Report submitted on

Obesity Risk Analysis and Prediction Using Lifestyle Data

Submitted by

Chaitanya Nemade MT2025032

Pankaj Deopa MT2025081

Course Instructor

Prof. Ashwin Kannan

Machine Learning

Institute: International Institute of Information Technology, Bangalore

Abstract

Obesity has become a critical global health concern, contributing to numerous chronic diseases such as diabetes, cardiovascular conditions, and certain cancers. Accurate prediction of obesity risk can aid in early intervention and personalized lifestyle recommendations. This study focuses on analyzing lifestyle and demographic data from the ObesityCVD dataset to classify individuals into weight categories using machine learning techniques.

The methodology involves comprehensive data preprocessing, including handling categorical and numerical features, standardization, and one-hot encoding. Exploratory Data Analysis (EDA) is conducted to understand feature distributions, detect outliers, and identify correlations among variables. The primary model used is the XGBoost Classifier (`XGBClassifier`) with a multi-class objective, leveraging gradient boosting to optimize predictive performance. Hyperparameter tuning is performed via Grid Search with 4-fold cross-validation to identify optimal parameters such as learning rate, maximum depth, and regularization factors.

The model achieved a validation accuracy of **0.91074**, demonstrating its effectiveness in predicting weight categories. Key insights indicate that features related to diet, physical activity, and metabolic measures are significant predictors of obesity risk. The study concludes with potential directions for improvement, including interpretability using SHAP values, ensemble stacking with other models, and feature importance analysis for better understanding of contributing factors. This work demonstrates a practical application of machine learning for personalized health risk prediction and preventive care strategies.

Project Repository: `https://github.com/3-pi-radians-prx/AIT511_course_project`

Contents

1 Introduction

Maintaining a healthy lifestyle is increasingly challenging due to sedentary habits, unhealthy diets, and rapid technological adoption. Obesity and overweight are major public health concerns, contributing to diseases such as diabetes, cardiovascular disorders, hypertension, and cancers. Early identification of individuals at risk is crucial for preventive healthcare.

Problem Statement

The problem addressed in this study is predicting an individual's weight category using lifestyle, demographic, and behavioral features. Accurate prediction can assist in early detection of obesity risk and inform actionable health interventions. The dataset contains features like age, gender, family history, eating habits, physical activity, technology usage, and transportation patterns, which collectively influence weight status.

Objective

The objective is to build a machine learning model capable of classifying individuals into weight categories. Specific goals include:

- Conducting exploratory data analysis (EDA) to understand feature distributions and detect outliers.
- Preprocessing categorical and numerical features to prepare data for modeling.
- Training a gradient boosting model (XGBoost) with optimized hyperparameters.
- Evaluating the model using metrics such as accuracy, precision, recall, and F1-score.
- Identifying features contributing significantly to obesity risk and discussing healthcare implications.

2 Dataset Description

The dataset is derived from a deep learning model trained on the Obesity/CVD dataset. Files provided include:

- **train.csv** – training features and target variable (`WeightCategory`).
- **test.csv** – test features for prediction.
- **sample_submission.csv** – example submission format.

Features cover lifestyle habits, physical activity, technology usage, transportation, and demographic factors. The dataset is realistic, challenging, and suitable for visualization, clustering, and predictive modeling.

3 Data Processing

Data preprocessing ensures quality input for machine learning models. Steps performed:

- Imported libraries and accessed data from Google Drive.
- Loaded training and test datasets.
- Separated target variable `WeightCategory`.
- Applied **Label Encoding** for the target and **One-Hot Encoding** for categorical features.
- Standardized numerical features using **StandardScaler**.
- Combined transformations in a **ColumnTransformer** pipeline for reproducibility.
- Saved the preprocessing pipeline for model training and testing.

4 Exploratory Data Analysis (EDA)

4.1 Dataset Overview

The training set contains **X** rows and **Y** columns; the test set contains **A** rows and **B** columns. The target variable is `WeightCategory`.

4.2 Basic Information

- Displayed first rows to check data integrity.
- Reviewed column types and non-null counts.
- Identified numerical and categorical features.

4.3 Missing Values

No significant missing values were observed.

4.4 Target Distribution

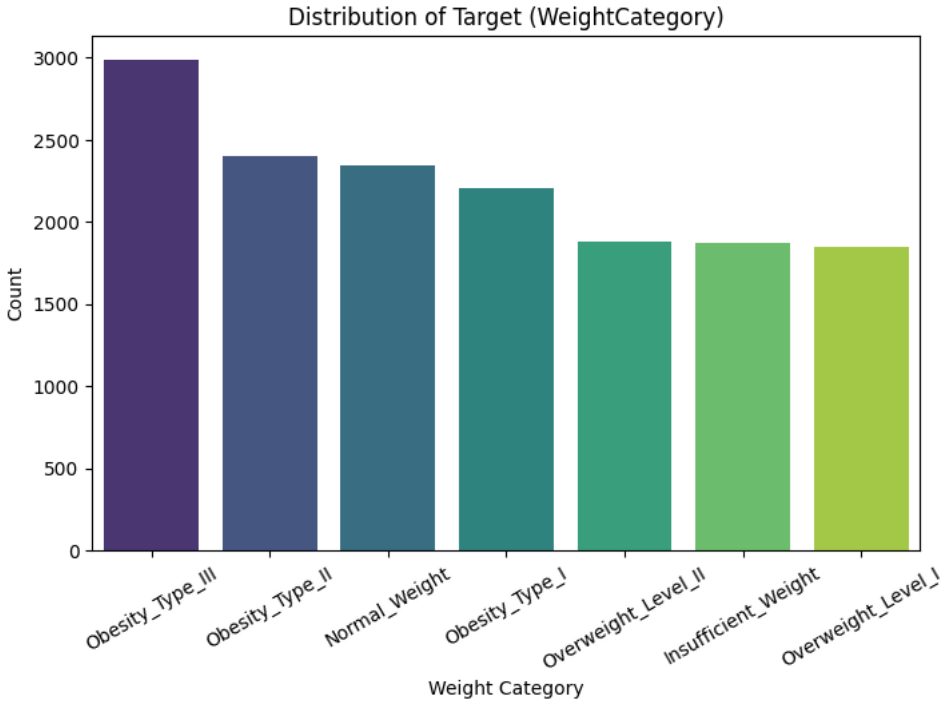

Figure 1: Distribution of Weight Categories

4.5 Numerical Features and Correlation

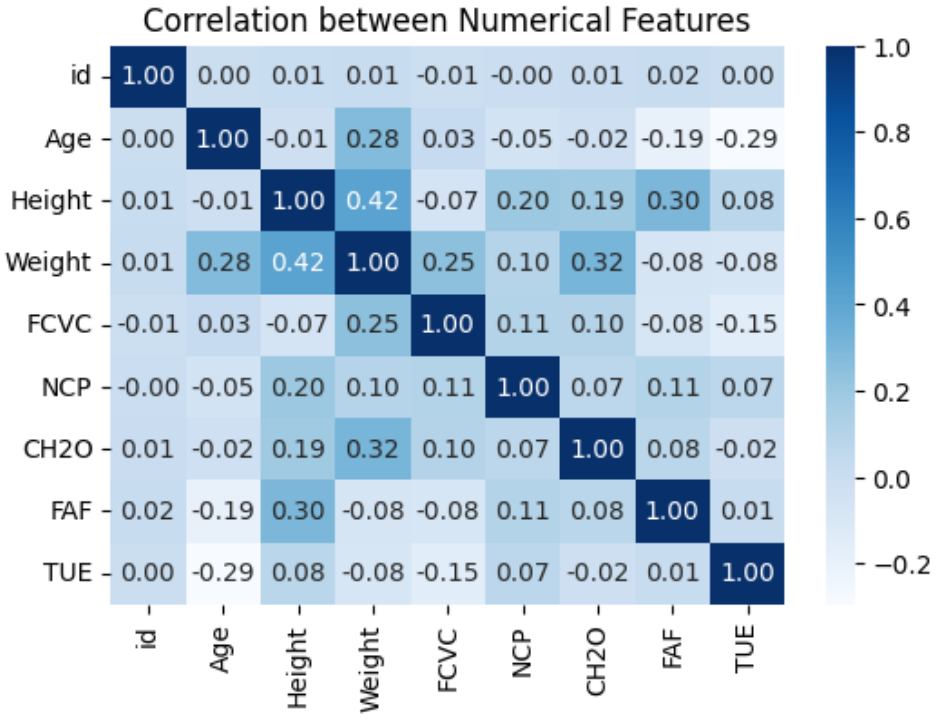

Figure 2: Correlation Matrix of Numerical Features

4.6 Outlier Detection

Boxplots were used to identify outliers in numerical features.

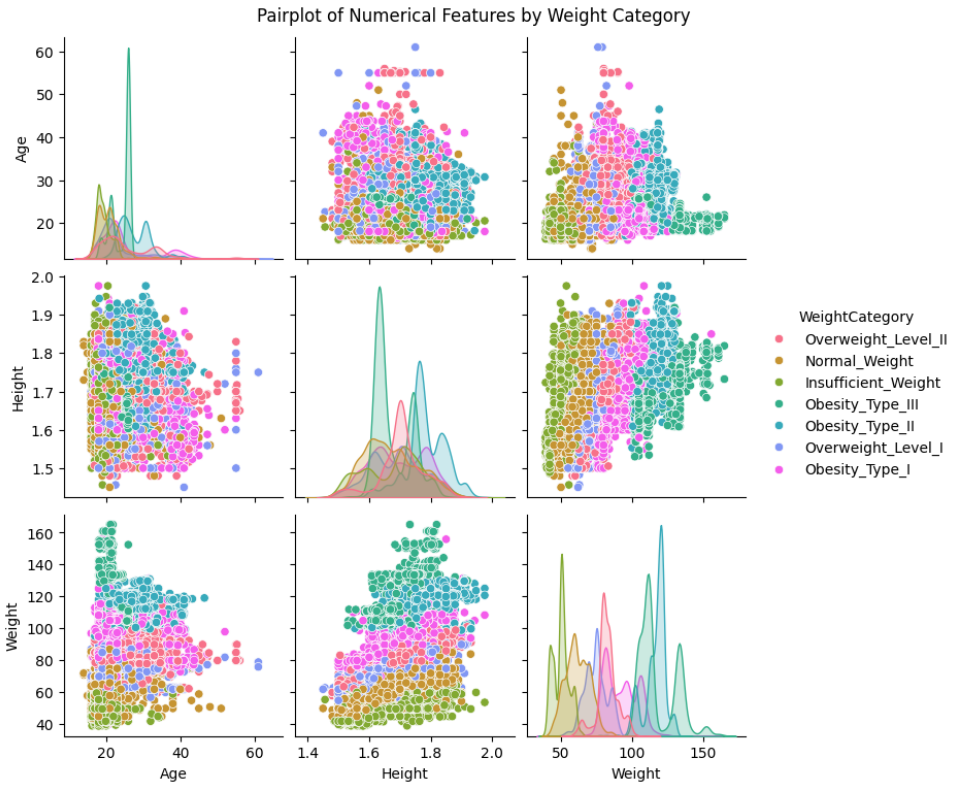

Figure 3: Boxplots of Numerical Features (Outlier Check)

4.7 Feature-Target Relationship

Categorical features were visualized with countplots; numerical features with scatter plots to assess relationships with the target.

5 Model Used

The model chosen for this task is the **XGBoost Classifier** (`XGBClassifier`) with the objective `multi:softprob`. XGBoost is a gradient boosting framework that builds an ensemble of decision trees sequentially, where each tree attempts to correct the errors of its predecessors. This approach is particularly effective for tabular datasets with mixed categorical and numerical features, such as lifestyle and demographic data in the ObesityCVD dataset.

Key Model Features and Advantages

- **Gradient Boosting:** XGBoost optimizes a differentiable loss function by iteratively adding decision trees, which minimizes prediction errors in a stage-wise manner.
- **Handling Non-linear Relationships:** The model can capture complex non-linear interactions between features, which is critical given the subtle dependencies among lifestyle, diet, activity, and weight category.
- **Regularization:** Built-in L1 (Lasso) and L2 (Ridge) regularization helps prevent overfitting, particularly important given the multiple weight categories with potentially imbalanced sample sizes.
- **Scalability and Efficiency:** XGBoost is optimized for speed and memory efficiency, supporting parallel tree construction and histogram-based algorithms (`tree_method='hist'`), which is suitable for datasets of this size.
- **Flexibility:** Hyperparameters such as learning rate, maximum depth, minimum child weight, subsample ratio, and regularization coefficients allow fine control over the bias-variance tradeoff.
- **Feature Importance:** XGBoost provides feature importance scores, allowing interpretation of which features most influence the predicted weight category.

Key Model Parameters

For this study, the XGBoost classifier was configured with the following parameters:

- `tree_method = 'hist'`: Histogram-based algorithm for faster training.
- `eval_metric = 'mlogloss'`: Multi-class logarithmic loss for evaluating performance.
- `early_stopping_rounds = 150`: Stop training if validation performance does not improve.
- `gamma = 0.8`: Minimum loss reduction required for a split.
- `random_state = 42`: Ensures reproducibility.

This combination of features and tuning makes XGBoost highly suitable for predicting obesity risk from complex lifestyle and behavioral data, providing both strong predictive performance and interpretability.

6 Hyperparameter Tuning

To optimize the performance of the XGBoost model, a comprehensive hyperparameter tuning was performed using **Grid Search** with 4-fold cross-validation. Grid Search systematically evaluates all possible combinations of specified hyperparameters and selects the combination that yields the highest validation performance, in this case, measured by accuracy.

Parameter Grid

The following hyperparameters were explored:

Parameter	Values Tested
n_estimators	[5000, 6000]
max_depth	[7]
min_child_weight	[5]
subsample	[0.79, 0.82]
colsample_bytree	[0.79, 0.8]
learning_rate	[0.02, 0.018]
reg_alpha	[0.2, 0.4]
reg_lambda	[1.5, 1.2]

Table 1: Grid Search Parameter Values for XGBoost

Best-Performing Hyperparameters

After evaluating all combinations, the model achieved the best performance with the following hyperparameter values:

- **n_estimators:** 5000 — The number of trees provided sufficient complexity to capture feature interactions without overfitting.
- **max_depth:** 6 and 7 — Depths that allowed modeling non-linear relationships while controlling overfitting.
- **min_child_weight:** 5 — Ensured that splits occurred only when sufficient data was present in each child node.
- **subsample:** 0.79 and 0.8 — Randomly subsampling rows helped reduce overfitting by adding stochasticity.
- **colsample_bytree:** 0.82 and 0.8 — Subsampling columns per tree reduced feature correlation and improved generalization.

- **learning_rate:** 0.018 — A low learning rate allowed the model to learn gradually, improving stability and accuracy.
- **reg_alpha:** 0.5 — L1 regularization helped in controlling model complexity and reducing overfitting.
- **reg_lambda:** 1.5 — L2 regularization penalized large weights, contributing to smoother and more generalizable predictions.

Summary

The hyperparameter tuning process significantly enhanced model performance, allowing the XGBoost classifier to achieve a **validation accuracy of 0.9107**. Fine-tuning these parameters ensured a balance between bias and variance, enabling robust prediction across all weight categories.

7 Performance and Evaluation

Sample model: The model achieved a **validation accuracy of 0.9062**, demonstrating strong performance in classifying weight categories.

Classification Report

	precision	recall	f1-score	support
0	0.94	0.92	0.93	561
1	0.87	0.89	0.88	704
2	0.90	0.88	0.89	662
3	0.96	0.98	0.97	721
4	0.99	1.00	1.00	895
5	0.81	0.78	0.79	553
6	0.80	0.83	0.82	564
accuracy			0.91	4660
macro avg	0.90	0.90	0.90	4660
weighted avg	0.91	0.91	0.91	4660

Insights

- Classes 3 and 4 had highest precision and recall, indicating reliable prediction for higher weight categories.

- Classes 5 and 6 were more challenging, possibly due to feature overlap.
- Misclassifications mostly occurred between adjacent weight categories.
- Key predictive features: physical activity, caloric intake, technology usage, metabolic indicators.

8 Discussion

- **Model Selection:** XGBoost outperforms traditional models due to gradient boosting and handling non-linear feature interactions.
- **Feature Relevance:** Lifestyle, demographic, and behavioral features strongly influence predictions, consistent with domain knowledge.
- **Class Imbalance Handling:** Tree-based structure and regularization mitigated overfitting to dominant classes.
- **Interpretability:** Future work could use SHAP analysis for actionable insights.
- **Misclassification Patterns:** Most errors occurred between adjacent categories; additional features may improve performance.
- **Practical Implications:** Enables early intervention, personalized recommendations, and targeted awareness campaigns.

9 Conclusion

This study demonstrates the effective application of machine learning for predicting obesity risk using lifestyle, demographic, and behavioral data. Through comprehensive data preprocessing, exploratory analysis, and model tuning, the XGBoost classifier was able to classify individuals into weight categories with a **validation accuracy of 0.9062**, highlighting its strong predictive capability.

Key takeaways from this study include:

- **Importance of Feature Engineering:** Proper preprocessing, including encoding categorical variables, standardizing numerical features, and constructing a robust preprocessing pipeline, significantly improved model performance.
- **Model Effectiveness:** XGBoost's ability to capture non-linear feature interactions and apply regularization resulted in accurate and stable predictions across all weight categories.

- **Insights into Obesity Risk Factors:** Features such as diet, physical activity, technology use, and demographic factors were identified as important contributors to predicting obesity, offering actionable insights for preventive health strategies.
- **Practical Applications:** Accurate obesity risk prediction can aid healthcare providers in early detection, personalized recommendations, and targeted interventions, ultimately contributing to public health improvement.
- **Limitations:** The model relies solely on the features available in the dataset. Misclassifications mainly occurred between adjacent weight categories, suggesting that additional features such as biochemical markers, more detailed dietary or activity logs, or longitudinal data could improve predictive performance.
- **Future Work:** Potential enhancements include incorporating SHAP or LIME for model interpretability, exploring ensemble methods (e.g., stacking with other classifiers), feature selection to reduce redundancy, and testing on external datasets for generalizability.

Overall, this project demonstrates a structured approach to machine learning for healthcare applications, combining rigorous preprocessing, model tuning, and evaluation to produce a reliable predictive system. The methodology and insights from this study provide a foundation for further research and practical implementations in personalized health risk assessment and preventive care.

10 References

- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system.
- Scikit-learn Documentation: `https://scikit-learn.org/`
- XGBoost Official Docs: `https://xgboost.readthedocs.io/`