Academic Year	Module	Project	Project Type
2	Concepts and Technologies of AI	Final	Individual Report

Predictive Analysis of Chronic Medical Conditions in Depressed Individuals Using Depression Dataset

Student Id	: 2408947
Student Name	: Rujan Maharjan
Section	: L5CG16
Module Leader	: Siman Giri
Tutor	: Ronit Shrestha
Submitted on	: 11/02/2025

Abstract:

Aim: This report focuses on predicting chronic medical conditions in individuals with depression by applying classification techniques.

Approach: The analysis utilizes the Depression dataset, which includes various attributes related to personal and lifestyle factors. The methodology involves Exploratory Data Analysis, building classification models with hyper-parameter optimization, and selecting the most significant features.

Results: The performance of the models was compared using accuracy and classification repport.

Introduction:

This project aims to predict chronic medical conditions in individuals with depression using the Depression Dataset, sourced from Kaggle (https://www.kaggle.com/datasets/anthonytherrien/depression-dataset). The dataset includes information on personal and lifestyle factors and aligns with UNSDG 3: Good Health and Well-Being. By providing data on various health, lifestyle, and socio-economic aspects, it supports deeper insights into mental illnesses such as depression and the development of potential preventive measures (https://sdgs.un.org/goals/goal3). The goal of this analysis is to create a predictive model that estimates chronic medical conditions in individuals with depression based on the features in the dataset.

Methodology:

Before model development, the dataset was checked for missing values using methods like isnull().sum() and info(), revealing no missing entries. As the dataset was predominantly categorical with only three numerical columns, outliers were not a concern. However, numerical data was scaled using StandardScaler to facilitate better correlation mapping. Data visualization included count plots to represent various features and a histogram for the income column. Correlation matrices were created after encoding categorical data through one-hot and label encoding. Additionally, irrelevant columns such as Name, Age Group, and Income Level were dropped.

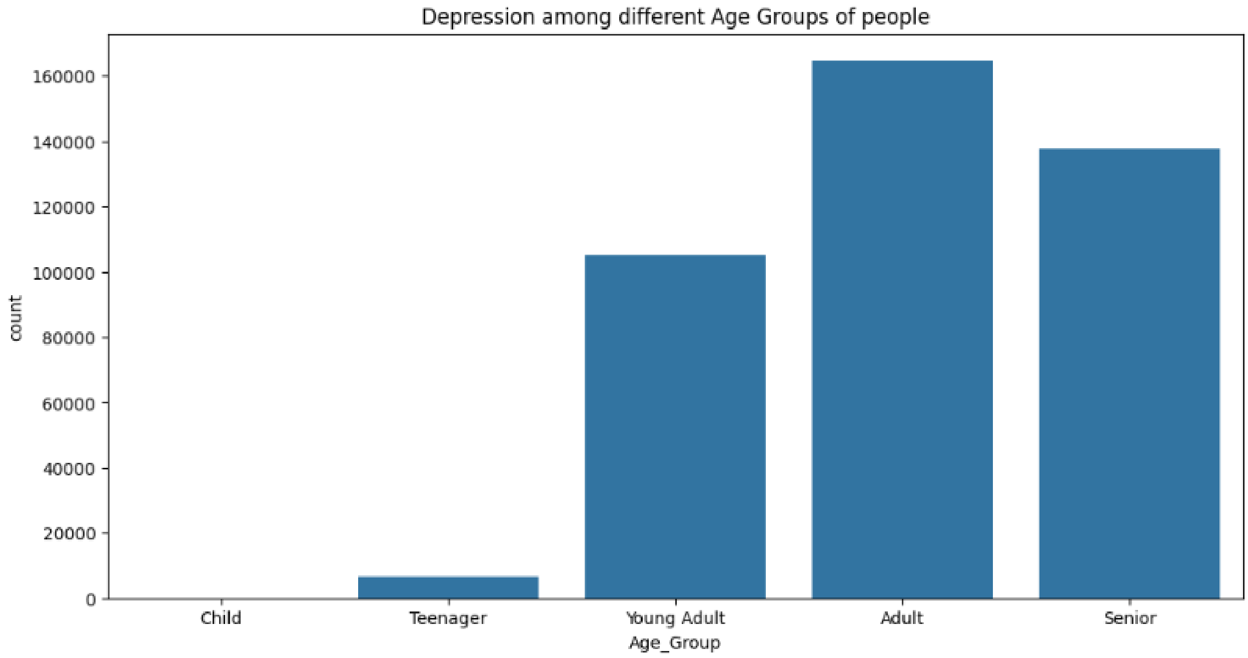

Figure-1: Count plot of depression among different Age groups

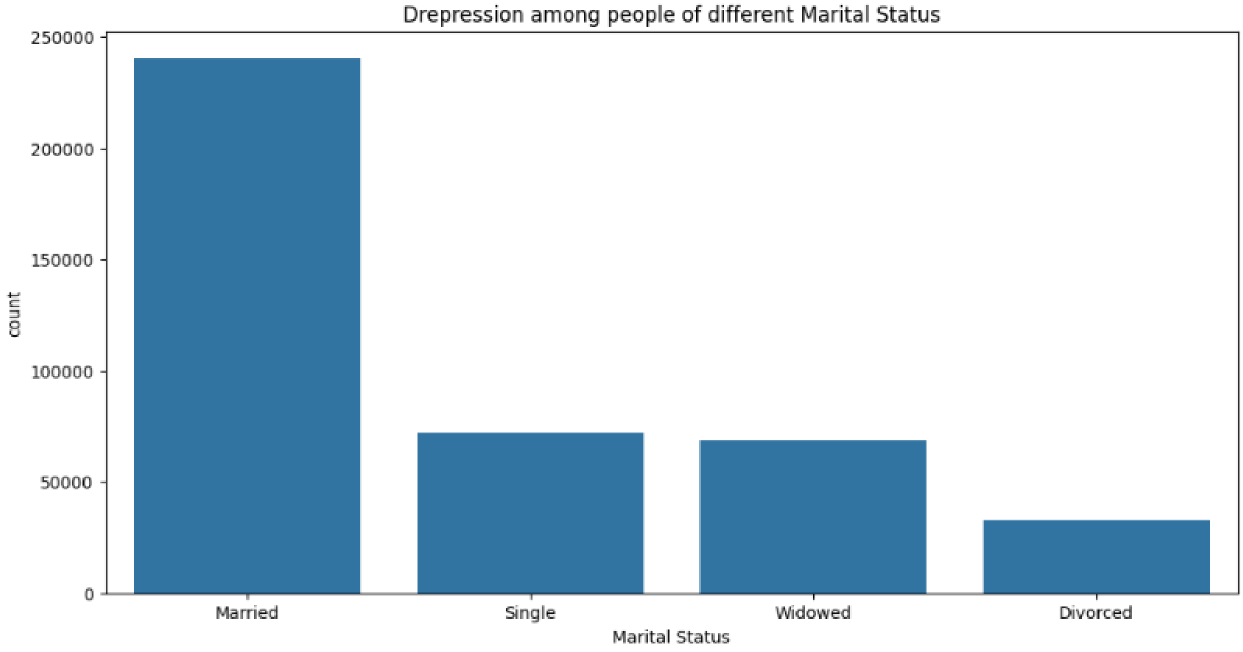

Figure-2: Count plot of Depression among different Marital status

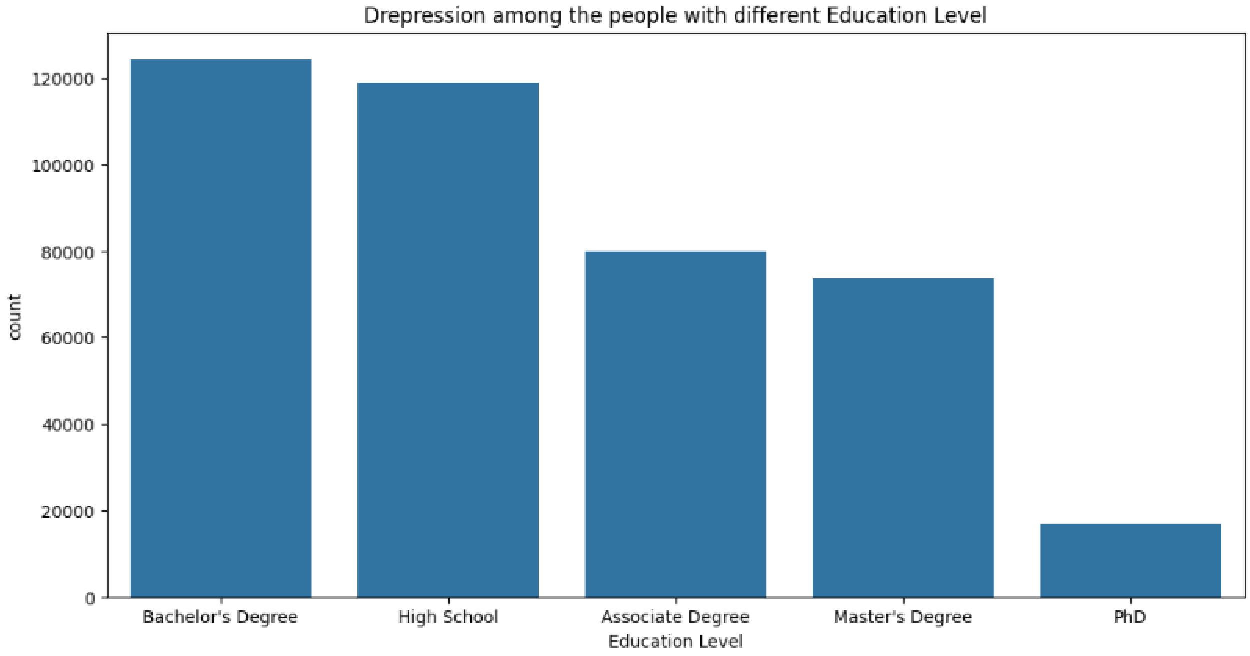

Figure-3: Count plot of Depression among different Education Level

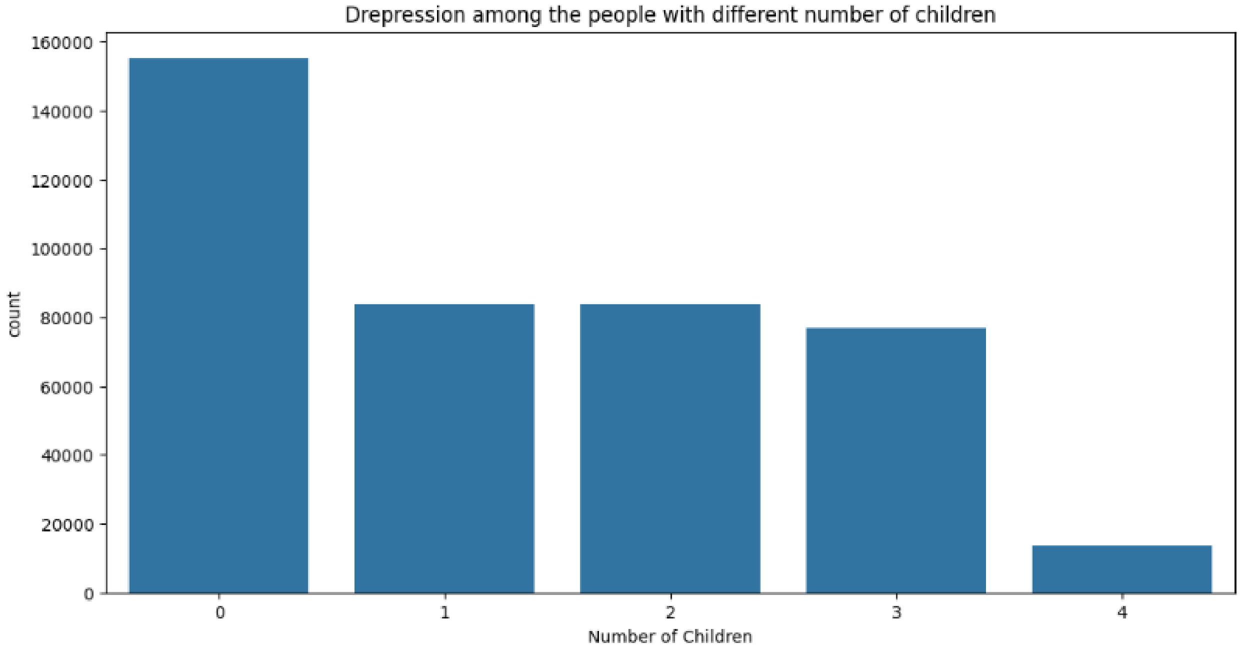

Figure-4: Count plot of Depression among different Number of children

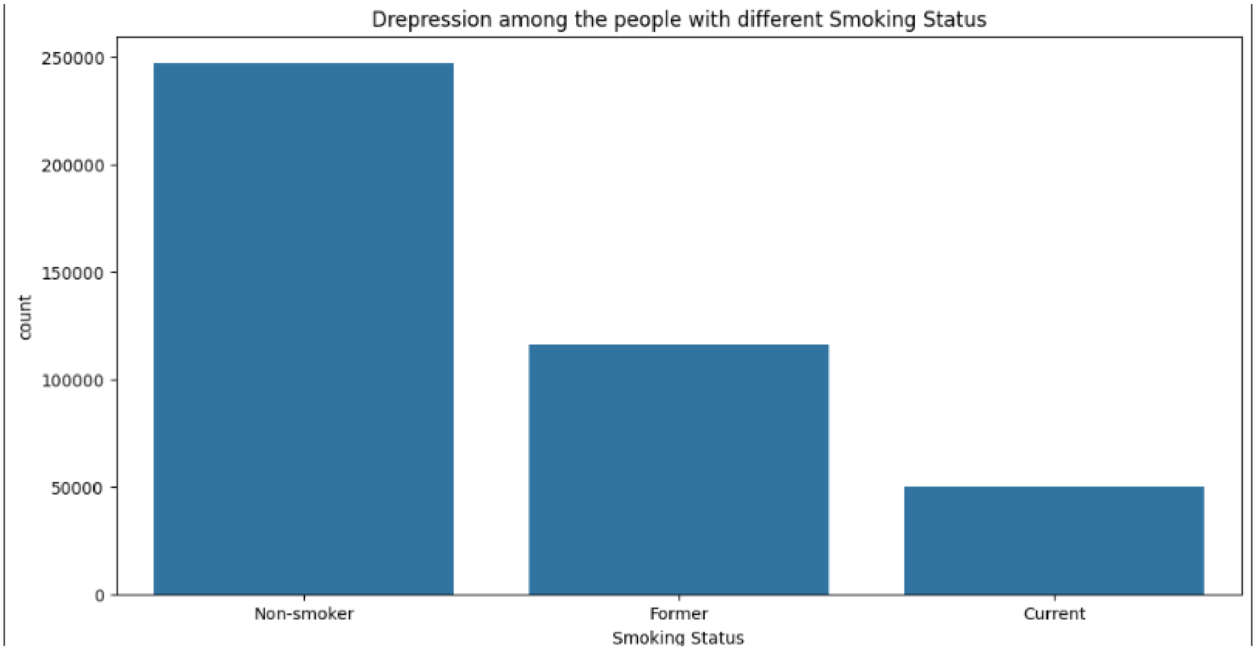

Figure-5: Count plot of Depression among different Smoking status

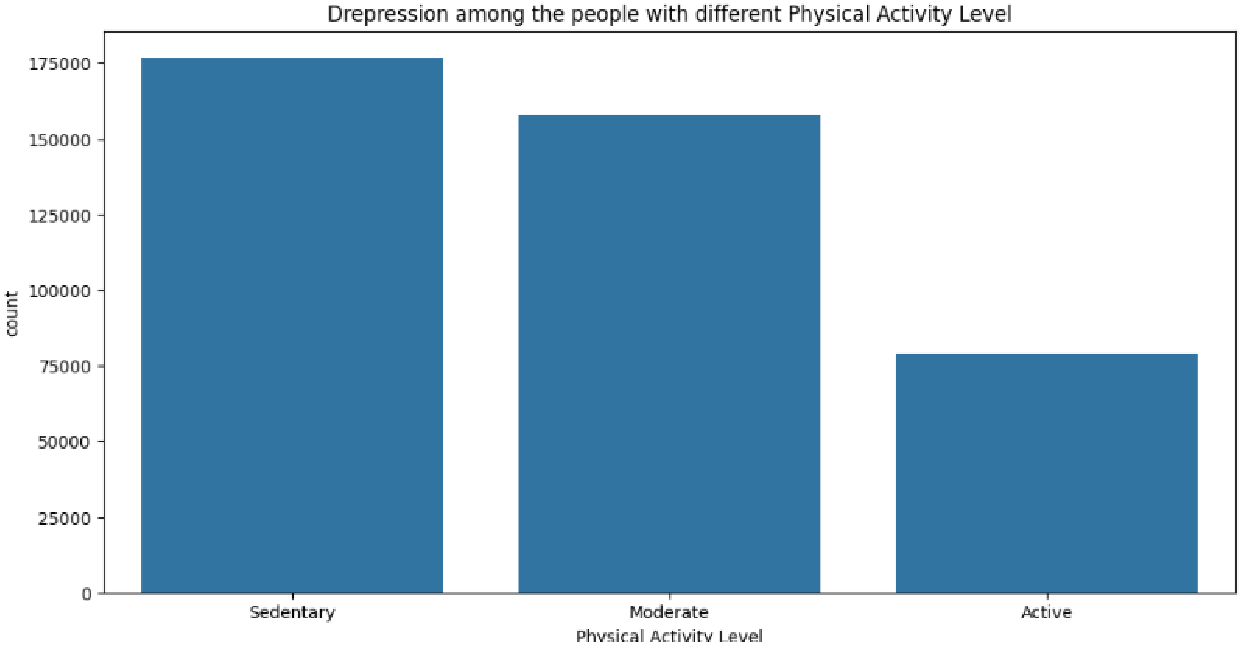

Figure-6: Count plot of Depression among different Physical Activity Level

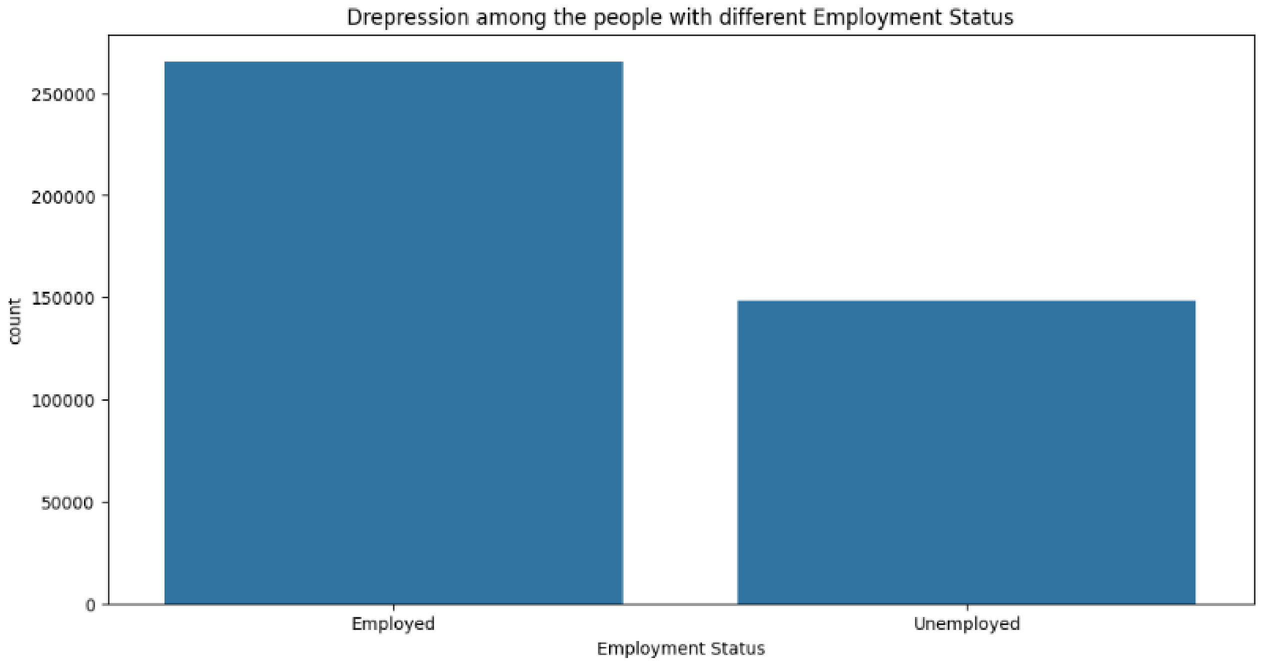

Figure-7: Count plot of Depression among different Employment status

\

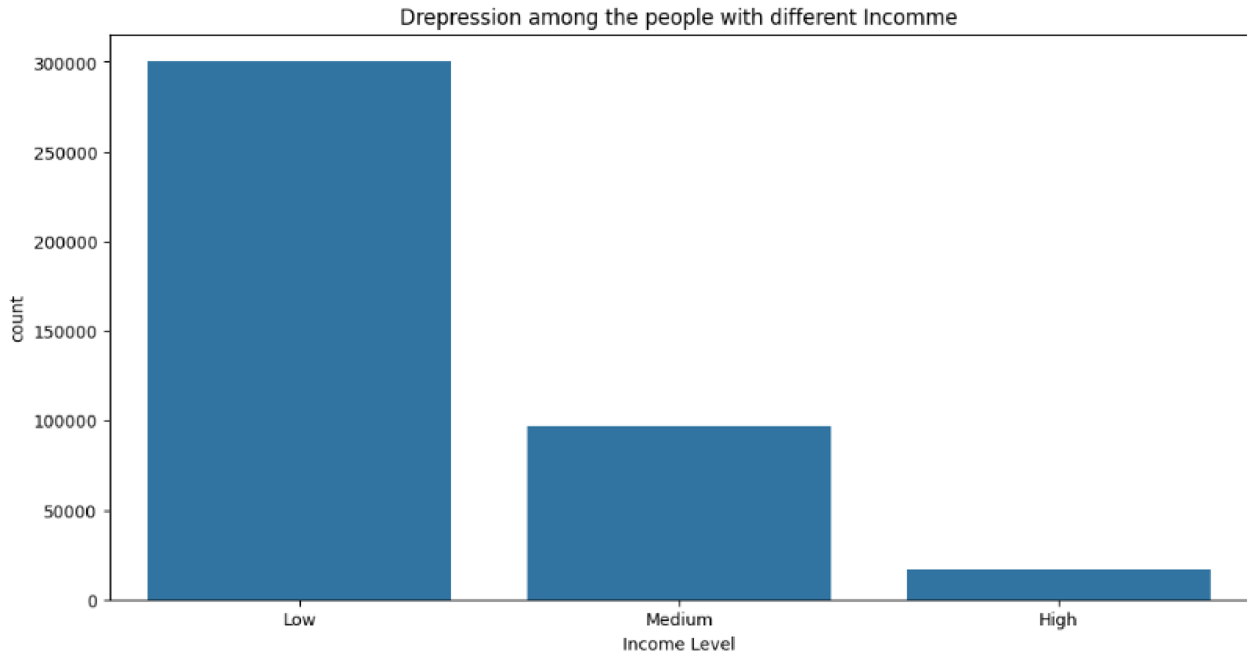

Figure-8: Count plot of Depression among different Income Level

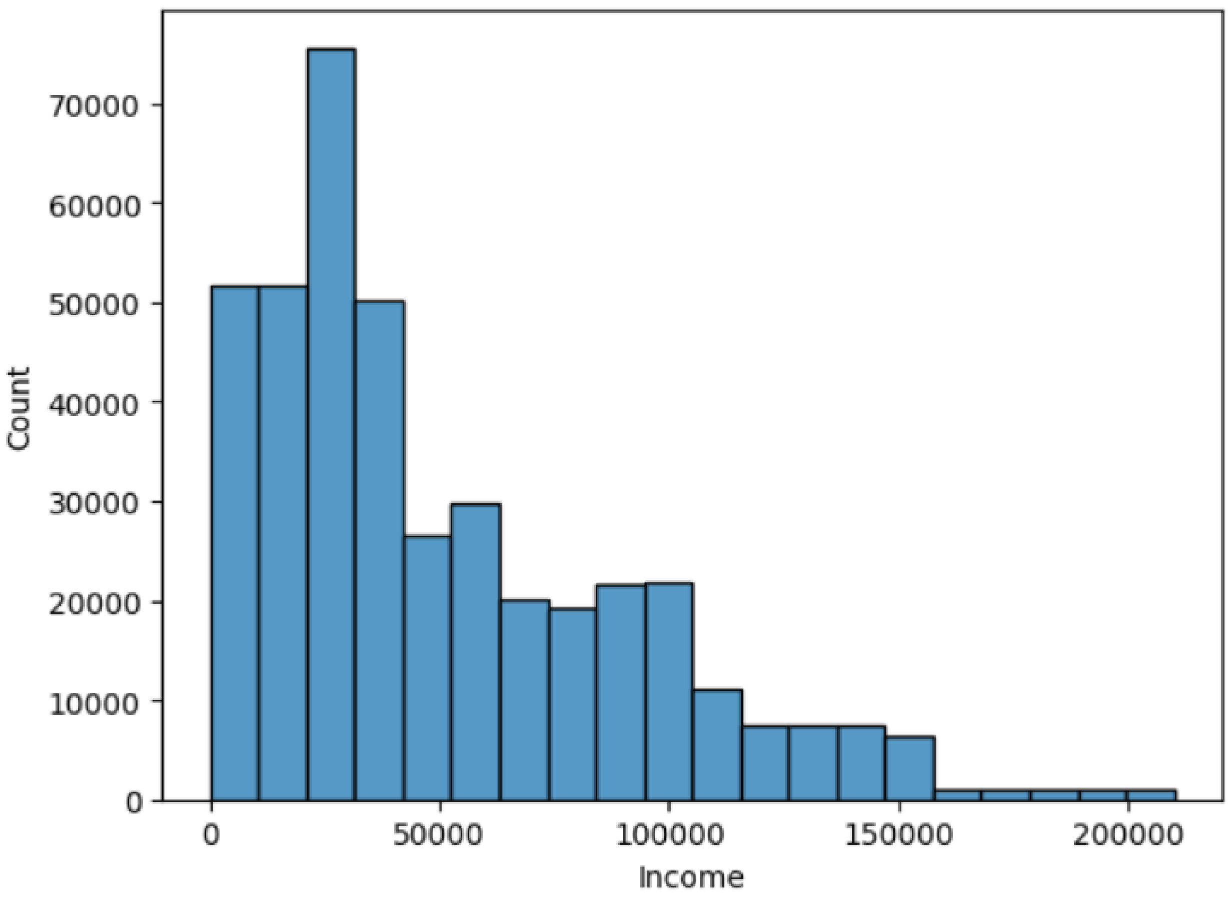

Figure-9: Histogram of Income in the dataset

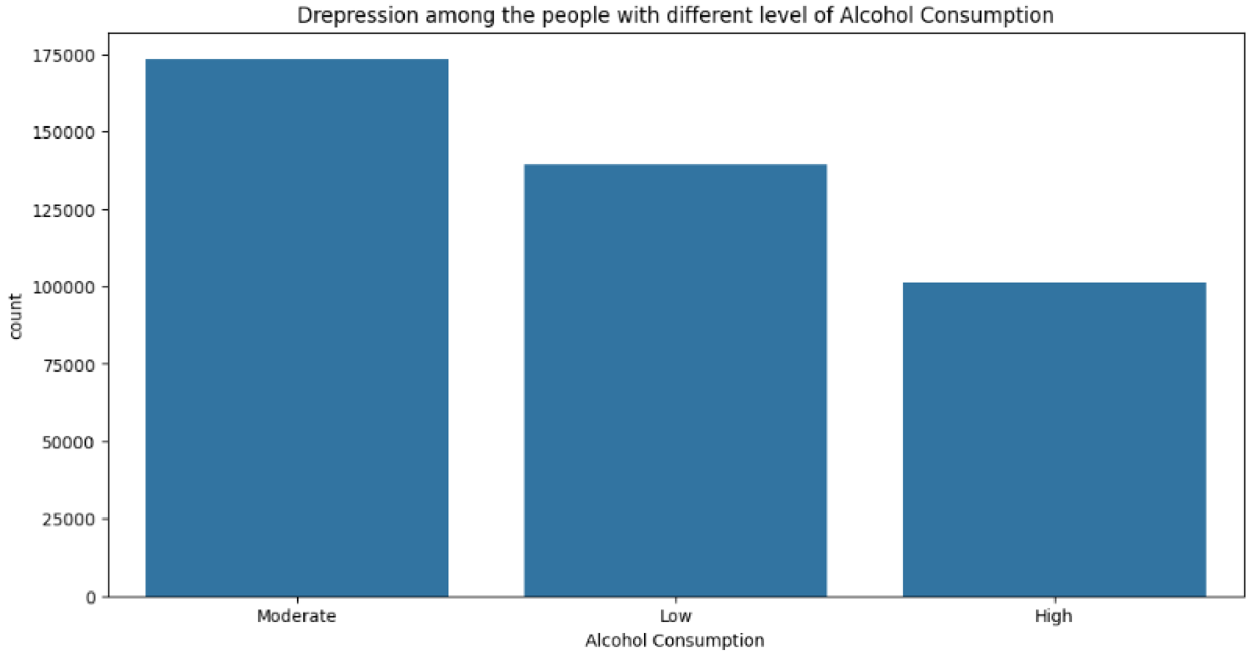

Figure-10 : Count plot of Depression among different Alcohol Consumption

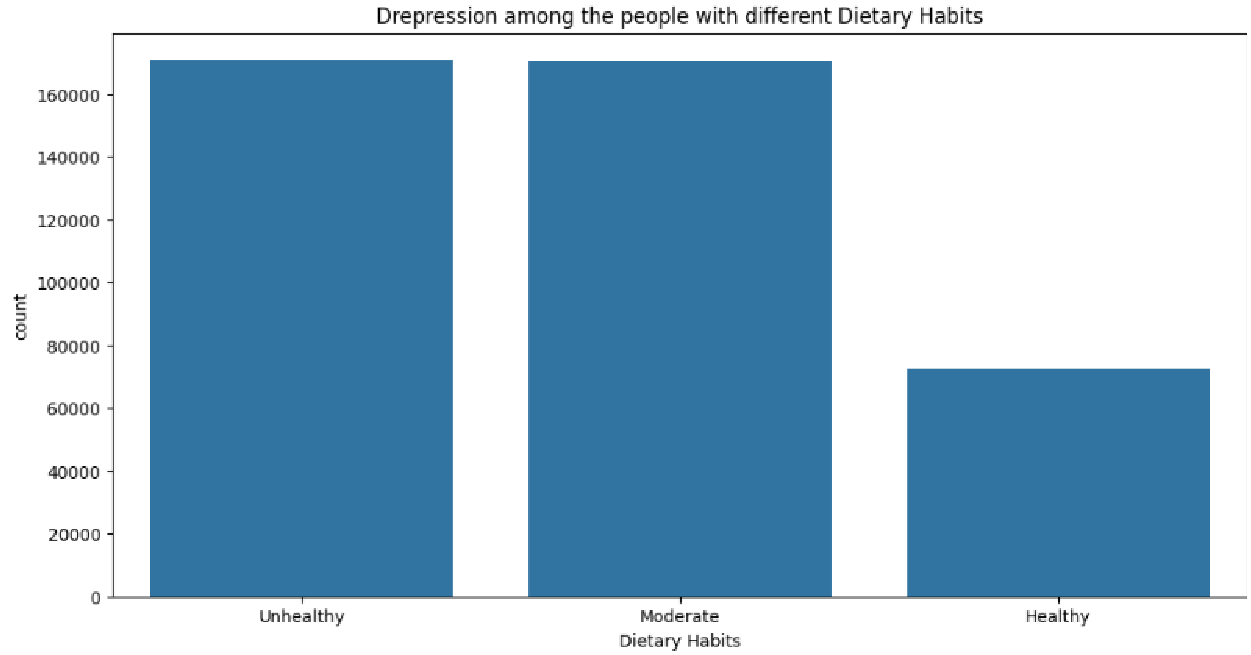

Figure-11: Count plot of Depression among different Dietary Habits

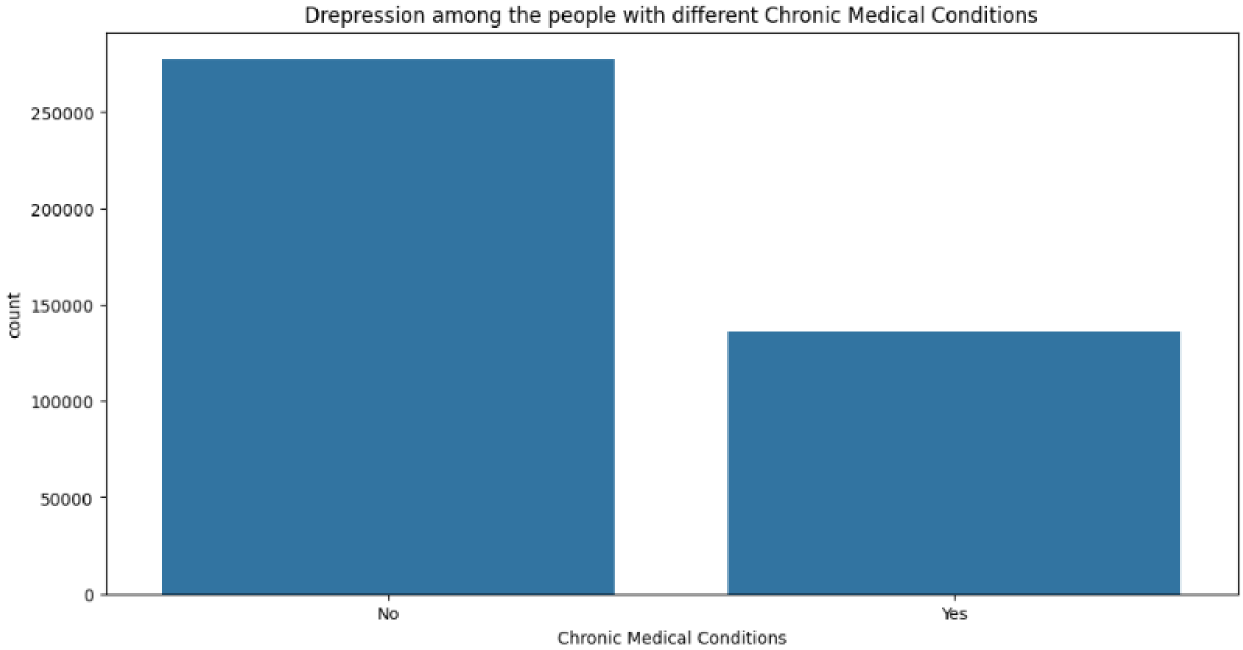

Figure-12: Count plot of Depression among different Medical condition

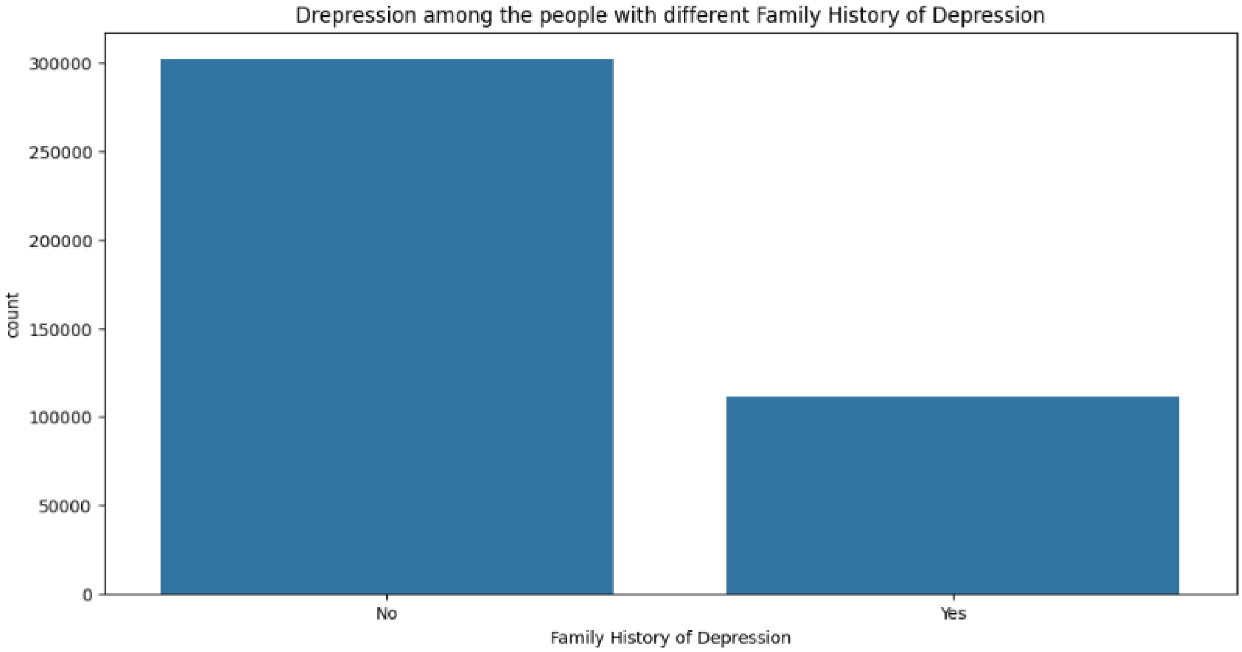

Figure-13: Count plot of Depression with different History of Depression

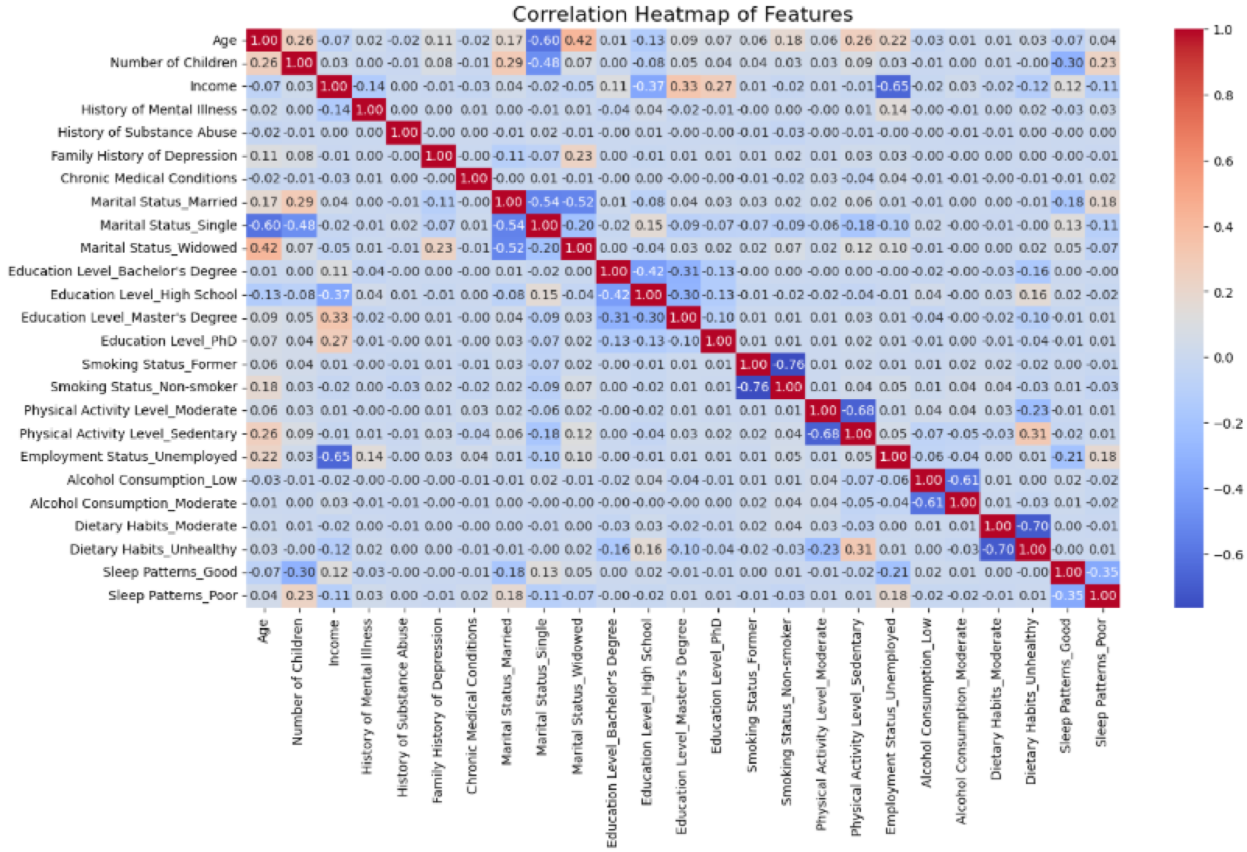

Figure-14: Correlation matrix of encoded data

Model Building:

Classification models that were considered for this task are logistic regression and decision tree. The models were built by using a train test split of 80 to 20 using all other columns as features to predict the target of Chronic Medical conditions.

Model Evaluation:

The model’s performance was evaluated in accuracy which for logistic regression was 0.57 with a precision score of 0.7 on precision, 0.6 on recall, and 0.6 on f1-score. For the decision tree, the accuracy score was 0.56 with a precision score of 0.6, recall of 0.6, and f1-score of 0.6. These metrics were used as they are commonly used to evaluate classification models.

Hyper-parameter Optimization:

To upgrade the model’s performance, hyper parameter optimization was performed using Grid-Search CV. The best parameters identified for logistic regression were ‘C’: 0.1 and ‘solver’:’liblinear’. For the decision tree model, the optional parameters were ‘criterion’:’gini’, ‘max_features’ : ‘sqrt’ and ‘max-dept’ : ‘5’.

Feature Selection:

The Recursive Feature Elimination (RFE) was employed to determine the most relevant features for predicting the target variable. The selected features included the smoking status, physical activity level, employment statues and alcohol consumption.

Conclusion:

The model evaluated using the metrics showed a sub par result as it was about half of the predicitons were wrong and there were multiple problems with the data processing.

The final model chosen the logistic regression was improved using the feature chosen using the RFE but it showed little improvement showing the amount of despair it can induce in the learning capabilities. When making this project several challenges were faced which included labeling the data as one hot coding that just did not work for the categorical data and was kept as an object instead of an integer. As for suggestions for improvement, there are many suggestions as asking for professional help and being more into the process of data scaling and processing.

Discussion:

The model's performance was evaluated using an accuracy and classification report and it resulted in the model performing not satisfactory and did not work well with the dataset. The hyperparameter tuning and feature selection helped a little in the improvement of the model but the improvement was not much and was mostly negligible. The model did not fair well and it was not a satisfactory result with a split of half and even in those there were disparages. For limitations on this model, it is not built well enough to have another chance and needs further work if not a full makeover. For future research, a better build model and better encoding can be used to do this which can drastically increase the chances of a proper function model.