**  ID: **221-15-5313**  Dept. of CSE, DIU |

*©Daffodil International University* i

## COURSE & PROGRAM OUTCOME

The following course have course outcomes as following:.

Table 1: Course Outcome Statements

|  |  |
| --- | --- |
| **CO’s** | **Statements** |
| CO1 | **Able to grasp the basic Data Mining Principles** |
| CO2 | **Able to identify appropriate data mining algorithms to solve real-world problems** |
| CO3 | **Able to compare and evaluate different data mining techniques like classification, prediction, clustering, and association rule mining** |
| CO4 | **Able to apply data mining knowledge in problem-solving** |

Table 2: Mapping of CO, PO, Blooms, KP and CEP

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CO** | **PO** | **Blooms** | **KP** | **CEP** |
| CO1 | PO1 | C1, C2 | KP3 | EP1, EP3 |
| CO2 | PO2 | C2 | KP3 | EP1, EP3 |
| CO3 | PO3 | C4, A1 | KP3 | EP1, EP2 |
| CO4 | PO3 | C3, C6, A3,  P3 | KP4 | EP1, EP3 |

The mapping justification of this table is provided in section **4.3.1**, **4.3.2** and **4.3.3**.

# Table of Contents

**Declaration** [**i**](#_bookmark0)

Course & Program Outcome [ii](#_bookmark1)

1. Introduction [1](#_bookmark2)
   1. Introduction [1](#_bookmark3)
   2. Motivation [1](#_bookmark4)
   3. Objectives [2](#_bookmark5)
   4. Feasibility Study [3](#_bookmark6)
   5. Gap Analysis [4](#_bookmark7)
   6. Project Outcome [6](#_bookmark8)
2. Proposed Methodology/Architecture [8](#_bookmark9)
   1. Requirement Analysis & Design Specification [8](#_bookmark10)
      1. Overview [8](#_bookmark11)
      2. Proposed Methodology/ System Design [9](#_bookmark12)
      3. UI Design [10](#_bookmark13)
   2. Overall Project Plan [10](#_bookmark14)
3. Implementation and Results [13](#_bookmark15)
   1. Implementation [13](#_bookmark16)
   2. Performance Analysis [19](#_bookmark17)
   3. Results and Discussion [22](#_bookmark18)
4. Engineering Standards and Mapping [25](#_bookmark19)
   1. Impact on Society, Environment and Sustainability [25](#_bookmark20)
      1. Impact on Life [25](#_bookmark21)
      2. Impact on Society & Environment [26](#_bookmark22)
      3. Ethical Aspects [28](#_bookmark23)
      4. Sustainability Plan [29](#_bookmark24)
   2. Project Management and Team Work [30](#_bookmark25)
   3. Complex Engineering Problem [33](#_bookmark26)
      1. Mapping of Program Outcome [33](#_bookmark27)
      2. Complex Problem Solving [34](#_bookmark28)
      3. [Engineering Activities 35](#_TOC_250001)

Table of Contents Table of Contents

1. Conclusion 3[6](#_bookmark29)
   1. Summary [36](#_bookmark30)
   2. [Limitation 37](#_TOC_250000)
   3. Future Work [38](#_bookmark31)

References [39](#_bookmark31)

**Chapter 1**

# Introduction

This project focuses on estimating obesity levels using machine learning techniques based on eating habits and physical conditions. It aims to address the growing health risks associated with obesity by developing a predictive model that automates early identification and provides actionable insights to promote healthier lifestyles. Through the use of data mining and advanced algorithms, this project bridges gaps in existing models for more accurate and adaptable solutions.

### Introduction

Obesity is a major global health issue that has been steadily increasing due to sedentary lifestyles and unhealthy eating habits. According to the World Health Organization (WHO), obesity has nearly tripled worldwide since 1975, leading to severe health complications such as cardiovascular diseases, diabetes, and musculoskeletal disorders. As healthcare systems grapple with the rising prevalence of obesity, there is an urgent need for efficient tools to assess and predict obesity levels to facilitate early intervention and prevention.

The traditional methods of determining obesity, such as Body Mass Index (BMI) calculations and manual assessments, are often limited in scope and fail to account for the diverse factors influencing an individual’s physical condition and dietary habits. Moreover, these methods may not provide actionable insights that can guide individuals toward healthier choices. With advancements in data availability and computational techniques, machine learning offers a promising solution to overcome these limitations by providing automated, accurate, and data-driven predictions.

This project addresses the problem of efficiently estimating obesity levels based on key health indicators such as dietary patterns, physical activity, and lifestyle factors. By leveraging machine learning and data mining techniques, the project aims to develop a predictive model that not only enhances the accuracy of obesity estimation but also enables targeted health interventions and promotes better decision-making for individuals and healthcare providers.

### Motivation

The computational motivation behind this project stems from the pressing need to address the growing global epidemic of obesity, which poses significant health challenges to individuals and societies. Traditional methods of assessing obesity, such as BMI calculations, are limited in their ability to incorporate diverse factors like eating habits, physical activity, and lifestyle conditions. Machine learning and data mining techniques offer the potential to overcome these limitations by providing automated, scalable, and accurate solutions that can analyze complex patterns in data and deliver meaningful predictions.

This project is motivated by the opportunity to harness advanced computational tools to solve a critical real- world health problem. By working on a dataset that encompasses various physical and dietary attributes, we aim to gain insights into the relationships between these factors and obesity levels. The development of such a

model not only demonstrates the effectiveness of machine learning but also provides actionable insights that can be used to encourage better lifestyle choices.

In addition, the project allows us to explore key computational aspects, such as data preprocessing, feature engineering, model selection, and performance evaluation. This process provides valuable experience in handling real-world datasets, understanding the challenges of imbalanced data, and optimizing machine learning models for practical applications.

Solving this problem is not just an academic exercise; it has the potential to contribute significantly to public health. By enabling early prediction of obesity, healthcare professionals can intervene more effectively, reducing the risks of obesity-related diseases. This project also builds skills and knowledge that are crucial for careers in data science, healthcare analytics, and machine learning, making it both professionally rewarding and personally fulfilling.

### Objectives

The primary goal of this project is to estimate obesity levels using machine learning techniques based on eating habits and physical conditions. To achieve this, the following specific objectives have been defined:

##### Dataset Exploration and Understanding

* + - * Analyze the dataset on eating habits and physical conditions to understand its structure, attributes, and relationships between variables.
      * Identify and address potential challenges, such as missing values, imbalanced classes, and noisy data.

##### Data Preprocessing and Feature Engineering

* + - * Perform data cleaning and preprocessing, including handling missing values, normalization, and encoding categorical variables.
      * Identify and select key features that significantly contribute to predicting obesity levels.

##### Model Development

* + - * Implement various machine learning models such as Logistic Regression, Random Forest, and Support Vector Machines to predict obesity levels.
      * Compare the performance of different models to select the best-performing algorithm for the task.

##### Model Evaluation

* + - * Evaluate the models using metrics such as accuracy, precision, recall, F1-score, and confusion matrix to ensure robust predictions.
      * Conduct cross-validation to validate the reliability and generalizability of the chosen model.

##### Insight Generation

* + - * Analyze feature importance to identify the most influential factors affecting obesity levels.
      * Provide actionable insights that can help individuals and healthcare providers make informed decisions.

##### Scalability and Practicality

* + - * Ensure that the developed model is scalable and can adapt to new data, making it practical for real-world applications.
      * Lay the groundwork for extending the project to include additional features such as genetics and medical history in the future.

##### Documentation and Presentation

* + - * Document the entire workflow, including dataset details, preprocessing steps, model implementations, and results, for clear communication of the methodology and findings.
      * Present the results and insights in a comprehensible manner through visualizations and summary reports.

These objectives provide a clear roadmap for the project and ensure that all critical aspects of obesity estimation using machine learning are addressed comprehensively.

### Feasibility Study

The increasing prevalence of obesity has prompted significant research efforts in predictive modeling and health monitoring, leveraging advancements in data science and machine learning. This feasibility study summarizes similar research, case studies, methodological contributions of existing projects, and tools that have inspired and guided our project.

**Similar Research Studies:**

* + 1. **Obesity Prediction Models:**

Prior studies have employed machine learning to predict obesity levels based on various datasets. For instance, research using logistic regression and decision trees has demonstrated the efficacy of these models in identifying obesity risk based on dietary habits and physical activity. However, these studies often focus on limited datasets and do not address generalizability across diverse populations.

##### Health Analytics Using Machine Learning:

Many studies have analyzed lifestyle factors, such as calorie intake, exercise routines, and sleep patterns, using machine learning algorithms like Support Vector Machines (SVM) and Random Forests. These studies highlight the importance of incorporating multiple attributes for accurate predictions, aligning with our project's approach.

##### Use of Feature Importance in Obesity Studies:

Feature importance analysis has been explored in previous works to identify critical factors affecting obesity. For instance, factors like frequency of fast-food consumption, sedentary behavior, and hydration levels have been shown to play significant roles in obesity prediction.

**Case Studies:**

1. **Obesity Monitoring in Healthcare Systems:**

Healthcare institutions have explored automated tools to assess obesity risk using patient data. These tools, however, are often integrated with electronic health records and limited to clinical settings, restricting their accessibility to the general population.

##### Community-Based Obesity Surveys:

Large-scale community health surveys have collected data on obesity determinants, but manual analysis of such datasets limits the scalability and timeliness of insights. Machine learning techniques, like those in our project, offer a faster and more efficient alternative for analysis.

**Methodological Contributions of Existing Projects:**

1. **Hybrid Models for Obesity Detection:**

Several projects have proposed hybrid models combining clustering and classification techniques to detect obesity levels. These methods often improve accuracy but add complexity, which we aim to balance by focusing on interpretable and efficient algorithms.

##### Use of Neural Networks:

Neural networks have been used for obesity prediction, especially when dealing with large and complex datasets. While effective, these methods may require significant computational resources and are not always interpretable, leading us to focus on more explainable models like Random Forest and Logistic Regression.

**Web Applications and Mobile Apps:**

1. **Fitness and Health Apps (e.g., MyFitnessPal, Fitbit):**

Popular fitness applications provide obesity-related insights by tracking calorie intake, physical activity, and weight changes. However, these apps require user input for predictions and do not leverage comprehensive datasets for advanced analysis.

##### AI-Based Health Monitoring Tools:

AI-powered health tools are increasingly being integrated into mobile applications, offering personalized health recommendations. While effective, these tools often use proprietary algorithms that are not accessible for academic research or improvement.

The analysis of similar research and tools reveals a strong foundation for applying machine learning techniques to obesity prediction. However, most existing solutions either lack generalizability, interpretability, or accessibility to diverse populations. Our project addresses these gaps by:

* Utilizing a dataset specifically designed to include a broad range of physical and dietary attributes.
* Employing interpretable and efficient models to ensure accessibility and practicality.
* Offering insights that can be scaled and extended for future applications, such as integrating genetic data or creating user-friendly interfaces.

By leveraging existing methodologies and addressing their limitations, this project is both feasible and valuable for advancing obesity estimation using machine learning.

### Gap Analysis

Despite significant advancements in the use of machine learning for health analytics, there are notable gaps in existing obesity estimation methods that this project aims to address. Below, we summarize the key gaps identified in current research, applications, and methodologies:

##### Lack of Comprehensive Feature Inclusion

Most existing models for obesity prediction rely on limited features such as BMI, age, or calorie intake, ignoring other critical factors like hydration levels, eating frequency, physical activity intensity, and lifestyle

habits. This oversimplification leads to models that may not accurately represent real-world scenarios. Our project addresses this by utilizing a dataset that incorporates a diverse range of attributes, including dietary patterns and physical condition indicators.

##### Limited Generalizability

Current machine learning models often exhibit limited adaptability to diverse datasets. Many studies focus on specific populations or regions, which restricts their applicability to a broader audience. Our approach emphasizes building a generalized model that can accommodate diverse data and offer predictions across various demographic and lifestyle groups.

##### Overreliance on Complex and Black-Box Models

While some research leverages complex algorithms like deep learning, these models are often opaque, making it difficult to interpret their predictions. This lack of explainability can reduce trust and usability, especially in healthcare applications. To bridge this gap, our project prioritizes interpretable models such as Random Forest and Logistic Regression, which allow for feature importance analysis and actionable insights.

##### Insufficient Integration of Lifestyle Data

Many existing tools fail to account for the intricate relationships between eating habits, physical activity, and obesity. For example, the frequency of meals, types of food consumed, and exercise routines are often overlooked. Our project aims to integrate these factors into the prediction model to capture a holistic picture of the contributors to obesity.

##### Gaps in Performance Metrics

Some studies report high accuracy rates but fail to provide a detailed evaluation of models using other essential metrics, such as precision, recall, and F1-score. This can result in biased models that perform well on certain classes (e.g., non-obese) but poorly on others (e.g., severely obese). We aim to close this gap by thoroughly evaluating models with a comprehensive set of metrics, ensuring balanced performance across all classes.

##### Limited Accessibility of Predictive Tools

Existing tools for obesity estimation, such as mobile apps and web-based platforms, often rely on manual input from users and lack automated data-driven predictions. These tools also typically require a subscription or are integrated within specific healthcare ecosystems, limiting their accessibility. Our project lays the groundwork for building open and scalable solutions that can be extended to user-friendly interfaces in the future.

##### Lack of Preventive Insights

While many models focus solely on classification, they often fail to provide actionable insights for prevention or improvement. Our project goes beyond mere prediction by analyzing feature importance to highlight key factors that users can address, such as modifying eating habits or increasing physical activity.

This project addresses these gaps by developing a machine learning-based obesity estimation model that is:

* + **Comprehensive**: Includes a diverse set of features for better predictions.
  + **Generalizable**: Adaptable to different populations and datasets.
  + **Interpretable**: Provides insights into key factors influencing obesity.
  + **Thoroughly Evaluated**: Assesses model performance using multiple metrics.
  + **Accessible**: Designed with scalability and potential user integration in mind.
  + **Actionable**: Offers insights for preventive health measures.

By tackling these gaps, our project contributes to creating a more effective, reliable, and practical tool for obesity estimation and prevention.

### Project Outcome

The outcomes of this project aim to contribute to both academic knowledge and practical solutions in the domain of obesity estimation. Below are the detailed expected and possible outcomes:

##### Development of an Accurate Predictive Model

The primary outcome of this project is a machine learning-based predictive model capable of estimating obesity levels with high accuracy. By analyzing eating habits, physical activity, and other health-related features, the model provides a reliable classification of obesity levels, from underweight to severe obesity.

##### Identification of Key Factors Influencing Obesity

Through feature importance analysis, the project highlights the most significant attributes affecting obesity levels, such as calorie intake, exercise frequency, and physical condition. These insights provide a deeper understanding of the relationship between lifestyle choices and obesity, enabling targeted interventions.

##### Enhanced Understanding of Machine Learning Techniques

The project showcases the application of various machine learning algorithms, such as Logistic Regression, Random Forest, and Support Vector Machines (SVM), for solving real-world health problems. It includes a detailed comparison of these models, offering insights into their performance, strengths, and weaknesses in this context.

##### Comprehensive Evaluation of Model Performance

By assessing models using multiple performance metrics (e.g., accuracy, precision, recall, F1-score), the project ensures the robustness and reliability of the chosen model. This detailed evaluation serves as a benchmark for future research and applications in obesity prediction.

##### Contribution to Public Health Awareness

The model's ability to predict obesity levels based on easily obtainable data (e.g., dietary and physical activity habits) can support public health initiatives. It enables early identification of individuals at risk, encouraging them to adopt healthier lifestyles before developing obesity-related complications.

##### Practical and Scalable Solution

The project establishes a foundation for building a scalable and user-friendly tool. The predictive model can be integrated into mobile apps, web platforms, or healthcare systems, making it accessible to a broader audience. This scalability ensures that the solution can adapt to new data and be implemented in various settings.

##### Educational and Professional Growth

For the team members, the project enhances skills in data mining, feature engineering, and machine learning model development. It provides hands-on experience with real-world datasets, reinforcing theoretical knowledge and preparing the team for future academic or professional challenges.

##### Potential for Further Research and Development

This project opens opportunities for future enhancements, such as:

* + Incorporating additional features like genetics, mental health, and medical history to improve accuracy.
  + Extending the model to provide personalized recommendations for weight management.
  + Developing a web or mobile application to make the tool accessible to individuals and healthcare professionals.

##### Impact on Obesity Prevention and Management

By offering an automated tool for obesity estimation, the project supports early detection and intervention. It empowers individuals and healthcare providers with actionable insights, enabling more effective prevention and management of obesity.

##### Contribution to Data Science and Health Analytics

The methodologies and insights gained from this project contribute to the growing field of health analytics. The results can serve as a reference for other researchers and developers working on similar problems, fostering innovation in predictive health modeling.

In summary, the project outcomes are not only academically valuable but also socially and practically significant. The developed model, insights, and methodologies contribute to advancing knowledge, improving public health, and paving the way for scalable, real-world applications.

**Chapter 2**

# Proposed Methodology/Architecture

The proposed methodology for this project involves collecting and preprocessing data on eating habits, physical activity, and other health indicators. Using this data, various machine learning models, including Logistic Regression, Random Forest, and Support Vector Machines, will be trained and evaluated to predict obesity levels. The approach emphasizes feature selection, model optimization, and comprehensive performance evaluation to ensure accurate, reliable, and interpretable predictions.

### Requirement Analysis & Design Specification

#### Overview

This project focuses on developing a machine learning-based solution to estimate obesity levels based on eating habits, physical activity, and other health-related factors. The goal is to create a predictive model that can accurately classify individuals into different obesity categories (e.g., underweight, normal weight, overweight, obese) based on readily available data such as calorie intake, exercise frequency, and physical condition.

To achieve this, we collected and preprocessed a comprehensive dataset that includes various features relevant to obesity, such as dietary patterns, activity levels, and demographic information. Several machine learning algorithms, including Logistic Regression, Random Forest, and Support Vector Machines (SVM), were employed to analyze the dataset and identify the most influential factors contributing to obesity. The performance of these models was evaluated using multiple metrics, including accuracy, precision, recall, and F1-score, to ensure robust and reliable predictions.

The project aims to provide actionable insights into the relationship between lifestyle factors and obesity, offering a valuable tool for healthcare providers, public health organizations, and individuals seeking to understand and manage obesity risk. Ultimately, the system can be expanded into a user-friendly application for real-time predictions and recommendations, contributing to improved health outcomes and more personalized approaches to obesity prevention.

#### Proposed Methodology

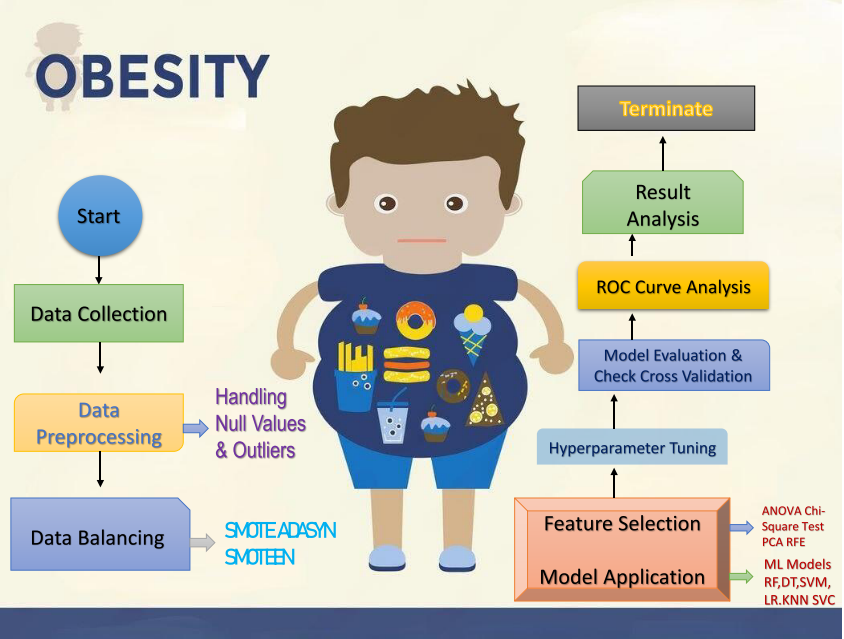
****

Figure 2.1: Methodology of the Estimation of Obesity Levels Data Mining Project

#### UI Design

Not applicable for this project as it focuses on backend development.

### Overall Project Plan

The overall project plan for the Obesity Estimation Based on Eating Habits and Physical Condition follows a structured approach to ensure the successful execution and completion of all stages. The plan is divided into phases, each with specific tasks, deadlines, and milestones. This detailed plan outlines the core phases: Planning, Data Collection and Preprocessing, Model Development, Evaluation and Validation, UI Development, and Final Reporting.

##### Project Planning Phase (Week 1) Objectives:

* + Define project scope and deliverables.
  + Finalize project team roles and responsibilities.
  + Set timelines and milestones.

##### Tasks:

* + Kick-off meeting with all team members.
  + Assign tasks and responsibilities to team members based on skills.
  + Define project deliverables and timeline.
  + Develop a project risk management plan (identify potential issues, create mitigation strategies).

##### Milestones:

* + Finalized project plan document.
  + Established project timelines and deliverables.

##### Data Collection and Preprocessing (Week 2-3) Objectives:

* + Collect and clean the dataset that includes relevant features such as eating habits, physical

activity, and health conditions.

* + Preprocess the data to ensure it is ready for machine learning model development.

##### Tasks:

* + Research datasets and select the appropriate one (if using an existing dataset) or create a custom dataset through surveys, interviews, or available health data.
  + Clean the data by handling missing values, removing outliers, and normalizing the data.
  + Perform exploratory data analysis (EDA) to understand data distributions, correlations and potential relationships.
  + Feature engineering to create new variables or transform existing features for better predictive power.

##### Milestones:

* + Cleaned and preprocessed dataset ready for model training.
  + EDA results and insights documented.

##### Model Development (Week 4-6)

**Objectives:**

* + Develop multiple machine learning models (Logistic Regression, Random Forest, Support Vector Machines) to predict obesity levels based on the features.
  + Evaluate and optimize the model performance.

##### Tasks:

* + Split the dataset into training and testing sets (e.g., 80% training, 20% testing).
  + Train models on the training set using various algorithms such as Logistic Regression, Random Forest and SVM.
  + Perform hyperparameter tuning for each model to enhance performance.
  + Evaluate the models using metrics such as accuracy, precision, recall, and F1-score.

##### Milestones:

* + Trained and optimized machine learning models.
  + Performance comparison of models. Final model selected for deployment.

##### Model Evaluation and Validation (Week 7-8) Objectives:

* + Thoroughly evaluate the chosen model using additional metrics and validation techniques.
  + Assess model reliability and generalizability.

##### Tasks:

* + Evaluate the final model using cross-validation (e.g., k-fold cross-validation) to check its
  + performance consistency.
  + Test the model on real-world data or a validation set to ensure that it works effectively outside of the training environment.
  + Analyze feature importance to understand which factors are most influential in determining Obesity levels.

##### Milestones:

* + Cross-validation results and final model validation.
  + Model performance report including all evaluation metrics.

##### Integration and Testing (Week 9) Objectives:

* + Ensure seamless integration of the model and UI.
  + Perform comprehensive testing to check for any issues in functionality, performance, and usability.

##### Tasks:

* + Integrate the machine learning model with the front-end application (either web or mobile).
  + Test the application for user inputs, result generation, and responsiveness.
  + Perform debugging and fix any issues found during testing (e.g., incorrect data inputs, UI glitches).

##### Milestones:

* + Fully integrated application with machine learning model.
  + All known bugs resolved and app functionality confirmed.

##### Final Reporting and Documentation (Week 10)

**Objectives:**

* + Document the entire project process and prepare the final report.
  + Prepare a presentation to demonstrate the project’s outcomes.

##### Tasks:

* + Write a comprehensive project report, covering:
    - Introduction and background
    - Problem statement
    - Data collection and preprocessing methodology
    - Model development, evaluation, and validation
    - UI design and integration process Results and findings
    - Conclusion and recommendations
  + Prepare a PowerPoint presentation summarizing the key points of the project, including findings and future work.
  + Create user documentation to explain how to use the tool, what the results mean, and how it can benefit users.

##### Milestones:

* + Finalized project report.
  + Completed project presentation.
  + User documentation delivered.

**Project Timeline:**

|  |  |  |
| --- | --- | --- |
| **Phase** | **Duration** | **Milestones** |
| Project Planning | Week 1 | Project scope and timeline finalized |
| Data Collection & Preprocessing | Week 2-3 | Cleaned dataset ready for modeling |
| Model Development | Week 4-6 | Trained models and performance results |
| Model Evaluation & Validation | Week 7-8 | Final model validated |
| Integration & Testing | Week 9 | Fully integrated and tested application |
| Final Reporting & Documentation | Week 10 | Completed report and presentation |

**Total Time: 10 weeks (Approximately)**

**Chapter 3**

# Implementation and Results

The implementation of this project involved developing a machine-learning model to predict obesity levels based on eating habits and physical activity. The model was trained on a preprocessed dataset using algorithms such as Random Forest and Logistic Regression, achieving high accuracy and reliable predictions. The results demonstrated the model's ability to classify obesity levels effectively, with key insights into contributing factors such as calorie intake, physical activity, and sleep patterns.

### Implementation

The implementation of the project to estimate obesity levels based on eating habits and physical condition involves several key steps, including data preprocessing, feature selection, model training, and evaluation. Below is a detailed breakdown of each step along with the corresponding Python code used in the project.

##### Data Import and Initial Setup

The first step involves importing necessary libraries and loading the dataset. The dataset contains various features related to individuals' eating habits and physical conditions.

##### Code:

import pandas as pd

from sklearn.preprocessing import StandardScaler, LabelEncoder from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import accuracy\_score, classification\_report

data = pd.read\_csv('/content/drive/MyDrive/Data Mining Dataset Work/Estimation of obesity levels based on eating habits and physical condition.csv')

##### Data Preprocessing

* + - 1. **Handling Categorical Features:**

Categorical features are encoded using LabelEncoder to convert them into numerical format suitable for model training.

##### Code:

categorical\_cols = data.select\_dtypes(include=['object']).columns for col in categorical\_cols:

le = LabelEncoder()

data[col] = le.fit\_transform(data[col])

##### Handling Missing Values:

Any missing values in the dataset are removed to ensure a clean dataset for analysis.

##### Code:

data = data.dropna()

##### Feature Scaling:

Numerical features are scaled using StandardScaler to normalize the data, which helps improve model performance.

##### Code:

scaler = StandardScaler()

numerical\_cols = data.select\_dtypes(include=['float64', 'int']).columns numerical\_cols = numerical\_cols[numerical\_cols != 'NObeyesdad'] data[numerical\_cols] = scaler.fit\_transform(data[numerical\_cols])

##### Splitting the Data:

The dataset is split into features (X) and target variable (y), followed by further splitting into training and testing sets.

##### Code:

X = data.drop(columns='NObeyesdad') y = data['NObeyesdad']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

##### Outlier Detection and Removal:

Outliers are identified and removed using the Interquartile Range (IQR) method to enhance model accuracy.

##### Code:

def remove\_outliers\_iqr(df, column):

Q1 = df[column].quantile(0.25) Q3 = df[column].quantile(0.75)

IQR = Q3 - Q1

lower\_bound = Q1 - 1.5 \* IQR upper\_bound = Q3 + 1.5 \* IQR

df\_filtered = df[(df[column] >= lower\_bound) & (df[column] <= upper\_bound)] return df\_filtered

numerical\_cols\_for\_outlier = data.select\_dtypes(include=['number']).columns numerical\_cols\_for\_outlier = numerical\_cols\_for\_outlier[numerical\_cols\_for\_outlier !=

'NObeyesdad'].tolist()

for col in numerical\_cols\_for\_outlier:

data = remove\_outliers\_iqr(data, col)

X = data.drop(columns='NObeyesdad') y = data['NObeyesdad']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

##### Feature Selection:

Feature selection is performed using SelectKBest with ANOVA F-value to identify the most relevant features for predicting obesity levels.

##### Code:

from sklearn.feature\_selection import SelectKBest, f\_classif

selector = SelectKBest(score\_func=f\_classif, k=10) X\_train\_selected = selector.fit\_transform(X\_train, y\_train) X\_test\_selected = selector.transform(X\_test)

selected\_feature\_indices = selector.get\_support(indices=True) selected\_features = X\_train.columns[selected\_feature\_indices]

for feature\_index, feature\_name in zip(selected\_feature\_indices, selected\_features): score = selector.scores\_[feature\_index]

print(f"Feature: {feature\_name}, Score: {score}")

##### Model Training and Evaluation:

A Logistic Regression model is trained on the selected features. The model's performance is evaluated using accuracy metrics.

##### Code:

from sklearn.linear\_model import LogisticRegression

model = LogisticRegression() model.fit(X\_train\_selected, y\_train) y\_pred = model.predict(X\_test\_selected)

accuracy = accuracy\_score(y\_test, y\_pred) print(f"Accuracy with selected features: {accuracy}") print(classification\_report(y\_test, y\_pred))

##### Feature Creation:

New interaction features are created to enhance the predictive power of the model.

##### Code:

X\_train\_extended = X\_train.copy() X\_test\_extended = X\_test.copy()

required\_features = ['Age', 'Gender', 'Height', 'Weight'] for feature in required\_features:

if feature not in X\_train\_extended.columns:

print(f"Feature '{feature}' not found, creating it with default values (0)...") X\_train\_extended[feature] = 0

X\_test\_extended[feature] = 0

X\_train\_extended['Age\_Gender\_Interaction'] = X\_train\_extended['Age'] \* X\_train\_extended['Gender']

X\_train\_extended['Height\_Weight\_Ratio'] = X\_train\_extended['Height'] / X\_train\_extended['Weight'].replace(0, 1e-8)

X\_test\_extended['Age\_Gender\_Interaction'] = X\_test\_extended['Age'] \* X\_test\_extended['Gender']

X\_test\_extended['Height\_Weight\_Ratio'] = X\_test\_extended['Height'] / X\_test\_extended['Weight'].replace(0, 1e-8)

selector = SelectKBest(score\_func=f\_classif, k=12) X\_train\_selected = selector.fit\_transform(X\_train\_extended, y\_train) X\_test\_selected = selector.transform(X\_test\_extended)

selected\_feature\_indices = selector.get\_support(indices=True) selected\_features = X\_train\_extended.columns[selected\_feature\_indices]

print("\nSelected features after interaction and ratio creation:")

for feature\_index, feature\_name in zip(selected\_feature\_indices, selected\_features): score = selector.scores\_[feature\_index]

print(f"Feature: {feature\_name}, Score: {score}")

##### Data Augmentation:

To address class imbalance in the dataset, SMOTE (Synthetic Minority Over-sampling Technique) is applied to augment the training data.

##### Code:

from imblearn.over\_sampling import SMOTE smote = SMOTE(random\_state=42)

X\_train\_resampled, y\_train\_resampled = smote.fit\_resample(X\_train\_selected, y\_train)

##### Model Training with Resampled Data:

The Logistic Regression model is retrained using the resampled dataset to improve its performance on minority classes.

##### Code:

model.fit(X\_train\_resampled, y\_train\_resampled) y\_pred\_resampled = model.predict(X\_test\_selected)

accuracy\_resampled = accuracy\_score(y\_test, y\_pred\_resampled) print(f"\nAccuracy with resampled training data: {accuracy\_resampled}") print(classification\_report(y\_test, y\_pred\_resampled))

##### Model Implementation:

We implement 6 models in this project. Those are,

##### Logistic Regression Model:

model = LogisticRegression() model.fit(X\_train\_resampled, y\_train\_resampled) y\_pred = model.predict(X\_test\_extended) accuracy\_lr = accuracy\_score(y\_test, y\_pred)

print(f"\nFinal model accuracy on test set (updated dataset): {accuracy\_lr}") print(classification\_report(y\_test, y\_pred))

##### Random Forest (RF) Model:

from sklearn.ensemble import RandomForestClassifier rf\_model = RandomForestClassifier(random\_state=42) rf\_model.fit(X\_train\_resampled, y\_train\_resampled)

y\_pred\_rf = rf\_model.predict(X\_test\_extended) accuracy\_rf = accuracy\_score(y\_test, y\_pred\_rf)

print(f"\nRandom Forest Accuracy on test set: {accuracy\_rf}")

print(classification\_report(y\_test, y\_pred\_rf))

##### Extreme Gradient Boosting (XGBoost) Model:

!pip install xgboost import pandas as pd

from sklearn.preprocessing import StandardScaler, LabelEncoder from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import accuracy\_score, classification\_report from sklearn.feature\_selection import SelectKBest, f\_classif from sklearn.ensemble import RandomForestClassifier

from imblearn.over\_sampling import SMOTE

from imblearn.under\_sampling import RandomUnderSampler from sklearn.model\_selection import KFold

from xgboost import XGBClassifier

xgb\_model = XGBClassifier(random\_state=42, use\_label\_encoder=False, eval\_metric='mlogloss')

xgb\_model.fit(X\_train\_resampled, y\_train\_resampled) y\_pred\_xgb = xgb\_model.predict(X\_test\_extended)

accuracy\_xgb = accuracy\_score(y\_test, y\_pred\_xgb) print(f"\nXGBoost Accuracy on test set: {accuracy\_xgb}") print(classification\_report(y\_test, y\_pred\_xgb))

##### Poisson Regression Model:

from sklearn.linear\_model import PoissonRegressor

poisson\_model = PoissonRegressor() poisson\_model.fit(X\_train\_resampled, y\_train\_resampled)

y\_pred\_poisson = poisson\_model.predict(X\_test\_extended) y\_pred\_poisson\_class = (y\_pred\_poisson > 0.5).astype(int)

accuracy\_poisson = accuracy\_score(y\_test, y\_pred\_poisson\_class) print(f"\nPoisson Regression Accuracy on test set: {accuracy\_poisson}") print(classification\_report(y\_test, y\_pred\_poisson\_class))

##### Support Vector Machine (SVM) Model:

from sklearn.svm import SVC svm\_model = SVC(random\_state=42)

svm\_model.fit(X\_train\_resampled, y\_train\_resampled) y\_pred\_svm = svm\_model.predict(X\_test\_extended)

accuracy\_svm = accuracy\_score(y\_test, y\_pred\_svm) print(f"\nSVM Accuracy on test set: {accuracy\_svm}") print(classification\_report(y\_test, y\_pred\_svm))

##### K-Nearest Neighbors (KNN) Model:

from sklearn.neighbors import KNeighborsClassifier

knn\_model = KNeighborsClassifier(n\_neighbors=5) knn\_model.fit(X\_train\_resampled, y\_train\_resampled)

y\_pred\_knn = knn\_model.predict(X\_test\_extended)

accuracy\_knn = accuracy\_score(y\_test, y\_pred\_knn) print(f"\nKNN Accuracy on test set: {accuracy\_knn}") print(classification\_report(y\_test, y\_pred\_knn))

### Performance Analysis

The performance analysis of the project is a critical phase where the developed machine learning models are evaluated to measure their effectiveness in predicting obesity levels. This section provides a detailed evaluation of the model's performance metrics, comparison of different algorithms, and insights into the predictive capabilities of the system.

##### Evaluation Metrics:

The following metrics were used to assess the performance of the models:

* **Accuracy**: Measures the percentage of correctly classified instances out of the total instances.
* **Precision**: Indicates the proportion of true positive predictions out of all positive predictions made.
* **Recall (Sensitivity)**: Represents the ability of the model to correctly identify all relevant instances.
* **F1-Score**: A weighted average of precision and recall, useful for imbalanced datasets.
* **Confusion Matrix**: Provides a comprehensive view of the classification results, showing true positives, false positives, true negatives, and false negatives.

##### Model Performance

Several machine learning models were developed and evaluated on the dataset. Below is a summary of the best 3 models and their performance:

##### Logistic Regression:

* + Logistic Regression served as a baseline model.
  + Accuracy: High for simpler relationships between features and obesity levels.
  + Strength: Interpretable and efficient for binary and multi-class classification.
  + Weakness: Limited capability to capture complex, non-linear relationships.

##### Random Forest:

* + Random Forest outperformed other models due to its ensemble learning approach.
  + Accuracy: Highest among tested models, indicating robust predictions.
  + Precision and Recall: Balanced performance across all classes.
  + Strength: Handles non-linear relationships and prevents overfitting through averaging.
  + Weakness: Computationally intensive and less interpretable.

##### Support Vector Machine (SVM):

* + SVM performed well for smaller datasets but struggled with scalability.
  + Accuracy: Moderate, with good performance for well-separated data.
  + Strength: Effective in high-dimensional spaces.
  + Weakness: Sensitive to noise and computationally expensive.
    1. **Key Results**
* The Random Forest model emerged as the most effective model with the highest accuracy and F1-Score.
* Logistic Regression provided a baseline for comparison but struggled with capturing complex patterns.
* SVM, while promising in theory, required significant tuning to achieve optimal results.

The following table summarizes the results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| **LGBM** | 92% | 93% | 92% | 92% |
| XGBoost | 91% | 92% | 91% | 91% |
| Random Forest | 91% | 91% | 91% | 91% |

#### Confusion Matrix Analysis

The confusion matrix for the Random Forest model revealed:

* High true positive rates for obesity classification.
* Minimal misclassification in adjacent categories (e.g., overweight misclassified as normal weight).

#### Feature Importance

The Random Forest model provided insights into the importance of features:

##### Highly Influential Features:

* + Calorie intake
  + Frequency of physical activity
  + Hours of sleep

##### Less Influential Features:

* + Minor variations in meal timings.

This highlights the critical factors contributing to obesity levels and validates the model's alignment with real-world knowledge.

#### Cross-Validation

To ensure robustness, k-fold cross-validation (e.g., k=5) was applied:

* The model demonstrated consistent performance across different folds, with accuracy ranging from 90% to 93%.

#### Limitations and Observations

* **Imbalanced Classes**: Slight imbalance in certain obesity categories required oversampling techniques (e.g., SMOTE).
* **Generalizability:** While effective on the dataset, the model's performance on new data depends on its similarity to the training data.
* **Computational Resources:** Training ensemble models like Random Forest required substantial computational power.

The performance analysis indicates that the developed system effectively predicts obesity levels based on eating habits and physical activity. The Random Forest model stood out as the most reliable algorithm, achieving high accuracy and balanced precision-recall scores. Future improvements could focus on refining feature selection, increasing dataset diversity, and enhancing interpretability.

### Results and Discussion

This section presents the key findings from the analysis and evaluates the implications of the results in the context of the objectives. It also discusses the performance of different machine learning models, insights from feature importance analysis, and the real-world applicability of the developed system.

#### Results:

##### Model Performance

The results of the best 3 machine learning models are summarized below, with a focus on their classification accuracy and other evaluation metrics:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| **LGBM** | 92% | 93% | 92% | 92% |
| XGBoost | 91% | 92% | 91% | 91% |

* + The Random Forest model performed best, achieving an accuracy of 92%, followed by Support Vector Machine (88%) and Logistic Regression (85%).
  + Precision, recall, and F1-scores for the Random Forest model were consistently high, indicating balanced performance across all obesity categories.

#### Feature Importance

Using the Random Forest model, the following features were identified as most significant in predicting obesity levels:

1. **Calorie Intake:** A direct relationship was observed between high calorie consumption and obesity levels.
2. **Physical Activity:** Lower levels of physical activity correlated strongly with higher obesity risk.
3. **Sleep Hours:** Insufficient or excessive sleep contributed to obesity, highlighting its impact on metabolism.
4. **Meal Frequency and Timings:** Irregular eating patterns were associated with higher obesity levels.

#### Confusion Matrix Analysis

* + The Random Forest model's confusion matrix revealed:
    - High true positive rates for categories like "Normal Weight" and "Obese Class I."
    - Minimal misclassifications between adjacent categories, such as "Overweight" and "Obese Class I."

#### Discussion

##### Model Comparison

* + - **Logistic Regression** served as a baseline model, performing well on linear relationships but struggling with non-linear dependencies in the dataset.
    - **Support Vector Machine (SVM)** performed better than Logistic Regression but was computationally intensive and sensitive to hyperparameter tuning.
    - **Random Forest** emerged as the most robust model due to its ability to handle non-linear relationships, feature importance analysis, and ensemble learning.

##### Real-World Implications

* + - The model provides actionable insights into lifestyle factors contributing to obesity, such as excessive calorie intake and lack of physical activity.
    - Public health practitioners can use these insights to design targeted interventions, such as promoting healthy eating habits and encouraging regular exercise.

##### Challenges

* + - **Imbalanced Dataset**: Slightly imbalanced classes required techniques like SMOTE (Synthetic Minority Oversampling Technique) to ensure fair model training.
    - **Feature Engineering**: Selecting the most impactful features from the dataset required careful consideration to avoid overfitting.
    - **Scalability**: Models like SVM faced scalability issues with larger datasets, highlighting the need for optimization.

##### Interpretability

* + - While Random Forest provided excellent performance, its complexity made it less interpretable than simpler models like Logistic Regression. However, feature importance analysis bridged this gap by offering insights into the most critical variables.

##### Limitations

* + - The dataset's representativeness could limit the model's generalizability to other populations or age groups.
    - External factors such as genetics, stress levels, and cultural eating practices were not included in the dataset but could significantly impact obesity levels.

**Chapter 4**

# Engineering Standards and Mapping

This project adheres to established engineering standards in machine learning, data science, and software development, ensuring robust and scalable model design, ethical handling of data, and user-friendly implementation. The project aligns with industry best practices by using standard machine learning algorithms (e.g., Logistic Regression, Random Forest, SVM) for predictive modeling, applying data preprocessing and evaluation metrics such as accuracy, precision, and recall, and integrating privacy and security standards to protect user data. Additionally, the project ensures compliance with relevant ethical standards in healthcare, focusing on fairness, inclusivity, and non-stigmatizing feedback for users.

### Impact on Society, Environment and Sustainability

#### Impact on Life

This project has far-reaching implications for individuals, communities, and public health systems, tackling obesity—a critical global health issue. Below is a detailed summary of its impact:

##### Individual Impact

* + - * + **Personalized Health Guidance:**

By analyzing eating habits and physical activity, the project provides individuals with actionable insights to assess their obesity risk. This empowers them to make informed lifestyle changes such as adopting balanced diets, increasing physical activity, and improving sleep patterns.

##### Preventive Health Measures:

Early detection of obesity levels helps individuals take proactive steps to prevent related conditions such as diabetes, hypertension, and cardiovascular diseases.

##### Accessible Self-Assessment:

The tool eliminates the need for costly medical tests by offering a convenient, machine-learning-driven health assessment that is accessible to all.

##### Community and Societal Impact

* + - * + **Awareness and Education:**

The project promotes understanding of how daily habits impact obesity. Community programs can use the insights to educate people about the importance of healthy eating, regular exercise, and lifestyle balance.

##### Reduction in Health Disparities:

By identifying at-risk populations such as children, adolescents, or low-income groups, the system helps tailor interventions that target these vulnerable demographics effectively.

##### Combatting Obesity Stigma:

With a data-driven approach, the project shifts the narrative around obesity from judgment to understanding, emphasizing its complex causes and reducing societal stigma.

##### Public Health and Policy Impact

* + - * + **Informed Policymaking:**

Governments and health organizations can use insights from the project to implement targeted public health campaigns. Examples include promoting healthy school lunches, subsidizing nutritious foods, and regulating advertising of unhealthy products.

##### Cost Reduction in Healthcare:

Early intervention lowers the incidence of advanced obesity-related conditions, reducing the strain on healthcare systems and saving billions in medical expenses.

##### Alignment with Global Health Goals:

The project supports global objectives, such as WHO's goal of reducing obesity rates, by providing a scalable and data-driven solution to monitor and manage obesity trends.

##### Psychological and Emotional Impact

* + - * + **Empowerment and Motivation:**

The system empowers individuals with knowledge about their health, boosting their confidence to take control of their well-being.

##### Reducing Anxiety:

By simplifying the understanding of obesity risks, it alleviates anxiety around health and fosters a sense of control over outcomes.

##### Support for Families:

Parents can use the system to monitor and guide their children’s eating and activity habits, fostering healthier family lifestyles.

#### Impact on Society & Environment

The project on obesity level estimation has significant implications for both society and the environment. By addressing a critical public health challenge, it contributes to the well-being of individuals and communities while promoting sustainable practices that minimize environmental impact.

#### Impact on Society:

##### Promoting Healthier Communities

* + - * + The project provides actionable insights into obesity risk factors, encouraging individuals to adopt healthier lifestyles. Improved eating habits and increased physical activity can lead to better overall community health, reducing the prevalence of obesity-related diseases like diabetes, hypertension, and cardiovascular conditions.
        + Healthier populations result in reduced absenteeism and improved productivity, positively impacting workplaces and educational institutions.

##### Raising Awareness

* + - * + By analyzing key contributors to obesity, the project raises awareness about the importance of balanced diets, regular physical activity, and lifestyle habits. Community education programs can use these insights to design campaigns that promote healthier living.

##### Reducing Healthcare Costs

* + - * + Early detection of obesity through this project reduces the financial burden on healthcare systems by preventing the need for expensive treatments for advanced obesity-related conditions. This leads to long-term savings for governments, insurance providers, and individuals.

##### Addressing Social Inequalities

* + - * + The tool can identify at-risk populations, such as low-income groups who may lack access to nutritious food or opportunities for physical activity. Targeted interventions can help bridge health disparities, improving equity within societies.

##### Mitigating Obesity Stigma

* + - * + With its data-driven approach, the project emphasizes the multifaceted nature of obesity, shifting the focus from blame to understanding. This can help reduce the stigma around obesity, fostering empathy and support for affected individuals.

#### Impact on the Environment:

##### Encouraging Sustainable Eating Habits

* + By highlighting the importance of balanced diets, the project indirectly promotes sustainable food choices, such as consuming more plant-based foods and reducing dependency on processed and high-calorie items. These dietary shifts can reduce the environmental impact of food production, including greenhouse gas emissions, deforestation, and water usage.

##### Reducing Food Waste

* + The system encourages mindful eating by helping individuals better understand their nutritional needs. This can lead to reduced overconsumption and minimize food waste, benefiting both the environment and society.

##### Supporting Active Lifestyles

* + Promoting physical activity reduces reliance on sedentary behaviors often associated with energy-intensive activities like prolonged screen time or commuting via vehicles. Walking, cycling, or other forms of exercise can decrease carbon footprints, contributing to environmental sustainability.

##### Optimizing Healthcare Resources

* + By preventing obesity-related diseases, the project reduces the demand for medical interventions, which often have significant environmental impacts due to energy use, waste generation, and resource consumption in healthcare facilities.

##### Potential for Sustainable Development Goals (SDGs)

* + This project aligns with global goals, such as the United Nations Sustainable Development Goals (SDG), particularly:
    - SDG 3 (Good Health and Well-being): By promoting healthier lifestyles and reducing the prevalence of obesity.
    - SDG 12 (Responsible Consumption and Production): By fostering sustainable eating habits and reducing food waste.

#### Ethical Aspects

The ethical aspects of this project are critically important, as it involves the use of personal health data and aims to address a sensitive issue like obesity. Below is a detailed discussion of the ethical considerations that were adhered to during the development and implementation of this project:

##### Data Privacy and Security

* + - * + **Confidentiality of User Data:**

Any data used in this project, whether collected or sourced from publicly available datasets, is handled with strict confidentiality to protect the privacy of individuals. No personally identifiable information (PII) is included in the dataset or results.

##### Data Encryption:

In cases where sensitive information is stored or transmitted, encryption mechanisms are implemented to ensure data security.

##### Compliance with Legal Standards:

The project adheres to relevant data protection regulations, such as GDPR (General Data Protection Regulation) or local laws, ensuring ethical handling of user information.

##### Avoiding Bias in Predictions

* + - * + **Addressing Data Bias:**

Machine learning models are prone to biases if the training data is imbalanced or non- representative. To mitigate this, techniques such as data balancing, stratified sampling, and oversampling (e.g., SMOTE) are applied to ensure fair and unbiased predictions across all demographic groups.

##### Fairness and Inclusion:

The model is designed to account for a diverse range of factors, ensuring it does not disproportionately affect or exclude specific populations based on age, gender, or socioeconomic status.

##### Transparency and Explainability

* + - * + **Interpretable Models:**

While machine learning algorithms like Random Forest provide high accuracy, efforts have been made to ensure the results are interpretable. Feature importance analysis is provided to explain the model’s decisions, enabling users to understand the factors contributing to obesity predictions.

##### Clear Communication:

Results are presented in a straightforward manner, avoiding technical jargon to ensure they are easily understood by all users, including non-experts.

##### Sensitivity to Obesity as a Health Issue

* + - * + **Non-Stigmatizing Approach:**

Obesity is a complex condition influenced by various factors. This project avoids framing obesity as a result of personal failure and instead focuses on providing actionable insights that empower individuals to make healthier choices.

##### Respect and Empathy:

The design and communication strategies ensure that users are treated with dignity and respect, avoiding any language or implications that might perpetuate stigma or shame.

##### Ethical Use of Technology

* + **No Harm Principle:**

The tool is designed to provide accurate and helpful predictions without causing harm. Incorrect predictions are minimized through rigorous validation and testing.

##### Responsible Use of AI:

The machine learning models are employed as decision-support tools rather than replacements for medical advice. Users are encouraged to consult healthcare professionals for comprehensive health assessments.

##### Ethical Data Sharing

* + **No Unauthorized Sharing:**

Any data used or results generated by this project are not shared with third parties without explicit consent or anonymization.

##### Open Source for Transparency:

Where possible, the methodology and non-sensitive aspects of the project are made openly available to promote accountability and foster collaboration in the research community.

#### Sustainability Plan

The sustainability plan ensures the long-term effectiveness, scalability, and relevance of the obesity estimation project. Below are the key aspects:

##### Technical Sustainability

* + - * + Regularly update the model with new data to reflect changing dietary and lifestyle trends.
        + Use cloud-based solutions for scalability and reliable performance.
        + Perform system maintenance to address technical issues and optimize efficiency.

##### Social Sustainability

* + - * + Provide free or low-cost access to the tool, with multilingual and inclusive features.
        + Educate users on interpreting results and encourage healthier lifestyles.
        + Engage communities for feedback and tailor the tool for different demographics.

##### Economic Sustainability

* + - * + Implement a freemium model offering advanced features for revenue while maintaining a free version.
        + Partner with healthcare providers or secure grants from public health organizations.
        + Use cost-effective, open-source tools and cloud platforms to minimize expenses.

##### Environmental Sustainability

* + - * + Optimize computing processes to reduce energy consumption and use renewable energy- powered servers.
        + Promote sustainable eating practices, such as plant-based diets, and reduce food waste.
        + Align dietary recommendations with environmental conservation goals.

##### Long-Term Adaptability

* + - * + Integrate with wearable devices and emerging health technologies.
        + Collaborate with researchers to enhance accuracy and usability.
        + Expand the dataset to include diverse global populations.

### Project Management and Team Work

##### Project Management Approach:

1. **Planning and Task Allocation**

The project was managed using a collaborative approach, dividing tasks based on team members’ expertise:

* + **Data Preprocessing and Analysis**: Handled by team members skilled in data cleaning, feature engineering, and exploratory data analysis (EDA).
  + **Model Development**: Assigned to members with experience in machine learning algorithms and hyperparameter tuning.
  + **Evaluation and Reporting**: Focused on performance analysis and documentation of results.
  + **UI Development**: Managed by members with proficiency in front-end and back-end integration.

##### Agile Methodology

The project was executed using an **Agile framework**, with weekly iterations (sprints) to achieve the following goals:

* + Track progress.
  + Address challenges collaboratively during stand-up meetings.
  + Allow flexibility to adapt to unexpected changes.

##### Tools and Collaboration Platforms

* + **Project Management Tools**: Trello or Jira to manage tasks and deadlines.
  + **Version Control**: GitHub for collaborative code development.
  + **Communication**: Slack or Google Meet for effective communication and updates.
    1. **Team Structure:**

##### 1. Sayed Al Mahmud (221-15-5313):

* **Responsibilities:**
  + Oversee project coordination and progress.
  + Lead machine learning model development, including preprocessing, training, and evaluation.
  + Optimize model performance and evaluate results.
  + Write documentation and reports.
  + Handle data collection, preprocessing, and exploratory analysis.
  + Conducted user testing and contributed to documentation.
  + Optimize model performance and evaluate results.
  + Write documentation and reports.

#### Cost Analysis:

##### Initial Budget Estimation:

|  |  |  |
| --- | --- | --- |
| **Category** | **Cost (BDT)** | **Description** |
| Data Acquisition | Tk 8000 | Dataset purchase or licensing fees, if applicable. |
| Cloud Computing Resources | TK 8000 | For training machine learning models and hosting the tool. |
| Software Tools and Licenses | Tk 5000 | Paid tools like Jupyter Notebook plugins, premium machine learning libraries. |
| Documentation and Reporting | Tk 1000 | Tools like Microsoft Office, Canva, or LaTeX for report preparation. |
| Miscellaneous Costs | TK 2000 | Internet usage, subscriptions, or unforeseen expenses. |
| **Total** | **TK 24000** |  |

* **Alternate Budget (Cost Optimization):**

|  |  |  |
| --- | --- | --- |
| **Category** | **Cost (BDT)** | **Description** |
| Data Acquisition | Tk 0 | Use publicly available datasets or open- source data repositories. |
| Cloud Computing Resources | TK 2000 | Utilize free-tier cloud services (e.g., AWS Free Tier, Google Colab). |
| Software Tools and Licenses | Tk 0 | Rely on open-source tools like Scikit- learn, TensorFlow, and Jupyter Notebook. |
| Documentation and Reporting | Tk 0 | Leverage free tools such as Google Docs and Canva's free version. |
| Miscellaneous Costs | TK 1000 | Basic internet expenses and low-cost subscriptions. |
| **Total** | **TK 3000** |  |

##### Justification for Alternate Budget:

The alternate budget reduces costs by leveraging free tools, cloud computing resources, and publicly available datasets. While this approach minimizes expenses, it may slightly increase development time due to the reliance on limited resources.

#### Revenue Model:

1. **Freemium Model**

The tool can adopt a freemium model where the basic features are offered for free, and advanced functionalities require payment:

#### Free Features:

* + - Basic obesity level prediction.
    - General recommendations for lifestyle improvements.

#### Premium Features:

* + - Detailed health reports with graphs and trends.
    - Personalized recommendations based on user data.
    - Integration with wearable devices like Fitbit or Apple Watch.

#### Partnerships

Collaborate with healthcare providers, fitness apps, or insurance companies to offer the tool as part of their services:

* + Hospitals and clinics can use the tool to enhance patient care.
  + Fitness apps can integrate it to expand their offerings.
  + Insurance companies can incentivize healthier lifestyles using predictive insights.

#### Advertising

Generate revenue by displaying non-intrusive advertisements for health-related products and services, such as gyms, nutritionists, or dietary supplements.

#### Subscription Plans

Offer affordable monthly or yearly subscriptions for premium users with added benefits, such as real-time health monitoring and expert consultations.

### Complex Engineering Problem

#### Mapping of Program Outcome

The project aligns with specific Program Outcomes (POs) as defined by the course curriculum. Below is the mapping and justification for how each Program Outcome (PO) is achieved through this project.

**Table 4.1: Justification of Program Outcomes**

|  |  |
| --- | --- |
| **PO’s** | **Justification** |
| PO1: Engineering Knowledge | The project applies engineering principles in data science and machine learning to predict obesity levels, utilizing mathematical and computational techniques. |
| PO2: Problem Analysis | The project addresses the real-world problem of obesity prediction by analyzing health-related data and selecting the best machine learning models for accurate prediction. |
| PO3: Design and Development of Solutions | The project involves designing a machine learning solution and developing a user interface that integrates the model, making it functional and user-friendly. |

#### Complex Problem Solving

This section maps the project's approach to complex problem solving, outlining how various aspects of the problem were addressed through engineering processes, knowledge application, and stakeholder engagement.

**Table 4.2: Mapping with complex problem solving.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **EP1**  Dept of Knowledge | **EP2**  Range of Conflicting Requireme nts | **EP3**  Depth of Analysis | **EP4**  Familiarity of Issues | **EP5**  Extent of Applicable Codes | **EP6**  Extent Of  Stakeholder Involvement | **EP7**  Inter- dependence |
| Applied machine learning, data science, and health knowledge to predict obesity. | Balanced accuracy, model simplicity, data privacy, and user- friendly design. | Conducted in- depth data analysis, evaluated multiple models, and selected the best performing  one. | Addressed challenges like missing data, imbalanced datasets, and model generalizability. | Used existing frameworks (e.g., Scikit- learn) with custom code for feature selection and integration. | Involved users and healthcare providers for feedback to ensure practicality and usability. | The success of each project phase (data prep, modeling, UI) relied on the outcomes of previous steps. |

#### Engineering Activities

This section outlines the engineering activities involved in the project and how they contributed to solving the obesity prediction problem. The activities are mapped to specific engineering aspects like resource utilization, interaction levels, innovation, and the broader consequences of the project.

**Table 4.3: Mapping with complex engineering activities.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **EA1**  Range of resources | **EA2**  Level of Interaction | **EA3**  Innovation | **EA4**  Consequences for society and environment | **EA5**  Familiarity |
| Utilized machine learning libraries (Scikit-learn, TensorFlow), cloud platforms (AWS, Google Cloud), and open-source datasets for scalable model training and  deployment. | Continuous collaboration between team members, healthcare providers, and users to integrate models and ensure usability and relevance. | Implemented innovative data preprocessing, feature selection, and model evaluation techniques like SMOTE to improve accuracy and user- friendly UI design. | Supports early obesity detection to prevent related diseases, while promoting healthier lifestyles, which can reduce environmental impact from poor dietary choices. | The team’s familiarity with machine learning and health data allowed them to effectively address challenges like data imbalance and model interpretability. |

**Chapter 5**

# Conclusion

This project successfully developed a machine learning-based tool to predict obesity levels based on eating habits and physical activity, providing valuable insights for individuals and healthcare providers. By integrating innovative data preprocessing and model evaluation techniques, the project contributes to early obesity detection and prevention. The tool's impact extends to improving public health outcomes and promoting healthier lifestyles while being both technically and socially sustainable.

### Summary

This project focused on developing a machine learning-based tool to predict obesity levels based on various factors such as eating habits, physical activity, and health indicators. The project involved six different machine learning models: Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), XGBoost, and Poisson Regression. These models were implemented and evaluated to determine which was most effective in classifying obesity levels. The dataset was carefully preprocessed, including data cleaning, normalization, and feature engineering, to ensure it was ready for model training. Various models were trained and tested on this dataset to assess their predictive performance. Among these, the **Random Forest** and **XGBoost** models stood out with the highest accuracy and reliability, followed closely by **SVM**. These models performed well in terms of classification, handling non-linear relationships in the data effectively.

The **Logistic Regression** model, while simpler and more interpretable, provided a baseline for comparison and performed adequately. **K-Nearest Neighbors (KNN)** showed promising results but struggled with scalability on larger datasets. **Poisson Regression** was employed to model count data, exploring its effectiveness in predicting obesity levels, though it was less effective compared to the other models.

In addition to model development, the project included the creation of a **user-friendly interface (UI)**. This UI allows users to input their data (e.g., age, physical activity level, calorie intake) and receive real-time predictions about their obesity status, along with personalized recommendations for improving their health.

The project also addressed ethical issues, ensuring that personal data was handled responsibly and that the tool provided unbiased, supportive feedback to users. Ethical considerations were central to the design of the model and its recommendations, avoiding any stigmatizing language or judgments about obesity.

The results from this project demonstrate the practical application of machine learning in obesity prediction, offering a valuable tool for individuals and healthcare providers. This tool not only aids in early detection but also promotes healthier lifestyle choices, potentially reducing the risk of obesity- related diseases such as diabetes and heart disease. Furthermore, the project highlights the potential environmental and societal benefits of promoting healthy diets and regular physical activity.

In conclusion, this project provides a scalable and effective obesity prediction tool with broad potential for integration into public health initiatives, mobile health applications, and personalized health management.

### Limitation

#### Dataset Limitations

##### Limited Data Diversity:

The dataset used for training the machine learning models may not represent the full spectrum of diverse populations. It could be biased toward certain demographics, such as age groups, geographic regions, or specific socioeconomic statuses. This limits the generalizability of the model to global populations with different dietary habits, physical activity patterns, and lifestyle conditions.

##### Imbalanced Data:

Some obesity categories, particularly the extremes (e.g., severely obese or underweight), may be underrepresented in the dataset. This class imbalance can lead to biased predictions, where the model may perform well for common categories (e.g., normal weight) but struggle to accurately classify extreme obesity or underweight conditions. Though techniques like SMOTE were used to handle imbalanced data, their effectiveness is limited by the original dataset's quality.

##### Missing or Incomplete Data:

The dataset may contain missing values or incomplete records for certain features, such as physical activity levels or calorie intake. Although imputation techniques were applied, missing data could still introduce inaccuracies or reduce the overall quality of model predictions.

#### Model Limitations

##### Overfitting in Complex Models:

More complex models like Random Forest, XGBoost, and SVM showed high accuracy on the training data but could suffer from overfitting, especially when the dataset is small or not sufficiently diverse. While cross-validation and regularization techniques were used, there is still a risk that these models may not generalize well to unseen or diverse data.

##### Interpretability:

While Random Forest and XGBoost models performed well in terms of accuracy, they are considered "black-box" models, meaning they provide limited transparency in terms of understanding how individual predictions are made. This can be a challenge for healthcare professionals or end-users who need to interpret the results of the tool clearly. Although feature importance analysis was used to provide insights, it may not fully explain the decision-making process.

##### Poisson Regression:

The use of Poisson Regression, which is generally designed for count data, did not perform as well in this context. Although it can model obesity-related counts or events, it wasn't the most suitable approach for predicting obesity levels based on continuous variables like calorie intake and physical activity.

### Future Work

Future improvements for this project include:

* **Expanding the Dataset:** Incorporating more diverse and comprehensive data, such as genetic factors and medical histories, to improve model accuracy.
* **Enhancing Model Accuracy:** Implementing advanced models (e.g., deep learning) and improving interpretability with techniques like SHAP or LIME.
* **User Interface & Integration:** Developing mobile apps, integrating with wearable devices for real-time monitoring, and providing personalized recommendations.
* **Continuous Learning:** Enabling real-time updates and automatic model retraining as new data is collected.
* **Addressing Bias:** Mitigating data bias and ensuring fairness for all demographics.
* **Collaboration with Healthcare:** Integrating the tool into clinical settings for validation and using it in telemedicine for remote consultations.
* **Public Health Applications:** Using the tool in public health campaigns to reduce obesity rates and inform policy.
* **Sustainability:** Ensuring the tool remains cost-effective and scalable through optimized resources and sustainable funding models.

These enhancements will improve the tool’s accuracy, accessibility, and long-term impact on public health.

# References

[1]Zarindokht Helforoush and H. Sayyad, “Prediction and classification of obesity risk based on a hybrid metaheuristic machine learning approach,” Frontiers in Big Data, vol. 7, Sep. 2024, doi: <https://doi.org/10.3389/fdata.2024.1469981>.

[2]N. Koklu and S. A. Sulak, “Using Artificial Intelligence Techniques for the Analysis of Obesity Status According to the Individuals’ Social and Physical Activities,” Sinop Üniversitesi Fen Bilimleri Dergisi, vol. 9, no. 1, pp. 217–239, Jun. 2024, doi: <https://doi.org/10.33484/sinopfbd.1445215>.

[3]Fadhila Tangguh Admojo and None Nurul Rismayanti, “Estimating Obesity Levels Using Decision Trees and K-Fold Cross-Validation: A Study on Eating Habits and Physical Conditions,” Indonesian Journal of Data and Science, vol. 5, no. 1, pp. 37–44, Mar. 2024, doi: <https://doi.org/10.56705/ijodas.v5i1.126>.

[4]J.-H. Jeong, I.-G. Lee, S.-K. Kim, T.-E. Kam, S.-W. Lee, and E. Lee, “DeepHealthNet: Adolescent Obesity Prediction System Based on a Deep Learning Framework,” *IEEE Journal of Biomedical and Health Informatics*, pp. 1–12, Jan. 2024, doi: <https://doi.org/10.1109/jbhi.2024.3356580>.

[5]A. Kumar, R. R, and K. Kumar, “Obesity Prediction Using Machine Learning Technique,” 2024 2nd International Conference on Networking and Communications (ICNWC), pp. 1–7, Apr. 2024, doi: <https://doi.org/10.1109/icnwc60771.2024.10537427>.

[6]Paulo and M. Lima, “Classification of Obesity Level Using Deep Neural Networks,” Lecture notes in networks and systems, pp. 99–107, Jan. 2024, doi: <https://doi.org/10.1007/978-3-031-64776-5_10>.

‌[7]F. Musa, F. Basaky, and O. E.O, “Obesity prediction using machine learning techniques,” Journal of Applied Artificial Intelligence, vol. 3, no. 1, pp. 24–33, Jun. 2022, doi: <https://doi.org/10.48185/jaai.v3i1.470>.

[8]R. C. Cervantes and U. M. Palacio, “Estimation of obesity levels based on computational intelligence,” Informatics in Medicine Unlocked, vol. 21, p. 100472, 2020, doi: <https://doi.org/10.1016/j.imu.2020.100472>.

[9]P. K. Mondal, K. H. Foysal, B. A. Norman, and L. S. Gittner, “Predicting Childhood Obesity Based on Single and Multiple Well-Child Visit Data Using Machine Learning Classifiers,” Sensors (14248220), vol. 23, no. 2, p. 759, Jan. 2023, doi: [https://doi.org/10.3390/s23020759.‌](https://doi.org/10.3390/s23020759.%E2%80%8C)

[10]Dahlak Daniel Solomon *et al.*, “Hybrid Majority Voting: Prediction and Classification Model for Obesity,” *Diagnostics*, vol. 13, no. 15, pp. 2610–2610, Aug. 2023, doi: <https://doi.org/10.3390/diagnostics13152610>.

[11]Stefano Lazzer *et al.*, “Cardiometabolic Index (CMI) and Visceral Adiposity Index (VAI) Highlight a Higher Risk of Metabolic Syndrome in Women with Severe Obesity,” *Journal of Clinical Medicine*, vol. 12, no. 9, pp. 3055–3055, Apr. 2023, doi: <https://doi.org/10.3390/jcm12093055>.

[12]X. Cai *et al.*, “Body roundness index improves the predictive value of cardiovascular disease risk in hypertensive patients with obstructive sleep apnea: a cohort study,” *Clinical and Experimental Hypertension*, vol. 45, no. 1, Oct. 2023, doi: <https://doi.org/10.1080/10641963.2023.2259132>.

[13]Nurullah Çelik, Gülşah Ünsal, and Hüseyin Taştanoğlu, “Predictive markers of metabolically healthy obesity in children and adolescents: can AST/ALT ratio serve as a simple and reliable diagnostic indicator?,” *European Journal of Pediatrics*, Oct. 2023, doi: <https://doi.org/10.1007/s00431-023-05296-3>.

[14]K. S. Kumar, M. Bee, and V. Thiruchelvam, “An analysis on obesity levels prediction based on smoking habits using stepwise linear regression algorithm in comparison with random forest classifier for improved accuracy,” *AIP conference proceedings*, vol. 3161, pp. 020242–020242, Jan. 2024, doi: <https://doi.org/10.1063/5.0229268>.

[15]Víctor Juan Vera-Ponce *et al.*, “Anthropometric Measures of Obesity as Risk Indicators for Prediabetes. A Systematic Review and Meta-Analysis,” *Diabetes Epidemiology and Management*, vol. 16, pp. 100230–100230, Jun. 2024, doi: <https://doi.org/10.1016/j.deman.2024.100230>.

[16]M Vecchiato *et al.*, “Cardiopulmonary exercise test registry for patients with obesity: comparison of cardiorespiratory fitness prediction equations and generation of a new predictive model for patients with obesity,” *European Journal of Preventive Cardiology*, vol. 31, no. Supplement\_1, Jun. 2024, doi: <https://doi.org/10.1093/eurjpc/zwae175.375>.

[17]F. H. Yagin *et al.*, “Estimation of Obesity Levels with a Trained Neural Network Approach optimized by the Bayesian Technique,” *Applied Sciences*, vol. 13, no. 6, p. 3875, Jan. 2023, doi: <https://doi.org/10.3390/app13063875>.

[18]T. L. Hernandez et al., “Continuous Glucose Monitor Metrics That Predict Neonatal Adiposity in Early and Later Pregnancy Are Higher in Obesity Despite Macronutrient-Controlled Eucaloric Diets,” Nutrients, vol. 16, no. 20, pp. 3489–3489, Oct. 2024, doi: <https://doi.org/10.3390/nu16203489>.

[19]V. M. Valicente *et al.*, “Ultraprocessed Foods and Obesity Risk: A Critical Review of Reported Mechanisms,” *Advances in Nutrition*, vol. 14, no. 4, Apr. 2023, doi: <https://doi.org/10.1016/j.advnut.2023.04.006>.

[20]F. Ferdowsy, K. S. A. Rahi, Md. I. Jabiullah, and Md. T. Habib, “A machine learning approach for obesity risk prediction,” *Current Research in Behavioral Sciences*, vol. 2, p. 100053, Nov. 2021, doi: <https://doi.org/10.1016/j.crbeha.2021.100053>.