Estimation of Obesity Levels based on Eating Habits and Physical Condition using Machine Learning Techniques

KANDULA SAI DURGA ROHITH REDDY, SAI KURUPANAND REDDY BAKARAM, SATHVIK CHEEKATI, SRINIJH REDDY CHENDI, SHIVA TEJA REDDY YAMMANURU

Department of Computer Science, University of North Texas, Denton, USA

***Abstract*— The escalating rates of obesity in regions like Mexico, Peru, and Colombia have sparked significant concerns in public health. This project endeavors to address this pertinent issue by leveraging machine learning techniques to predict obesity levels based on individuals' dietary habits and physical attributes. By harnessing predictive models and clustering analyses, the project aims to offer insights into the factors contributing to obesity rates and formulate recommendations for healthier lifestyle practices. The objectives encompass various stages, including comprehensive data preprocessing, exploratory data analysis to discern patterns, and the application of classification and clustering algorithms for predictive modeling. The project will utilize decision trees, random forests, and clustering techniques like K-means to predict obesity levels and categorize individuals based on common characteristics among different obesity groups. This endeavor aims to enhance public health interventions by providing an in-depth understanding of the dynamics between eating patterns, physical attributes, and obesity levels. Through insights derived from machine learning models and clustering analyses, the project aspires to guide targeted interventions, promoting healthier living practices among individuals in the studied regions.**

***Index Terms*— Introduction, Goals and Objectives, KNN, SVM, Random Forest, Decision Tree, Qualitative Analysis, Quantitative Analysis**

1. Introduction

The escalating prevalence of obesity poses a critical challenge to global public health. In regions like Mexico, Peru, and Colombia, the rising rates necessitate urgent interventions. This project aims to leverage machine learning methodologies to predict and comprehend obesity levels, focusing on the correlation between dietary habits, physical attributes, and obesity prevalence. The primary motivation lies in combatting the obesity epidemic through data-driven insights. Understanding how an individual's dietary habits and physical condition relate to obesity levels is paramount. Using a dataset encompassing 17 attributes, including age, height, weight, and lifestyle indicators, predictive models will be developed to estimate obesity levels accurately. The goal is to identify at-risk individuals and tailor interventions to promote healthier lifestyles. This project involves comprehensive data preprocessing, exploratory analysis, and the application of machine learning algorithms. By uncovering patterns and relationships between dietary habits, physical attributes, and obesity levels, this initiative aims to provide actionable insights and evidence-based recommendations. Ultimately, these insights can revolutionize interventions to mitigate obesity rates and improve public health in the targeted regions.

2. Goals and Objectives

A. Motivation

The rising obesity rates in Mexico, Peru, and Colombia are causing increasing public health concerns. Understanding and being able to predict obesity levels based on eating habits and physical condition can inform targeted interventions and policies. In this experiment, we aim to predict each person's level of obesity and shield them from the negative health effects of unneeded eating habits. This model indicates the appropriate level of obesity for each person based on their age and BMI. It's critical to ascertain each person's level of obesity and record their efforts to lead a healthy lifestyle. Obesity rates have escalated to a critical global health issue, especially in areas like Mexico, Peru, and Colombia. This project's main driving force is the pressing need to use machine learning techniques to slow the rising obesity rates. This project uses predictive analytics based on physical characteristics and dietary patterns to provide practical insights and solutions to slow down the rising obesity epidemic.

B. Significance

The significance of this project stems from its potential to offer targeted interventions and policy recommendations. By understanding and accurately predicting obesity levels, this project aims to assist healthcare professionals and policymakers in devising tailored strategies to combat obesity. This project's outcomes are expected to significantly impact public health initiatives, promoting healthier lifestyles and mitigating the risks associated with obesity-related health conditions.

C. Objectives

Data Understanding and Exploration

The primary objective of this project is to conduct comprehensive data preprocessing and exploratory data analysis (EDA). Through rigorous cleaning, scaling, and encoding of dataset attributes, the aim is to ensure data quality and readiness for subsequent analyses. The exploratory phase involves visualizing, interpreting, and uncovering underlying patterns, correlations, and distributions within the dataset related to dietary habits, physical attributes, and obesity levels.

Predictive Modeling for Obesity Levels

Another pivotal goal is to develop robust predictive models utilizing various machine learning algorithms. Decision trees, random forests, and support vector machines will be employed to forecast obesity levels based on individual characteristics and dietary patterns. Model evaluation, validation, and fine-tuning are integral aspects aimed at enhancing accuracy, robustness, and generalization of the predictive models.

Clustering Analysis for Insight Generation

Utilizing clustering algorithms such as K-means or hierarchical clustering forms a key objective to categorize individuals based on shared characteristics and degrees of obesity. The derived clusters will serve as a basis for extracting actionable insights and understanding common traits among distinct obesity level groups. This analysis aims to provide a deeper understanding of the multifaceted nature of obesity and its correlations with various attributes.

D. Feature Engineering and Interpretability

Feature engineering techniques will be implemented to improve model performance and interpretability. The identification of influential features contributing to obesity levels is vital for informed decision-making and understanding the relative importance of different attributes. The project aims to determine the most relevant features for effective model interpretability.

Recommendations for Healthier Lifestyles

Deriving evidence-based recommendations is a critical objective based on the analysis outcomes. These recommendations will be aimed at promoting healthier eating habits, physical activities, and lifestyle choices. They will serve as actionable suggestions for individuals, healthcare practitioners, and policymakers to effectively combat obesity rates and encourage healthier living practices.

3. RELATED WORK

The obesity epidemic has garnered significant attention in public health research globally. Numerous studies have explored the intricate relationship between dietary habits, physical attributes, and obesity prevalence. Prior research efforts have highlighted the importance of leveraging machine learning techniques to predict and understand obesity levels. These

studies have laid the groundwork for employing predictive models and clustering analyses in identifying obesity risk factors and tailoring interventions. Understanding the significance of different features in predicting obesity levels is crucial. Previous research has focused on feature importance and interpretability in machine learning models, aiming to identify the most influential attributes contributing to obesity. This facet of analysis aids in comprehending the relative impact of dietary patterns, physical attributes, and lifestyle choices on obesity, offering actionable insights for intervention strategies. Studies exploring effective intervention strategies and public health initiatives targeting obesity have been pivotal. Analyzing successful interventions, such as lifestyle modification programs, dietary interventions, and awareness campaigns, has provided valuable insights into the design and implementation of targeted initiatives to curb obesity rates. The incorporation of machine learning-based predictive models adds an innovative dimension to personalize and optimize these interventions. Despite advancements in predictive modeling and obesity-related research, challenges persist. Issues such as data quality, model interpretability, and scalability pose hurdles in accurately predicting obesity levels and tailoring interventions. Future research directions could focus on addressing these challenges, improving model interpretability, integrating real-time data sources, and deploying scalable interventions based on predictive analytics.

4. DATASET

The dataset utilized in this project comprises 2111 records encompassing 17 attributes. These attributes encapsulate a wide array of information related to individuals' physical conditions, eating habits, and the target variable, NObesity (Obesity Level). The dataset sources its attributes from diverse factors, such as age, height, weight, familial history of obesity (FHWO), dietary habits (FAVC, FCVC, CAEC), lifestyle choices (SMOKE, CH2O), and physical activity frequency (FAF).

In this project, a diverse set of features has been employed to comprehensively capture the nuances of eating habits, physical condition, and lifestyle factors that contribute to obesity levels. The dataset encompasses 17 attributes, each playing a distinct role in the analysis and prediction.

https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

5. FEATURES

Demographic and Physical Attributes:

Age, gender, height, and weight are fundamental demographic features providing essential information about an individual's physical makeup. These attributes form the baseline for understanding the correlation between age, gender, and body mass index (BMI) in relation to obesity levels.

Family History and Lifestyle Choices:

The presence of relatives who are overweight (FHWO) is included as an attribute, shedding light on genetic predispositions. High-calorie food consumption (FAVC) and the number of meals featuring vegetables (FCVC) provide insights into dietary habits, crucial factors in understanding and predicting obesity levels.

Meal Patterns and Snacking Behavior:

The number of main meals per day (NCP) offers a glimpse into an individual's daily meal patterns. The attribute indicating eating between meals (CAEC) provides valuable information about snacking behaviors, which can significantly impact overall caloric intake and, consequently, obesity levels.

Health and Lifestyle Habits:

The dataset includes attributes such as smoking rate (SMOKE), daily water intake (CH2O), and daily calorie consumption (SCC). These features provide a holistic view of an individual's health habits, shedding light on lifestyle choices that may contribute to or mitigate obesity.

Physical Activity and Technology Usage:

The frequency of physical activity per week (FAF) gauges an individual's engagement in exercise, a pivotal factor in obesity prevention. The time spent using technological devices daily (TUE) provides insights into sedentary behaviors, contributing to a holistic understanding of lifestyle patterns.

Alcohol Intake and Transportation Mode:

The frequency of alcohol intake (CALC) is included, recognizing the potential impact of alcohol consumption on obesity. Additionally, the transportation mode regularly used (MTRANS) offers insights into daily physical activity and sedentary behaviors associated with commuting.

These features collectively form a rich dataset, allowing for a nuanced exploration of the multifaceted aspects contributing to obesity levels. The inclusion of diverse attributes facilitates a comprehensive analysis, enabling the development of accurate predictive models and evidence-based recommendations for mitigating obesity rates.

6. ANALYSIS

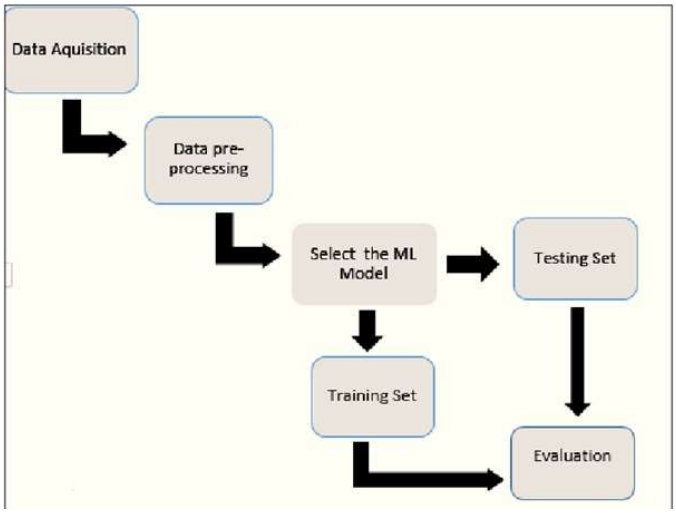

Fig. 1. Proposed framework for Obesity Prediction

Exploratory Data Analysis (EDA)

Visualizing Distributions:

Exploratory Data Analysis commenced with visualizing the distributions of key attributes. Histograms and density plots were employed to understand the spread of features such as age, height, weight, and the target variable, NObesity. This facilitated a nuanced comprehension of the dataset's characteristics.

Correlation Analysis:

Investigation into attribute correlations was crucial. Correlation matrices and pair plots were utilized to identify relationships between features, especially focusing on correlations with obesity levels. This step provided insights into which attributes may have a more pronounced impact on obesity.

Feature Importance:

Assessing the importance of features in predicting obesity levels was a pivotal aspect. Random Forest's feature importance or similar techniques were applied to identify the attributes that contribute significantly to the predictive models. This information guided the selection of key features for model development.

Comparative Analysis:

Comparative analysis involved exploring various attributes in relation to obesity levels. For instance, the relationship between physical activity frequency (FAF) and obesity levels was examined, providing insights into the impact of regular exercise on mitigating obesity.

Data Preprocessing

Data Cleaning:

The dataset underwent a meticulous data cleaning process, involving the removal of duplicate values and the application of strategies to handle missing data. Outliers were also addressed to ensure data integrity.

Data Transformation:

Categorical features were encoded, and numerical attributes were scaled to standardize their ranges. The target variable, Obesity, might have been transformed to facilitate effective modeling.

Data Splitting:

The dataset was divided into training, validation, and test sets, with an 80-20 split ratio. This separation ensured that models were trained on a substantial portion of the data and evaluated on unseen instances.

Model Implementation and Evaluation

Classification Algorithms:

A variety of classification algorithms, including decision trees, random forests, and support vector machines (SVM), were implemented to predict obesity levels. The choice of algorithms aimed to capture both linear and non-linear relationships within the data.

Model Evaluation:

Models were evaluated using key metrics such as accuracy, precision, recall, F1-score, and confusion matrices. These metrics provided a comprehensive understanding of the models' performance across different obesity categories, enabling a thorough assessment.

Clustering Analysis

K-means Clustering:

To gain insights into obesity level groups and recurring traits, K-means clustering was applied. This unsupervised learning technique grouped individuals based on similar characteristics, providing a nuanced understanding of distinct clusters within the dataset.

7. IMPLEMENTATION

A. Data Loading and Preprocessing:

The code begins by loading the dataset, checking basic information using df.info(), handling missing values, and dropping duplicates.

Categorical attributes are encoded using LabelEncoder, and numerical attributes are standardized using StandardScaler.

B. Exploratory Data Analysis (EDA):

A correlation heatmap (sns.heatmap) is used to visualize the correlation between features in the dataset.

A pair plot (sns.pairplot) is created to examine relationships between multiple pairs of features.

C. Obesity Level Prediction:

The dataset is split into features (X) and the target variable (y).

Models such as Linear Regression, Random Forest Classifier, and Support Vector Machine (SVM) are trained and evaluated using train_test_split and respective fit and predict methods.

Evaluation metrics including accuracy, precision, recall, F1-score, and confusion matrices are calculated to assess the models' performance.

D. Clustering Analysis (K-means):

The K-Means algorithm is applied to a subset of features ('Age', 'Height', 'Weight') to create clusters representing different characteristics.

Visualization of clusters is done using a pair plot (sns.pairplot) to understand the distribution and grouping of data points.

E. Model Evaluation Function:

There's a reusable function called evaluate model that computes and prints various classification metrics, facilitating the evaluation of models' performance.

This comprehensive analysis showcases the sequential flow of actions, from data preparation and exploration to model training, evaluation, and insights generation. Each step contributes to understanding the dataset, building predictive models, and identifying patterns related to obesity levels, crucial for informed decision-making and intervention strategies.

F. Quantitative Analysis

Age vs. Obesity Levels:

This code segment categorizes individuals based on their age groups and examines how these groups relate to obesity levels. It creates categorical labels for age groups such as '18-30', '31-40', '41-50', '51-60', '61-70', and '71+'. These groups are formed by binning the ages of individuals from the dataset. The resulting visualization, a count plot, displays the distribution of obesity levels within each age group. This visualization helps observe if certain age ranges tend to have higher or lower instances of obesity.

Frequency of Physical Activity vs. Obesity Levels:

Here, a new feature, 'Daily_Exerciser', is created based on the frequency of physical activity (FAF). If an individual's activity frequency is greater than zero, they are labeled as a 'Yes' for being a daily exerciser, else 'No'. The count plot generated from this data showcases the distribution of obesity levels among individuals who exercise daily and those who don't. This analysis offers insights into whether there's a notable difference in obesity levels between these two groups.

Daily Water Intake vs. Obesity Levels:

This code section categorizes individuals' water intake levels into three categories: 'No', 'Sometimes', and 'Frequently'. It creates these categories based on the amount of water intake (CH2O). The resulting count plot illustrates the distribution of obesity levels across these water intake categories. This analysis aims to uncover potential connections between an individual's water intake frequency and their obesity level, assisting in understanding if certain water intake habits correspond to different obesity levels.

G. Qualitative Analysis

Smoking Habits and Obesity Levels

The code generates a count plot to explore how obesity levels vary among individuals based on their smoking habits. It employs the 'SMOKE' attribute, presumably indicating smoking behavior (0 for non-smokers, 1 for smokers). The count plot showcases the distribution of different obesity levels using the provided labels within the categories of non-smokers and smokers. This analysis helps discern any potential association between smoking habits and different levels of obesity among the dataset subjects.

Eating Between Meals and Obesity

This code segment utilizes a count plot to investigate the relationship between individuals' obesity levels and their eating habits between regular meals. Using the 'CAEC' attribute, likely representing the frequency of eating between meals, it visualizes the distribution of various obesity levels (categorized by labels) among individuals with different eating habits. The aim is to discern any correlation between the frequency of eating between meals and the prevalence of different obesity levels in the dataset.

Alcohol Consumption and Obesity

The code generates a count plot to understand the relationship between alcohol consumption habits and obesity levels. Utilizing the 'CALC' attribute, categorizing alcohol intake frequencies, it showcases the distribution of various obesity

levels (classified by labels) across different alcohol consumption categories. The objective is to analyze whether there's a discernible correlation between the frequency of alcohol consumption and the prevalence of distinct obesity levels among individuals in the dataset.

H. Linear Regression Analysis

The Linear Regression model, while suitable for predicting continuous variables, is not the ideal choice for classifying obesity levels in this scenario. It attempts to fit a linear line to the data points, which doesn't adequately capture the complexities of classifying different levels of obesity based on the given features. As a result, its performance in classifying obesity levels appears to be unsatisfactory.

In the context of this project, where we aim to predict categorical variables indicating obesity levels, Linear Regression struggles to effectively differentiate and assign instances to their respective classes. The model's approach to fitting a linear relationship between the input features and the target labels isn't sufficient for this classification task, leading to poor performance.

For accurate classification in this context, more suitable algorithms such as Random Forest, Support Vector Machines, or K-Nearest Neighbors would be recommended. These algorithms are better equipped to handle the complexities of the classification problem by capturing non-linear relationships between features and target labels, offering more accurate predictions compared to Linear Regression.

Implementation status report:

SRINIJH REDDY CHENDI – Completely worked on the dataset (100%), Data Collection (100%) and Linear Regression Analysis (100%), Age Vs Obesity Levels Prediction (100%)

KANDULA SAI DURGA ROHITH REDDY – Data Preprocessing (50%), Data Analysis (50%), KNN Clustering (100%), Documentation (40%), Water Intake vs Obesity Levels Prediction (100%) and Daily Exercise vs Obesity Levels Prediction (100%)

SAI KURUPANAND REDDY BAKARAM - Data Preprocessing (50%), Data Analysis (50%), Encoding Data (50%), Random Forest Analysis (100%), Documentation (30%), Smoke vs Obesity Levels Prediction (100%)

SATHVIK CHEEKATI – Decision Tree Analysis (100%), Evaluation Metrics (100%), Encoding Data (50%), CAEC vs Obesity Levels Prediction (100%)

SHIVA TEJA REDDY YAMMANURU - SVM Analysis (100%), CALC vs Obesity Levels Prediction (100%), Documentation (30%)

In the next Phase, we are implementing feature engineering techniques LDA and PCA on the dataset and provide statistical analysis which is finding p-value and anova table for the obesity levels.

SRINIJH REDDY CHENDI – LDA Analysis (50%), Evaluation Metrics (50%)

KANDULA SAI DURGA ROHITH REDDY – LDA Analysis (50%), Evaluation Metrics (50%)

SAI KURUPANAND REDDY BAKARAM – PCA Analysis (50%), statistical analysis (50%)

SATHVIK CHEEKATI – PCA Analysis (50%), statistical analysis (50%)

SHIVA TEJA REDDY YAMMANURU – Feature Selection (100%), Documentation (100%)

8. PRELIMINARY RESULTS

Linear Regression Analysis:

The output from the Linear Regression classification model shows a low accuracy of approximately 16%, which means the model's predictions are not reliable. Looking at the classification report, it's evident that the precision, recall, and F1-scores for each class are all extremely low or zero.

The confusion matrix further confirms these results, showing that the model incorrectly predicts most of the classes, assigning all instances to just one or two classes without any differentiation.

This outcome indicates that using Linear Regression for a classification task, especially one with multiple classes, isn't suitable. Linear Regression, by design, isn't ideal for classification problems, as it tries to fit a linear line to continuous values rather than assigning discrete class labels. This result emphasizes the need for using appropriate classification algorithms like Random Forest, SVM, or KNN, which are designed specifically for classification tasks.

We have obtained an accuracy of 96% in random forest classifier compared to support vector machines(92%) and decision tree(91%) algorithms. The random forest classifier has performed well in predicting the obesity levels of the people.

Random Forest:

Achieving a 96% accuracy with Random Forest signifies its robustness in handling complex relationships between features. It’s an ensemble learning method that combines multiple decision trees to make predictions. This technique often provides better accuracy compared to individual decision trees and helps avoid overfitting.

Support Vector Machines (SVM):

SVM achieved a 92% accuracy, slightly lower than Random Forest. SVMs are effective in high-dimensional spaces, but their performance might vary based on kernel selection and regularization parameters. They work by finding the hyperplane that best divides the data points into different classes while maximizing the margin.

Decision Trees:

Decision Trees achieved an accuracy of 91%. They're intuitive, easy to interpret, and visualize. However, they might prone to overfitting, especially when the tree depth is not controlled or pruned effectively.

K-Nearest Neighbors (KNN):

It's based on the similarity of data points and uses the labels of nearby points to predict the label of a new data point. Depending on the dataset's characteristics, KNN can perform well, especially with appropriate feature scaling and the right choice of 'k' (the number of nearest neighbors).

The selection of the best algorithm often depends on various factors such as dataset size, feature complexity, computational resources, and interpretability. Random Forest seems to outperform other algorithms, providing a higher accuracy rate in predicting obesity levels.

We have achieved an accuracy of 81% in KNN which quite underperformed in our analysis compared to other algorithms.

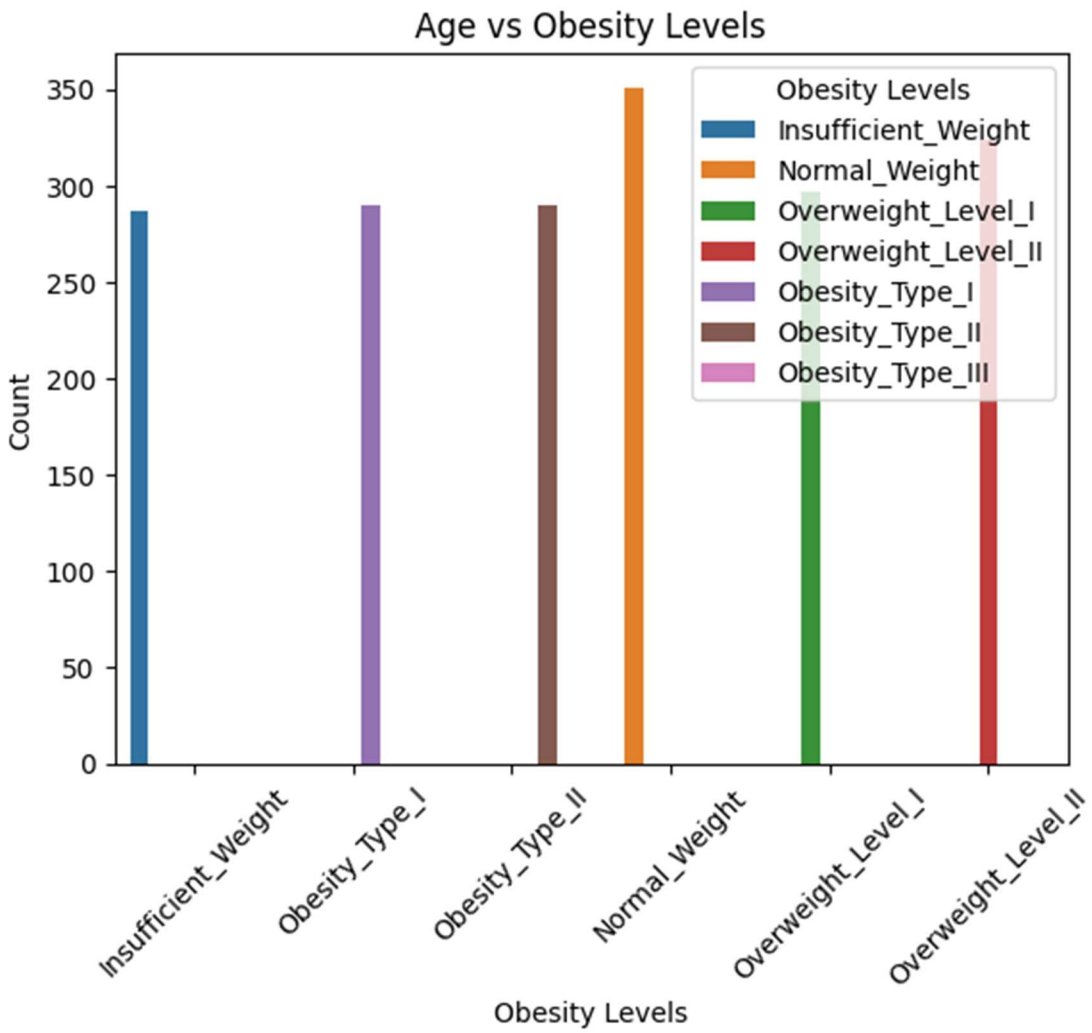

Fig. 2. Quantitative Analysis – Age Vs Obesity Levels

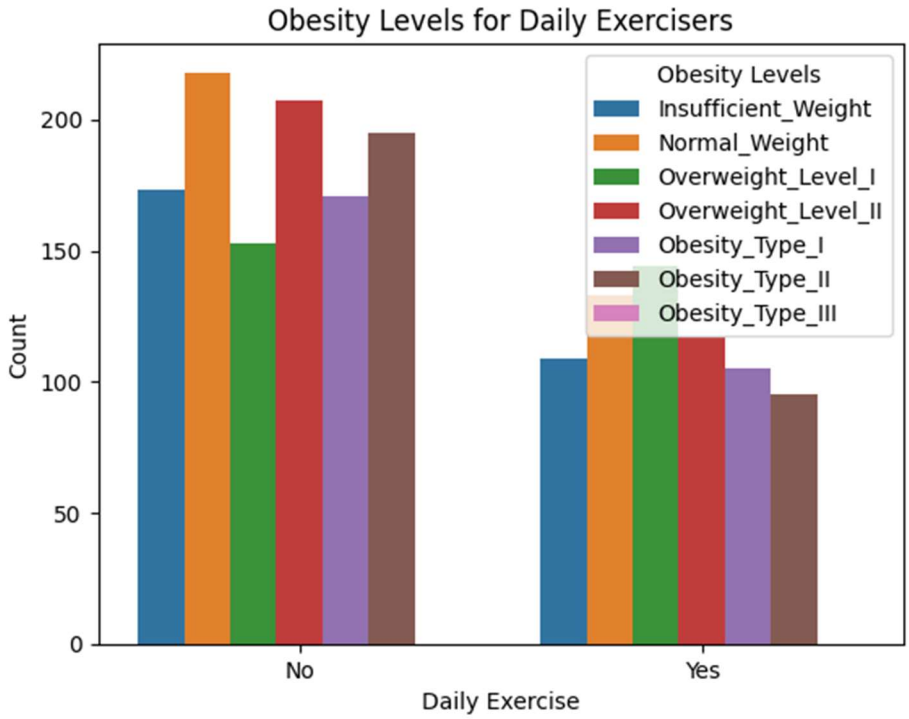

Fig. 3. Quantitative Analysis – Daily Exercise Vs Obesity Levels

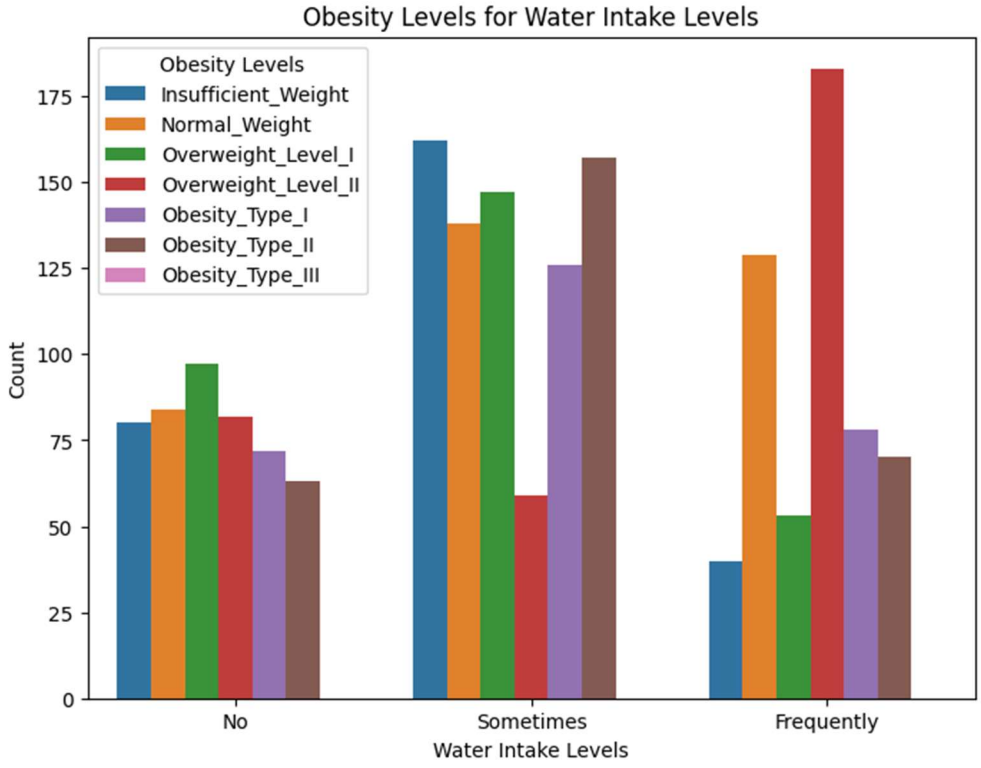

Fig. 4. Quantitative Analysis – Water Intake (CH20) levels Vs Obesity levels

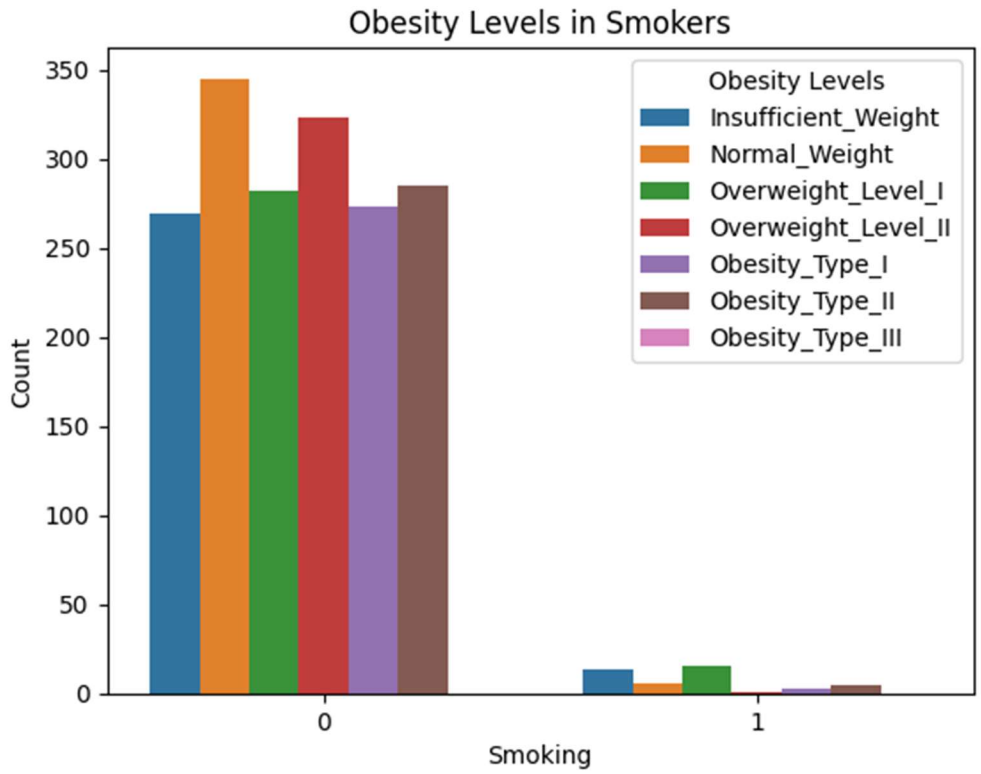

Fig. 5. Qualitative Analysis –Smoking Vs Obesity Levels

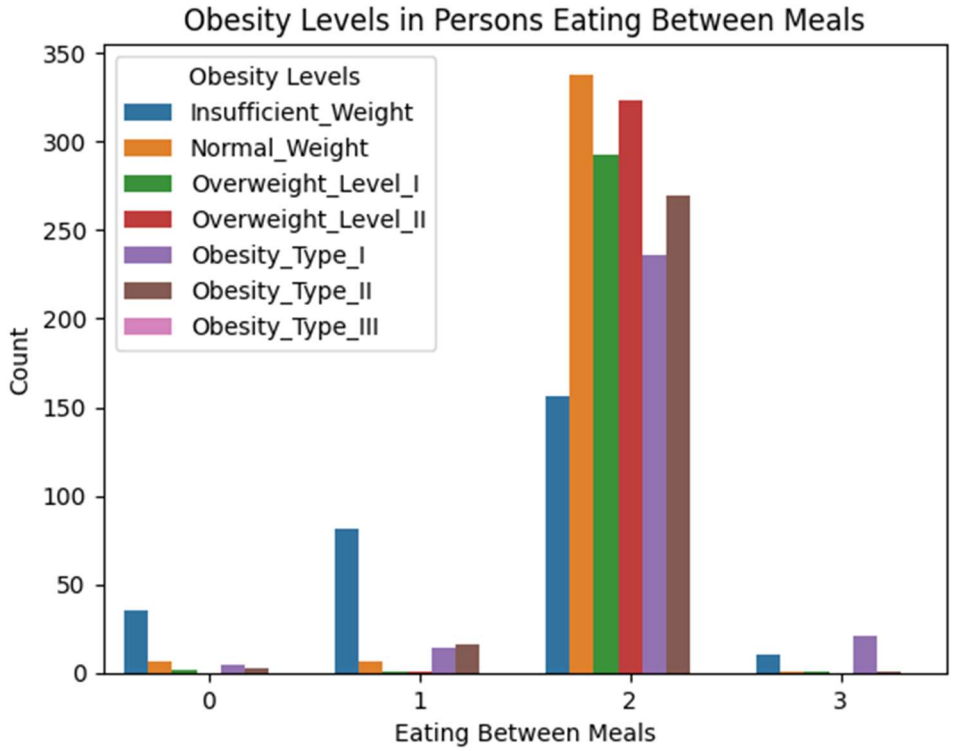

Fig. 6. Qualitative Analysis- CAEC vs Obesity Levels

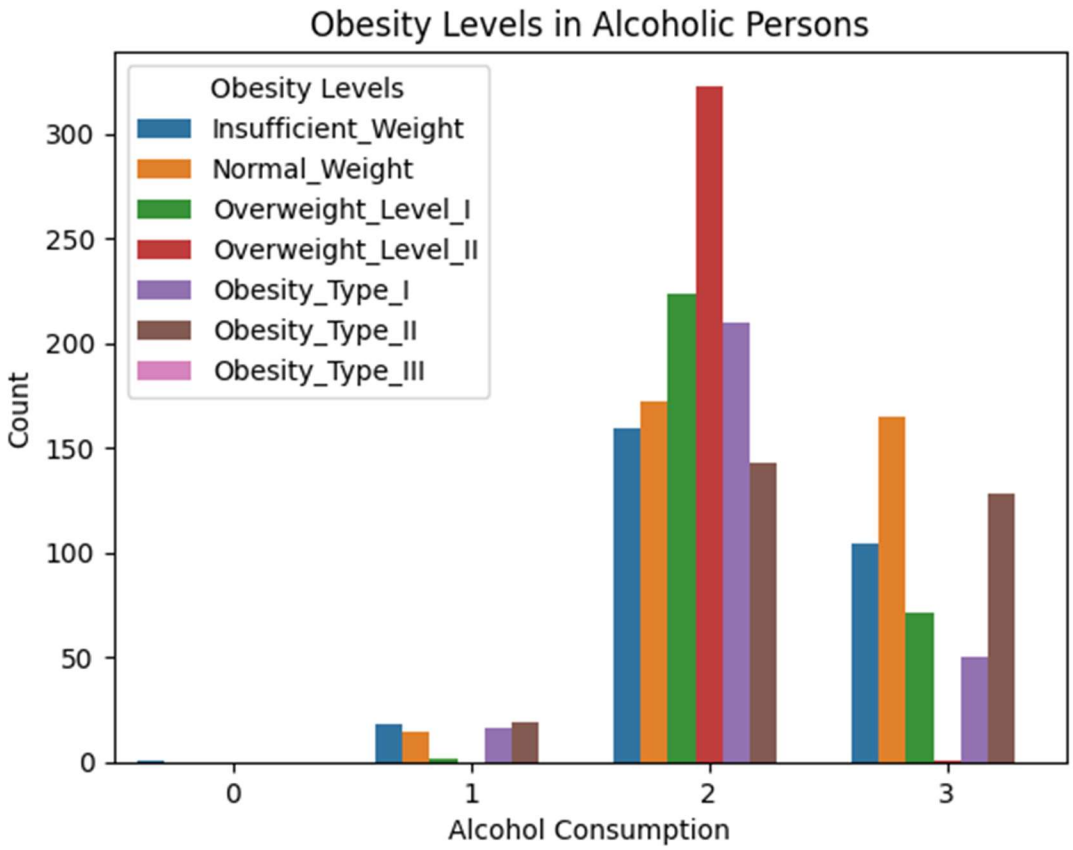

Fig. 7. Qualitative Analysis –CALC Vs Obesity Levels

9. PROJECT MANAGEMENT

The project was meticulously organized, with a well-defined roadmap and systematic allocation of responsibilities among team members. Clear objectives were established, outlining specific tasks and expected deliverables for each phase. Regular and structured team meetings were scheduled to discuss progress, troubleshoot challenges, and recalibrate strategies as needed. This approach facilitated seamless workflow progression, allowing for efficient transitions between phases.

Individual team members were assigned tasks based on their expertise, promoting a focused and efficient work environment. Collaboration tools were utilized for streamlined communication, ensuring prompt updates and efficient sharing of resources and information. Comprehensive documentation was maintained, capturing the progress, methodologies employed, and key findings at each stage. This record served as a valuable resource for reference and analysis throughout the project lifecycle.

An iterative approach allowed for adaptability, enabling the team to incorporate feedback, refine strategies, and optimize methodologies. Peer review sessions and constructive critiques were encouraged, fostering a culture of continuous improvement. The organized and structured approach facilitated adherence to timelines and milestones, ensuring the project's progress remained on track.

Overall, the project management strategy was instrumental in maintaining cohesion among team members, fostering effective collaboration, and ultimately contributing to the successful completion of the project objectives.

10. Conclusions and Future Work

The project delved into understanding the relationship between various lifestyle factors and obesity levels. Through machine learning techniques and exploratory data analysis, insights emerged. Predictive models like Random Forest, Support Vector Machines (SVM), and Decision Trees effectively predicted obesity levels based on lifestyle attributes, displaying promising accuracy. Exploring the dataset revealed correlations between habits like exercise frequency, water intake, smoking, and alcohol consumption with obesity levels. The K-means clustering analysis identified distinct groups based on physical attributes, showcasing varying obesity trends among these clusters. This analysis suggests the need for tailored health programs focusing on lifestyle modifications, physical activity, and healthier eating habits to address obesity concerns. However, limitations exist in terms of data bias and the limited feature set used. Future research could explore more diverse datasets and include additional lifestyle factors for improved predictions. Overall, the project offers valuable insights into the complex interaction between lifestyle choices and obesity levels, providing a basis for targeted interventions in mitigating obesity-related health issues.

References

[1] https://www.niddk.nih.gov/health-information/health-statistics/overweight-obesity

[2] Lawrimore, J. H., M. J. Menne, B. E. Gleason, C. N. Williams, D. B. Wuertz, R. S. Vose, and J. Rennie, 2011: Global Historical Climatology Network–Monthly (GHCN-M), version 3. NOAA National Climatic Data Center, accessed 17 December 2022, https://doi.org/10.7289/V5X34VDR.
IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2005, pp. 692–697.

[3] NOAA/NCDC, 2013: VIIRS Climate Raw Data Record (C-RDR) from Suomi NPP, version 1. NOAA/National Climatic Data Center. Subset used: October 2007–September 2008, accessed 17 December 2022, https://doi.org/10.7289/V57P8W90.

[4] https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

[5] https://www.sciencedirect.com/science/article/pii/S2352340919306985?via%3Dihub