**MACHINE LEARNING AND PATTERN RECOGNITION(B9DA109)**

**KARTHIK LAKKIMSETTI**

REPORT:

**“Estimation of Obesity Levels Based on Eating Habits and Physical Condition" dataset from the UC Irvine Machine Learning Repository”.**

**DATA COLLECTION:**

For the data collection phase, three group members including myself researched and compared various potential datasets. After a thorough evaluation, we finalized the "Estimation of Obesity Levels Based on Eating Habits and Physical Condition" dataset from the UC Irvine Machine Learning Repository. This dataset contained 17 relevant variables related to factors like demographics, lifestyle, diet, and physical activity that could impact obesity levels. In this stage, I searched through different sources to find datasets, checked if they fit our project needs, and led discussions to pick the right one.

**DATA PREPROCESSING:**

In the data preprocessing stage, I carefully examined the dataset to fix problems like outliers and missing data. I plotted boxplots for numerical variables such as age and number of daily meals to identify outliers. Upon careful examination, I determined that these outliers represented valid data points from the population distribution and should be retained. Additionally, I explored bivariate relationships between variables and the target through visualizations like count plots. Notable findings included the link between high-calorie food consumption and obesity risk, greater obesity levels among public transport users, and the influence of family history on obesity.

**EXPLORATORY DATA ANALYSIS:**

**BIVARIATE ANALYSIS:**

In the exploratory phase, I conducted a bivariate analysis to examine relationships between variables and obesity risk. Through visualizations, I identified key factors like gender, high-calorie food intake, and lack of physical activity as potential predictors of obesity levels. These insights guided the feature selection process by highlighting the most relevant variables to include for accurate obesity risk prediction modelling.

**MODEL SELECTION:**

**GRADIENT BOOSTING:**

My contribution involved explaining the concept of gradient boosting, emphasizing its underlying mechanism of building decision trees on weak learners. I clarified how gradient boosting constructs decision trees based on weak learners, which are sequentially trained to minimize errors. I highlighted key techniques employed in gradient boosting, such as controlling tree depth, setting minimum samples per leaf, and implementing regularization methods.

**DATA MODELLING:**

**GRADIENT BOOSTING:**

My contribution to the Gradient Boosting phase involved running the gradient boost algorithm with our selected features and analyzing the results. I observed an accuracy score of 90.41%, indicating the effectiveness of our model in predicting obesity levels. I examined the distribution of predictions across different obesity classes and noted a high F1 score of 1 for obesity type III, suggesting precise classification. Furthermore, I conducted a thorough examination of the confusion matrix, identifying areas of improvement such as reducing false positives.

My active involvement throughout the project enabled me to gain invaluable insights across various stages. From selecting datasets to preprocessing data and conducting exploratory analysis, I gained skills in handling missing values, outliers, and categorical variables, and employing advanced modelling techniques like Gradient Boosting. This involvement facilitated a deeper understanding of data analysis and modelling, contributing significantly to the project's outcomes and my self-learning as well.