Abstract

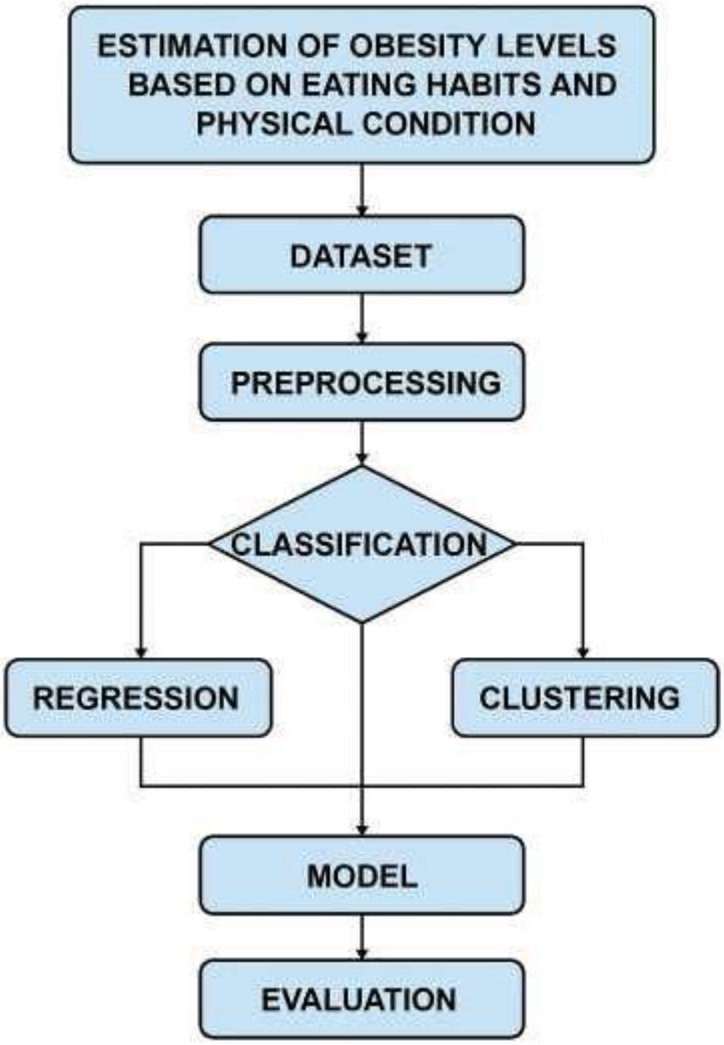
Obesity is a major global health concern caused by a combination of dietary, behavioral, and physical activity factors. This study applies machine learning techniques to classify and predict obesity levels based on individuals’ lifestyle data. The dataset from the UCI repository via Kaggle includes features such as age, gender, food habits, water intake, physical activity, smoking, and alcohol consumption. We implemented multiple supervised and unsupervised algorithms including Logistic Regression, Random Forest, SVM, Linear Regression, K-Means. The classification models achieved an accuracy of over 90%, while regression predicted BMI values with a high R^2 score. Clustering revealed distinct lifestyle groups correlating with obesity tendencies. This project demonstrates the power of data-driven approaches in health risk assessment and lays the foundation for personalized health recommendations.

Keywords:

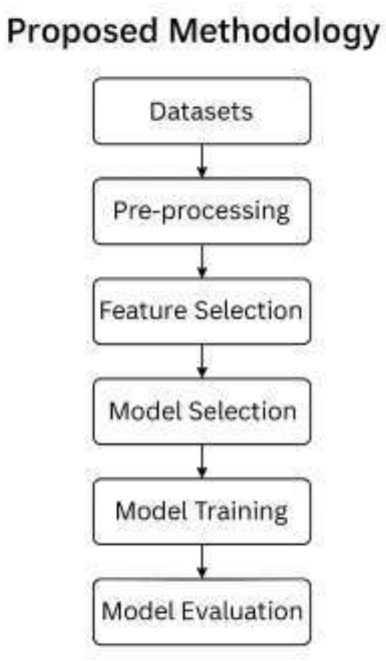
Obesity Prediction, Classification, Regression, Clustering, Machine Learning

1. Introduction:

Obesity is a major health issue affecting individuals across all age groups. Traditional methods for determining obesity levels often rely on BMI or clinical diagnoses, which may not consider a person's lifestyle comprehensively. With the growing availability of data, machine learning provides a powerful approach to model and predict obesity levels from behavioral and physiological indicators. This project utilizes various machine learning algorithms to classify and predict obesity levels from data collected on dietary habits, physical activity, and other lifestyle factors.


ESTIMATION OF OBESITY LEVELS BASED ON EATING HABITS AND PHYSICAL CONDITION
DATASET
PREPROCESSING
CLASSIFICATION
REGRESSION
CLUSTERING
MODEL
EVALUATION

2. Proposed Methodology:



a. Dataset:

The dataset used in this project includes individual lifestyle attributes such as eating habits, physical activity, and body measurements. It was sourced from Kaggle and includes both categorical and numerical features.

- Source: Kaggle (Obesity Dataset)
- Features: 17 variables including age, weight, height, eating habits, activity levels.
- https://www.kaggle.com/code/mpwolke/obesity-levels-life-style/input

b. Preprocessing:

- Label encoding for categorical variables
- BMI calculation: BMI = Weight / (Height^2)
- Standardization using StandardScaler
- Splitting data into training and testing (80/20)

c. Model Used

Task	**Model Used**
Classification	Logistic Regression, KNN Classifier, Decision Tree Classifier, Random Forest Classifier, SVM Classifier
Regression	Linear Regression, Lasso Regression, Ridge Regression, Decision Tree, Random Forest
Clustering	K- Means

3. Result

Model - Classification	**Accuracy**
Logistic Regression	72%
K-Nearest Neighbors	81%
Random Forest Classifier	**91%**
Gradient Boosting	88%
Support Vector Machine(SVM)	86%

Model - Regression	**RMSE / R^2 Score**
Linear Regression	12.68 / 0.77
Ridge Regression	12.68 / 0.77
Lasso Regression	12.92 / 0.76
Random Forest	**7.18 / 0.93**

Clustering (KMeans - Unsupervised)

Total clusters: 5

Cluster sizes:

- Cluster 3: 668
- Cluster 1: 535
- Cluster 0: 493
- Cluster 4: 371
- Cluster 2: 44

Insight: Cluster 2 is significantly underrepresented, indicating a potentially rare profile or outlier behavior. Consider examining its characteristics.

Key Observations

1. Model Performance

🟩Random Forest performed best for both:

- Classification (91% accuracy)
- Regression ($R^2 = 0.93$, RMSE = 7.18)

This suggests that obesity levels are influenced by non-linear relationships and interactions between features, which Random Forest handles well.

2. Classification Insights

- Logistic Regression gave the lowest accuracy (72%), showing that linear models are not ideal for this multi-class problem.
- KNN, SVM, and Gradient Boosting performed well (81%–88%) but not as good as Random Forest.

Certain classes like class 4 (Obesity Type II) achieved very high precision and recall, meaning these categories are easier to classify based on the input features.

Classes 1 and 6 showed lower precision/recall, indicating overlap or ambiguity in their feature patterns.

3. Regression Insights

- Linear models (Linear, Ridge, Lasso) showed moderate performance ($R^2 \sim 0.76–0.77$).
- Ensemble models (Random Forest, Gradient Boosting) gave significantly better prediction of obesity scores ($R^2 > 0.90$).

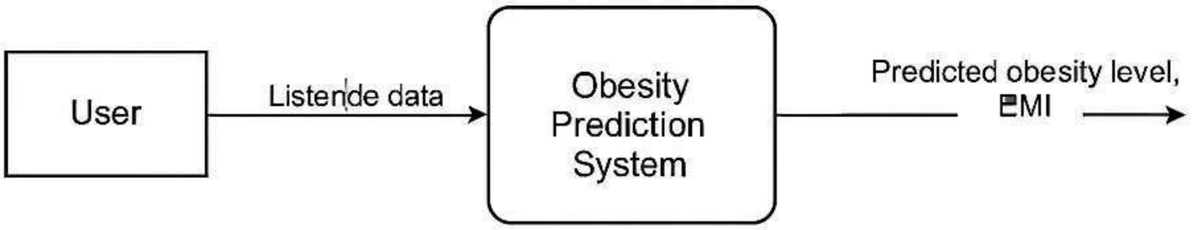
RMSE was lowest for Random Forest, indicating more accurate and stable predictions.

4. Clustering (KMeans)

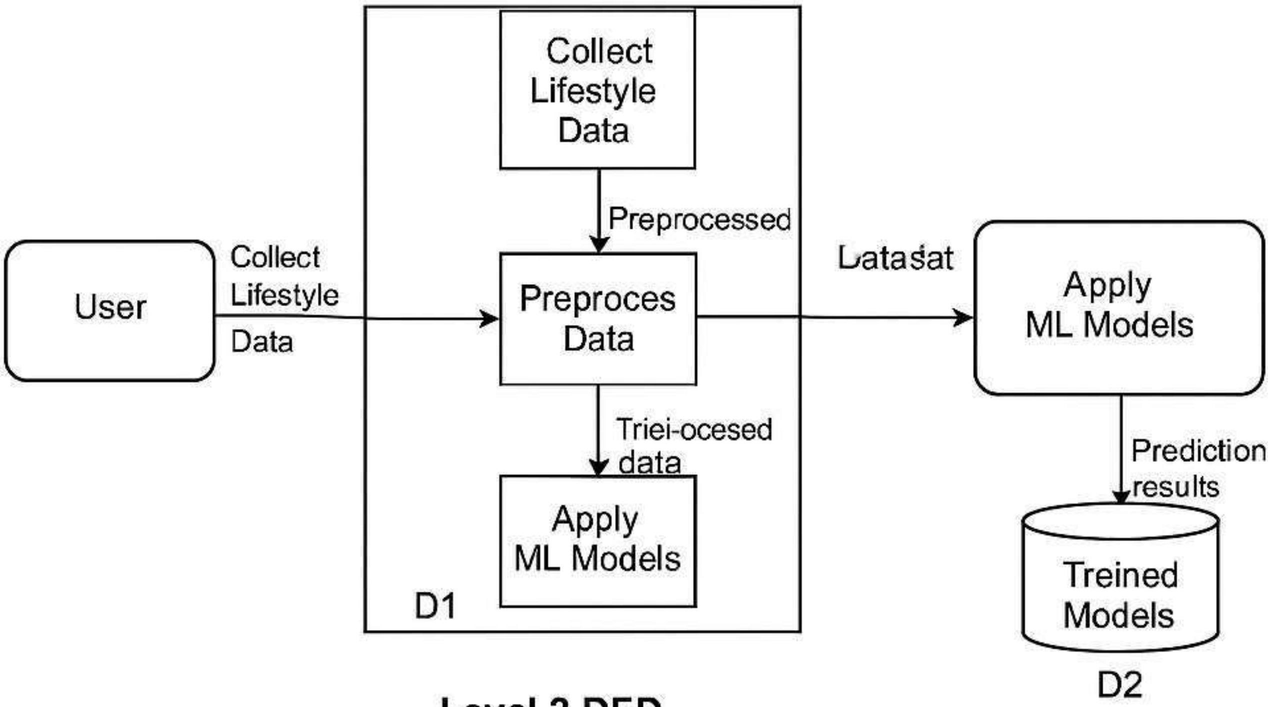
- KMeans grouped the population into 5 clusters, with the largest cluster having 668 individuals.

These clusters can represent hidden lifestyle patterns, but need further interpretation or labeling.

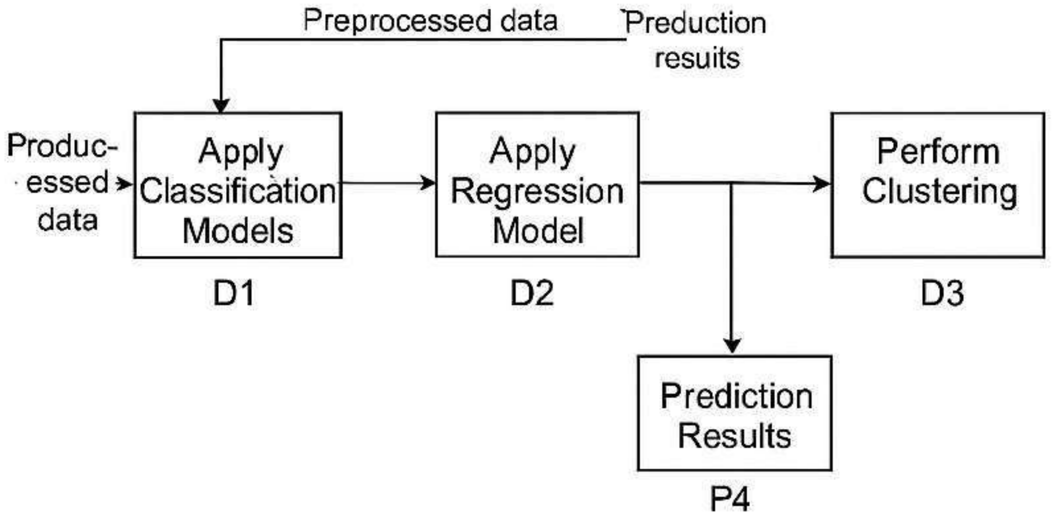
Context Level DFD (Level 0)



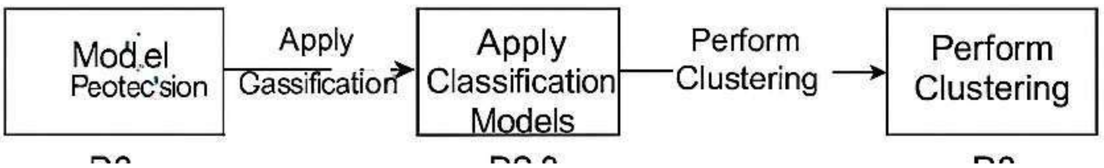
Level 1 DFD



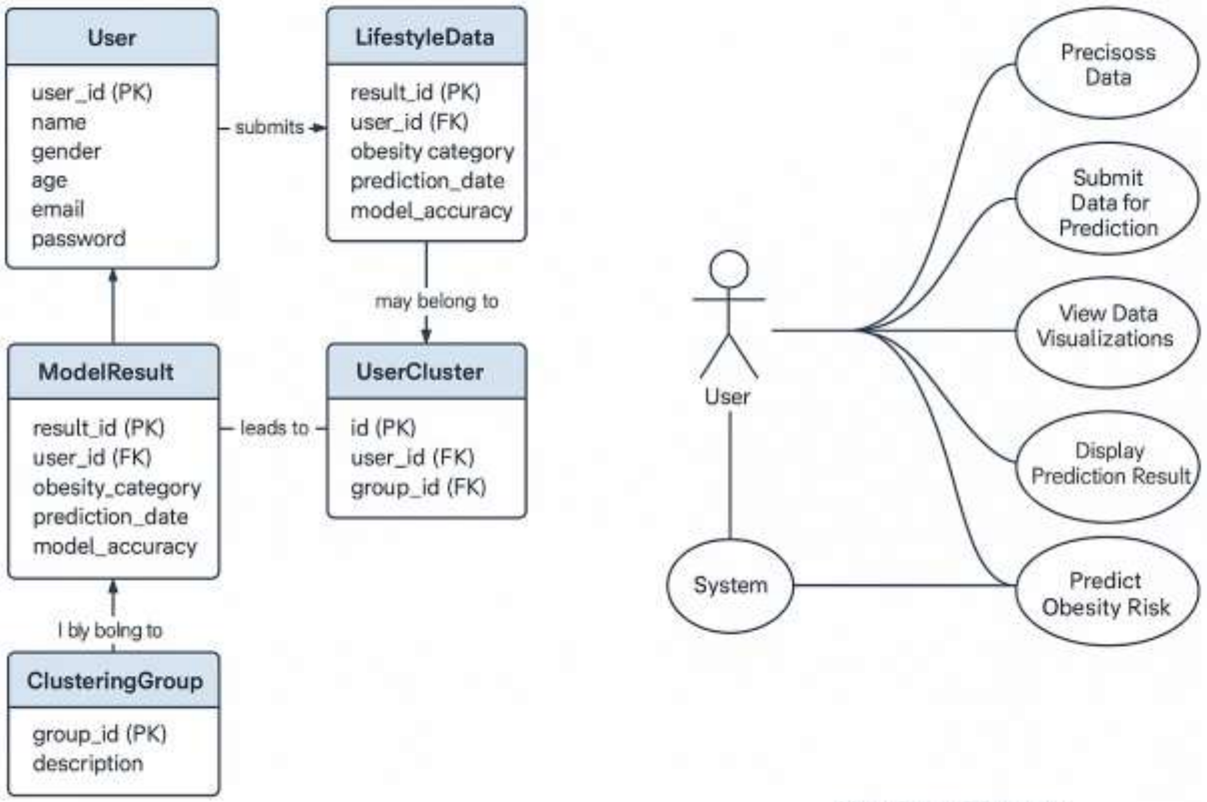
Level 2 DFD



Level 2 DDF

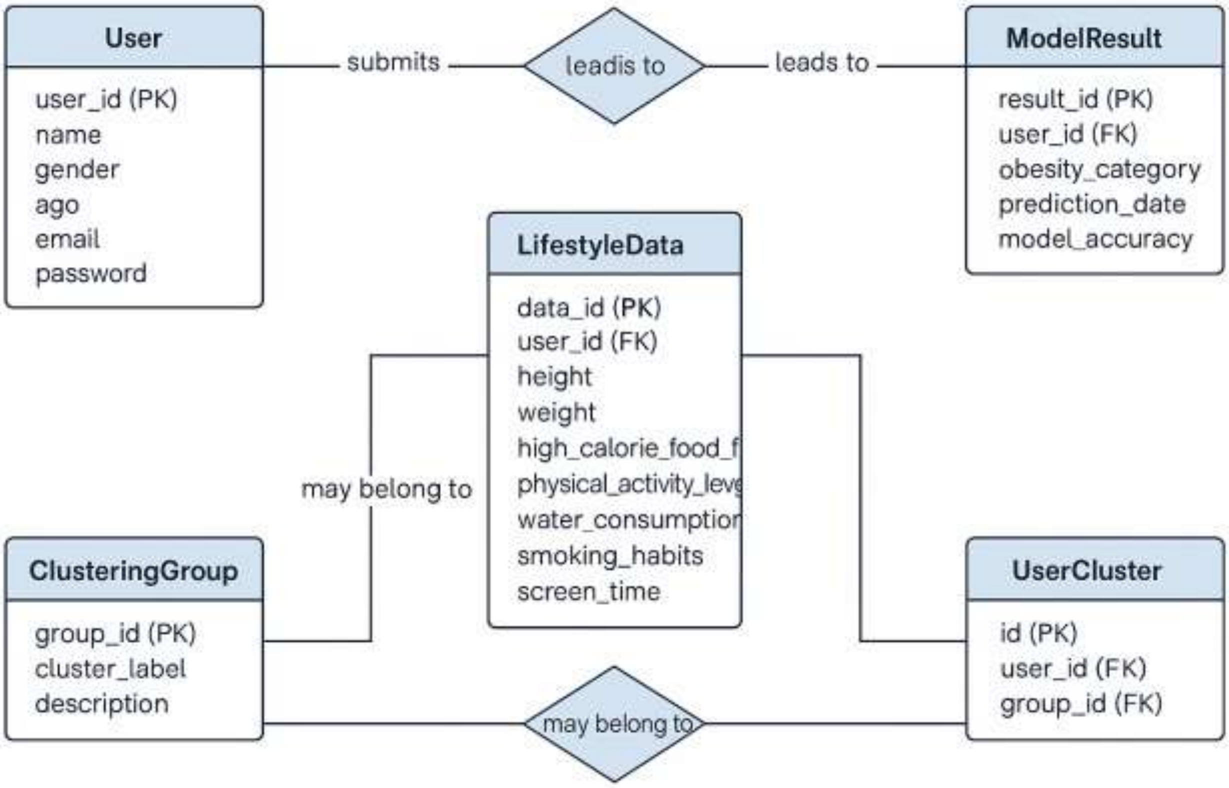


Use Case Diagram



Use Case Diagram

Entity- Relationship Diagram



4. Conclusion & Future Work

This study demonstrated that machine learning models can accurately estimate obesity levels using lifestyle data. Among classification techniques, Random Forest and Gradient Boosting provided the highest accuracy. In regression tasks, Random Forest Regressor gave the lowest RMSE and highest R^2. Clustering allowed for unsupervised discovery of behavioral patterns. Future work can focus on integrating time-series lifestyle data and applying deep learning models for enhanced performance and also work on Application and Website.

5. References

1. https://www.kaggle.com/datasets
2. https://scikit-learn.org/stable/
3. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7682147/
4. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight
5. https://towardsdatascience.com
6. https://scholar.google.com/
7. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning.
8. Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.
9. Obesity Classification Dataset – Kaggle
10. Journal of Obesity and Metabolic Syndrome