BRAC UNIVERSITY

Inspiring Excellence

Name	ID	Section
Md. Azmain Adib	24141112	07
Mohammed Al Wasee	22301199	02

Table of Contents

Submitted to

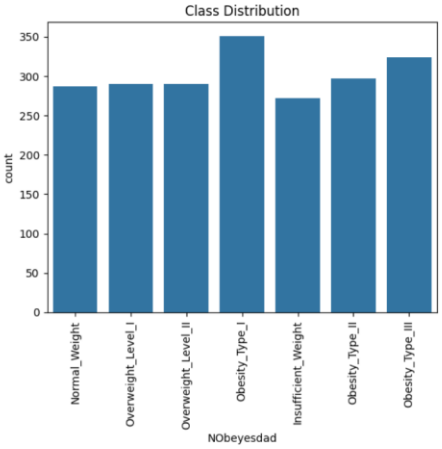
Nafiz Imtiaz Rafin, Nafiz Siddiqui Adnan

Lecturer

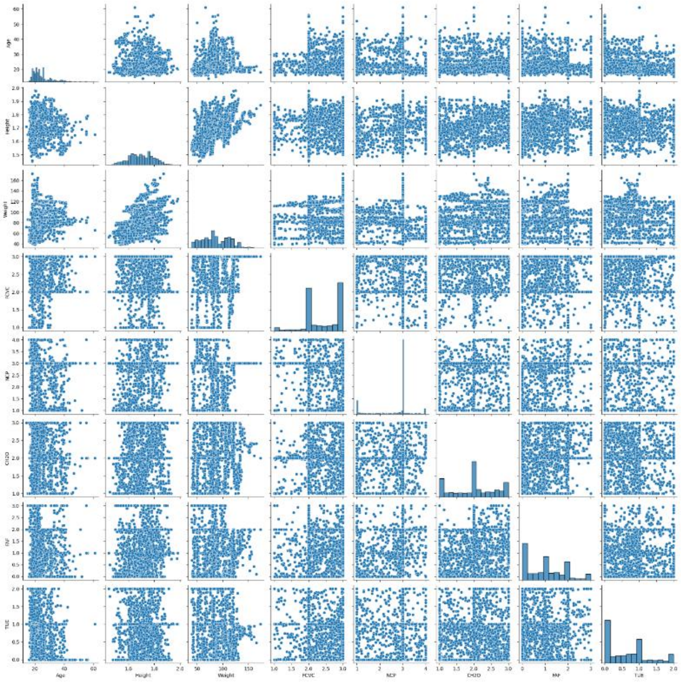
BRAC University

1. Introduction: One of the major effects on physical and mental health is obesity which is a serious growing public concern. Our project is concerned about the level of obesity which we took into account by various amount of information like age, gender, weight height and many more. The goal of our project is to Annalyse the person and find out the risk of being obese which might also help the healthcare system. Our motivation on this project is to survey the risk of being obese through the machine learning approach and eventually the doctors can take steps to prevent this health-related issues before the obesity takes into effect.

2. Dataset Description: This dataset is a classification problem as the targeted variable "NObeyesdad" has discreate classes. The dataset contains up to 16 features and about 2111 data records from there few of them are categorical features like Gender, family history, FAVC, CAEC, SMOKE and few are qualitative features like Age, height, weight, FCVC, NCP. There is also an imbalance in the distribution of classes.



Correlation:



3. Dataset Pre-processing: It's important to inspect the dataset before training machine learning models to avoid some potential issues. In the dataset there might be some null values. Now, from the dataset we have to separate the features and target value or label. To do this we analyzed the dataset and selected a label. Then we selected the features in x variable and similarly label in the y variable. We dropped the label column in the feature dataset. By running this command df.info(), we saw that there are no null values in our dataset. So, we don't have to drop any columns or impute any value. From this same command we also noticed that there are some Dtype = object values. So, there are some categorical features in our dataset. Machine learning algorithms, such as logistic regression and KNN, require numerical input. Therefore, categorical variables must be transformed into numerical formats to ensure compatibility with the models. To solve this problem, we used One-Hot Encoding to convert categorical features into numeric format. One-hot encoding creates new binary columns for each category in a categorical feature. The target variable (y) is also categorical and needs to be encoded into numeric format for multiclass classification. We used Label Encoding, which assigns a unique integer to each class in the target variable.

4. Feature Scaling: We applied feature scaling to ensure that all features have a comparable scale, which is essential for distance-based models like KNN and gradient-based methods like Logistic Regression. We used Standardization to scale the features. This method transforms the data to have a mean of 0 and a standard deviation of 1.

5. Dataset Splitting: The dataset was split into training and testing sets. We used an 70-30 split, where 70% of the data was used for training and 30% for testing. This was implemented using the train_test_split function from sklearn. We also used random start and stratify parameters to improve the reliability of the model evaluation.

6. Model Training and Testing: Three models were trained and tested: Logistic Regression, K-Nearest Neighbors (KNN), and Decision Tree. We implemented these models to compare the performance of these models on a multiclass classification problem and determine the best-performing one.

- **K-Nearest Neighbors (KNN):** A distance-based algorithm where predictions are made based on the closest k neighbors in the feature space.
- **Decision Tree:** A tree-based algorithm that splits data based on feature values to make predictions.
- **Logistic Regression:** A linear model for classification problems.

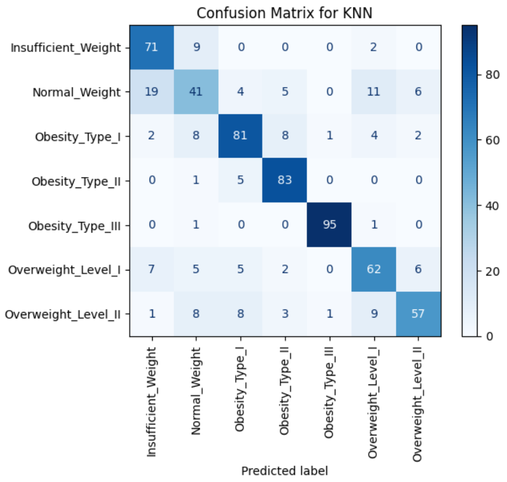
Each model was trained on the training set (X_train, y_train) and tested on the test set (X_test, y_test).

7. Comparison:

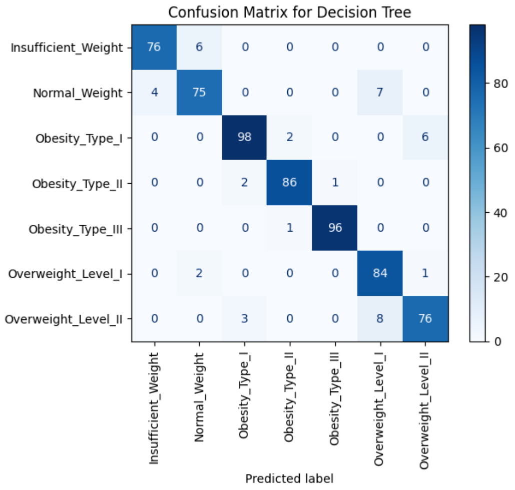
Model Name	Accuracy	Precision	Recall
KNN	77.29%	0.77	0.77
Decision Tree	93.22%	0.93	0.93
Logistic Regression	86.12%	0.86	0.86

Confusion Matrix:

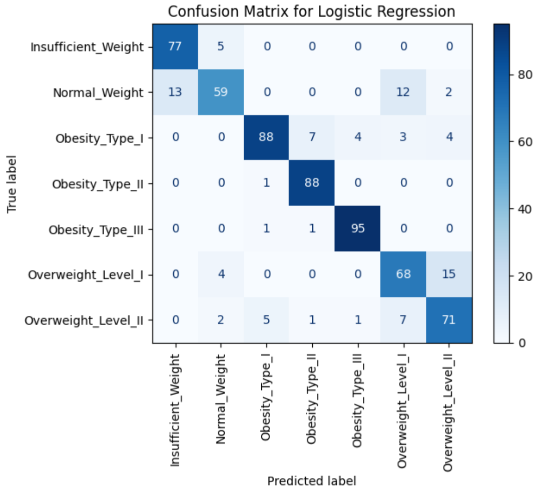
KNN:



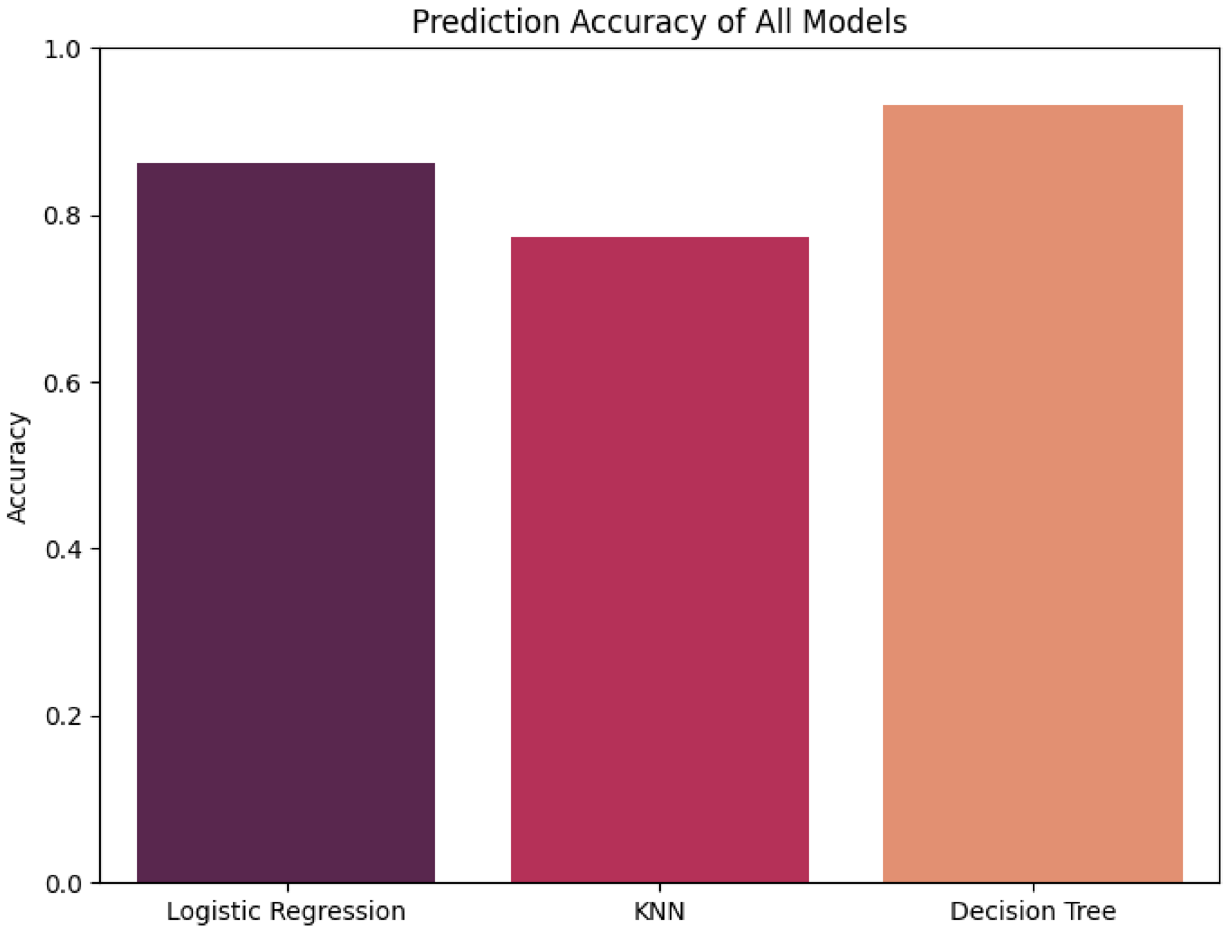
Decision Tree:



Logistic Regression:



Accuracy:



8. Conclusion: After evaluating three different machine learning models — K-Nearest Neighbors (KNN), Decision Tree, and Logistic Regression — on the multiclass classification problem, the following accuracies were obtained: K-Nearest Neighbors (KNN): 77.29%, Decision Tree: 93.22% and Logistic Regression: 86.12%.

The Decision Tree was the best model for this classification task, with the highest accuracy. However, the results might change with other datasets, and fine-tuning the settings could improve all the models.

Reference:

Estimation of Obesity Levels Based On Eating Habits and Physical Condition [Dataset]. (2019). UCI Machine Learning Repository. https://doi.org/10.24432/C5H31Z.