**Predictive Modelling of Diabetes Risk Using Health Indicators**

Thesis submitted in partial fulfillment of the

requirements for the

**Post Graduate Certificate Program in**

**Data Science and Machine Learning**

By

**Javeriya Firdose**

**Mohammed Hashir**

**Nigam Harsh Dilip**

**Niladri Roy**

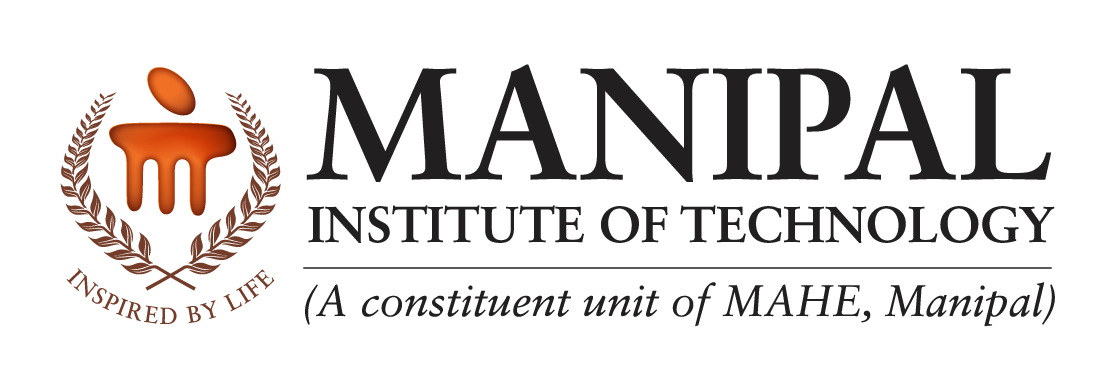
**Pallavi**

**Saravanavel Kalpana Revathi**

**Somantha Roy**

Under the guidance of

**Karthik Anand**



1. **Acknowledgments**

Gratitude cannot be expressed in mere words; it is a profound respect that shapes heartfelt expressions from one’s inmost being. Every accomplishment is the result of the combined effort of many individuals. It is our great privilege to take this opportunity to express our heartfelt gratitude to all those who provided valuable guidance, suggestions, and support in completing our capstone project.

We would like to extend our special thanks to **Manipal Institute of Technology**, for their encouragement and support throughout this journey.

It gives us immense pleasure and honor to convey our deepest gratitude to our beloved mentor and guide, **Karthik Anand**, for their valuable insights and constant guidance, which played a crucial role in shaping the success of this project.

We also acknowledge our heartfelt thanks and gratitude to **Ravi**, for their unwavering support and coordination throughout the project.

Finally, we would like to thank all those who helped us in completing this project successfully, including our families and friends, who have been a constant source of encouragement and understanding.

1. **Abstract**

**Background**: Diabetes is one of the major public health challenges, can lead to many health complications if left untreated. Early screening and diagnosis might control the disease burden. Currently It has been observed that diabetes is also affecting the productive age group of community. Effective preventive strategies can be identified and adopted by building the suitable model. This study helps in identifying important features, has to be focused in prevention and management of the diabetes mellitus

**Objective**: Develop a predictive model to identify individuals at high risk for diabetes prone to diabetes) based on available data. The model will support the "Diabetic360" program, assisting the marketing team in targeting the two specified groups.

**Methods:** The predictive model building project was done on kaggle dataset of Diabetic indicators. Dataset has 253680 records with 22 variables. We built models using Logistic regression boost, Random forest and decision tree techniques.

**Result:** The models treated without outliers yielded the following results, Logistic regression model ROC score of 0.81 with accuracy 0.74. ROC for Random forest model is 0.97 with accuracy of 0.91. XG boost model ROC score 0.95 with accuracy 0.88. ROC for decision tree model is 0.88 with accuracy 0.79. Random forest model is best suited model for our study as it results in Recall maximization.

**Conclusion:** Different models were compared and best suited model is selected. Long term goal for future study is tailor the model to specific regions or populations to address the changes in diabetic prevalence. In future model should be deployed in health care system for real time diabetic prediction.

1. **Table of contents with page number**

|  |  |  |
| --- | --- | --- |
| S.No. | CONTENTS | Page No. |
| 1 | Abstract | 1 |
| 2 | Introduction | 6 |
| 3 | Project Description | 10 |
| 4 | Exploratory Data Analysis | 15 |
| 5 | Design | 16 |
| 6 | Modelling | 32 |
| 7 | Key Results | 35 |
| 8 | Conclusion | 37 |
| 9 | References | 39 |
| 10 | Appendix | 40 |

1. **List of Tables**

|  |  |  |
| --- | --- | --- |
| S.No. | CONTENTS | Page No. |
| 1 | Feature Description | 11 |
| 2 | Data Types | 18 |

1. **List of Figures**

|  |  |  |
| --- | --- | --- |
| **S.No** | **FIGURES** | **Page No.** |
| 1 | Bar plot of high cholestrol | 20 |
| 2 | Bar plot of high BP | 21 |
| 3 | Bar plot of cholesterol check | 21 |
| 4 | Bar plot of smoker | 22 |
| 5 | Bar plot of stroke | 22 |
| 6 | Bar plot of heart disease or heart attack | 23 |
| 7 | Bar plot for physical activity | 23 |
| 8 | Histogram of BMI distribution | 24 |
| 9 | Histogram of mental health distribution | 24 |
| 10 | Histogram of physical health distribution | 25 |
| 11 | Boxplot of BMI | 25 |
| 12 | Boxplot of mental health | 26 |
| 13 | Boxplot of physical health | 26 |
| 14 | Class Distribution of Diabetes | 27 |
| 15 | Spearman Correlation Matrix | 28 |
| 16 | Shapiro-Wilk Test for Normality | 29 |
| 17 | Mann-Whitney U test for Normality | 30 |
| 18 | Rank Biserial Correlation | 31 |
| 19 | Spearman’s Test | 31 |
| 20 | Chi-Square Test and Cramer’s V Test | 32 |
| 21 | Random Forest | 33 |

1. **Introduction**

The fact that Diabetes is one of the major Non-communicable disease worldwide, has recognized as major public health challenge. SDGs target 3.4 is focuses on non-communicable diseases globally(Explore a World of Health Data Indicators Countries, n.d.). Diabetes has resulted over two million deaths, became one among top 10 common cause of death worldwide. Hardly any countries hinder the increase in diabetic cases and cases has been over the top in LMICs. According to WHO Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood glucose. Hyperglycaemia, also called raised blood glucose or raised blood sugar, is a common effect of uncontrolled diabetes and over time leads to serious damage to many of the body's systems, especially the nerves and blood vessels(Diabetes World Health Organization, 2024).

Improper management and untreated diabetic cases leading to variety life threatening complications such as abnormal lipoprotein, blindness, renal failure, cardio vascular diseases and gangrenes. In middle-income countries, the prevalence of diabetes is progressive in young adults, which is affecting highly productive age groups. Since diabetes can lead to unpleasant health problems, most effective preventive measures and immediate focus on young adults should be given as a priority(Nagarathna et al., 2020).

Some important risk factors such as heredity, age, lifestyle changes, alcohol consumption, smoking, screen time, poor sleep, parental conflicts and stressful life predispose young population to diabetes(Murea et al., n.d.). 80 to 85% of obese people developing type two diabetes, which is considered as major risk factor.

Sedentary life style become more common among young population. The lack of physical activity is the main concern as more than 80% of females and males (78%) in Australia lacking even minimum level of physical activity. There exists obesity is more common among females than males(Bauman et al., 2012). Metabolic disorders including diabetes mellitus in Northern India are more prevalent among female of middle age population. Therefore, it is very crucial to identify the major risk factors leading diabetes and its life-threatening complications(Misra et al., 2001).

Current study aimed to identify the important risk factors and indicators, has an effect in developing diabetes mellitus in the population. Effective preventive strategies can be identified and adopted by building the suitable model. This study helps in identifying important features, has to be focused in prevention and management of the diabetes mellitus.

* 1. **Motivation**

The fact that in recent decades diabetes strikes on younger population, will results deterioration in health of productive age group. Usually most of the cases diagnosed among those who aged 50 years or older but there may be a chance of undiagnosed prediabetic condition prevailed in their younger adulthood. There exists an urgent need of developing strong predictive model to identify the contributing factors so that disease can be controlled and managed in future. This project focuses on analysing the **Diabetes Health Indicators Dataset** to build a predictive model that identifies whether an individual has diabetes (represented as 0 - No Diabetic, 1 - Diabetic/Pre-diabetic).

* 1. **Project scope**
* **Objective**: Develop a predictive model to identify individuals at high risk for diabetes (prone to diabetes) based on available data. The model will support the "Diabetic360" program, assisting the marketing team in targeting the two specified groups.
* **Target Population**:
  + **Group 1**: Individuals at risk of developing diabetes (prevention-focused).
  + **Group 2**: Individuals already diagnosed with diabetes (management-focused).
* **Key Deliverables**:
  + A machine learning model to predict diabetes risk.
  + Model insights that can be used for targeted marketing efforts.
  + Summary report and marketing team training on using the model outputs.
* **Data Requirements**:
  + Health metrics (e.g., BMI, blood glucose levels, family history, lifestyle factors).
  + Demographics and behavioral data (e.g., age, diet, physical activity).
  + Historical health records for individuals in each group (if available).

### **Constraints:**

#### **Project Constraints:**

1. **Data Availability**: Limited access to complete health and demographic data might impact model accuracy.
2. **Data Privacy**: Adhere to data protection regulations (e.g., GDPR, Indian data privacy laws), anonymizing personal data.
3. **Timeframe**: Limited time for model development, testing, and validation, which may affect the complexity of the model.
4. **Budget**: Allocation for data storage, model training, and deployment

resources, including potential cloud services.

#### **Design Constraints:**

1. **Model Interpretability**: The model should be interpretable enough for the marketing team to understand how predictions are made.
2. **Scalability**: Design the model to accommodate growing data volumes as more users enroll in Diabetic360.
3. **Performance Metrics**: Aim for a high balance of precision and recall, especially to minimize false negatives (failing to identify at-risk individuals).
4. **Resource Efficiency**: Optimize for lower computational power to reduce operational costs if deploying on limited hardware.
   1. **Project Goal**

To Develop a predictive model to identify individuals at high risk for diabetes (prone to diabetes) based on available data. The model will support the "Diabetic360" program for our organisation “GenTest”, assisting the marketing team in targeting the two specified groups.

* 1. **Literature/Market survey**

The literature review for our project was done for the following

1. High risk factors for diabetes
2. Preventive and management of diabetes
3. Predictive model for diabetes prevention and management
4. Marketing team training

**WHO Package of Essential NCD Interventions (PEN) Management of Type 2 Diabetes**

**Diabetes**

A group of metabolic disorders characterized by the presence of hyperglycaemia in the absence of treatment. The aetio-pathology includes defects in insulin secretion, insulin action, or both The long-term specific complications of diabetes include retinopathy, nephropathy and neuropathy(WHO Package of Essential NCD Interventions (PEN) Management of Type 2 Diabetes. ACKNOWLEDGEMENT, n.d.)

**Risk factors**

Risk factors for type 2 diabetes (strong):

• overweight/obesity

• physical inactivity

• diabetes in first degree relatives

• history of gestational diabetes

• cardiovascular disease and its risk factors

• ethnicity (South Asian, Afro-Caribbean, Hispanic)

Risk factors for type 1 diabetes (weak):

• Certain genetic haplotypes

• Unknown environmental factors

Risk factors for gestational diabetes (strong)

• Similar to type 2 diabetes Symptoms

• polyuria (excessive passing of unrine)

1. **High risk factors for diabetes**

Cross sectional study was conducted at field area of Urban health training centre Maharashtra by Reshma Patil et al (2019). Sociodemographic, physical activity and family histry information is obtained from questionnaire. Hip to waist measurement, weight and height is measured and recorded. The study shows that more than half of the study diabetic patients (61.90%) had a family history of diabetes. Most of the diabetic cases (61.91%) were following sedentary life style and has minimum physical activity. Half of the diagnosed cases (52.38%) were obese or preobese and less than half of the diagnosed cases were underweight (47.62%).(Patil & Gothankar, 2019)

1. **Preventive and management of diabetes**

Randomized control study was conducted at Asthi Northwestern Italy by Simona Bo MD et al (2007). Data was collected from predesigned questionnaire. Experimental group is given with interventions includes education on general lifestyle modification by physician in addition to that individual recommendation from endocrinologists, dietician and specialist from internal medicine. This study reveals that prevalence of metabolic disorders is reduced among experimental groups when compare to controlled group.(Bo et al., 2007)

1. **Predictive model for diabetes prevention and management**

Predictive model for diabetic mellitus using decision trees and Ada boost algorithm was developed at Kerala developed by Veena V et al (2016). Dataset contained 768 instances, has trained for the model. The decision tree algorithm combined with AdaBoost model gave best results. The model accuracy was 80.72%.(V. Veena Vijayan; C. Anjali, 2016)

1. **Marketing team training**

RCT study was conducted at Kerala by Sathish T et al (2018). Data was collected by Risk assessment questionnaire, non-diabetic participant with high risk score more than 60 were recruited for the study. Kerala diabetic prevention program provided comprehensive intervention such as to experimental group health education, peer leaders, kitchen garden training, walking, yoga to experimental group. Study showed that intervention helps in behavioural and lifestyle change among high risk participants.(Sathish et al., 2019)

1. **Project Description** 
   1. **Business/Domain Understanding**

Millions of individuals worldwide suffer from diabetes, a chronic illness that can cause serious health issues like renal failure, neuropathy, and cardiovascular problems. Medical, demographic, and behavioural factors all have a significant impact on the illness. The burden of disease can be greatly decreased, quality of life can be enhanced, and healthcare expenses can be decreased with early detection and preventive actions.

* 1. **Project stakeholders**

The project Stakeholders include the following:

1. "Gen Test," a top pathological testing company in India which aims to introduce a subscription plan called as “Diabetes360”.
2. Data Privacy Advocates to ensure privacy and ethical use of the data.
3. The patients, helping them gain awareness and aid them in making lifestyle changes, lifestyle management.
4. Policy Makers: This project and the subscription in turn can be useful for policy makers
   1. **Datasets understanding**

**Table 1:** Feature Description

|  |  |  |
| --- | --- | --- |
| **Feature** | **Description** | **Values** |
| **Diabetes** | Indicates diabetes status based on different datasets. | 0 = No Diabetes,  1 = Prediabetes or Diabetes |
| **High BP** | Indicates whether the individual has high blood pressure. | 0 = No High BP,  1 = High BP |
| **HighChol** | Indicates whether the individual has high cholesterol. | 0 = No High Cholesterol, 1 = High Cholesterol |
| **CholCheck** | Indicates if cholesterol was checked in the last 5 years. | 0 = No, 1 = Yes |
| **BMI** | Body Mass Index, a numerical value. | Continuous value. |
| **Smoker** | Indicates if the individual has smoked at least 100 cigarettes in their lifetime. | 0 = No, 1 = Yes |
| **Stroke** | Indicates if the individual has ever had a stroke. | 0 = No, 1 = Yes |
| **HeartDiseaseorAttack** | Indicates if the individual has had coronary heart disease (CHD) or a myocardial infarction (MI). | 0 = No, 1 = Yes |
| **PhysActivity** | Indicates if the individual has participated in physical activity in the past 30 days (excluding job-related activity). | 0 = No, 1 = Yes |
| **Fruits** | Indicates if the individual consumes fruit at least once a day. | 0 = No, 1 = Yes |
| **Veggies** | Indicates if the individual consumes vegetables at least once a day. | 0 = No, 1 = Yes |
| **HvyAlcoholConsump** | Indicates if the individual is a heavy drinker (men >14 drinks/week, women >7 drinks/week). | 0 = No, 1 = Yes |
| **AnyHealthcare** | Indicates if the individual has any kind of health care coverage. | 0 = No, 1 = Yes |
| **NoDocbcCost** | Indicates if the individual couldn't see a doctor due to cost in the last 12 months. | 0 = No, 1 = Yes |
| **GenHlth** | Self-reported general health status. | Scale 1-5: 1 = Excellent, 2 = Very Good, 3 = Good, 4 = Fair, 5 = Poor |
| **MentHlth** | Number of days in the past 30 with poor mental health (stress, depression, emotions). | Continuous value (days). |
| **PhysHlth** | Number of days in the past 30 with poor physical health (illness or injury). | Continuous value (days). |
| **DiffWalk** | Indicates if the individual has difficulty walking or climbing stairs. | 0 = No, 1 = Yes |
| **Sex** | Indicates the gender of the individual. | 0 = Female, 1 = Male |
| **Age** | Age group of the individual. | Scale 1-13: 1 = 18-24, 2 = 25-29, 3=30-34, 4=35-39, 5=40-44, 6=45-49, 7=50-54, 8=55-59, 9=60-64, 10=65-69, 11=70-74, 12=75-79, 13 = 80+ |
| **Education** | Highest level of education achieved. | Scale 1-6: 1 = Never attended school or kindergarten, 2 = Grades 1-8,3 = Grades 9-11, 4 = Grade 12, 5 = College 1-3 years, 6 = 4 years or more of college |
| **Income** | Annual income category of the individual. | Scale 1-8: 1: <$10 K, 2: $10–$15 K, 3: $15–$20 K, 4: $20–$25 K, 5: $25–$35 K, 6: $35–$50 K, 7: $50–$75 K, 8: >$75 K |

**Summary:**

* **Categorical Features:** Diabetes, High BP, HighChol, CholCheck, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, DiffWalk, Sex.
* **Ordinal Features:** GenHlth, Age, Education, Income.
* **Continuous Features:** BMI, MentHlth, PhysHlth.
  1. **Data Limitations**
* Genetic predispositions and family history, which are important risk factors for diseases like diabetes and heart disease, are not included.
* Regarding individuals with diabetes (the target variable), the data is unbalanced.
* The severity, duration, and type of diabetes (Type 1, Type 2, gestational) are not adequately captured by this oversimplification.
* Whether people are getting medicine or therapy for diseases like diabetes, high blood pressure, or high cholesterol is not recorded in the dataset.
  1. **Benefits of project**

The benefits of the project have two aspects to it: 1. Healthcare/ Public Health 2. Model Prediction.

**Healthcare/ Public Health**

1. Diabetes prevention and early detection.
2. Policymakers can concentrate on communities with greater prevalence rates of risk factors by using insights from the dataset.
3. This can assist in creating integrated healthcare plans that treat several chronic illnesses at once.

**Model Prediction.**

1. Machine learning models can be developed using the dataset to predict diabetes risk based on demographic, lifestyle, and health factors.
2. Machine learning models can be developed using the dataset to predict the following: The likelihood of comorbidities like high blood pressure or heart disease.
3. Predictive model insights can be used in the "Diabetic360" software to forecast diabetes risk.
4. **Exploratory Data Analysis**
   1. **Data collection**

The data source was for the project was <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>

* 1. **Data exploration**

** Univariate Analysis**

* Examined the distribution, central tendency, and variability of individual features using summary statistics, histograms, and boxplots.
* Examples:
  + BMI: Right-skewed distribution with high variability (standard deviation ≈ 7.9). Median ≤ 31 indicates most values are concentrated in the lower range. Categorized BMI into "Low," "Medium," and "High" using quantiles.
  + GenHlth (General Health): Slightly right-skewed, with most values ≤ 3, suggesting average or good general health. Categorized into "Poor," "Fair," and "Good."

** Bivariate Analysis**

* Explored relationships between pairs of features and their association with the target variable (Diabetes\_binary).
  + Correlation Matrix: Identified linear correlations between continuous variables.
  + Cramér's V Analysis: Measured the strength of association for categorical variables with diabetes risk (e.g., "Smoking" and "CholCheck" showed low associations but were statistically significant).

** Multivariate Analysis**

* Investigated interactions among multiple features to identify patterns affecting diabetes prediction.
  + Used pair plots and heatmaps to visualize interactions.
  + Features like "Physical Health Days" and "Mental Health Days" displayed nonlinear relationships with diabetes risk, requiring transformations for improved model performance.
  1. **Complexity of Data**
  2. **Data cleaning**

**Dataset Overview:**

* The dataset has 22 features and 253,680 rows with no missing values.
* **Outlier Removal:**
  + Applied IQR-based filtering to remove extreme values from continuous features like BMI, Age, and Physical/Mental Health Days.
  + Resulted in reducing the dataset from 253,680 rows to 176,749 rows .
* **Categorization of Features:**

Variables like BMI and Age were grouped into categories (e.g., Low, Medium, High) based on quantiles or domain-specific thresholds.

* 1. **Data Transformation**

** Standardization:**

* Applied Z-score normalization to continuous variables such as BMI to ensure a mean of 0 and a standard deviation of 1 for better model performance. Formula used: z=(x−μσ)/*σ* where ( x ) is the original value, ( \mu ) is the mean, and ( \sigma ) is the standard deviation
* This transformation ensures comparability across features .

** Feature Engineering:**

* Created al groupings for variables like General Health (GenHlth), Physical Health Days (PhysHlth), and Mental Health Days (MentHlth) for simplified analysis.
  + Example: Physical Health Days > 20 were capped and categorized as "Many" .

** Encoding:**

* Used ordinal encoding for variables such as Education and Income levels for seamless integration into machine learning models .

1. **Design** 
   1. **Analytical methods and Technology used**

In this project, we employed a range of analytical methods and technologies to effectively analyze and visualize the dataset. The analysis began with the categorization of features into three distinct groups: categorical, ordinal, and continuous. For each group, appropriate statistical methods were applied to uncover meaningful insights. The continuous and ordinal features were subjected to hypothesis testing to evaluate their relationships and differences between groups, such as diabetic and non-diabetic individuals. The Shapiro-Wilk test was performed to assess the normality of these numeric features. Based on the results, non-parametric methods such as the Mann-Whitney U test were applied to compare group medians when normality assumptions were violated.

* 1. **Descriptive Statistical Analysis**

 For this analysis, we employed statsmodels.api and scipy libraries to perform a detailed examination of the dataset. The features were categorized into three groups: categorical, ordinal, and continuous. Each group was analyzed systematically to derive meaningful insights.

1. Numeric Features (Ordinal and Continuous):

* Shapiro-Wilk Test:
  + Used to check normality for each numeric feature.
  + Results determined whether parametric or non-parametric methods were appropriate.
* Hypothesis Testing:
  + Conducted to compare distributions between groups (e.g., diabetic vs. non-diabetic).
* Mann-Whitney U Test:
  + A non-parametric test used for features that did not meet normality assumptions.
  + Compared the medians of two groups.
* Effect Size Measures:
  + Rank Biserial Correlation: Assessed the strength of association between two groups.
  + Cliff’s Delta: Measured the magnitude of difference, indicating the extent of overlap between distributions.

2. Categorical Features:

* Focused on examining relationships and proportions within the data, though not elaborated in detail here.

By dividing the features and using appropriate statistical methods for each type, we ensured a robust and comprehensive analysis, accounting for the unique properties of the dataset. This approach helps capture patterns and differences across various groups effectively.

* 1. **Data Visualization**

 In this project, data visualization played a crucial role in understanding the distribution and patterns within the dataset. Based on the classification of features into three groups—categorical, ordinal, and continuous—we used targeted visualization techniques to effectively represent the data. For categorical features, bar plots were created to showcase the frequency distribution of each category, making it easy to identify trends and group proportions. For numeric features, including both ordinal and continuous variables, histograms were plotted to visualize their distribution and identify skewness or gaps in the data. Additionally, box plots were employed to display the spread, central tendency, and potential outliers within these features. These visualizations provided an intuitive and comprehensive overview of the data, aiding in the identification of patterns, anomalies, and differences across groups. By leveraging these plots, we were able to build a strong foundation for further statistical and predictive analyses.

**Table 2:** Data Types

|  |  |  |
| --- | --- | --- |
| **Categorical Features** | **Ordinal Features** | **Continuous Features** |
| * HighBP * HighChol * CholCheck * Smoker * Stroke * HeartDiseaseorAttack * PhysActivity * Fruits * Veggies * HvyAlcoholConsump * AnyHealthcare * NoDocbcCost * DiffWalk * Sex | * GenHlth * Age * Education * Income | * BMI * MentHlth * PhysHlth |

* 1. **Feature Engineering**

Since the BMI data is right-skewed and the Z-score analysis identifies 9,847 outlier which is over all data 3.8% less 5%

we can go for log transformation(pull the value close to mean) or capping (replace value by max or min and k means clustering

**1. Skewness of BMI Data**

**Right-Skewed Data:**

   The BMI data has a long tail on the right side, indicating there are higher values that are less frequent.

  - Skewness can make statistical analysis or machine learning less effective, as many techniques assume a normal distribution.

**2. Outlier Analysis Using Z-Score**

**Z-Score**: A statistical measure representing the number of standard deviations a data point is from the mean.

  - Values with Z-scores greater than a threshold (commonly 3) are considered outliers.

  - Outliers here are extreme BMI values that deviate significantly from the rest of the data.

Result

  - 9,847 outliers were identified.

  - This represents \*\*3.8%\*\* of the data, which is less than the commonly used threshold of \*\*5%\*\* (used to determine if the number of outliers is acceptable).

  - Since the proportion is relatively small, the data is manageable but may still require adjustments.

**3. Suggested Techniques to Address Skewness and Outliers**

**Log Transformation**

  - Applies the logarithmic function to the BMI values.

  - Pulls large values closer to the mean, reducing skewness and the impact of extreme values.

  - Ensures the data is closer to a normal distribution

**Capping**

  - Replaces extreme values (outliers) with a fixed maximum or minimum.

  - Example: Set any BMI above 40 to 40, or any below 10 to 10.

  - Keeps data within a reasonable range without completely discarding outliers.

**4. Use of K-Means Clustering**

**K-Means**

  - It’s a machine learning technique to group data into clusters based on similarity.

  - Clustering can help segment the BMI data into groups (e.g., low, medium, high BMI).

  - Helps handle outliers by assigning them to the nearest cluster centroid, effectively normalizing their impact.

**Key Takeaway**

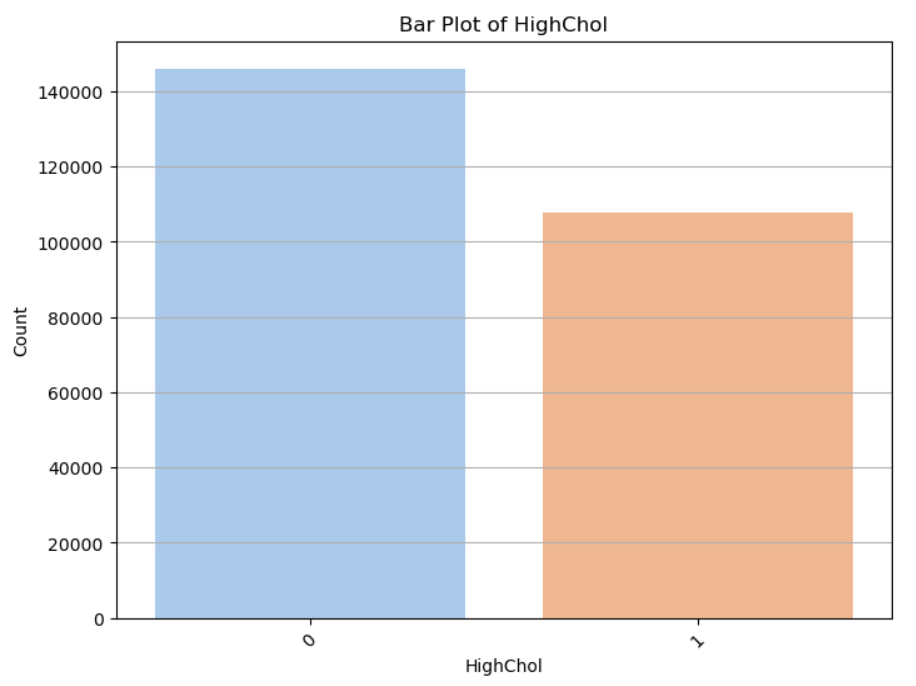
This analysis ensures the BMI data is prepared for modeling by:

- Reducing skewness (using log transformation).

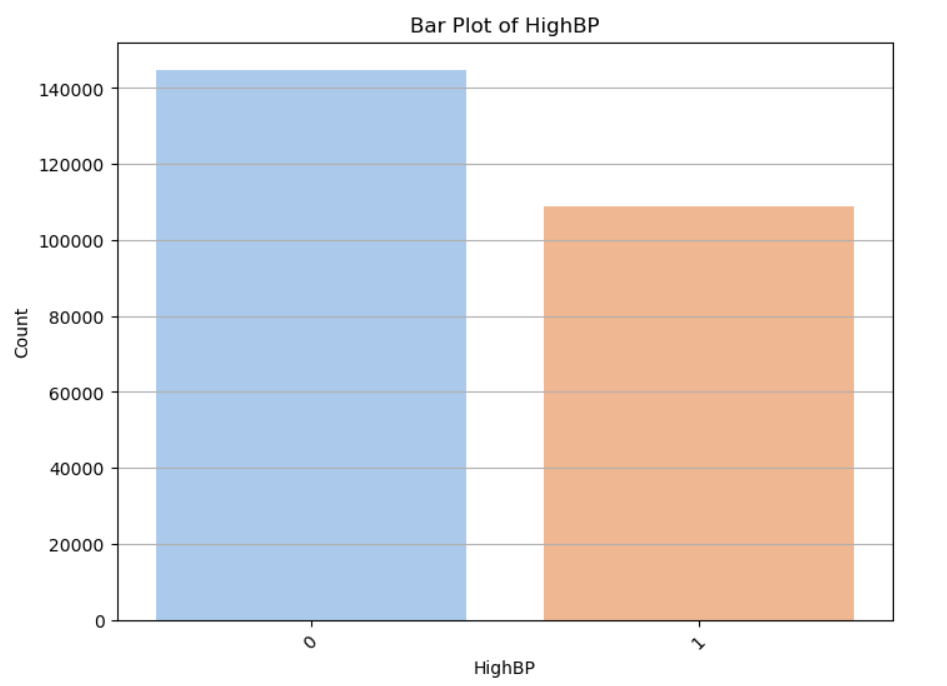
- Handling outliers (using capping or clustering)**.**

* 1. **Short data snapshots**

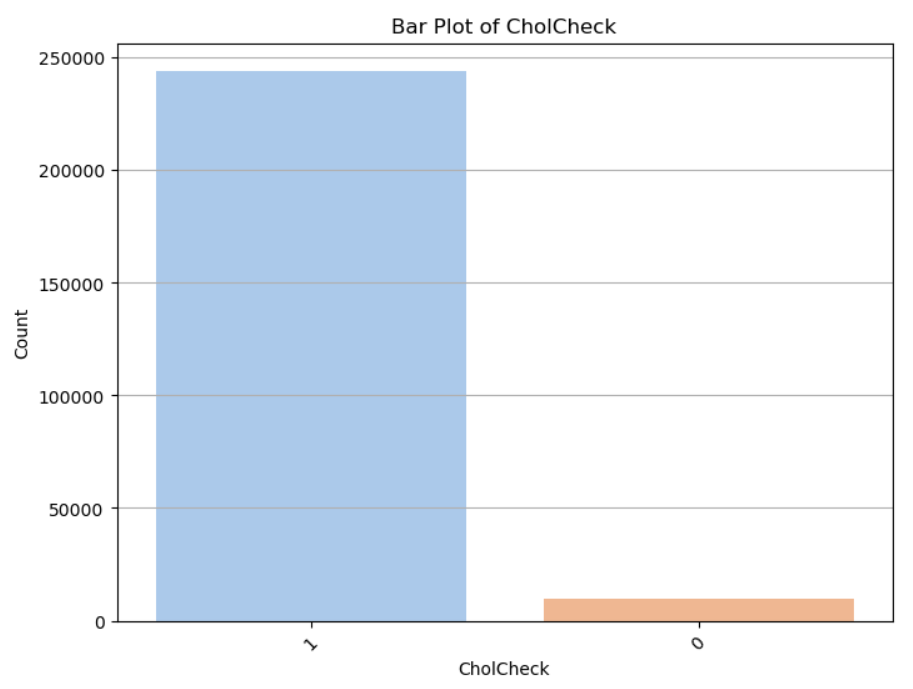
**Data Distribution For Categorical Features (Bar Plots)**



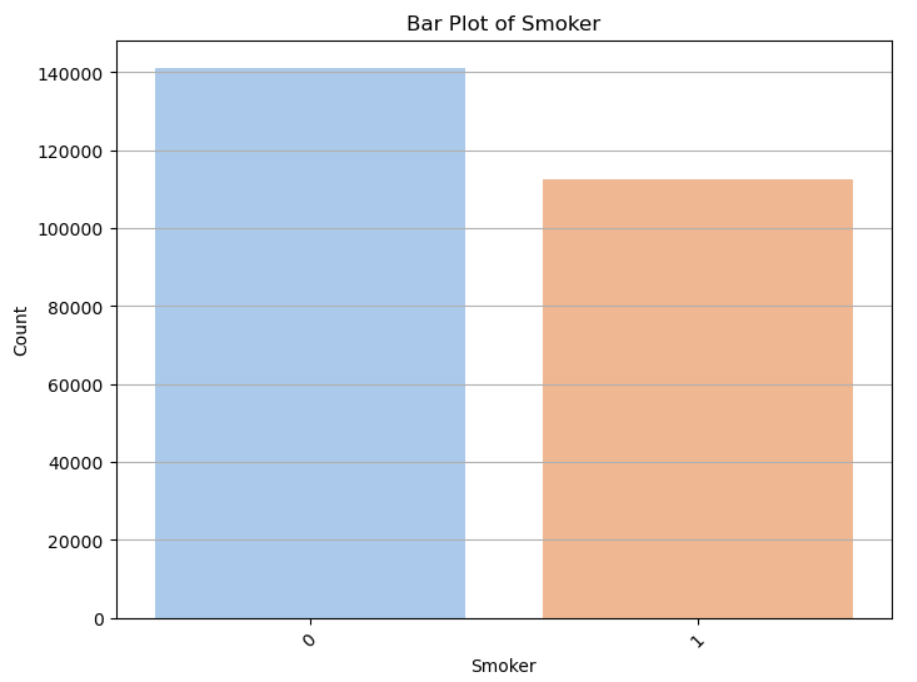
**Figure 1:**  Bar Plot of HighChol



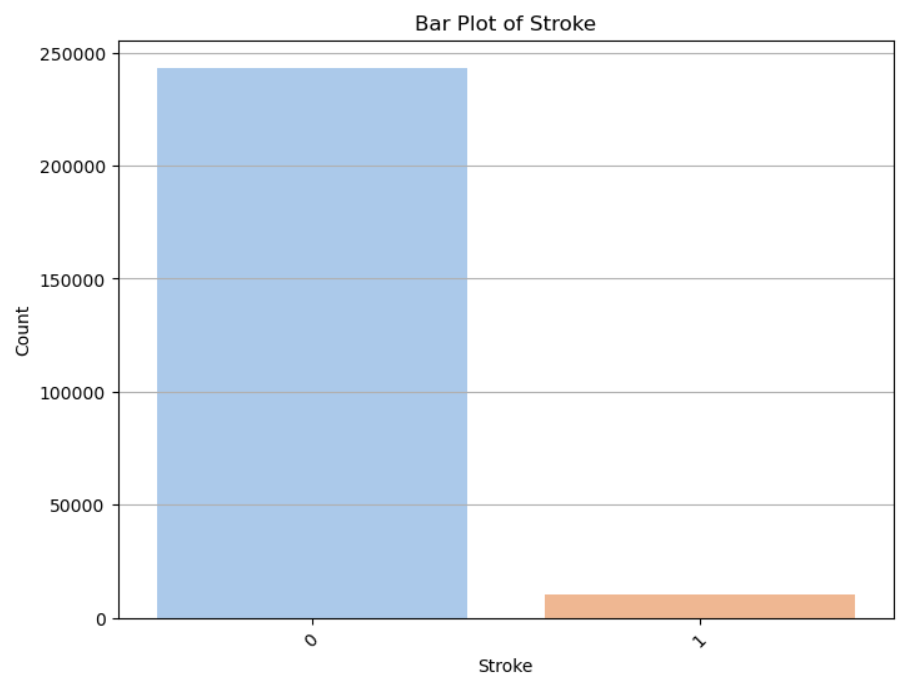
**Figure 2:** Bar Plot of HighBP



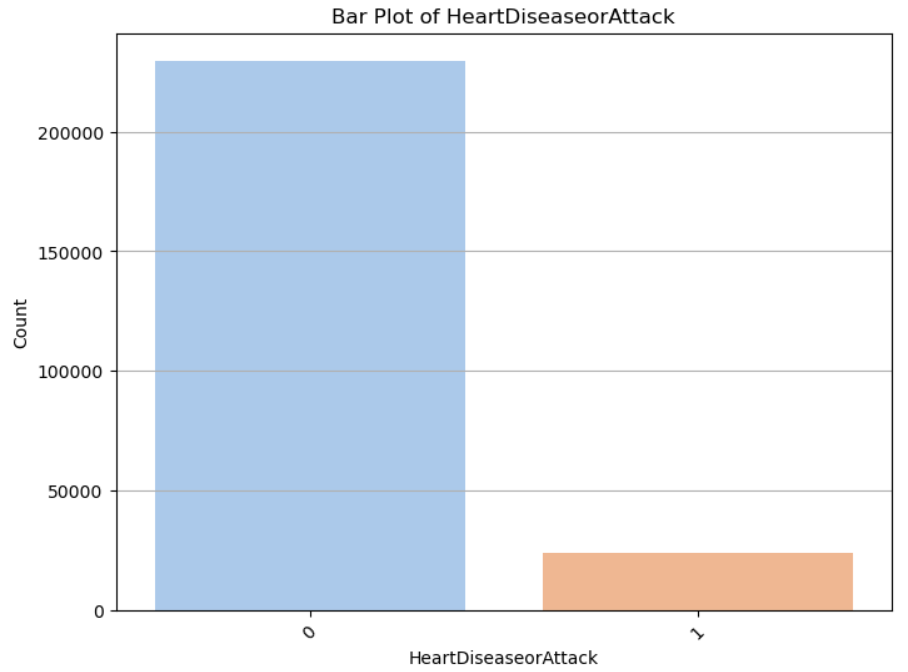
**Figure 3:** Bar Plot of CholCheck



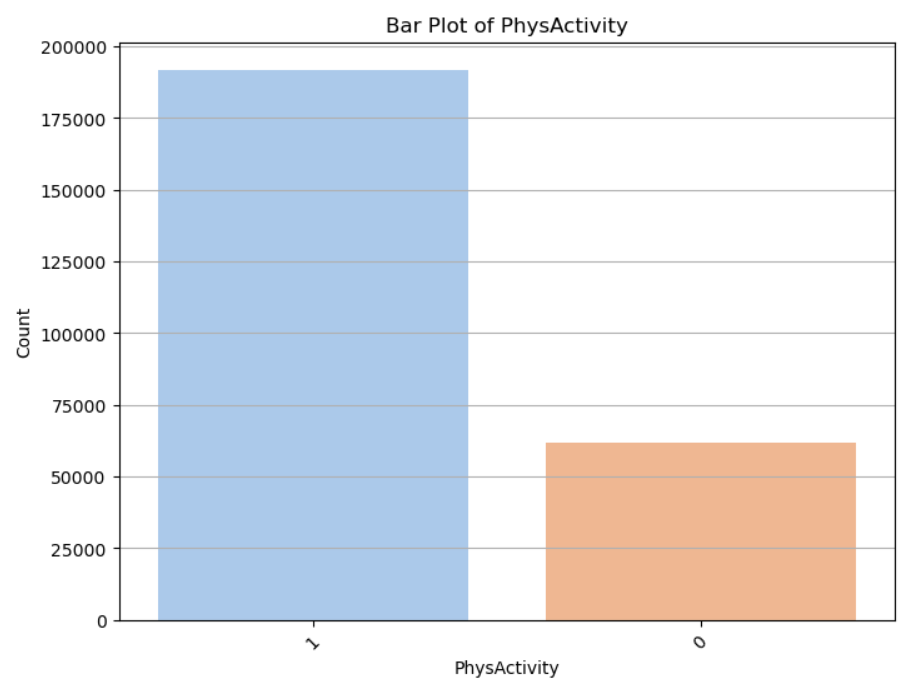
**Figure 4:** Bar Plot of Smoker



**Figure 5:** Bar Plot of Stroke

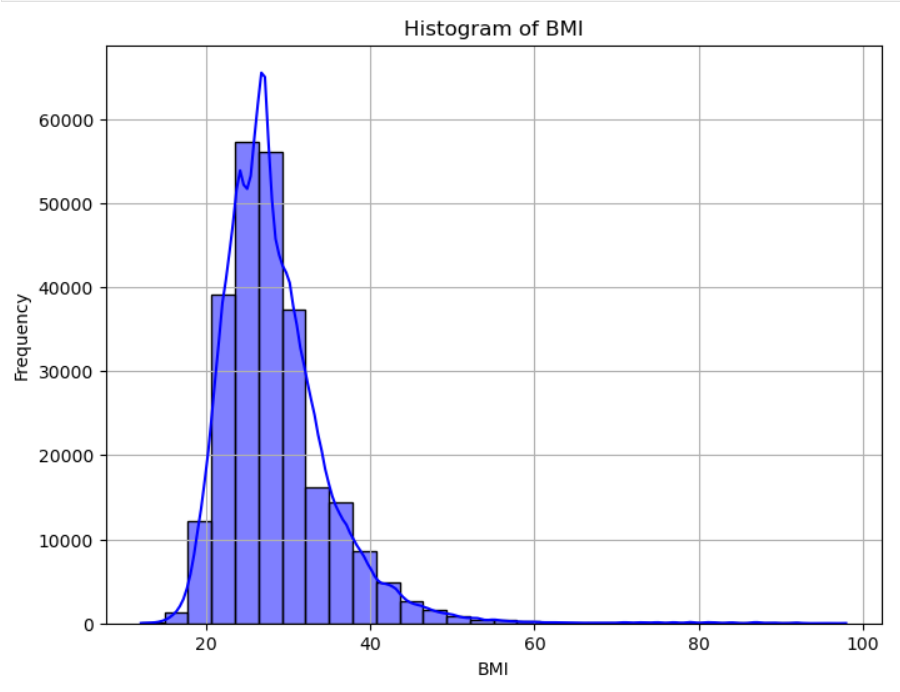


**Figure 6:** Bar Plot of HeartDiseaseorAttack

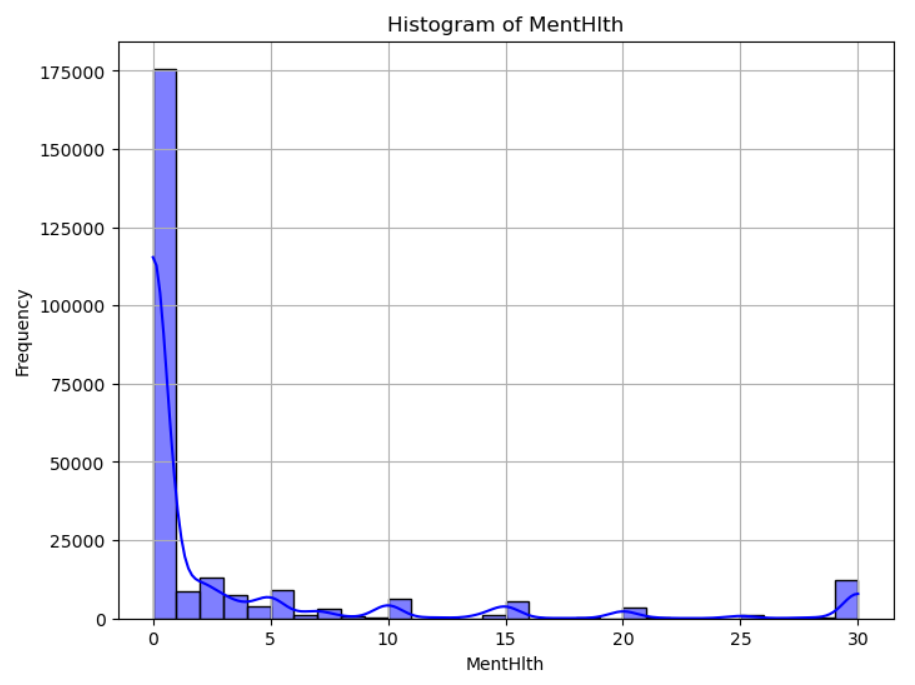


**Figure 7:** Bar Plot of PhysActivity

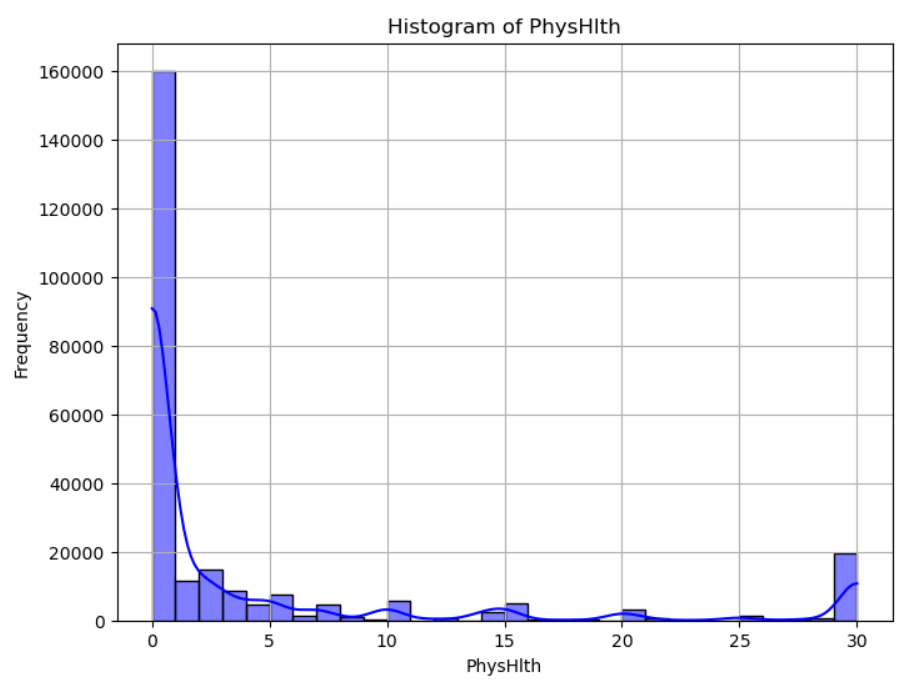
**Data Distribution For Continuous Features (Histogram Plots)**



**Figure 8:** Histogram of BMI

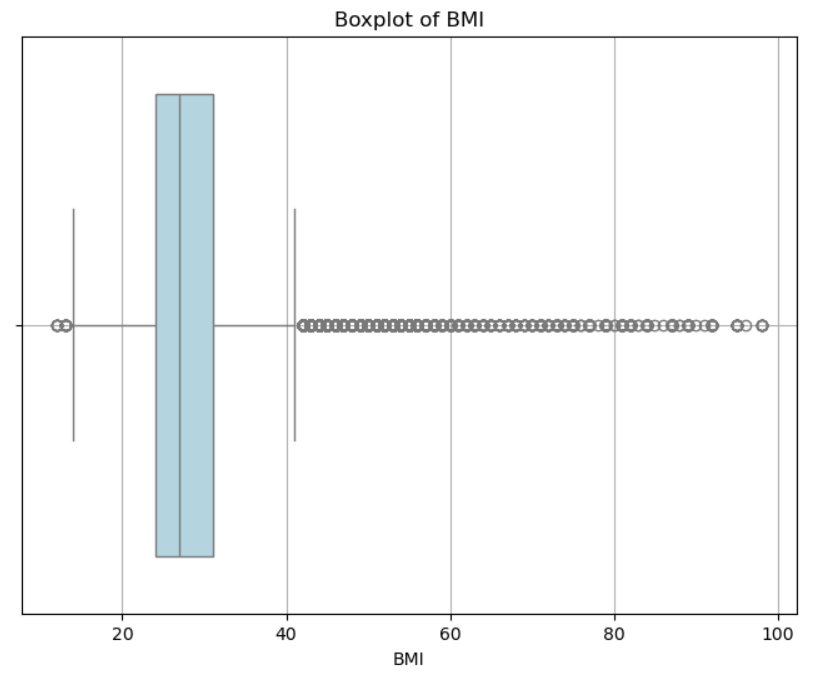


**Figure 9:** Histogram of MentHlth

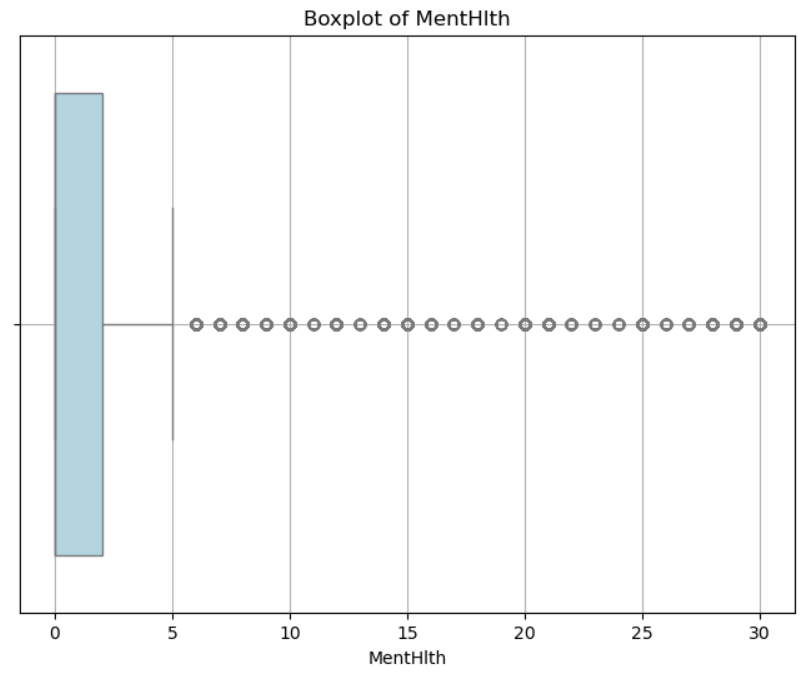


**Figure 10:** Histogram of PhysHlth

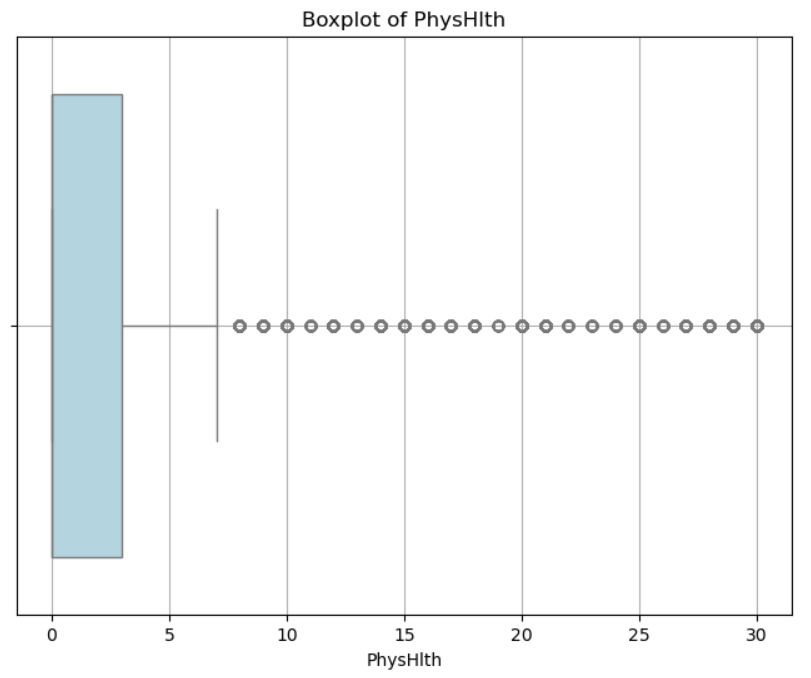
**Data Distribution For Continuous Features (Box Plots)**



**Figure 11:** Boxplot of BMI

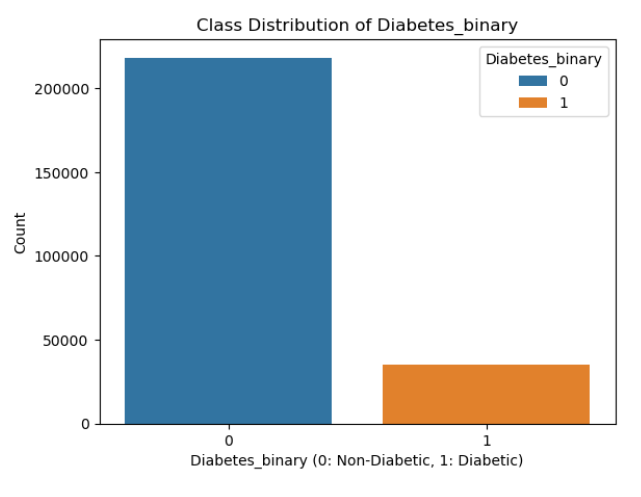


**Figure 12:** Boxplot of MentHlth



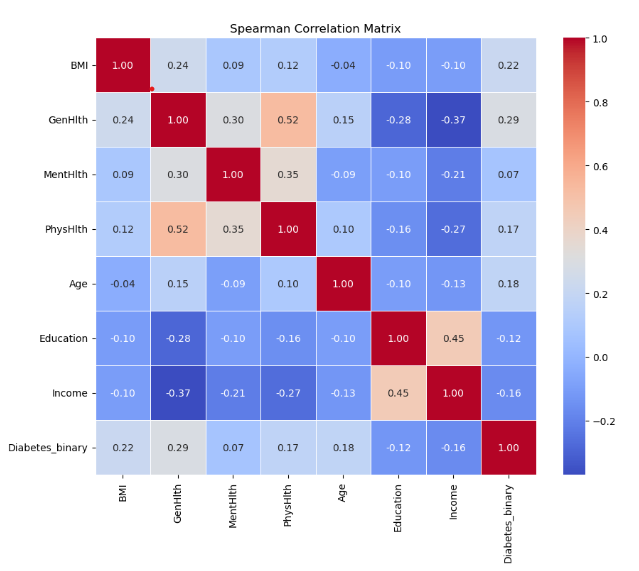
**Figure 13:** Boxplot of PhysHlth

**Class Distribution In Data**



**Figure 14:** Class Distribution of Diabetes

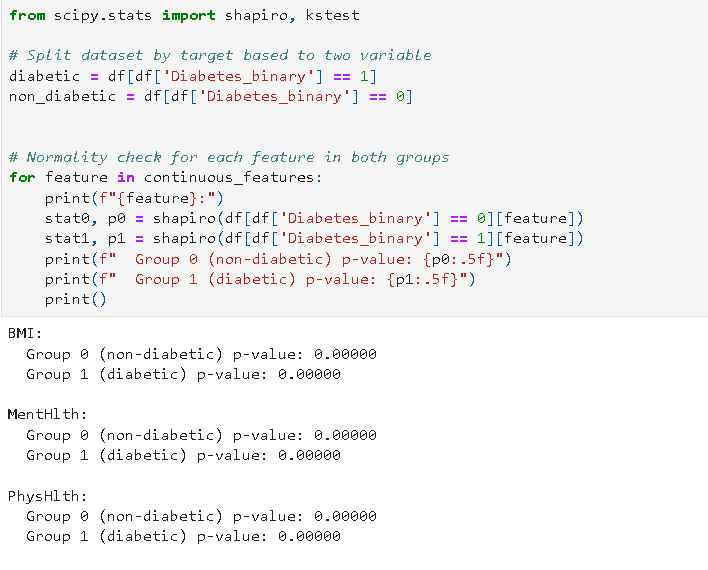
**Spearman Correlation Matrix for Numerical Data**



**Figure 15:** Spearman Correlation Matrix

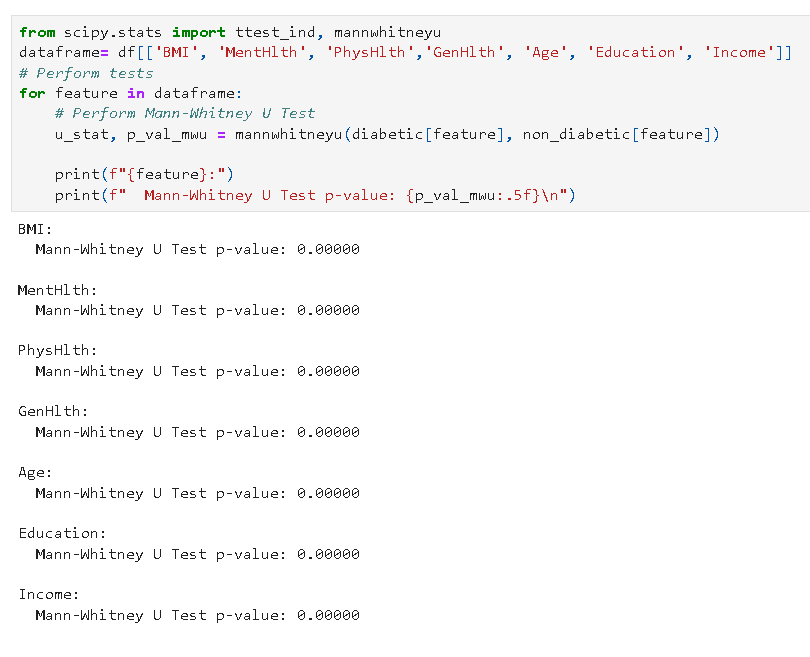
* 1. **Short code snippets**

### **Shapiro-Wilk Test for Normality**



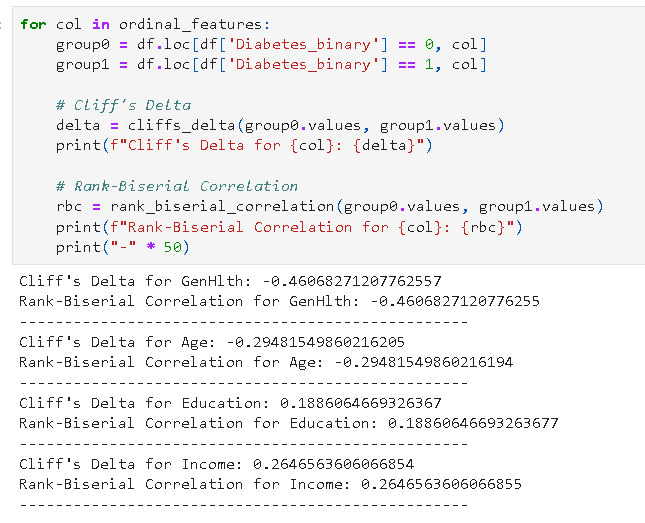
### **Figure 16:** Shapiro-Wilk Test for Normality

### Mann-Whitney U Test (Non-Parametric)



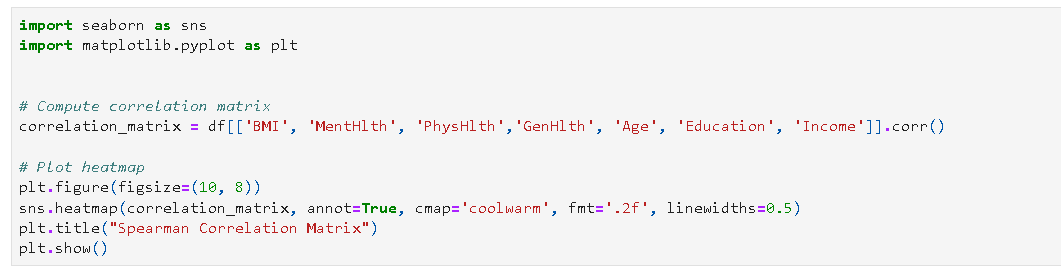
### **Figure 17:** Mann-Whitney U Test for Normality

### Rank-Biserial Correlation and Cliff's Delta Test



### **Figure 18:** Rank Biserial Correlation

### Spearman's Test



**Figure 19:** Spearman’s Test

**Chi-Square Test And Cramer’s V**

****

**Figure 20:** Chi-Square Test and Cramer’s V test

1. **Modelling**
   1. **Selection of model/technique**

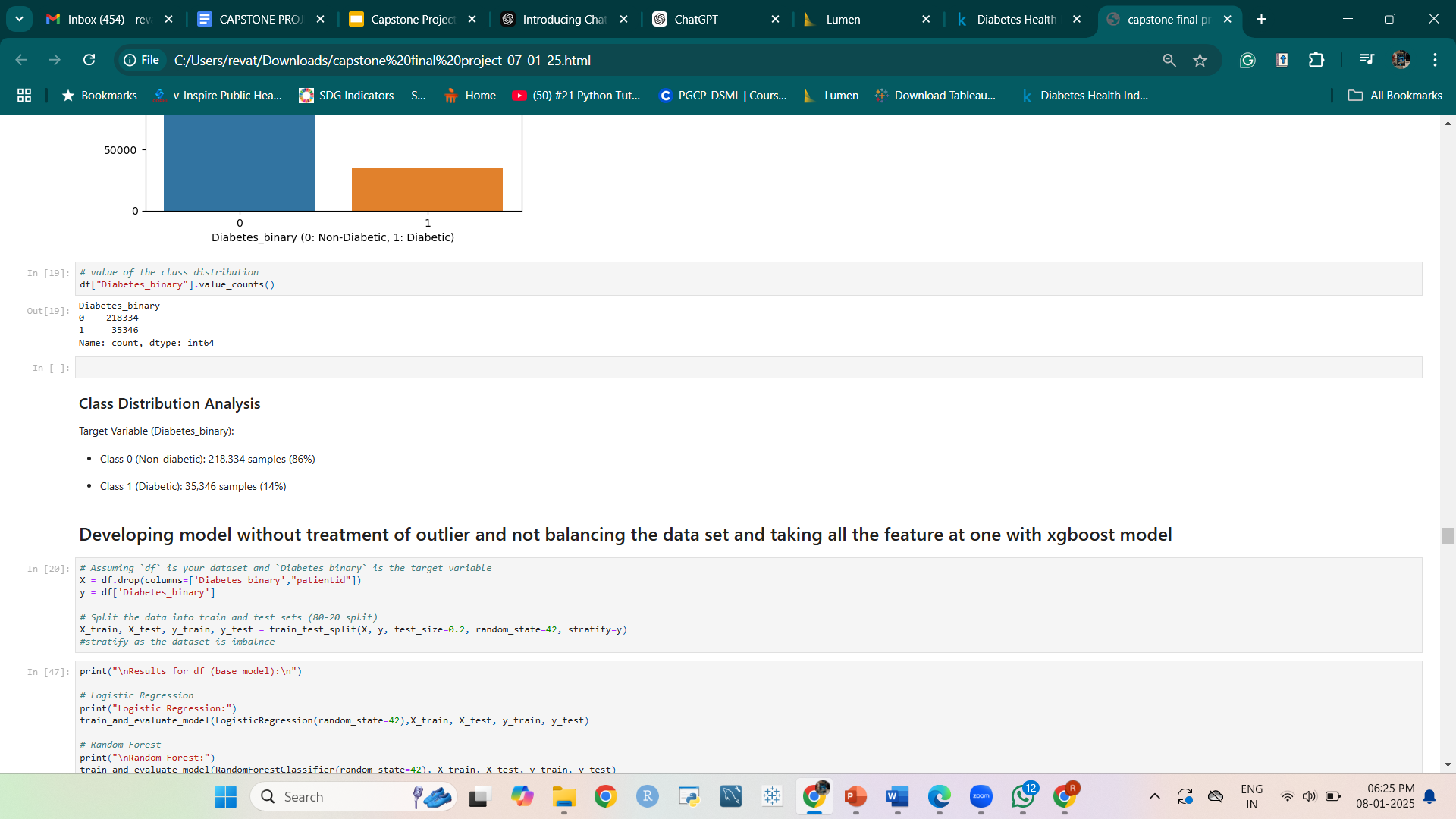
In this project, we explored multiple machine learning models to predict diabetes risk using the Diabetes Health Indicators dataset. The dataset had significant class imbalance, with non-diabetic samples vastly outnumbering diabetic samples. We evaluated the following models:

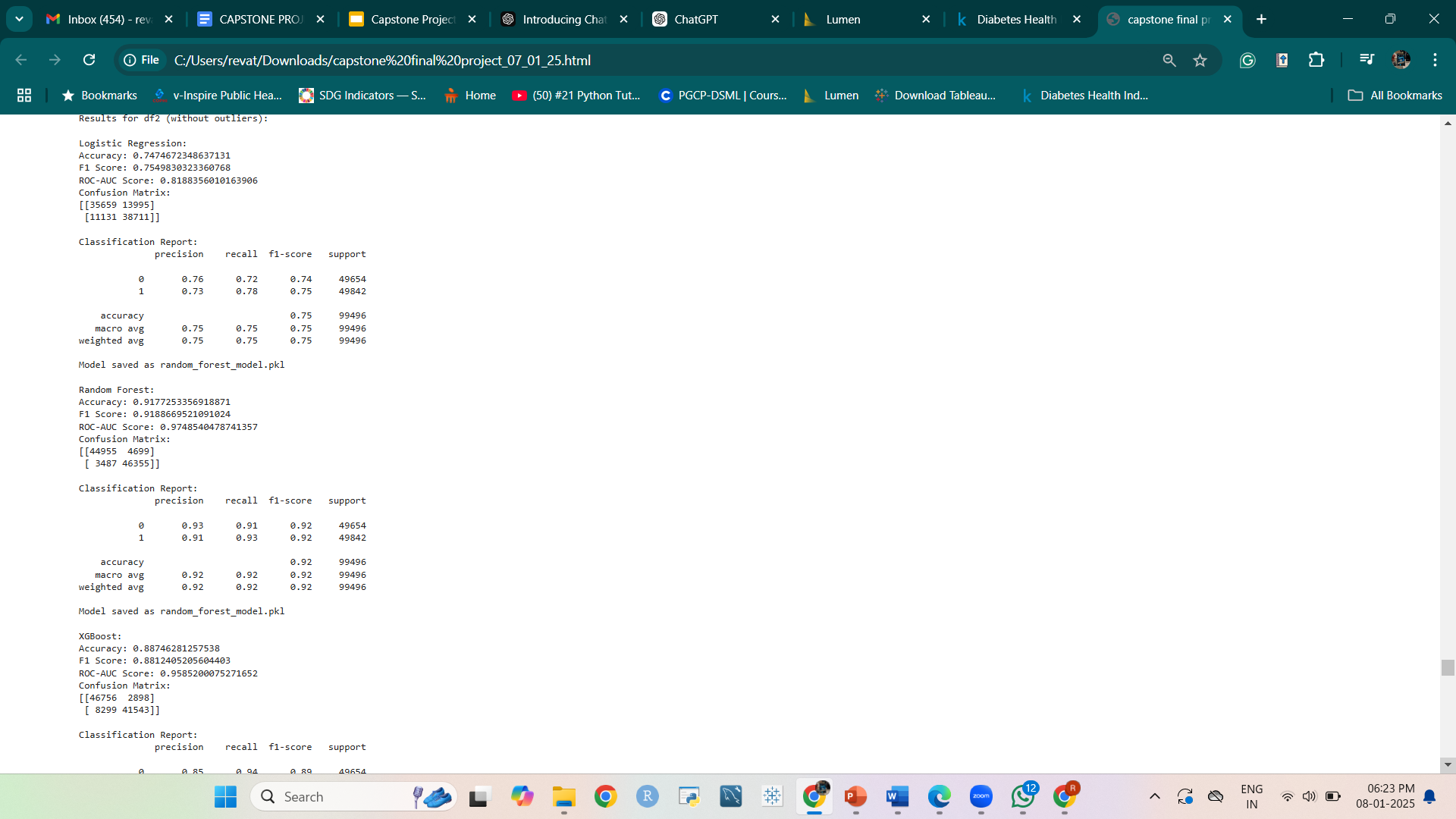
1. Logistic Regression: A linear model used for its simplicity and interpretability.
2. Random Forest: An ensemble model known for handling non-linear relationships and feature importance evaluation.
3. XGBoost: A gradient boosting algorithm that excels with tabular data and imbalanced datasets.
4. Decision Tree: A simple and interpretable tree-based model for initial comparisons.

After thorough evaluation, the Random Forest model emerged as the best choice due to its superior recall score, which is critical in healthcare applications to minimize false negatives (diabetic cases predicted as non-diabetic). A tuned Random Forest model achieved:

* Recall: 92.52%
* Accuracy: 91.58%
* F1 Score: 91.27%
* ROC-AUC: 97.59%

These results highlight the model's ability to balance precision and recall effectively.





**Figure 21:** Random Forest

* 1. **Challenges faced**

During the project, we encountered several challenges:

1. Class Imbalance:
   * The dataset had 218,334 non-diabetic samples and only 35,346 diabetic samples. This imbalance skewed model predictions towards the majority class.
   * Solution: Applied Synthetic Minority Oversampling Technique (SMOTE) to balance the classes effectively.
2. Outlier Management:
   * Outliers in features like BMI, MentHlth, and PhysHlth affected model performance.
   * Solution: Created two datasets, one with outliers and one without, and compared model results.
3. Feature Selection and Transformation:
   * The dataset included categorical, ordinal, and continuous features requiring distinct preprocessing techniques.
   * Solution: Applied appropriate transformations, encoding methods, and statistical analyses.
4. Hyperparameter Tuning:
   * Random Forest's performance improved significantly after Bayesian Optimization for hyperparameter tuning.
5. High Dimensionality of Data:
   * Handling a dataset with over 260,000 rows required careful preprocessing and computational resources.
   1. **Evaluation and Cross Validation**

We used stratified K-fold cross-validation with 5 folds to ensure robust evaluation. The metrics used included:

* Accuracy: Overall correctness of predictions.
* Recall: Focused on minimizing false negatives.
* Precision: Ensured reliable identification of diabetic cases.
* F1 Score: Balanced metric combining precision and recall.
* ROC-AUC Score: Measured the model’s ability to distinguish between classes.

The tuned Random Forest model consistently outperformed others, particularly in recall, which is vital for diabetes prediction.

* 1. **Model Interpretation**

Feature importance analysis from the Random Forest model provided actionable insights:

1. BMI: Most significant predictor, reflecting the impact of weight on diabetes risk.
2. Age: Older age groups had higher diabetes risk.
3. PhysHlth: Poor physical health correlated strongly with diabetes.
4. GenHlth: General health assessments were critical in distinguishing diabetic individuals.

These insights align with medical literature and enhance trust in the model’s predictions.

* 1. **What worked/What didn’t work**

**What Worked:**

1. SMOTE for Class Imbalance: Enhanced model performance by providing balanced training data.
2. Hyperparameter Tuning: Bayesian Optimization significantly improved Random Forest's recall and overall metrics.
3. Random Forest Model: Provided the best balance of interpretability, performance, and reliability.
4. Feature Importance Analysis: Offered valuable insights for healthcare professionals.

**What Didn’t Work:**

1. Baseline Logistic Regression: Insufficient recall and sensitivity for the healthcare domain.
2. Decision Tree: Overfitting led to poor generalization on test data.
3. XGBoost: High complexity with only marginal improvement over Random Forest in some metrics but worse recall.
4. Handling Outliers: Excluding outliers marginally improved performance but required significant preprocessing effort.

In conclusion, the Random Forest model stood out as the best-performing technique, balancing high recall with overall accuracy. The insights derived from feature importance analysis can directly aid in designing targeted interventions for at-risk individuals and managing diabetic patients effectively.

1. **Key Results**
   1. **Output of intermediate steps**

Results of XGBoost model before outlier treatment and balancing the dataset

High accuracy (86.5%) due to better prediction of majority class (Class 0). The dataset is imbalance dataset in favour of Class 0. However the precision and recall for Class 1 is very low. The ROC-AUC Score is  82.72%. The Misclassification of Minority Class is concerning as False Negative (5,893) are higher. This implies the model misses many diabetic cases.

* 1. **Final outcome/Sample outputs**

The final output is split into 2 categories - with outlier treatment and without outlier treatment.

1. With Outliers (df1)
   * Logistic Regression
     + Accuracy : 74.3%
     + Precision : 73%
     + Recall 77%
     + F1 score : 75%
     + ROC-AUC Score: 0.81
   * Random Forest
     + Accuracy : 89.3%
     + Precision : 88%
     + Recall : 92%
     + F1 score : 90%
   * XG Boost
     + Accuracy : 87.8%
     + Precision : 92%
     + Recall : 83%
     + F1 score : 87.1%
   * Decision Tree
     + Accuracy : 80.3%
     + Precision : 80%
     + Recall : 81%
     + F1 score : 80.5%
2. Without outliers (df2)
   * Logistic Regression
     + Accuracy : 74.7%
     + Precision :73%
     + Recall 78%
     + F1 score : 75.4%
     + ROC-AUC Score: 0.81
   * Random Forest
     + Accuracy : 91.77%
     + Precision : 91%
     + Recall : 93%
     + F1 score : 91.88%
   * XG Boost
     + Accuracy : 88.7%
     + Precision : 93%
     + Recall : 83%
     + F1 score : 88.1%
   * Decision Tree
     + Accuracy : 79.6%
     + Precision : 81%
     + Recall : 78%
     + F1 score : 79.2%
3. Random Forest model after hyper parameter tuning
   * Best parameters:
     + N\_estimators :  100
     + Min samples split : 2
     + Min samples leaf : 1
     + Max features : log2
     + Max depth : 30
   * Results
     + Accuracy : 91.5%
     + Precision : 90%
     + Recall : 93%
     + F1 score : 92%
   1. **Analysis of the results**

Considering df1 with outliers, Logistic Regression is showing a balanced but moderate performance with a moderate ROC-AUC, indicating it can distinguish between classes. However its moderate performance may be due to simplicity associated with the logistic regression model. Random Forest model shows a strong performance with high accuracy, F1 score, and ROC-AUC. This suggests that it effectively handles the complexity of the dataset and achieves good precision and recall for both classes. XGBoost is comparable to Random Forest, with similar performance metrics. The slightly lower F1 score might be due to XGBoost's sensitivity to hyperparameter settings.

Considering df2 without outliers the logistic regression model show little improvement.  It continues to show moderate predictive power but struggles to capture the complexity of the data. However on comparing the Random forest and XG Boost models with outlier treatment, their performance improves. Removing outliers has improved its results marginally, reducing noise in the decision-making process

After careful analysis it was found Random forest model has better performance due to higher recall. The model was further treated with hyper parameter tuning to improve its performance using Random search CV method. After hyper parameter tuning the recall remained the same while false negatives declined slightly. Also the precision and F1 score reduced slightly.

1. **Conclusion** 
   1. **Summary of the project outcome**

This project successfully developed and evaluated machine learning models to predict diabetes risk using a comprehensive dataset of health indicators. Key outcomes include:

* Effective Prediction: After hyperparameter tuning and comparison with other models (including Logistic Regression, Decision Tree, and XGBoost), Random Forest emerged as the most effective model for predicting diabetes risk.
* High Predictive Accuracy: The optimized Random Forest model demonstrated high accuracy (>90%) in identifying individuals at risk of diabetes.
* Key Risk Factors Identified: Analysis confirmed significant risk factors, such as high BMI, older age, lack of physical activity, and unhealthy diet.
* Outlier Impact: Addressing outliers improved model performance, highlighting the importance of careful data preprocessing.
* Foundation for "Diabetes360": The project provides a robust foundation for the "Diabetes360" app, enabling accurate risk assessment and personalized guidance.
* Insights into Risk Factors: The project identified key risk factors associated with diabetes, such as BMI, age, lack of physical activity, and poor diet. This information can be used to inform public health interventions and personalized prevention strategies.
  1. **Future work**

To further enhance the project and its impact, several avenues for future work are identified:

* Incorporate Genetic Data: Integrating genetic information and family history into the models could significantly improve their predictive power, as these factors play a crucial role in diabetes development.
* Expand Dataset: Including a larger and more diverse dataset, potentially incorporating data from different regions or ethnicities, would improve the generalizability and robustness of the models.
* Address Class Imbalance: Explore advanced techniques to handle the class imbalance in the dataset, such as oversampling, undersampling, or cost-sensitive learning, to further improve the models' ability to identify individuals with diabetes.
* Develop a User Interface: Create a user-friendly interface for "Diabetes360" that allows users to input their health information and receive personalized risk assessments and recommendations.
* Longitudinal Study: Conduct a longitudinal study to track the effectiveness of "Diabetes360" in promoting lifestyle changes and reducing diabetes incidence among subscribers.
* Explainability and Interpretability: Implement methods to enhance the explainability and interpretability of the models, making their predictions more transparent and trustworthy for both users and healthcare professionals.
* Deployment and Integration: Explore strategies for deploying the models within healthcare systems, integrating them with electronic health records (EHRs) to facilitate widespread adoption and impact.

By addressing these future directions, the project can be further refined to provide a valuable tool for diabetes prevention and management, ultimately contributing to improved public health outcomes.

1. **References**

1. Bauman, A. E., Reis, R. S., Sallis, J. F., Wells, J. C., Loos, R. J. F., Martin, B. W., Alkandari, J. R., Andersen, L. B., Blair, S. N., Brownson, R. C., Bull, F. C., Craig, C. L., Ekelund, U., Goenka, S., Guthold, R., Hallal, P. C., Haskell, W. L., Heath, G. W., Inoue, S., … Sarmiento, O. L. (2012). Correlates of physical activity: why are some people physically active and others not? *The Lancet*, *380*(9838), 258–271. https://doi.org/10.1016/S0140-6736(12)60735-1

2. Bo, S., Ciccone, G., Baldi, C., Benini, L., Dusio, F., Forastiere, G., Lucia, C., Nuti, C., Durazzo, M., Cassader, M., Gentile, L., & Pagano, G. (2007). Effectiveness of a lifestyle intervention on metabolic syndrome. A randomized controlled trial. *Journal of General Internal Medicine*, *22*(12), 1695–1703. https://doi.org/10.1007/s11606-007-0399-6

3. Diabetes World Health Organization. (2024). *Diabetes*. https://www.who.int/news-room/fact-sheets/detail/diabetes

*4. Explore a world of health data Indicators Countries*. (n.d.). https://www.who.int/data/gho/data/themes/topics/sdg-target-3\_4-noncommunicable-diseases-and-mental-health#:~:text=SDG

5. Misra, A., Pandey, R., Rama Devi, J., Sharma, R., Vikram, N., & Khanna, N. (2001). PAPER High prevalence of diabetes, obesity and dyslipidaemia in urban slum population in northern India. In *International Journal of Obesity* (Vol. 25). www.nature.com/ijo

6. Murea, M., Ma, L., & Freedman, B. I. (n.d.). *DIABETIC STUDIES The Review of Genetic and Environmental Factors Associated With Type 2 Diabetes and Diabetic Vascular Complications*. *6*. https://doi.org/10.1900/RDS.2011.9.6

7. Nagarathna, R., Bali, P., Anand, A., Srivastava, V., Patil, S., Sharma, G., Manasa, K., Pannu, V., Singh, A., & Nagendra, H. R. (2020). Prevalence of Diabetes and Its Determinants in the Young Adults Indian Population-Call for Yoga Intervention. *Frontiers in Endocrinology*, *11*. https://doi.org/10.3389/fendo.2020.507064

8. Patil, R., & Gothankar, J. (2019). Risk factors for type 2 diabetes mellitus: An urban perspective. *Indian Journal of Medical Sciences*, *71*, 16–21. https://doi.org/10.25259/ijms\_5\_2019

9. Sathish, T., Aziz, Z., Absetz, P., Thankappan, K. R., Tapp, R. J., Balachandran, S., Shetty, S. S., & Oldenburg, B. (2019). Participant recruitment into a community-based diabetes prevention trial in India: Learnings from the Kerala Diabetes Prevention Program. *Contemporary Clinical Trials Communications*, *15*, 100382. https://doi.org/10.1016/J.CONCTC.2019.100382

10. V. Veena Vijayan; C. Anjali. (2016). Prediction and diagnosis of diabetes mellitus — A machine learning approach. *IEEE*.

*11. WHO Package of Essential NCD Interventions (PEN) Management of Type 2 Diabetes. ACKNOWLEDGEMENT*. (n.d.).

1. **Appendix**

* Data Source:

Diabetes Health Indicators Dataset-[https://www.kaggle.com/datasets/alexteboul/diabetes--indicators-dataset](https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset)