UNIVERSITY OF WOLVERHAMPTON

HERALD COLLEGE
KATHMANDU

CLASSIFICATION ANALYSIS REPORT

Student ID: 2417739

Student Name: Preeyanshu Singh

Section: L5CG20

Module Leader: Siman Giri

Tutor: Bibek Kanal

Abstract:

Introduction

In this coursework I have implemented classification model to predict obesity level of a person based on his/her various lifestyle, health conditions. This classification model analyzes the features data and predicts the level obesity in body. This classification project aligns with United Nations Sustainable Development Goal “UNSDG”. Gole 3 “Good Health and Well-being “In which we identify some of the major key factor that cause or may cause obesity.

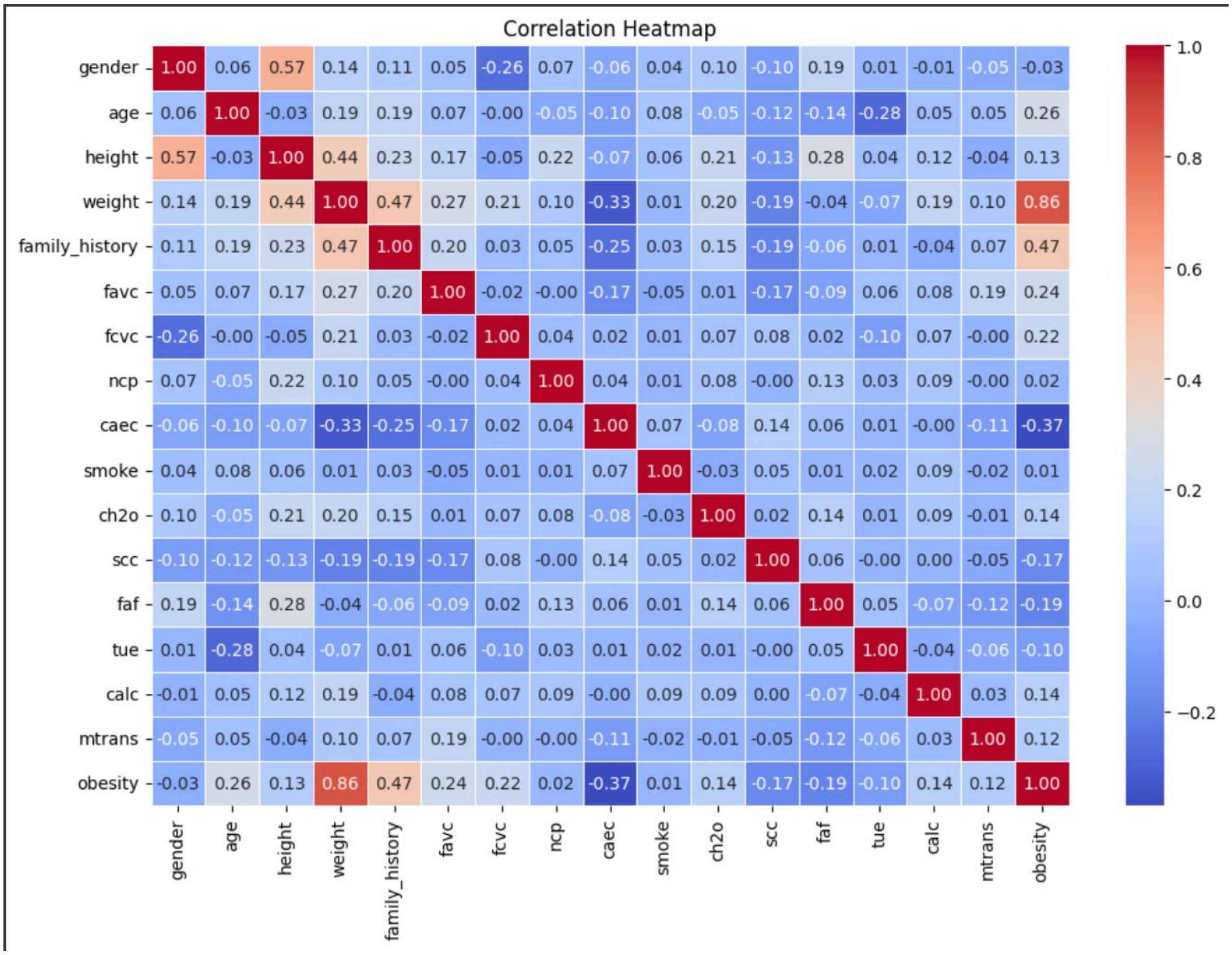
In this classification model I have used Obesity Prediction Dataset which is CVS format dataset and it was sourced from Kaggle it has many features which contains health and lifestyle attributes in which many of them are categorical data and some are numerical data like age, weight, height, fcvc, faf, ncp and target variable obesity level. The goal of this task is to build a prediction classification model that predicts the target variable which is obesity level with the help of features in the dataset.

Methodology

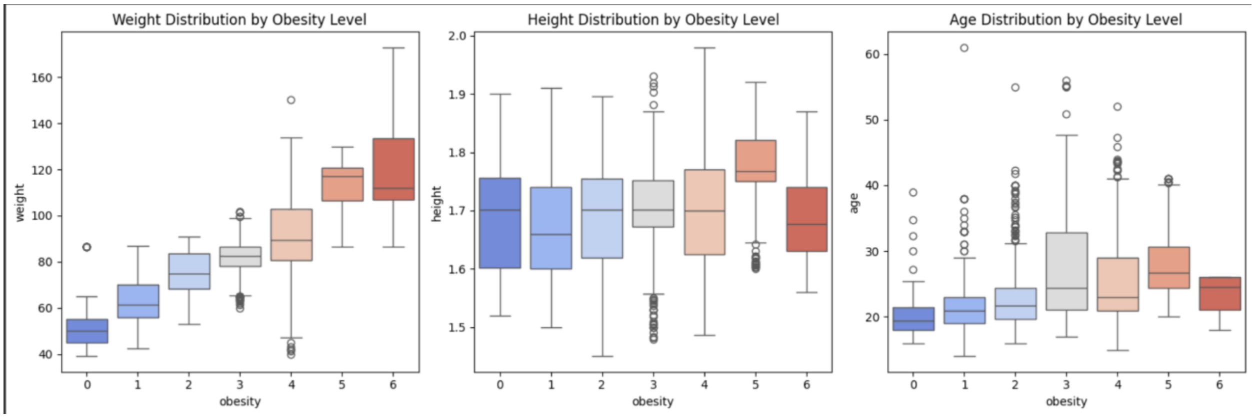
Data Preprocessing & Exploratory Data Analysis (EDA):

As for prediction propose we needed data could give us as accurate result as possible so we need to handle some missing values we also had to encode some of the categorical variables into numerical format and duplicates were removed.

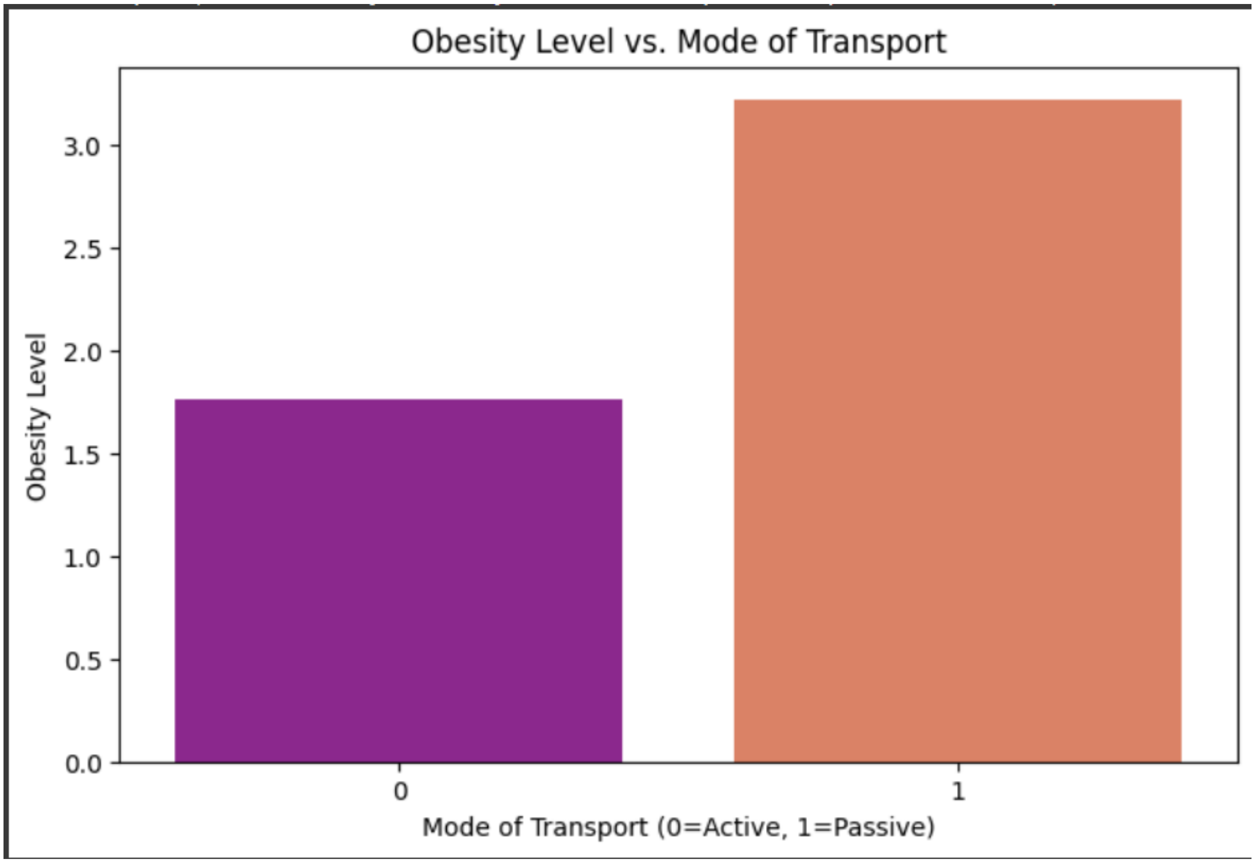
We also used many visualization plots to understand the relationship of features in dataset. we used heatmap correlation matrix to identify highly correlated features:



boxplots were used to analyze how age, weight, and height affect obesity:



while bar plots were utilized to understand the impact of lifestyle factors such as eating habits and transportation.



Model Building:

In total we implemented three model as asked by the coursework among which first model was Logistic Regression implemented from scratch, using the Sigmoid function and Binary Cross-Entropy Loss, with gradient descent for weight updates. And in the second models I used Scikit-Learn libraries for comparison purposes of a standard Logistic Regression model and third model is a Random Forest Classifier, which is a tree-based ensemble model designed to capture non-linear relationships.

Hyperparameter Tuning & Feature Selection

After implementing models, we optimize the models by performing hyperparameter tuning. Logistic Regression model was tuned using GridSearchCV by selecting the best values for parameters like C and Solver and the optimal parameter found where 'C': 1, 'penalty': 'l1', 'solver': 'liblinear' and Random Forest Classifier was tuned using RandomizedSearchCV determining the best parameters for n_estimators, max_depth, and others and optimal parameters found for this were 'n_estimators': 100, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'log2', 'max_depth': 20

Feature selection was carried out using SelectKBest with ANOVA F-score for Logistic Regression, identifying the top 5 features. For the Random Forest Classifier, feature importance was assessed, and the top 5 features were selected.

Model Evaluation

Assessment of the models are done based on some metrics like precision, accuracy, recall, and others .Accuracy is the part of accurate predictions in relation to total predictions made. Precision is the ratio of true positives to predicted positives. Recall measures the percentage of positive cases that are correctly identified. F1 is the harmonic mean of precision and recall. All these metrics were used as they are mostly adopted ones in evaluating a classification model, particularly when the classes are imbalanced.

Further, the pre-tuning and post-tuning accuracies of Logistic Regression were recorded as well as the accuracy after feature selection. In the same fashion, a similar exercise was done for the Random Forest Classifier where accuracies were found out pre- and post-hyperparameter tuning and feature selection. A comparison analysis was done to present the best model and its accuracy. Moreover, the effect of feature selection on model performance was also analyzed to get insights.

The evaluation of the models was made comprehensive by using confusion matrices for various false positives and false negatives.

Conclusion

Key insights from the project include the identification of the best model for obesity prediction, which was Logistic Regression with an accuracy of 0. 904.The analysis revealed whether feature selection improved the model's performance and highlighted the most important features influencing obesity

In conclusion, this coursework successfully implemented classification models for obesity prediction, enhancing accuracy through feature selection and hyperparameter tuning. Further improvements can be pursued by exploring advanced models and increasing dataset diversity.

Discussion

Model Performance:

The results suggest that the **Random Forest Classifier** performed better compared to Logistic Regression. The Random Forest model captured complex patterns and non-linear relationships better, leading to higher classification accuracy.

Impact of Hyperparameter Tuning & Feature Selection:

Hyperparameter tuning, feature selection improved model performance. The parameters which were optimized helped enhance predictive accuracy, and selecting the most relevant features reduced computation time while maintaining model effectiveness.

Interpretation of Results:

The study identified key predictors of obesity, including **dietary habits, physical activity levels, and body measurements**. These insights can help develop targeted strategies for obesity prevention and public health interventions.

Limitations:

Despite its effectiveness, the model has certain limitations:

- **Class imbalance issues** – Some obesity categories may have fewer samples, affecting model predictions.

- **Potential biases in data collection** – The dataset may not be fully representative of the general population.
- **Feature assumptions** – Some relationships between features may not be fully captured by the models.

Suggestions for Future Research:

Future research could explore:

- Improving classification accuracy.
- Handling class imbalance using weighted loss functions.
- Expanding the dataset to more features such as diverse populations and other health indicators.