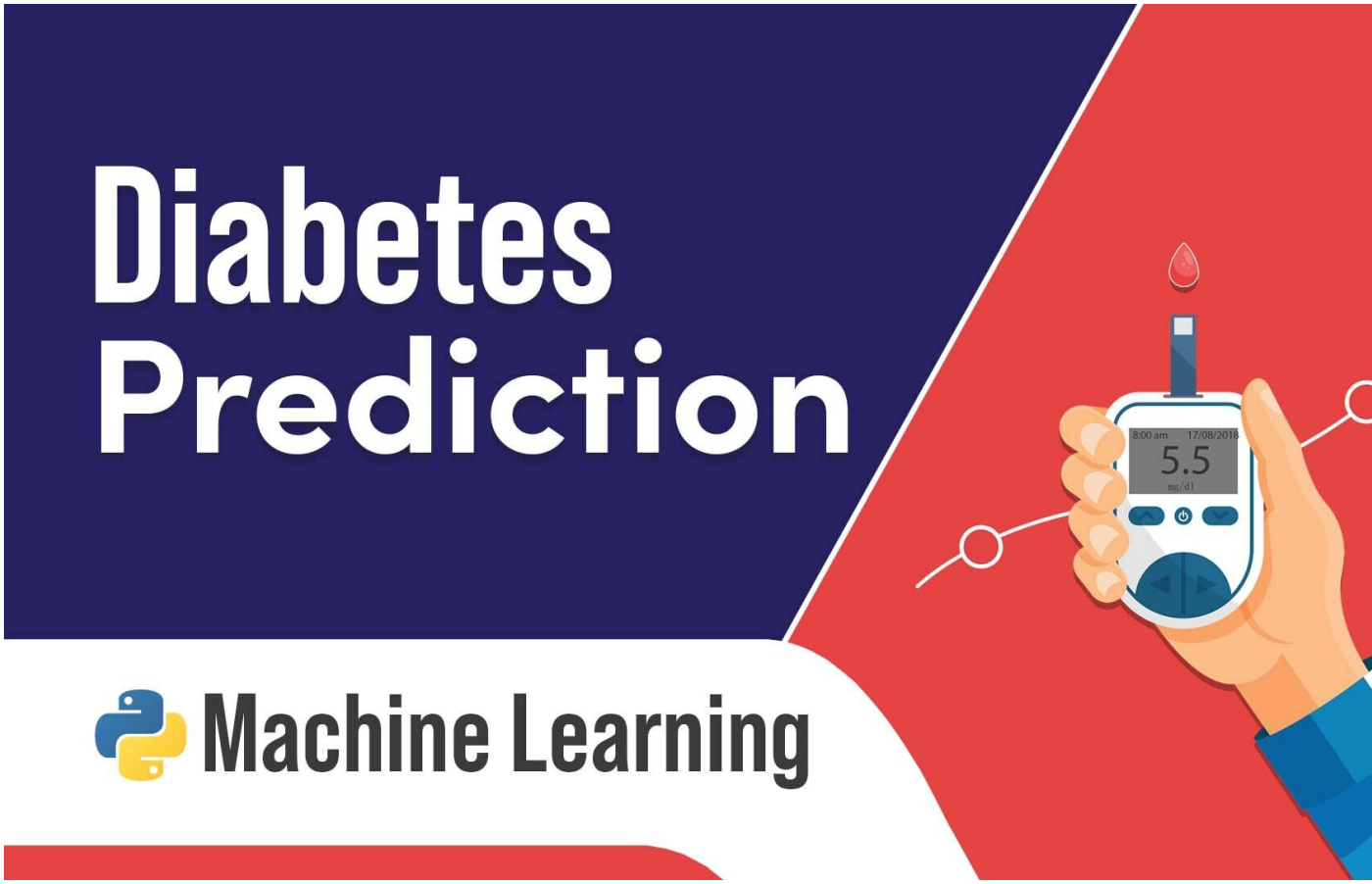


Predictive Modeling for Diabetes Risk

sobon menghorng

February 27, 2024

Contents

1 Main Objective **5**
1.1 Benefits of the Analysis 5

2 Brief Description of the Dataset **6**
2.1 Summary of Attributes 6
2.2 Outline of Analysis Objectives 8

3 Data Exploration and Prepossessing **8**
3.1 Descriptive Statistics 8
3.1.1 Overall Information 8
3.1.2 Features 9
3.1.3 Lifestyle and Health 9
3.1.4 Healthcare and General Health 9
3.1.5 Demographics 10
3.2 Prepossessing 10
3.2.1 Transformations 10
3.3 Data Cleaning 10
3.3.1 Handling Duplicates 10
3.3.2 Handling Missing Values 10
3.3.3 Handling the Outliers 12
3.3.4 Uni-variate Analysis 12
3.3.5 Bi-variate Analysis 13
3.4 Bivariate Analysis 14
3.4.1 Diabetic Characteristics by Past Medical Conditions 15
3.4.2 Education, Income, Age and Diabetes Risk 16
3.4.3 Body Mass Index (BMI) and Diabetes Risk 17
3.4.4 Mental and Physical Health (MentHlth, PhysHlth) vs Diabetes Risk 17
3.4.5 General Health (GenHlth) vs Diabetes Risk 18
3.5 Assessment of Statistical Significance and Association 18
3.5.1 Contingency Table (Cross-tabulation) 19
3.5.2 Chi-Square Test of Independence 19
3.6 Model Recommendations 20
3.7 Results 20
3.8 Conclusion 21

4 Machine Learning Model **21**
4.1 Overview of Imbalanced Dataset 21
4.2 Train-Test Split 21
4.3 Logistic Regression 22
4.4 Decision Trees 22
4.5 Naive Bayes 23
4.6 Random Forest 24
4.7 Oversample/Undersample Strategy 24
4.7.1 Cross-validation 25
4.7.2 Stratified K-Fold CV 25
4.7.3 Oversample/Undersample and Cross-Validation 25
4.7.4 Optimising for Recall 27

4.8 Baseline (No Oversampling) . . . 28
4.8.1 Model Evaluation . . . 29
4.9 Random Resampling Imbalanced Datasets . . . 30
4.9.1 Random Oversampling . . . 30
4.9.2 Imbalanced-Learn Pipeline . . . 30
4.9.3 Model Evaluation . . . 31
4.10 Random Undersampling Imbalanced Datasets . . . 33
4.11 SMOTE (Synthetic Minority Oversampling Technique) . . . 33
4.11.1 Cross Validation Results . . . 33
4.12 Performance Metrics . . . 34
4.13 Undersampling using Tomek Links . . . 35
4.13.1 Combining SMOTE and Tomek Links . . . 36
4.13.2 Model Evaluation . . . 36

5 Performance Comparison **38**
5.1 Model Performance Summary . . . 38
5.1.1 Logistic Regression . . . 38
5.1.2 Naive Bayes . . . 38
5.1.3 Random Forest and Decision Tree . . . 38
5.2 Best Model Selection . . . 39
5.3 Hyperparameter Tuning . . . 39
5.4 Model Evaluation Metrics . . . 40
5.5 AUC-ROC Curve . . . 40

6 Results **41**
6.1 Prediction Accuracy . . . 41
6.2 Identifying Risk Factors . . . 41
6.3 Feature Subset Evaluation . . . 41
6.4 Short Form Development . . . 41

7 Conclusion **42**

8 Referrals **43**

9 Appendix: Python Notebook **44**

List of Figures

1 Combined Countplot and Pie Chart of Diabetes_binary 9
2 Visualization of Missing Values 11
3 Box Plot Visualization of Numerical Features 12
4 Box Plot Visualizations of Bi-variate Analysis 13
5 Relationship Between High Blood Pressure (HighBP), High Cholesterol (HighChol), and Diabetes Risk 15
6 Education, Income, Age Distribution, and Diabetes Risk 16
7 BMI Distribution and Diabetes Risk 17
8 Mental and Physical Health vs Diabetes Risk 17
9 General Health and Diabetes Risk 18
10 Traditional and emerging risk factors that explain the increased risk of adverse events in patients with HF and associated diabetes 18
11 Pie chart showing the distribution of diabetic and non-diabetic individuals 21
12 Illustration of Train-Test Split with Stratified Sampling 22
13 Illustration of Logistic Regression Model 22
14 Illustration of Decision Tree Model 23
15 Illustration of Naive Bayes Model 23
16 Illustration of Random Forest Model 24
17 Illustration of Oversampling and Undersampling Techniques 24
18 Illustration of Cross-validation 25
19 Illustration of Stratified K-Fold Cross-Validation 25
20 Illustration of Data Leakage in Cross-Validation 26
21 Optimising for Recall 27
22 Baseline Performance without Oversampling 28
23 Confusion matrix of each model 29
24 Performance with Random Oversampling 31
25 Confusion matrix and scores 32
26 SMOTE illustration 33
27 Bar graph of model performance 34
28 confusion matrix for each model 35
29 Tomek Links illustration 36
30 Performance comparison of models 37
31 Hyper-Parameter Optimization 39
32 AUC-ROC Curve for logistic regression model with Class weights: 0.7311 40
33 Bar plot showing the coefficients of features in the logistic regression model for predicting diabetes risk. 41
34 Python Notebook 44

List of Tables

1 Contingency Table for Diabetes and Other Variables 19
2 Model Evaluation Metrics for Baseline (No Oversampling) 29
3 Model Evaluation Metrics with Random Oversampling 33
4 Performance comparison of different models with various sampling methods 38

Predictive Modeling for Diabetes Risk

SOBON MENGHORNG

February 27, 2024

1 Main Objective

The main objective of the analysis outlined in this analytics is to develop predictive models for diabetes risk. The focus is on creating models that can identify individuals at risk of developing diabetes, allowing for early diagnosis and intervention. This predictive modeling aims to assess the likelihood of an individual developing diabetes based on various factors such as age, education, income, location, race, and other social determinants of health.

1.1 Benefits of the Analysis

The benefits of this analysis are multifaceted and cater to both individuals and broader stakeholders in the healthcare system:

- **Early Intervention and Lifestyle Changes:** Predictive models enable early identification of individuals at risk, allowing for timely lifestyle interventions. Early diagnosis can lead to proactive measures such as weight management, healthy eating, and increased physical activity, which are known to mitigate the harms of diabetes.
- **Effective Resource Allocation:** Healthcare resources can be efficiently allocated based on predicted risks. High-risk individuals can receive targeted interventions and closer monitoring, optimizing the use of healthcare resources and potentially reducing the overall economic burden of diabetes.
- **Public Health Planning:** Public health officials can use the predictive models to plan and implement targeted public health campaigns. By understanding the specific demographics and social determinants associated with higher diabetes risk, officials can tailor interventions and education programs to reach populations most in need.
- **Cost Reduction:** Early diagnosis and intervention can lead to cost savings by preventing or delaying the onset of complications associated with diabetes. This, in turn, can reduce the economic burden on both individuals and the healthcare system.
- **Awareness and Education:** The analysis helps raise awareness about diabetes risk factors and encourages individuals to seek preventive measures. It also provides an opportunity for educational initiatives aimed at informing the public about the importance of regular health check-ups and lifestyle choices.

2 Brief Description of the Dataset

The dataset used in this analysis is derived from the Behavioral Risk Factor Surveillance System (BRFSS) 2015 survey conducted by the CDC. It consists of three clean CSV files, each containing responses from participants. The target variable varies among the datasets, with one having three classes (0 for no diabetes or only during pregnancy, 1 for prediabetes, and 2 for diabetes), and the other two having binary classes (0 for no diabetes and 1 for prediabetes or diabetes). The dataset includes 21 feature variables, capturing information related to health behaviors, chronic conditions, and preventive services utilization.

2.1 Summary of Attributes

Demographic Information:

- **Sex:** (0 for Female, 1 for Male)
- **Age:** (Fourteen-level age category)
    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 or older
- **Education:** (Highest grade or year of school completed)
    1. Never attended school or only kindergarten
    2. Grades 1 through 8 or Elementary
    3. Junior High School
    4. Senior High School
    5. Undergraduate Degree
    6. Magister
- **Income:** (Annual household income from all sources)
    1. less than $10,000

2. less than $16,000
3. less than $22,000
4. less than $28,000
5. less than $35,000
6. less than $52,000
7. less than $70,000
8. $75,000 or more

Health-Related Variables:

- **HighBP:** High blood pressure diagnosis (0, 1)
- **HighChol:** Ever been told of high blood cholesterol (0, 1)
- **CholCheck:** Cholesterol check within the past five years (0, 1)
- **BMI:** Body Mass Index
- **Smoker:** Ever smoked at least 100 cigarettes (0, 1)
- **Stroke:** Ever told of having a stroke (0, 1)
- **HeartDiseaseorAttack:** Ever reported having coronary heart disease or myocardial infarction (0, 1)
- **PhysActivity:** Physical activity or exercise in the past 30 days (0, 1)
- **Fruits:** Consume fruit 1 or more times per day (0, 1)
- **Veggies:** Consume vegetables 1 or more times per day (0, 1)
- **HvyAlcoholConsump:** Heavy alcohol consumption (0, 1)
- **AnyHealthcare:** Health care coverage (0, 1)
- **NoDocbcCost:** Unable to see a doctor due to cost in the past 12 months (0, 1)
- **GenHlth:** Would you say that in general your health is?
    1. Excellent
    2. very good
    3. Good
    4. Fair
    5. Poor
- **MentHlth:** Number of days with poor mental health in the past 30 days (0 30)
- **PhysHlth:** Number of days with poor physical health in the past 30 days (0 30)
- **DiffWalk:** Serious difficulty walking or climbing stairs (0, 1)

2.2 Outline of Analysis Objectives

Prediction of Diabetes:

- Evaluate the dataset's ability to accurately predict diabetes status based on survey responses.
- Assess the performance of predictive models using features such as demographics, health behaviors, and chronic conditions.

Identification of Key Predictive Factors:

- Identify risk factors most predictive of diabetes by analyzing feature importance.
- Understand the relationship between specific variables and the likelihood of diabetes.

Subset Selection for Accurate Prediction:

- Explore whether a subset of risk factors can be used to accurately predict diabetes.
- Utilize feature selection techniques to identify the most informative variables.

Creation of a Short Form for Predictive Purposes:

- Investigate the feasibility of creating a condensed set of questions (short form) that maintains predictive accuracy.
- Use feature selection methods to identify a subset of questions that are most indicative of diabetes risk.

3 Data Exploration and Prepossessing

3.1 Descriptive Statistics

3.1.1 Overall Information

- The dataset comprises 229,474 observations.
- The target variable `Diabetes_binary` has a mean of approximately 0.15, indicating that around 15% of the population has diabetes.

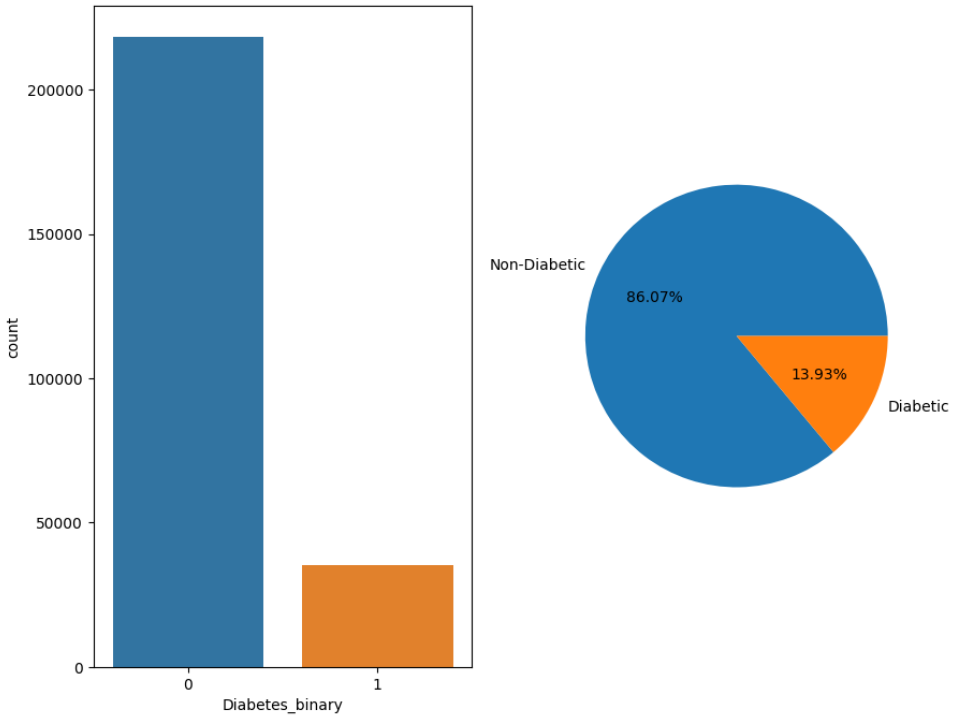


Figure 1: Combined Countplot and Pie Chart of Diabetes_binary

3.1.2 Features

- `HighBP` and `HighChol` are binary variables, with approximately 45% and 44% of the population having high blood pressure and high cholesterol, respectively.
- `CholCheck` is almost universally checked (95.9% of the population).
- `BMI` has a mean of 28.69 with a standard deviation of 6.79.
- `Smoker` is binary, with around 46.6% of the population being smokers.
- `Stroke` has a mean of 4.5%, and `HeartDiseaseorAttack` has a mean of 10.3%.

3.1.3 Lifestyle and Health

- `PhysActivity` has a mean of 73.3%, suggesting that a significant portion of the population engages in physical activity.
- `Fruits` and `Veggies` have means of 61.3% and 79.5%, respectively, indicating that a substantial portion of the population consumes fruits and vegetables regularly.
- `HvyAlcoholConsump` is relatively low, with 6.1% of the population being heavy alcohol consumers.

3.1.4 Healthcare and General Health

- `AnyHealthcare` shows that 94.6% of the population has some form of healthcare.
- `NoDocbcCost` indicates that 9.3% of the population does not incur any healthcare costs.

- `GenHlth` (General Health) has a mean of 2.60, suggesting that, on average, people rate their general health as good to excellent.
- `MentHlth` (Mental Health) has a mean of 3.51, indicating that, on average, people report a few mentally unhealthy days.
- `PhysHlth` has a mean of 4.68, suggesting that, on average, people report a few physically unhealthy days.
- `DiffWalk` indicates that roughly 18.6% of the population reports difficulty walking.

3.1.5 Demographics

- `Sex` shows a balanced distribution between males and females.
- `Age` has a mean of 40.3 years, indicating a relatively young population.
- `Education` has a mean of 4.98, on average people have an Undergraduate Degree.
- `Income` has a mean of 5.89, on average, people earn less than $28,000 per year.

3.2 Prepossessing

3.2.1 Transformations

- Data was transformed to integers.

3.3 Data Cleaning

3.3.1 Handling Duplicates

Upon examination, it was found that the dataset contained a total of 24,206 duplicate rows.

Action Taken:

To maintain the integrity of the data and ensure unbiased analysis, the duplicate rows were systematically removed. This decision was based on the understanding that duplicate entries offer no additional insights and could potentially skew the results.

Dataset Shape:

- Before dropping duplicates: 253,680 rows.
- After dropping duplicates: 229,474 rows.

3.3.2 Handling Missing Values

Detecting missing values in a dataset is facilitated by pandas through functions like `isna()`, `isnull()`, and `notna()`. For comprehensive details on handling missing data in pandas, refer to the documentation[1].

To provide an overview of missing values in our dataset, we utilized the `isnull()` function. The total count of missing values was then calculated using the `sum()` function, followed by sorting the results with `sort_values()`. A bar plot was generated to visualize

[1] $https://pandas.pydata.org/pandas-docs/stable/user_guide/missing_data.html$

the distribution of missing values across the first 20 columns, considering that the majority of missing values are concentrated within this range.

Fortunately, no missing values were detected in the dataset.

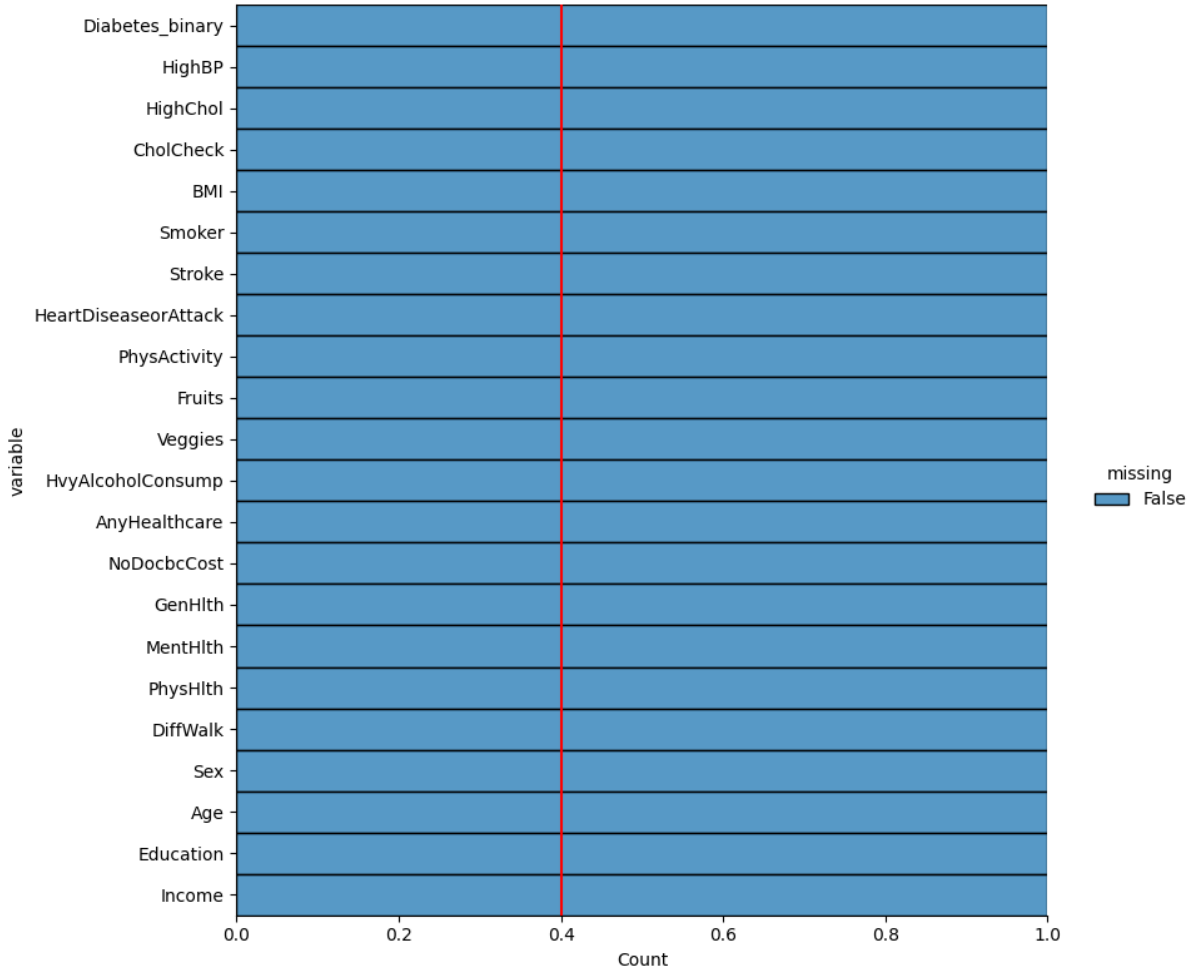


Figure 2: Visualization of Missing Values

3.3.3 Handling the Outliers

In statistics, an outlier is an observation point that is distant from other observations. An outlier can be due to some mistakes in data collection or recording, or due to naturally high variability of data points. How to treat an outlier highly depends on our data or the type of analysis to be performed. Outliers can markedly affect our models and can be a valuable source of information, providing us insights about specific behaviors.

There are many ways to discover outliers in our data. We can do Uni-variate analysis (using one variable analysis) or Multi-variate analysis (using two or more variables). One of the simplest ways to detect an outlier is to inspect the data visually, by making box plots or scatter plots.

3.3.4 Uni-variate Analysis

Uni-variate analysis involves examining the distribution of individual variables in a dataset. In this section, we utilize box plots to visualize the distribution of numerical features, excluding binary variables.

A box plot provides a graphical representation of the distribution of data through quartiles, with whiskers extending to indicate variability beyond the upper and lower quartiles. Outliers, if present, are displayed as individual points.

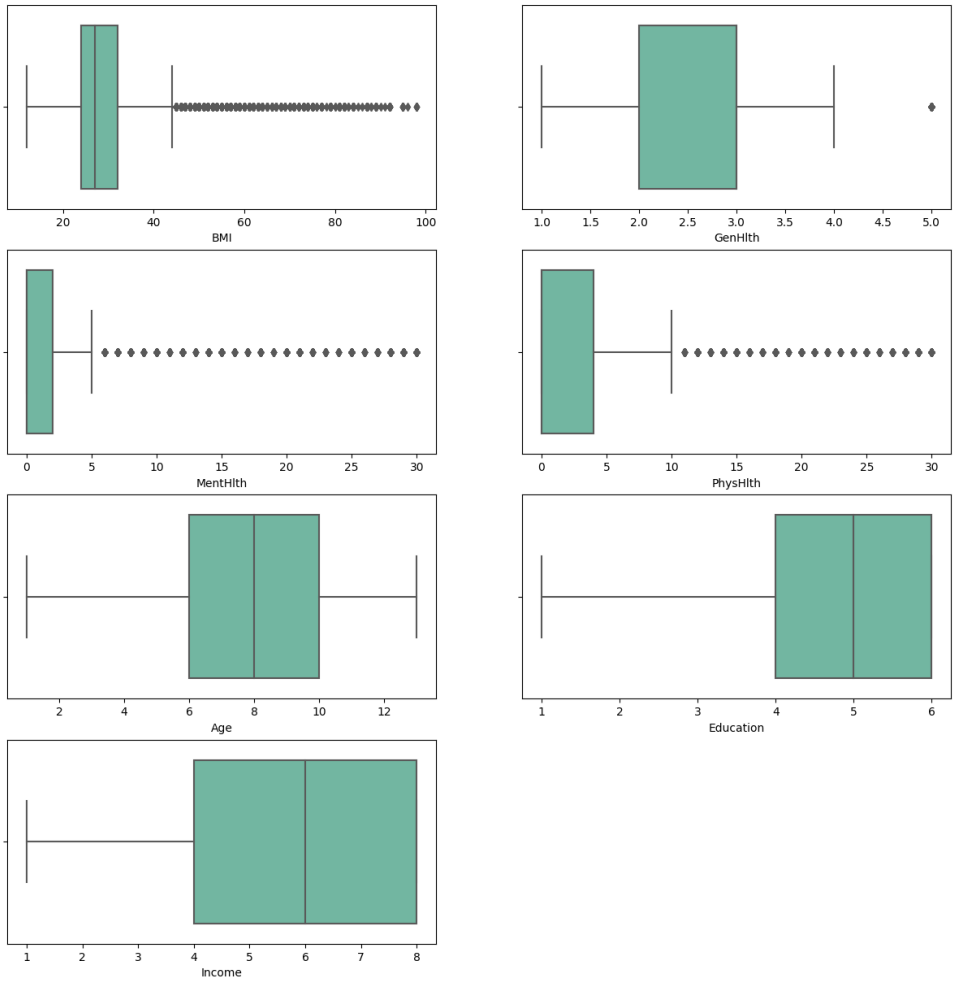
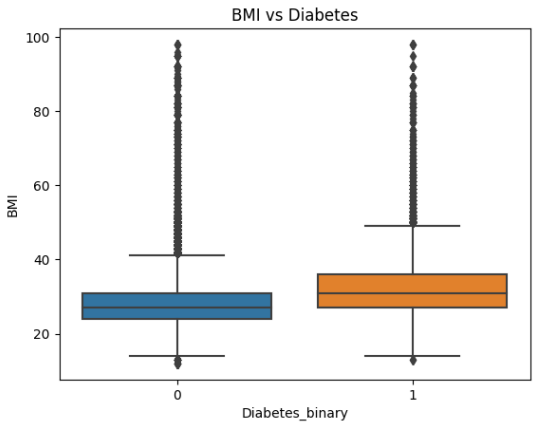


Figure 3: Box Plot Visualization of Numerical Features

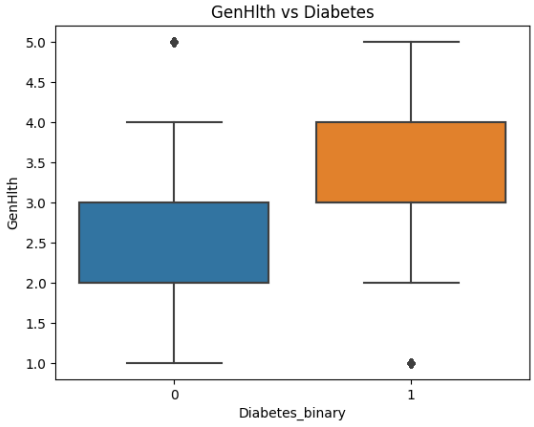
As depicted in Figure 3, we observe outliers in some of the box plots, indicating data points that significantly deviate from the rest of the population. The decision to retain or remove these outliers depends on a thorough understanding of the dataset and the specific analysis objectives. Notably, outliers in features such as 'BMI', 'GenHlth', 'MentHlth', and 'PhysHlth' may represent genuine data points and may not warrant removal, particularly if they provide valuable insights into the underlying population distribution.

3.3.5 Bi-variate Analysis

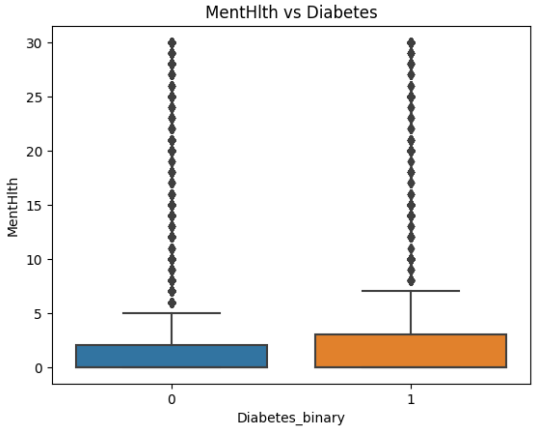
In the bivariate analysis, we examine the relationship between pairs of features, including `BMI`, *GenHlth*, *MentHlth*, and *PhysHlth*, with *Diabetes_binary*. Box plots are utilized to visualize the relationship between these parameters.



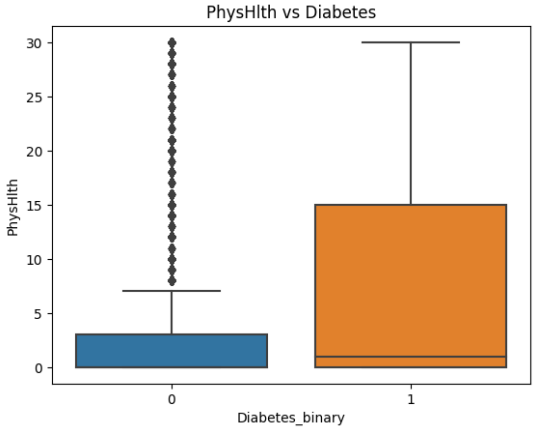
(a) `BMI` vs `Diabetes`



(b) `GenHlth` vs `Diabetes`



(c) `MentHlth` vs `Diabetes`



(d) `PhysHlth` vs `Diabetes`

Figure 4: Box Plot Visualizations of Bi-variate Analysis

BMI vs Diabetes: The box plot analysis reveals differences in BMI distribution between non-diabetic and diabetic individuals. Non-diabetic individuals tend to have BMI values clustered around the median, with a few outliers indicating higher BMI values. In contrast, diabetic individuals exhibit a wider spread of BMI values, suggesting greater

variability among them. In simpler terms, non-diabetic individuals generally have similar BMI values, while diabetic individuals display a more diverse range of body weights.

GenHlth vs Diabetes: Examining the relationship between GenHlth and diabetes reveals only one outlier point in the box plot. Interestingly, neither diabetic nor non-diabetic individuals exhibit a clear median value within the box plot. This suggests potential variability in the self-reported general health ratings across both groups.

MentHlth vs Diabetes: The box plot analysis indicates the presence of numerous outlier data points for both diabetic and non-diabetic groups in the MentHlth category. These outliers represent individuals with significantly higher or lower numbers of days with poor mental health, contributing to the wide variation observed. However, the absence of a visible median line within the box plot makes it challenging to determine the central tendency of MentHlth values for both groups.

PhysHlth vs Diabetes: For PhysHlth, the non-diabetic group exhibits extreme outlier values without a visible minimum point, suggesting a subset of individuals experiencing significantly more days with poor physical health. Conversely, the diabetic group does not display any visible outliers, indicating a more consistent distribution of PhysHlth values. The larger box size for the diabetic group suggests greater variability in the number of days with poor physical health among diabetic individuals, with the median line positioned closer to the lower range of values compared to the non-diabetic group.

Overall, the box plot analyses provide insights into the relationship between various health-related features and diabetes status, highlighting differences in distributions and variability between diabetic and non-diabetic individuals.

3.4 Bivariate Analysis

The bivariate analysis aims to explore relationships between variables, particularly focusing on how different factors relate to the risk of diabetes.

3.4.1 Diabetic Characteristics by Past Medical Conditions

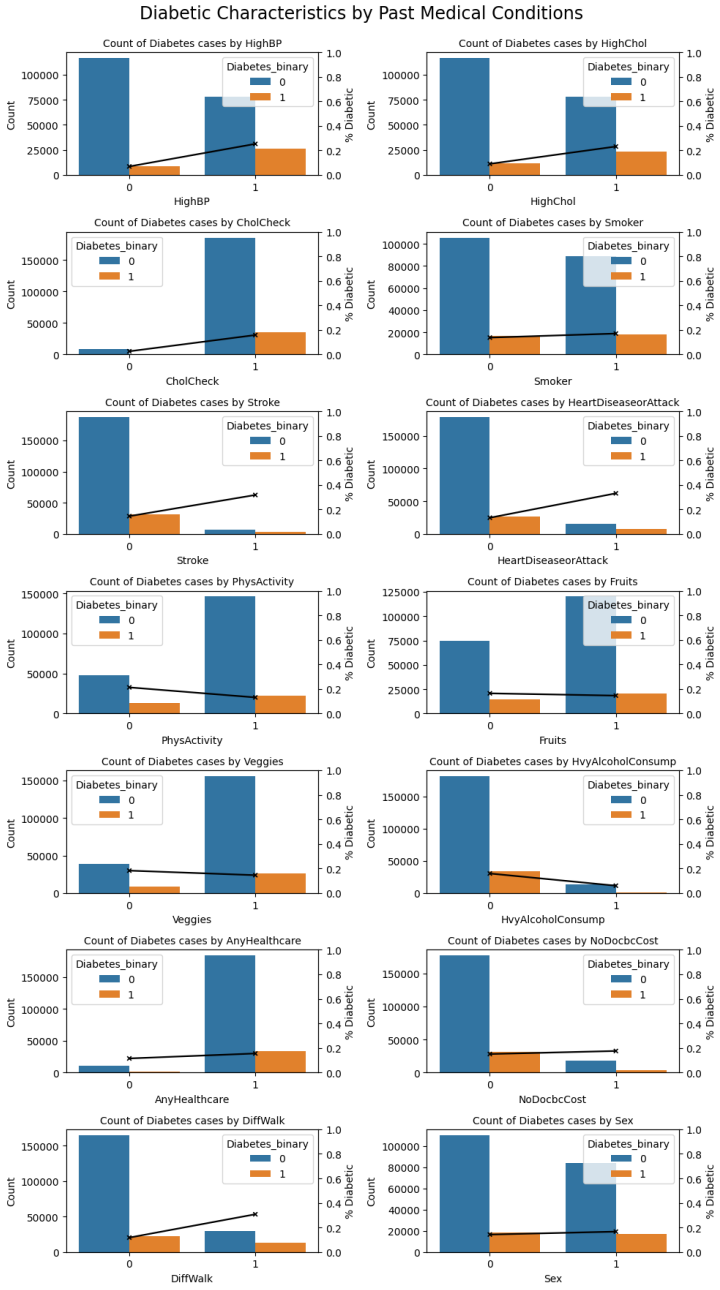


Figure 5: Relationship Between High Blood Pressure (HighBP), High Cholesterol (HighChol), and Diabetes Risk

Key Insights:

- There is a noticeable increase in diabetic risk among individuals with high blood pressure and high cholesterol levels.
- Cholesterol check seems to be associated with a higher risk of diabetes.
- Smoking does not appear to significantly increase the risk of diabetes according to the data.

- A history of stroke and heart disease indicates a greater likelihood of being diabetic.
- Heart disease attacks are more likely to occur in individuals with diabetes.
- People who engage in regular physical activity and consume fruits and vegetables daily appear to have a lower risk of diabetes, based on the data.
- The cost of doctor visits and access to healthcare does not seem to affect the likelihood of diabetes.
- Individuals who have difficulty walking are more likely to have diabetes.
- Gender does not appear to influence the likelihood of diabetes.

3.4.2 Education, Income, Age and Diabetes Risk

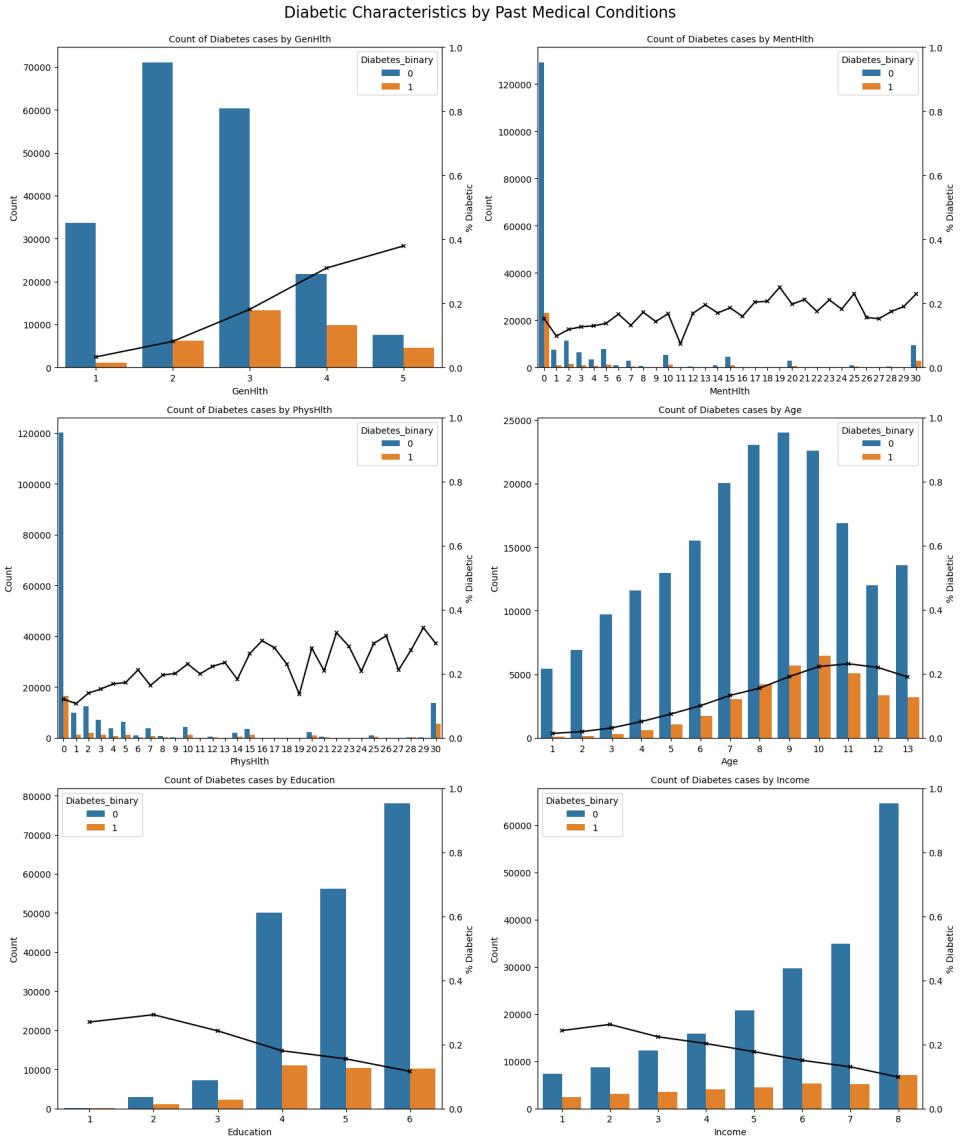


Figure 6: Education, Income, Age Distribution, and Diabetes Risk

Insights:

1. People with unhealthy lifestyles are likely to increase their risk of diabetes.
2. Increasing exposure to stress or unhealthy mental habits is associated with a higher risk of diabetes.
3. Data indicates that increasing age is associated with a higher risk of diabetes.
4. People with higher education and income are found to be at lower risk.
5. People who are physically active and eat fruits and vegetables every day are found to be at lower risk according to the data.

3.4.3 Body Mass Index (BMI) and Diabetes Risk

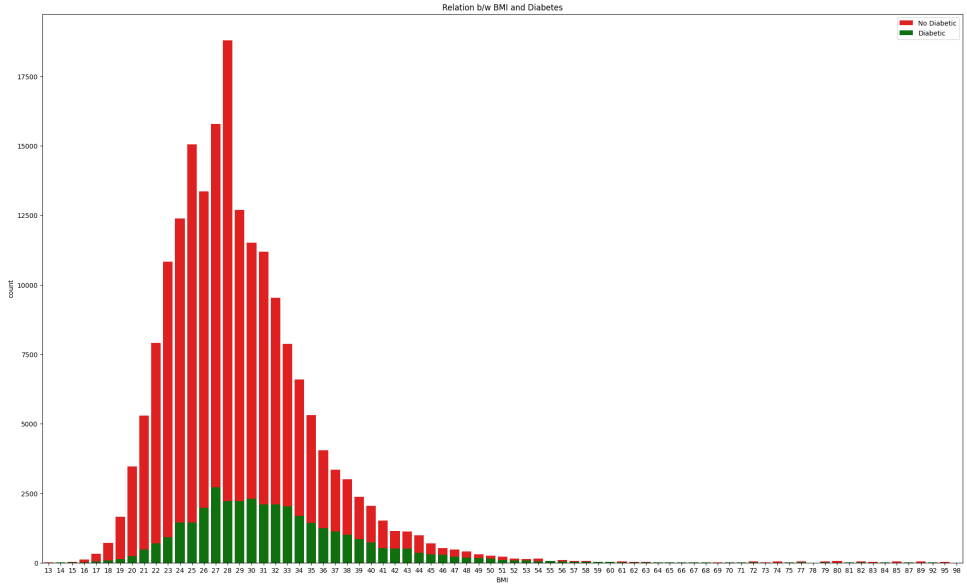


Figure 7: BMI Distribution and Diabetes Risk

Insights:

- Individuals with BMI values ranging between 24-33 are more likely to have diabetes.

3.4.4 Mental and Physical Health (MentHlth, PhysHlth) vs Diabetes Risk

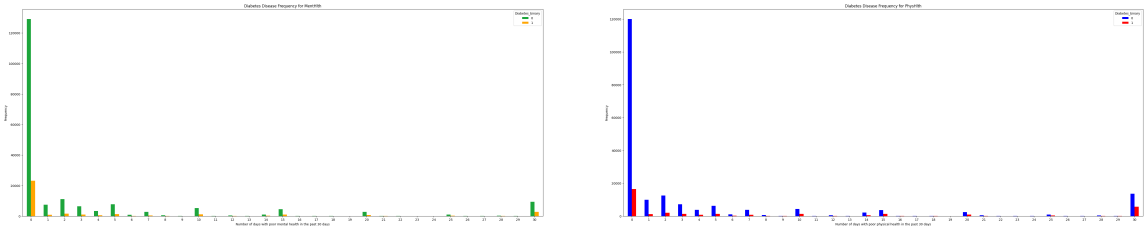


Figure 8: Mental and Physical Health vs Diabetes Risk

Insights:

- Individuals reporting poorer mental and physical health (MentHlth and PhysHlth) have a higher likelihood of diabetes.

3.4.5 General Health (GenHlth) vs Diabetes Risk

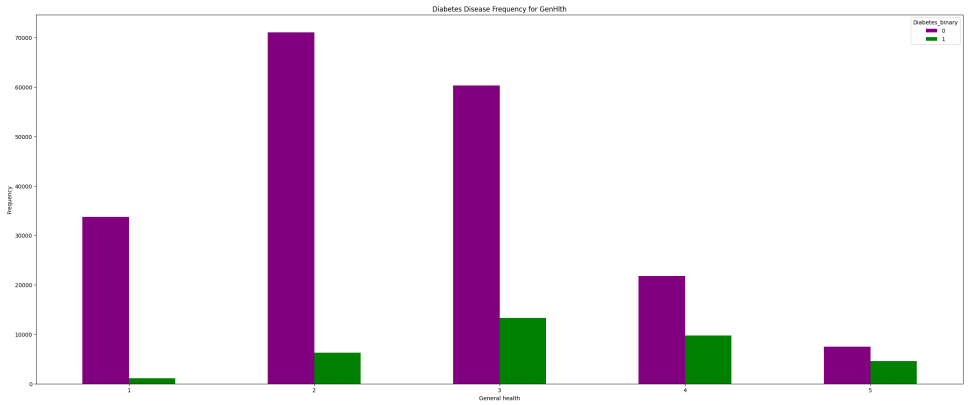


Figure 9: General Health and Diabetes Risk

Insights:

- Individuals with lower general health ratings (GenHlth) tend to have a higher risk of diabetes.

3.5 Assessment of Statistical Significance and Association

In this section, we assess the statistical significance and association between various variables and the presence of diabetes.

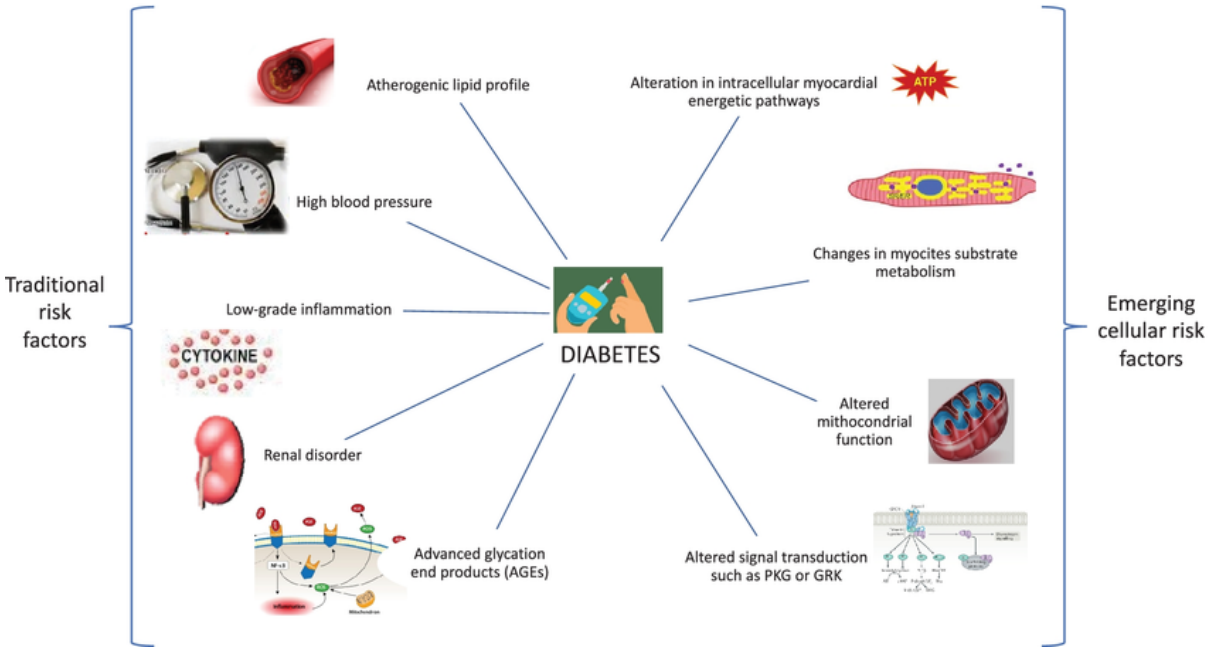


Figure 10: Traditional and emerging risk factors that explain the increased risk of adverse events in patients with HF and associated diabetes

3.5.1 Contingency Table (Cross-tabulation)

Diabetes_binary	**0**	**1**
HighBP	14	15
HighChol	1	0
CholCheck	0	0
BMI	0	1
Smoker	1	0
Stroke	0	1
HeartDiseaseorAttack	1	3
PhysActivity	1	0
Fruits	1	29
Veggies	0	0
HvyAlcoholConsump	0	2
AnyHealthcare	1	0
NoDocbcCost	0	7
GenHlth	3	5
MentHlth	4	2
PhysHlth	4	1
DiffWalk	0	0
Sex	1	1
Age	11	5
Education	6	5
Income	8	1

Table 1: Contingency Table for Diabetes and Other Variables

Interpretation: The contingency table provides a summary of the frequency distribution of diabetes (Diabetes_binary) across various factors such as HighBP, HighChol, CholCheck, BMI, and others.

3.5.2 Chi-Square Test of Independence

The chi-square test of independence is conducted to determine whether there is a significant association between the presence of diabetes and other categorical variables.

Results:

- The p-values for all tested variables are extremely low, indicating a rejection of the null hypothesis (H0) of independence.
- Therefore, we conclude that there is a significant relationship between diabetes and each of the tested variables.
- The selected features for further analysis include HighBP, HighChol, CholCheck, BMI, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk, Sex, Age, and Education.
- The variable Income is excluded from further analysis due to its lack of statistical significance.

3.6 Model Recommendations

Based on the dataset characteristics and project goals, the following machine learning models were recommended for initial exploration:

Logistic Regression:

Suitability: Ideal for binary classification tasks with categorical predictors.

Strengths: Provides interpretable results, handles categorical data efficiently, and offers insights into feature importance.

Recommendation: Logistic regression served as a baseline model, offering simplicity and interpretability. It helped identify significant risk factors associated with diabetes.

Decision Trees:

Suitability: Effective for capturing complex relationships and interactions between categorical predictors and the outcome variable.

Strengths: Handles nonlinear relationships well and provides clear feature importance.

Recommendation: Decision trees offered flexibility and could uncover patterns in the data. They helped identify key survey questions influencing diabetes risk.

Naive Bayes:

Suitability: Suitable for classification tasks with categorical predictors, assuming feature independence.

Strengths: Simple and efficient, especially with large datasets.

Recommendation: Naive Bayes provided a probabilistic approach to classification. While its assumption of feature independence may not hold strictly, it offered insights into conditional probabilities.

Random Forest:

Suitability: Robust to noise, overfitting, and handles complex relationships in the data.

Strengths: Provides high predictive performance, handles categorical data effectively, and requires minimal preprocessing.

Recommendation: Random forests were versatile and powerful, offering insights into feature importance and capturing intricate patterns in the data.

3.7 Results

- All tested variables exhibit extremely low p-values, rejecting the null hypothesis of independence.
- There is a significant relationship between diabetes and each tested variable.
- Selected features for further analysis include `HighBP`, `HighChol`, `CholCheck`, `BMI`, `Smoker`, `Stroke`, `HeartDiseaseorAttack`, `PhysActivity`, `Fruits`, `Veggies`, `HvyAlcoholConsump`, `AnyHealthcare`, `NoDocbcCost`, `GenHlth`, `MentHlth`, `PhysHlth`, `DiffWalk`, `Sex`, `Age`, and `Education`.
- Income is excluded due to its lack of statistical significance.
- Exploring logistic regression, decision trees, naive Bayes, and random forests as initial models for diabetes prediction offered a comprehensive approach. These models helped gain insights into the relationship between survey responses and diabetes risk, guiding further analysis and model refinement.

3.8 Conclusion

The analysis revealed significant findings related to diabetes risk factors, including lifestyle habits, medical history, demographic characteristics, and health-related behaviors. Each model contributed valuable insights, with logistic regression providing interpretable results, decision trees identifying complex relationships, naive Bayes offering a probabilistic perspective, and random forests delivering high predictive performance.

4 Machine Learning Model

In this section, we explore the application of various machine learning models to predict diabetes based on a dataset containing health-related features. We will discuss the performance and characteristics of logistic regression, decision trees, naive Bayes, and random forest algorithms.

4.1 Overview of Imbalanced Dataset

The dataset consists of health-related features, including indicators such as blood pressure, cholesterol levels, lifestyle habits, and demographic information. The target variable, *"Diabetes_binary"*, indicates whether an individual has diabetes or not.

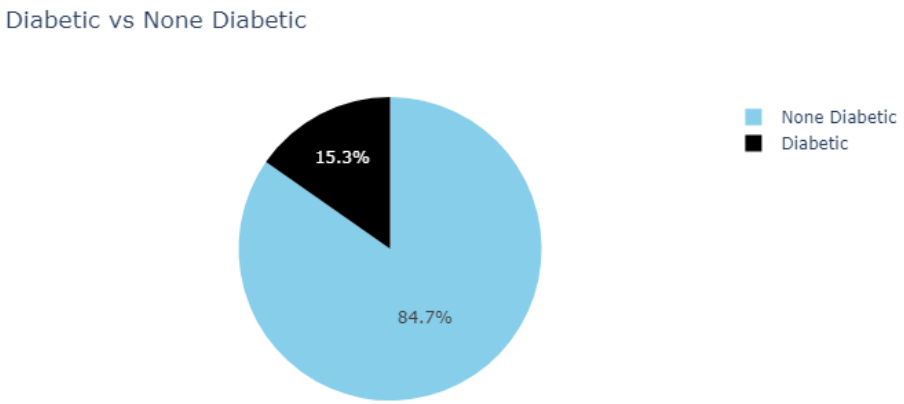


Figure 11: Pie chart showing the distribution of diabetic and non-diabetic individuals

The dataset is imbalanced, with 84.71% of individuals labeled as non-diabetic and only 15.29% as diabetic. This imbalance poses a challenge for traditional machine learning algorithms, as they may exhibit biased behavior towards the majority class. Blindly predicting the majority class would yield an accuracy of 84.71%, which is misleading.

4.2 Train-Test Split

To ensure an unbiased evaluation of our models, we employ stratified splitting during the train-test split. This method maintains the same class distribution in both the training and validation datasets, mitigating the risk of introducing bias due to class imbalance.

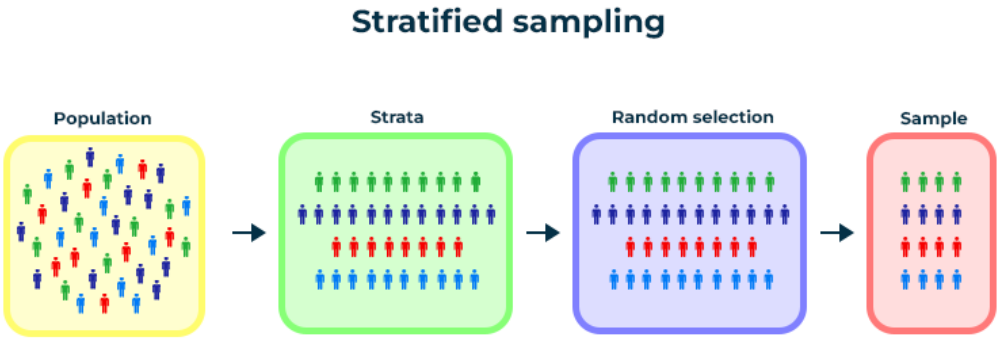


Figure 12: Illustration of Train-Test Split with Stratified Sampling

4.3 Logistic Regression

Logistic regression is a widely used algorithm for binary classification tasks. It models the probability of an outcome based on one or more predictor variables. Despite its simplicity, logistic regression can provide valuable insights into the relationship between features and the likelihood of diabetes.

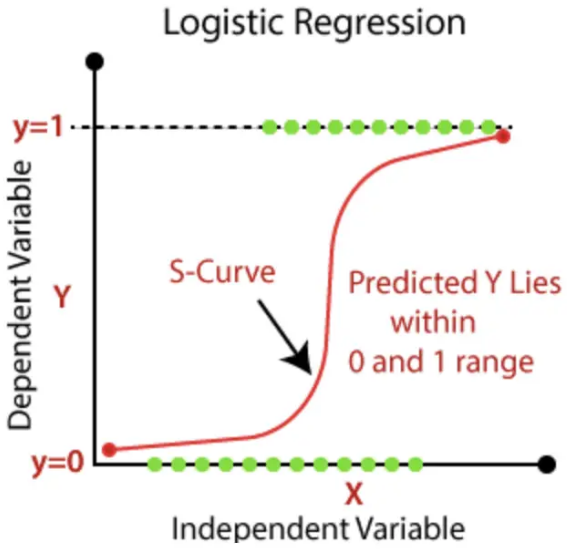


Figure 13: Illustration of Logistic Regression Model

4.4 Decision Trees

Decision trees are non-parametric models that partition the feature space into hierarchical decision rules. They are intuitive and easy to interpret, making them suitable

for understanding the factors contributing to diabetes risk. However, decision trees may suffer from overfitting, especially with complex datasets.

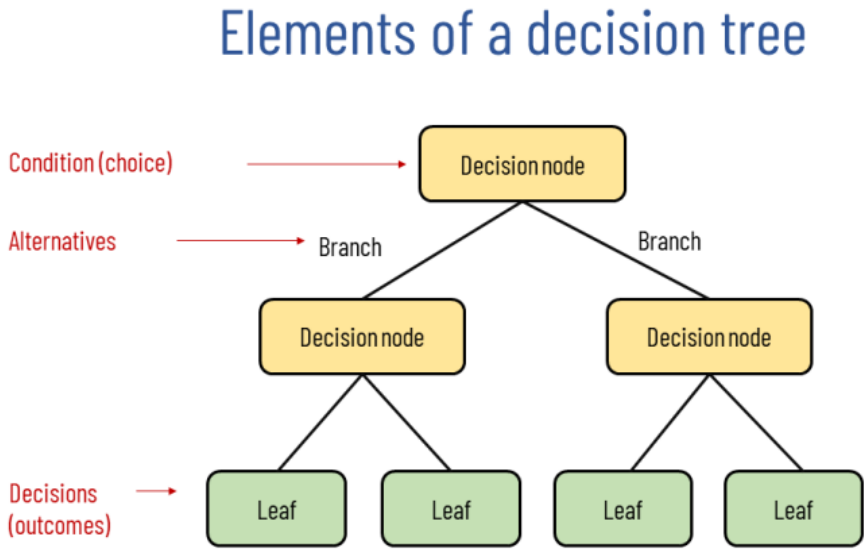
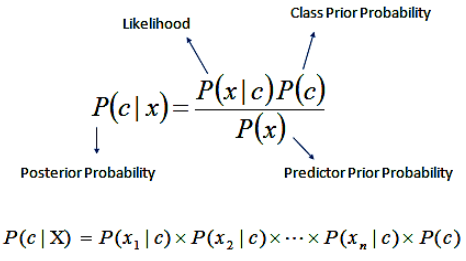


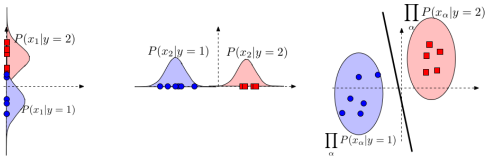
Figure 14: Illustration of Decision Tree Model

4.5 Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' theorem with strong independence assumptions between features. Despite its simplicity and computational efficiency, naive Bayes can perform well, especially with high-dimensional data. However, its assumption of feature independence may not hold in real-world scenarios.



(a) Bayes' theorem represents the probability of a label (y) resulting from a set of features (x) as a conditional probability. It is computed using the joint probability of each label occurring with the set of features and the marginal probability of the features occurring across all possible labels.

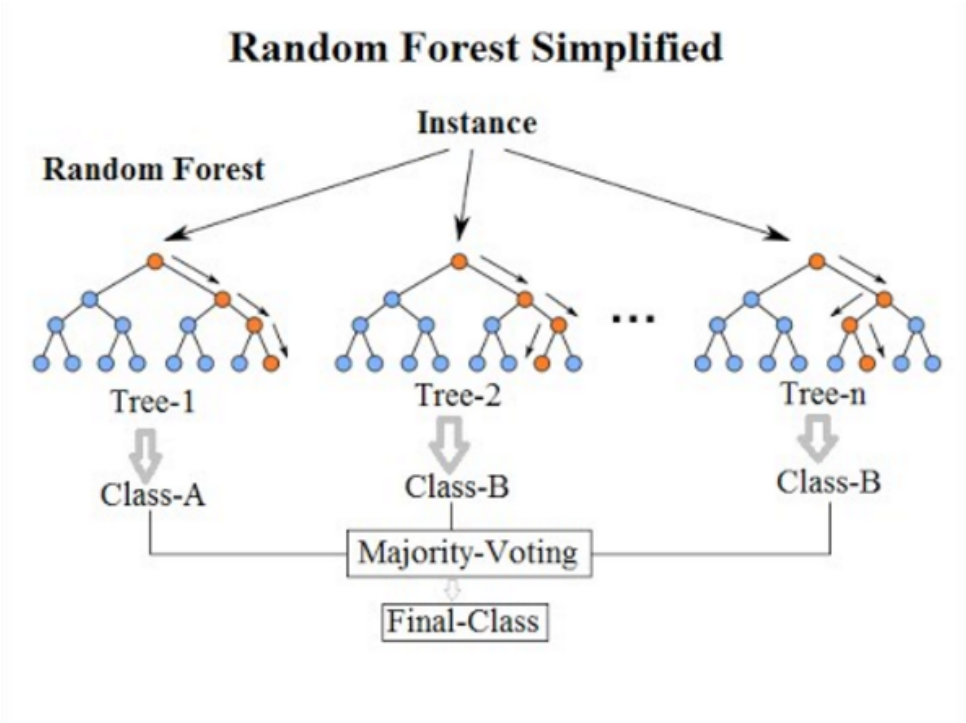


(b) Illustration behind the Naive Bayes algorithm. We estimate $P(x_\alpha|y)$ independently in each dimension (middle two images) and then obtain an estimate of the full data distribution by assuming conditional independence $P(x|y) = \prod_\alpha P(x_\alpha|y)$ (very right image).

Figure 15: Illustration of Naive Bayes Model

4.6 Random Forest

Random forest is an ensemble learning method that combines multiple decision trees to improve predictive performance and reduce overfitting. By averaging the predictions of multiple trees, random forests can capture complex relationships in the data and handle noisy or correlated features effectively.



(a) Random Forest Model

Figure 16: Illustration of Random Forest Model

4.7 Oversample/Undersample Strategy

When dealing with imbalanced datasets, it's crucial to handle oversampling or undersampling techniques properly. The general rule is to split the data into training and testing sets before applying any resampling techniques to avoid data leakage.

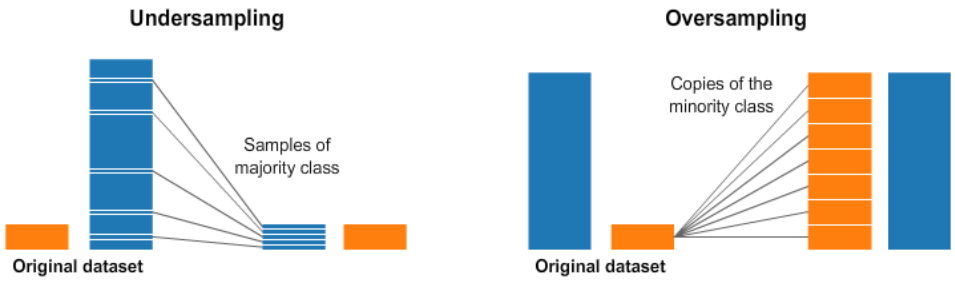


Figure 17: Illustration of Oversampling and Undersampling Techniques

4.7.1 Cross-validation

The best model is not the one that gives accurate predictions on the training data, but the one which gives good predictions on the new data and avoids overfitting and underfitting.

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation.

The purpose of cross-validation is to test the ability of a machine-learning model to predict new data.

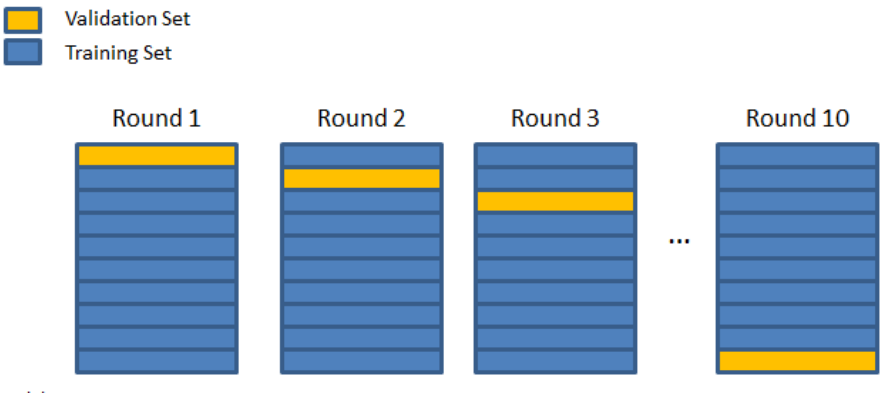


Figure 18: Illustration of Cross-validation

4.7.2 Stratified K-Fold CV

Stratified K-Fold cross-validation preserves class distributions in individual folds, making it suitable for imbalanced datasets. This approach ensures that each fold represents the dataset's class distribution accurately, providing a more reliable assessment of model performance.

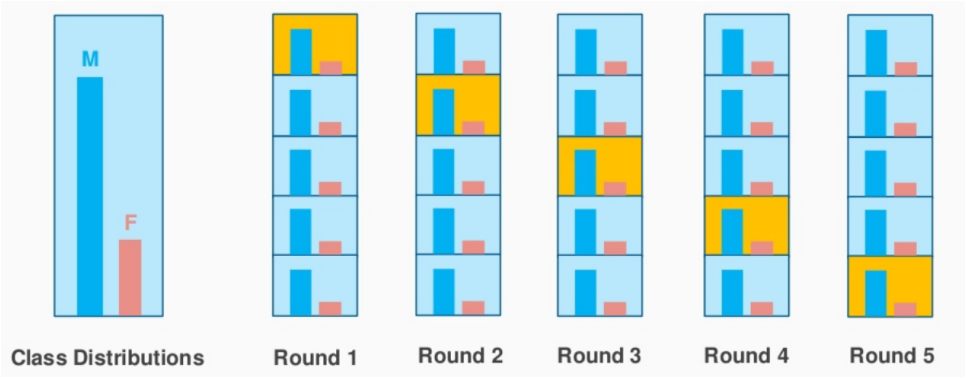


Figure 19: Illustration of Stratified K-Fold Cross-Validation

4.7.3 Oversample/Undersample and Cross-Validation

It's essential to perform oversampling or undersampling within the cross-validation loop to prevent data leakage. By resampling within each fold, we avoid biasing the validation

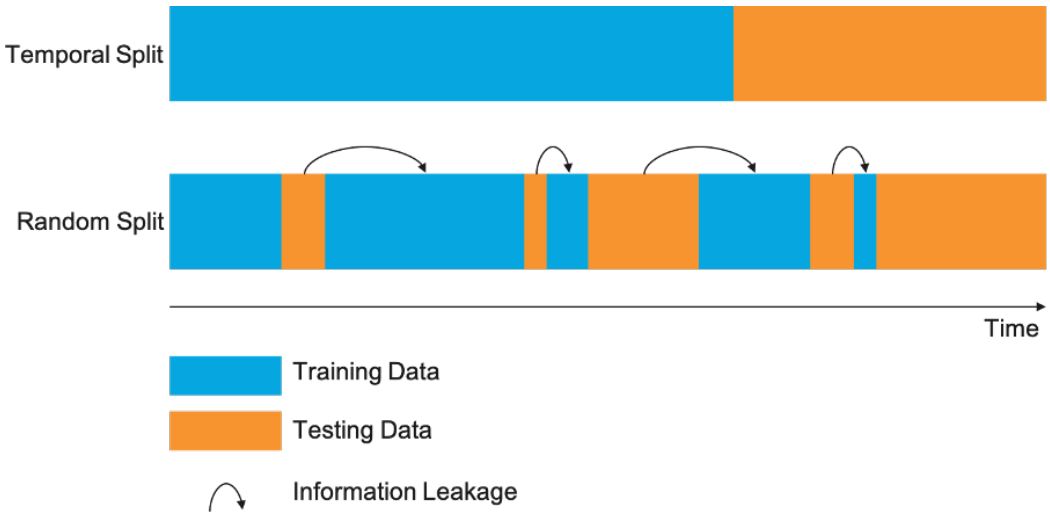


Figure 20: Illustration of Data Leakage in Cross-Validation

set and maintain the integrity of the evaluation process. Data leakage occurs when information from outside the training dataset is inadvertently used to create the model, leading to overly optimistic performance estimates. This can happen if the validation set is influenced by the training data or if features are extracted from the target variable.

4.7.4 Optimising for Recall

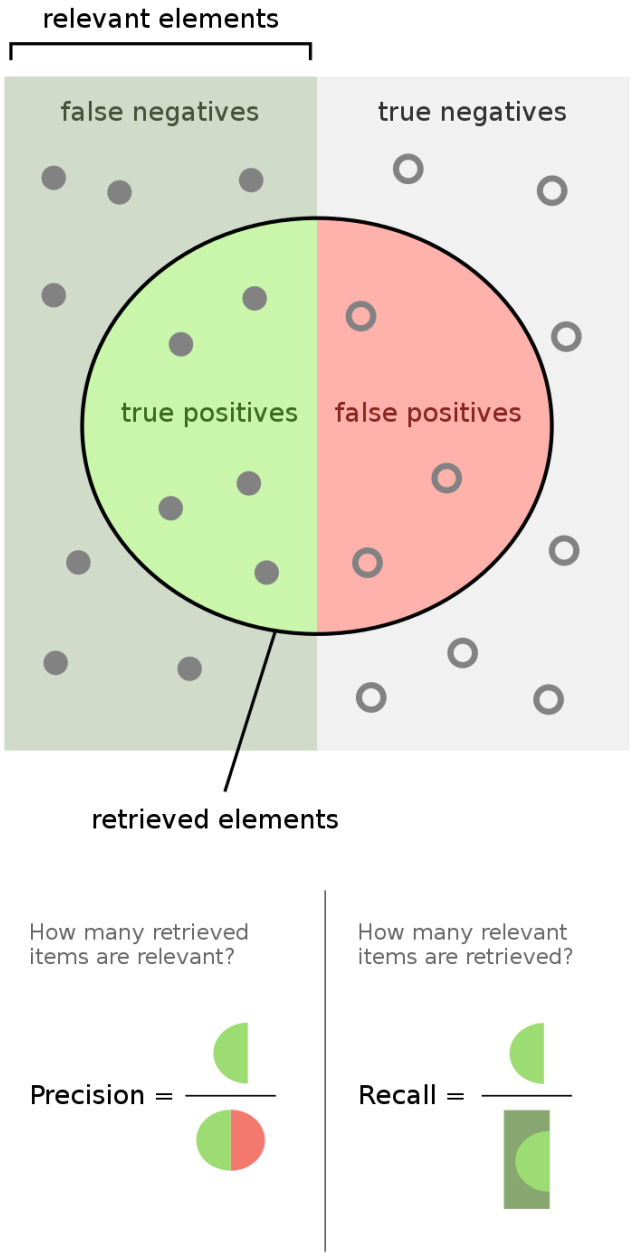


Figure 21: Optimising for Recall

Recall: **The ability of a model to correctly identify all diabetic individuals within a dataset.** It is calculated as the number of true positives (correctly identified diabetics) divided by the sum of true positives and false negatives (diabetics incorrectly classified as non-diabetic).

In the context of diabetic prediction, maximizing recall is crucial as it ensures that the model identifies as many diabetic individuals as possible, minimizing the risk of false negatives. False negatives, where a diabetic individual is incorrectly classified as non-diabetic, can have serious consequences as it may lead to delayed diagnosis and treatment.

4.8 Baseline (No Oversampling)

The baseline performance of the models without oversampling is evaluated using cross-validation recall scores.

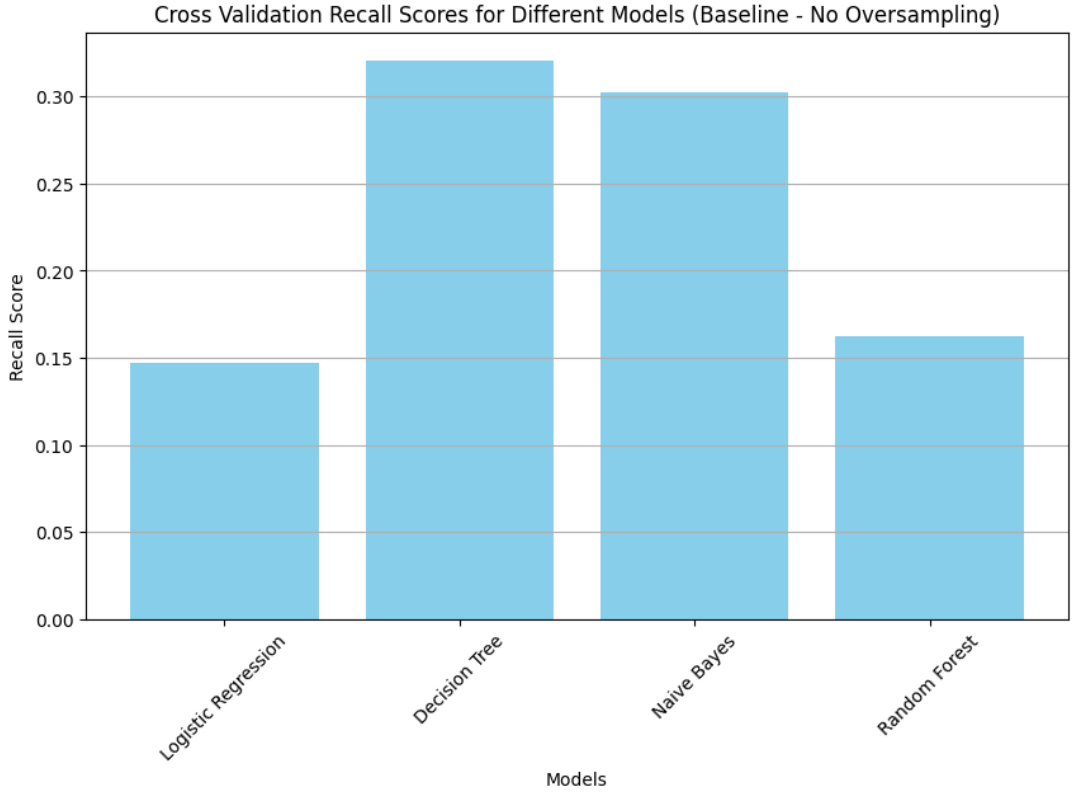


Figure 22: Baseline Performance without Oversampling

- Logistic Regression:
  - Cross Validation Recall scores: [0.1457, 0.1500, 0.1506, 0.1500, 0.1396]
  - Average Cross Validation Recall score: 0.1472
- Decision Tree:
  - Cross Validation Recall scores: [0.3168, 0.3214, 0.3210, 0.3193, 0.3256]
  - Average Cross Validation Recall score: 0.3208
- Naive Bayes:
  - Cross Validation Recall scores: [0.2971, 0.3163, 0.3051, 0.3036, 0.2890]
  - Average Cross Validation Recall score: 0.3022
- Random Forest:
  - Cross Validation Recall scores: [0.1585, 0.1732, 0.1598, 0.1673, 0.1528]
  - Average Cross Validation Recall score: 0.1623

Comparing the average cross-validation recall scores, the decision tree model performs the best with an average recall score of 0.3208, followed by logistic regression with an average recall score of 0.1472, naive Bayes with 0.3022, and random forest with 0.1623.

Model	Recall	Precision	F1 Score	Accuracy	Imbalance Method
Logistic Regression	0.1553	0.5452	0.2417	0.8510	Baseline (no oversampling)
Decision Tree	0.3222	0.2866	0.3033	0.7737	Baseline (no oversampling)
Naive Bayes	0.3081	0.4061	0.3504	0.8253	Baseline (no oversampling)
Random Forest	0.1723	0.4841	0.2541	0.8453	Baseline (no oversampling)

Table 2: Model Evaluation Metrics for Baseline (No Oversampling)

4.8.1 Model Evaluation

Confusion matrices provide a comprehensive view of the performance of a classification model. Each row of the matrix represents the instances in a predicted class, while each column represents the instances in an actual class. The diagonal elements of the matrix represent the instances that were correctly classified, while off-diagonal elements represent misclassifications.

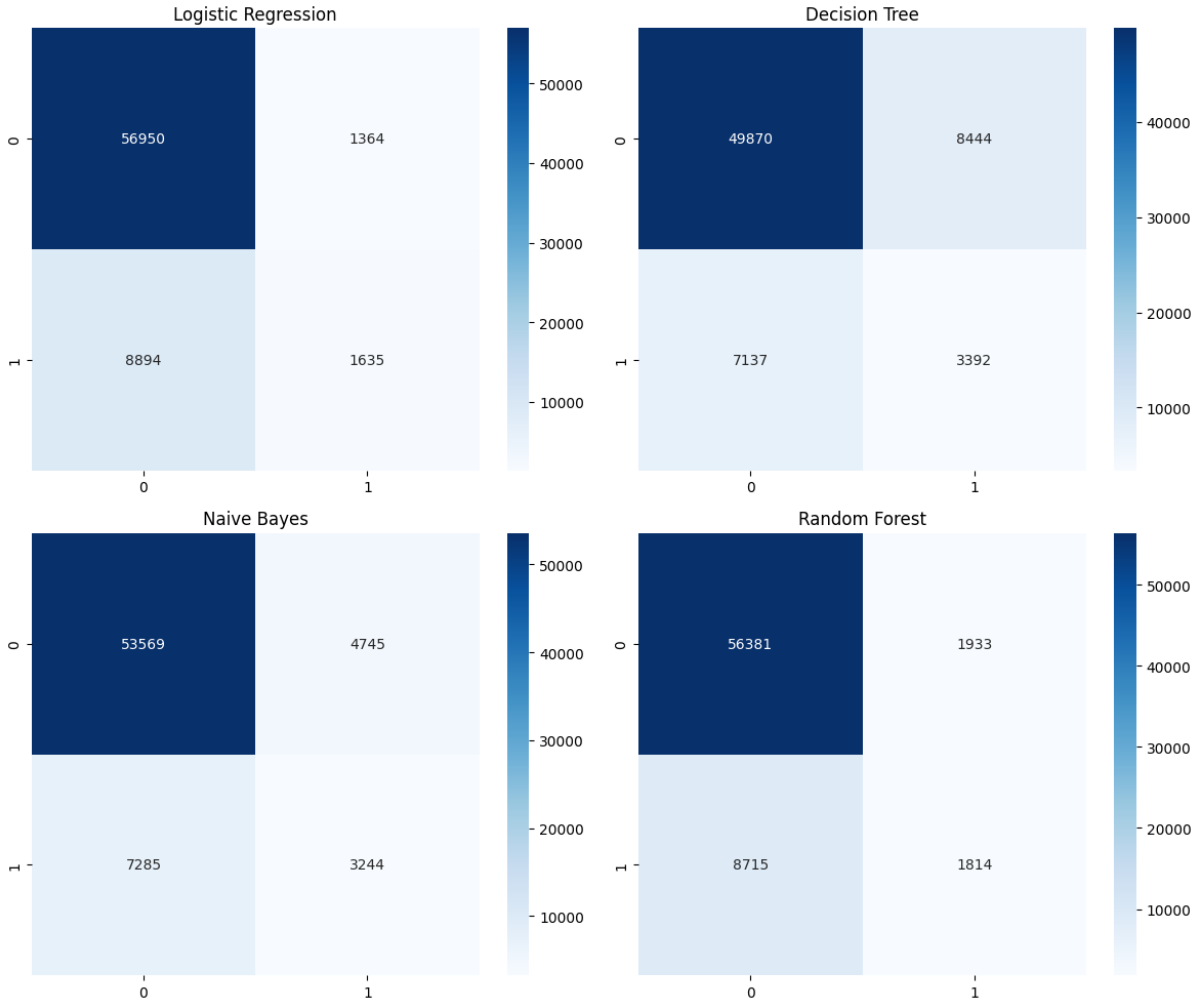


Figure 23: Confusion matrix of each model

4.9 Random Resampling Imbalanced Datasets

Resampling involves creating a new transformed version of the training dataset in which the selected examples have a different class distribution.

There are two main approaches to random resampling for imbalanced classification; they are oversampling and undersampling.

- Random Oversampling: Randomly duplicate examples in the minority class.
- Random Undersampling: Randomly delete examples in the majority class.

They are referred to as "naive resampling" methods because they assume nothing about the data and no heuristics are used. This makes them simple to implement and fast to execute, which is **desirable for very large and complex datasets.**

Important

Change to the class distribution should be only applied to the training dataset. The intent is to influence the fit of the models. The resampling is not applied to the test or holdout dataset used to evaluate the performance of a model.

4.9.1 Random Oversampling

Random oversampling may increase the likelihood of overfitting since it creates exact copies of minority class examples. For example, if every data point from the minority class is copied six times before splitting, and a three-fold validation is performed, each fold would, on average, contain two copies of each point. Consequently, a classifier might construct rules that are accurate but specific to replicated examples.

Class Distribution:

- Non-diabetic: 136,063 / 50.0% of the dataset
- Diabetic: 136,063 / 50.0% of the dataset

4.9.2 Imbalanced-Learn Pipeline

The purpose of the pipeline is to assemble several steps that can be cross-validated together while setting different parameters.

During the cross-validation process, we should split into training and validation segments. Then, on each segment, we should:

1. Oversample the minority class.
2. Train the classifier on the training segment.
3. Validate the classifier on the remaining segment.

The pipeline is a great way to do this smartly.

The `imblearn` package contains a lot of different samplers for oversampling and undersampling. These samplers cannot be placed in a standard `sklearn` pipeline.

To allow for using a pipeline with these samplers, the `imblearn` package also implements an extended pipeline which has a bunch of extra functions to do with transforming and sampling.

4.9.3 Model Evaluation

- Logistic Regression
  - **Cross Validation Recall scores:** [0.75356125, 0.77182984, 0.74211276, 0.75457875, 0.74216524]
  - **Average Cross Validation Recall score:** 0.7528495703134416
- Decision Tree
  - **Cross Validation Recall scores:** [0.28184778, 0.29757785, 0.29167515, 0.28652829, 0.27879528]
  - **Average Cross Validation Recall score:** 0.28728486988206103
- Naive Bayes
  - **Cross Validation Recall scores:** [0.68294668, 0.69082027, 0.69082027, 0.68884819, 0.68762719]
  - **Average Cross Validation Recall score:** 0.6882125209827225
- Random Forest
  - **Cross Validation Recall scores:** [0.3040293, 0.30246285, 0.3026664, 0.2964998, 0.29263329]
  - **Average Cross Validation Recall score:** 0.2996583284187599

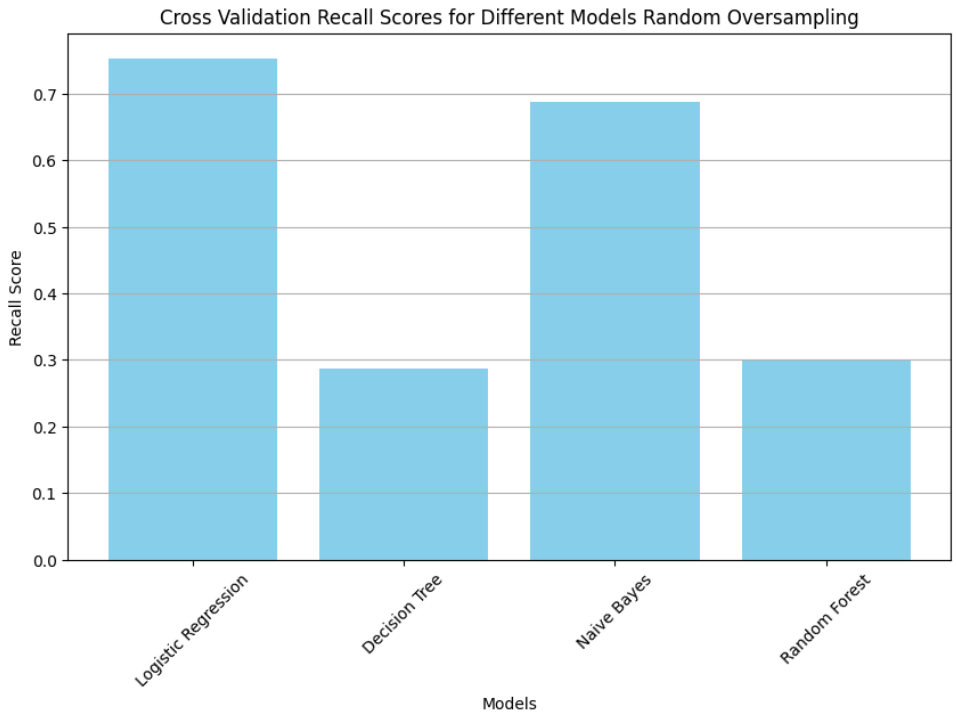


Figure 24: Performance with Random Oversampling

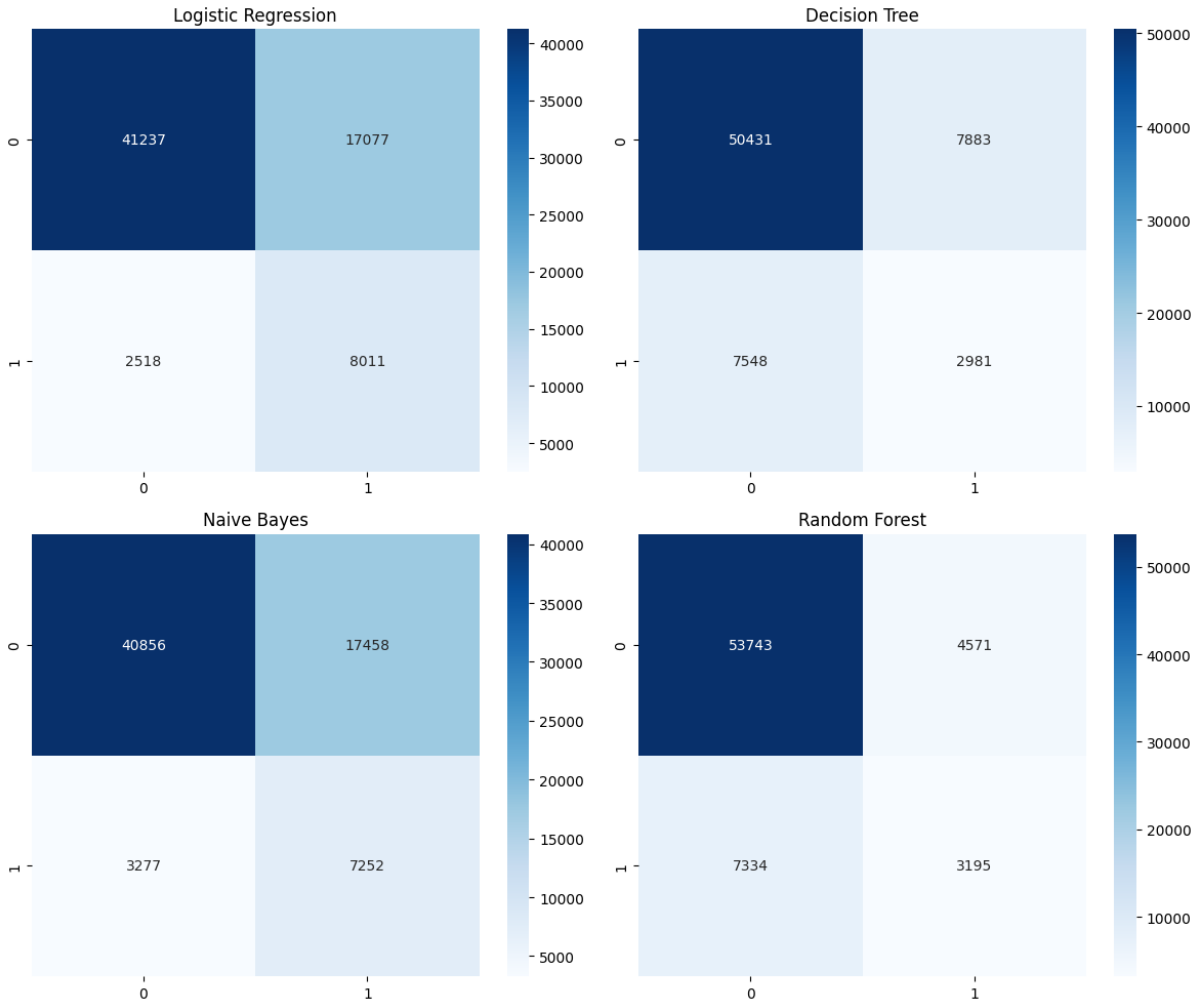


Figure 25: Confusion matrix and scores

- Among the models evaluated, logistic regression and Naive Bayes show relatively higher recall scores, indicating better performance in identifying diabetic cases.
- Random Forest has the highest accuracy but a lower recall score compared to logistic regression and Naive Bayes. This indicates a potential trade-off between accuracy and the ability to identify diabetic cases.
- Decision Tree performs the worst among the evaluated models, with lower recall, precision, and F1 score.

Model	Recall	Precision	F1 Score	Accuracy
Logistic Regression	0.760851	0.319316	0.449841	0.715367
Naive Bayes	0.688764	0.293484	0.411589	0.698807
Random Forest	0.303448	0.411409	0.349276	0.82707
Decision Tree	0.283123	0.274392	0.278689	0.775852

Table 3: Model Evaluation Metrics with Random Oversampling

4.10 Random Undersampling Imbalanced Datasets

Random undersampling involves randomly selecting examples from the majority class to delete from the training dataset. This approach may be more suitable for those datasets where there is a class imbalance although a sufficient number of examples in the minority class, such a useful model can be fit:

- Non-Diabetic: 24568 / 50.0% of the dataset
- Diabetic: 24568 / 50.0% of the dataset

With our dataset after undersampling, we have only 49136 records, so it's not the best idea to take advantage of that technique.

4.11 SMOTE (Synthetic Minority Oversampling Technique)

SMOTE (Synthetic Minority Oversampling Technique) synthesizes elements for the minority class. SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space, and drawing a new sample at a point along that line.

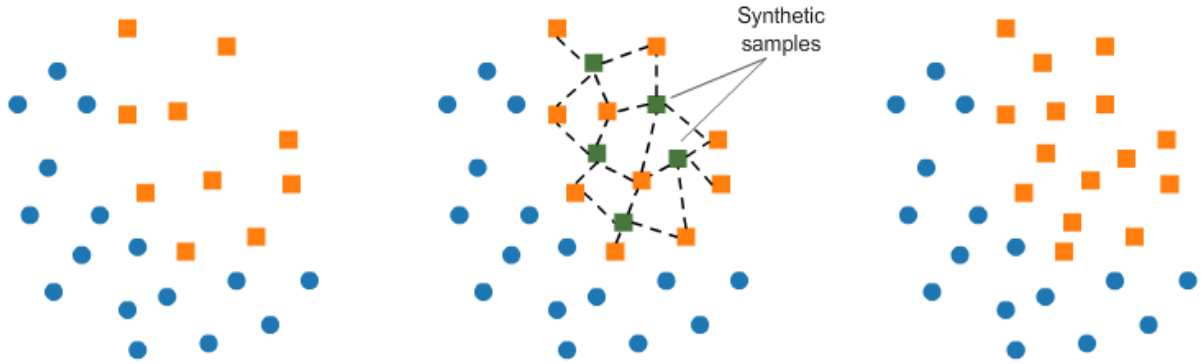


Figure 26: SMOTE illustration

4.11.1 Cross Validation Results

- **Model: Logistic Regression**
  - Cross Validation Recall scores are: [0.67236467 0.68776715 0.65703236 0.65486365 0.65506716]
  - Average Cross Validation Recall score: 0.6654189987591168

- **Model: Decision Tree**
  - Cross Validation Recall scores are: [0.4019129 0.42316304 0.41441075 0.40822141 0.40720391]
  - Average Cross Validation Recall score: 0.4109824002354025
- **Model: Naive Bayes**
  - Cross Validation Recall scores are: [0.65282865 0.66293507 0.64665174 0.64387464 0.64570615]
  - Average Cross Validation Recall score: 0.650399250582438
- **Model: Random Forest**
  - Cross Validation Recall scores are: [0.44403744 0.44657032 0.44799512 0.42490842 0.43243793]
  - Average Cross Validation Recall score: 0.4391898480032004

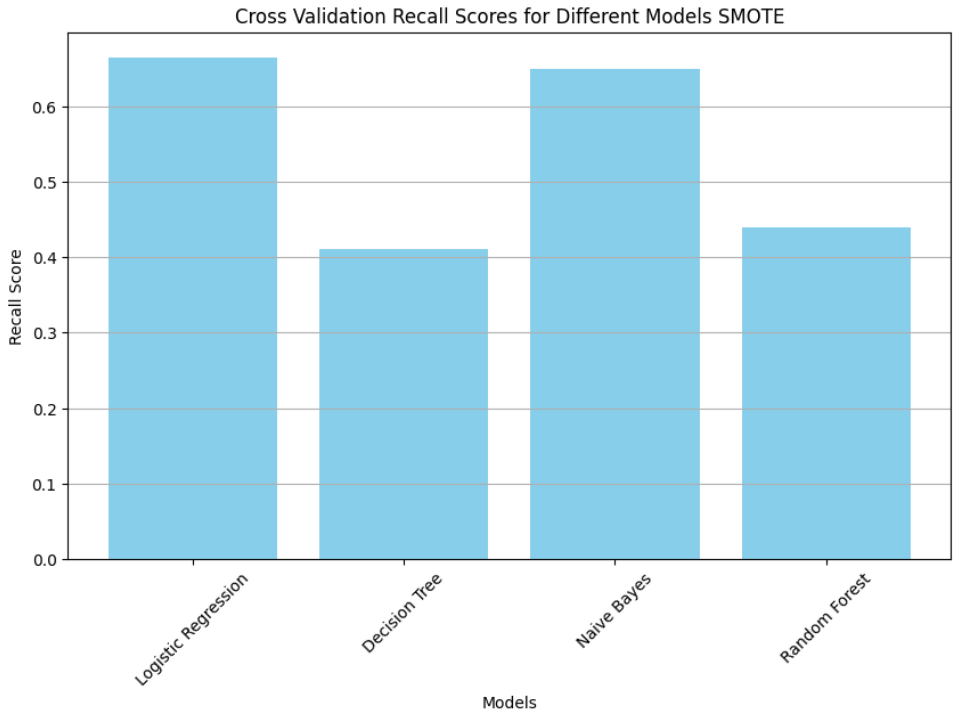


Figure 27: Bar graph of model performance

4.12 Performance Metrics

Model	Recall	Precision	F1 Score	Accuracy	Imbalance Method
Logistic Regression	0.680122	0.295336	0.411836	0.702889	SMOTE
Naive Bayes	0.662171	0.276864	0.390468	0.683817	SMOTE
Random Forest	0.444487	0.321054	0.372819	0.771277	SMOTE
Decision Tree	0.420268	0.243963	0.308717	0.712142	SMOTE

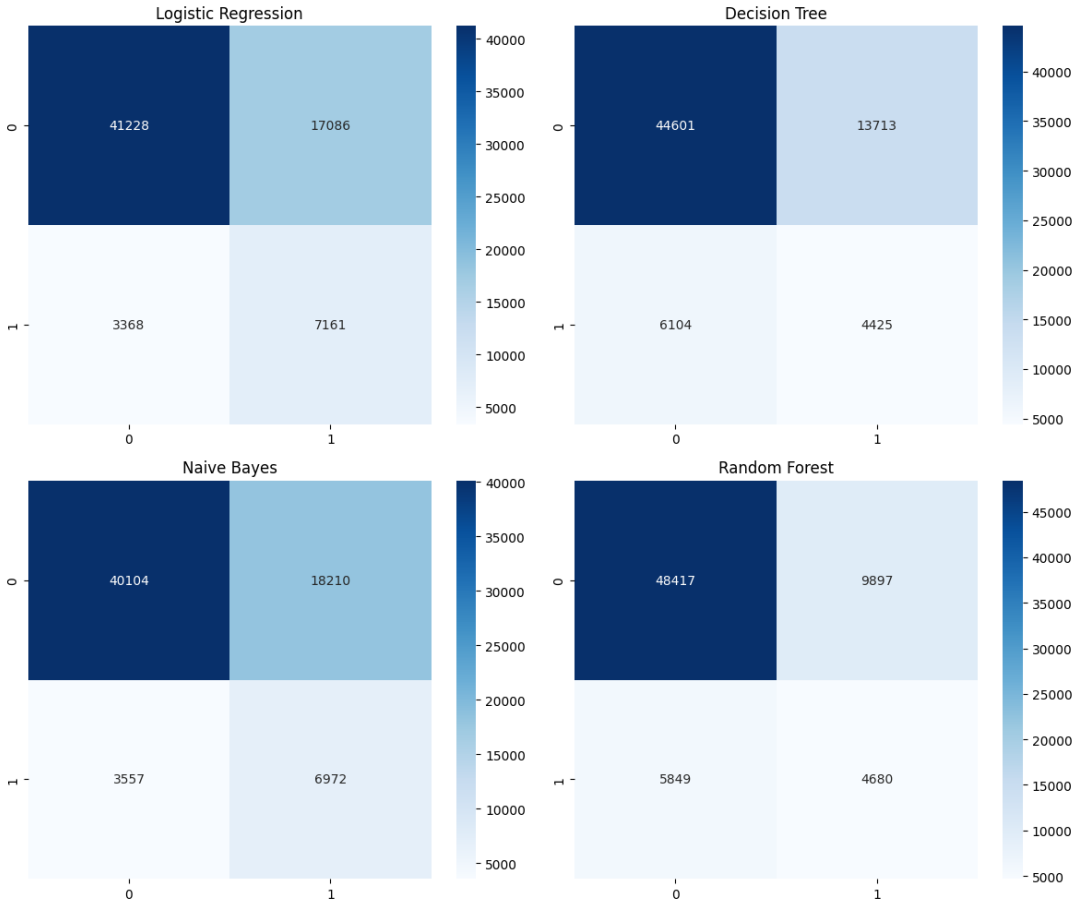


Figure 28: confusion matrix for each model

- The Random Forest model outperforms the other models in terms of recall and accuracy, while the Logistic Regression and Naive Bayes models show moderate performance across all metrics.
- All models exhibit relatively low precision, suggesting a high rate of false positives, which could be further optimized.
- The choice of model should consider the specific requirements and constraints of the application, balancing factors such as interpretability, computational complexity, and the importance of correctly identifying diabetic cases.

4.13 Undersampling using Tomek Links

Tomek Links is an under-sampling technique that was developed in 1976 by Ivan Tomek. It is one of the modifications from Condensed Nearest Neighbors (CNN). It can be used to find desired samples of data from the majority class that has the lowest Euclidean distance from the minority class data and then remove it.

Non-Diabetic: 130189 / 84.12% of the dataset
Diabetic: 24568 / 15.88% of the dataset

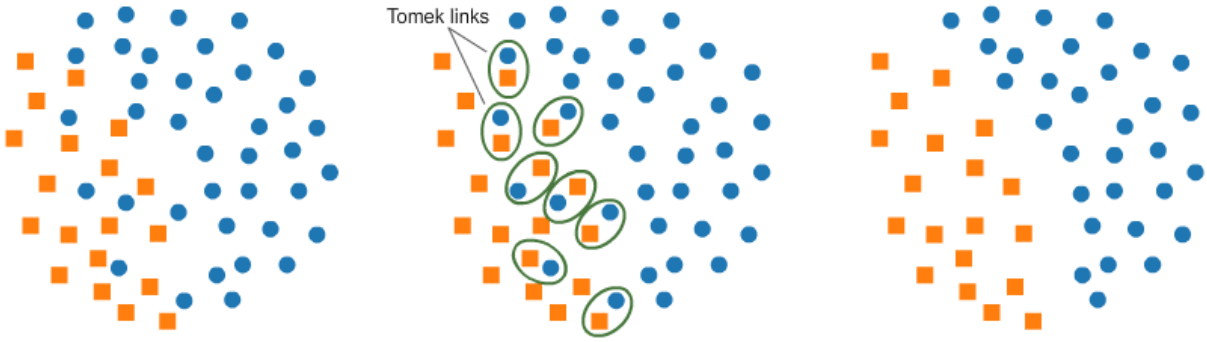


Figure 29: Tomek Links illustration

4.13.1 Combining SMOTE and Tomek Links

A combination of over-sampling the minority (abnormal) class and under-sampling the majority (normal) class can achieve better classifier performance than only under-sampling the majority class. This method was first introduced by Batista et al. (2003).

The process of SMOTE-Tomek Links is as follows:

1. Start of SMOTE: choose random data from the minority class.
2. Calculate the distance between the random data and its k nearest neighbors.
3. Multiply the difference with a random number between 0 and 1, then add the result to the minority class as a synthetic sample.
4. Repeat steps 2–3 until the desired proportion of minority class is met (End of SMOTE).
5. Start of Tomek Links: choose random data from the majority class.
6. If the random data's nearest neighbor is the data from the minority class (i.e. create the Tomek Link), then remove the Tomek Link.

4.13.2 Model Evaluation

- **Model: Logistic Regression**
  - Cross Validation Recall scores are: [0.68091168 0.70038673 0.66639528 0.66564917 0.66218966]
  - Average Cross Validation Recall score: 0.6751065031341847
- **Model: Decision Tree**
  - Cross Validation Recall scores are: [0.42551893 0.43578262 0.43211887 0.41717542 0.41269841]
  - Average Cross Validation Recall score: 0.4246588482493225
- **Model: Naive Bayes**
  - Cross Validation Recall scores are: [0.65689866 0.66537757 0.64909424 0.65323565 0.65282865]

- Average Cross Validation Recall score: 0.6554869544896006

- **Model: Random Forest**
  - Cross Validation Recall scores are: [0.46113146 0.45898636 0.46142886 0.44973545 0.44037444]
  - Average Cross Validation Recall score: 0.45433131523096926

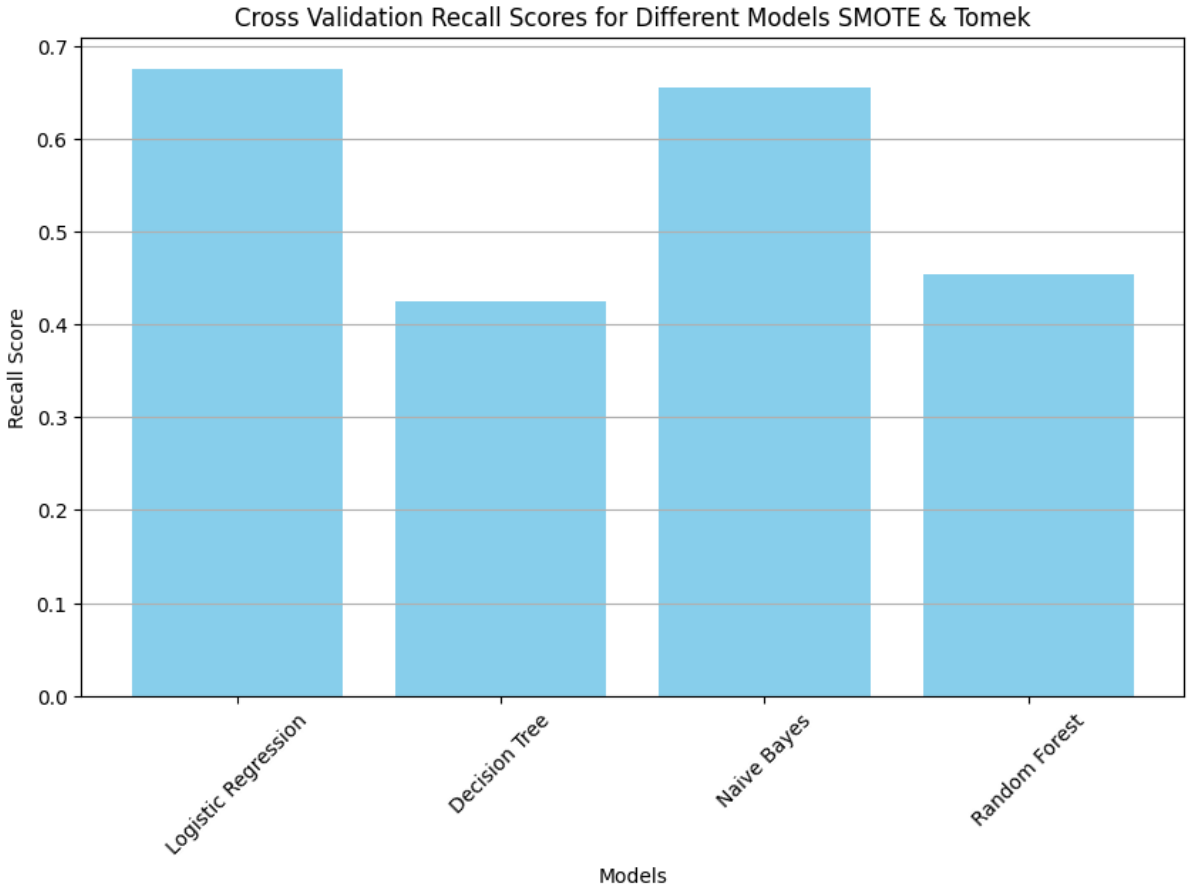


Figure 30: Performance comparison of models

5 Performance Comparison

Table 4: Performance comparison of different models with various sampling methods

Model	Recall	Precision	F1 Score	Accuracy	Imbalance Method
Logistic Regression	0.760851	0.319316	0.449841	0.715367	Random Oversampling
Naive Bayes	0.688764	0.293484	0.411589	0.698807	Random Oversampling
Logistic Regression	0.680122	0.295336	0.411836	0.702889	SMOTE
Naive Bayes	0.662171	0.276864	0.390468	0.683817	SMOTE
Random Forest	0.444487	0.321054	0.372819	0.771277	SMOTE
Decision Tree	0.420268	0.243963	0.308717	0.712142	SMOTE
Decision Tree	0.322158	0.286583	0.303331	0.773673	Baseline (no oversampling)
Naive Bayes	0.308101	0.406058	0.350362	0.825255	Baseline (no oversampling)
Random Forest	0.303448	0.411409	0.349276	0.82707	Random Oversampling
Decision Tree	0.283123	0.274392	0.278689	0.775852	Random Oversampling
Random Forest	0.172286	0.484121	0.254133	0.845329	Baseline (no oversampling)
Logistic Regression	0.155285	0.545182	0.241721	0.850994	Baseline (no oversampling)

5.1 Model Performance Summary

We evaluated several models using different sampling methods, including random oversampling, SMOTE, and baseline (no oversampling). The evaluation metrics considered include recall, precision, F1 score, and accuracy.

5.1.1 Logistic Regression

- Achieved the highest recall, precision, and F1 score across different sampling methods.
- Demonstrated consistent performance, particularly when using random oversampling.
- Recommended for its effectiveness in identifying individuals at risk of diabetes.

5.1.2 Naive Bayes

- Performed relatively well, especially in terms of recall, but with lower precision and F1 score compared to logistic regression.
- Showed consistency across different sampling methods.

5.1.3 Random Forest and Decision Tree

- Random Forest generally outperformed Decision Tree but showed varying performance across sampling methods.
- Both models exhibited lower recall compared to logistic regression and Naive Bayes.

5.2 Best Model Selection

Considering our goals of prediction accuracy and identifying risk factors, **Logistic Regression** emerges as the most suitable model. It consistently achieved high performance across different metrics and sampling methods. However, further analysis, including feature importance and interpretability, is recommended to optimize the model for specific objectives.

5.3 Hyperparameter Tuning

Hyperparameter tuning, performed using GridSearchCV, aimed to identify optimal parameter values for the logistic regression model. The best parameters obtained from the search were as follows: $C = 0.001$, $class\_weight = None$, $penalty = l1$, $random\_state = 42$, and $solver = liblinear$. These parameters were chosen based on their ability to maximize the recall score, which was found to be approximately 0.756. The recall score represents the proportion of true positive predictions among all actual diabetic cases, indicating the model's effectiveness in correctly identifying individuals with diabetes.

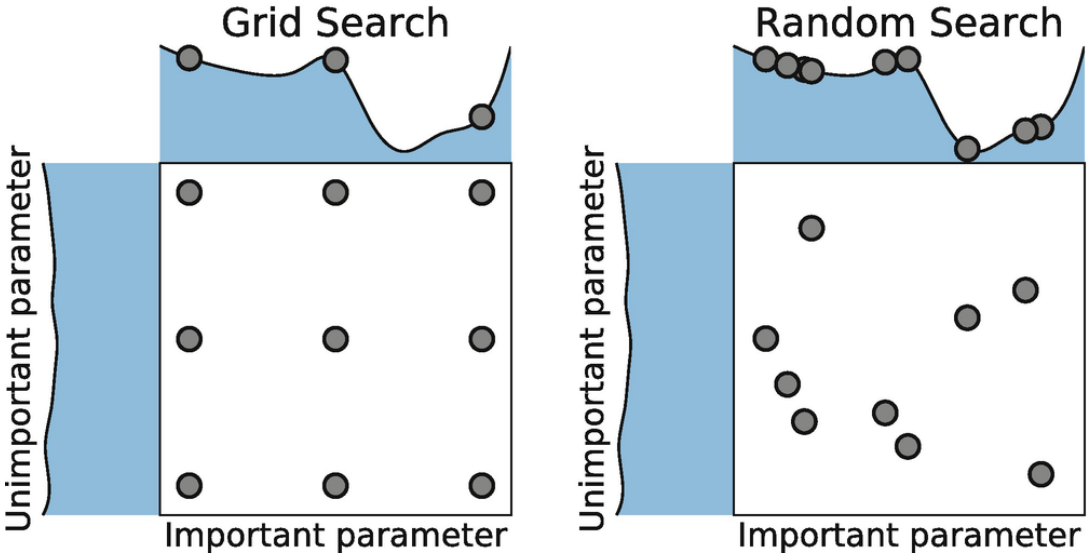


Figure 31: Hyper-Parameter Optimization

5.4 Model Evaluation Metrics

5.5 AUC-ROC Curve

The AUC-ROC (Area Under The Curve - Receiver Operating Characteristics) curve is a graphical representation of the performance of a classification model at various threshold settings. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) for different threshold values.

The AUC score measures the degree of separability between the classes. Specifically, for the logistic regression model with class weights, the AUC-ROC curve yielded a score of 0.7311. This score indicates that the model has a 73.11% chance of correctly distinguishing between the positive and negative classes.

- An AUC of 0.7 implies a good ability of the model to distinguish between positive and negative classes.
- An AUC of approximately 0.5 indicates that the model performs no better than random chance.
- An AUC close to 0 suggests that the model is predicting the opposite class.

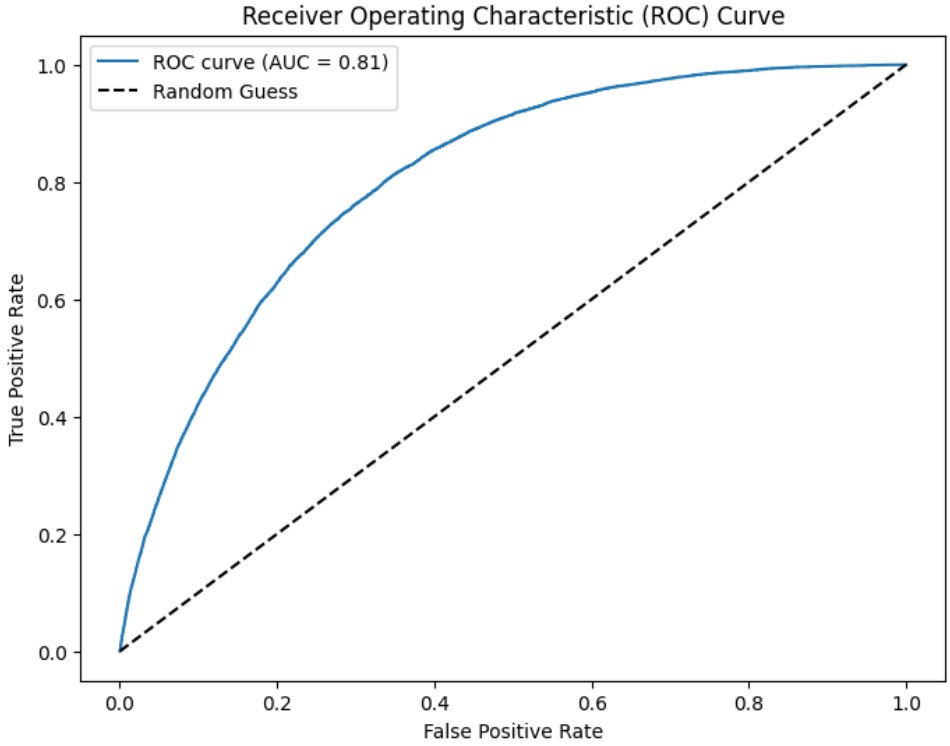


Figure 32: AUC-ROC Curve for logistic regression model with Class weights: 0.7311

6 Results

6.1 Prediction Accuracy

The model achieved an accuracy of approximately 71.00%, indicating that it correctly predicts whether an individual has diabetes around 71.00% of the time.

6.2 Identifying Risk Factors

Our analysis identified several significant predictors of diabetes risk. These include High Blood Pressure (HighBP), High Cholesterol (HighChol), General Health Status (GenHlth), Age, and BMI, among others.

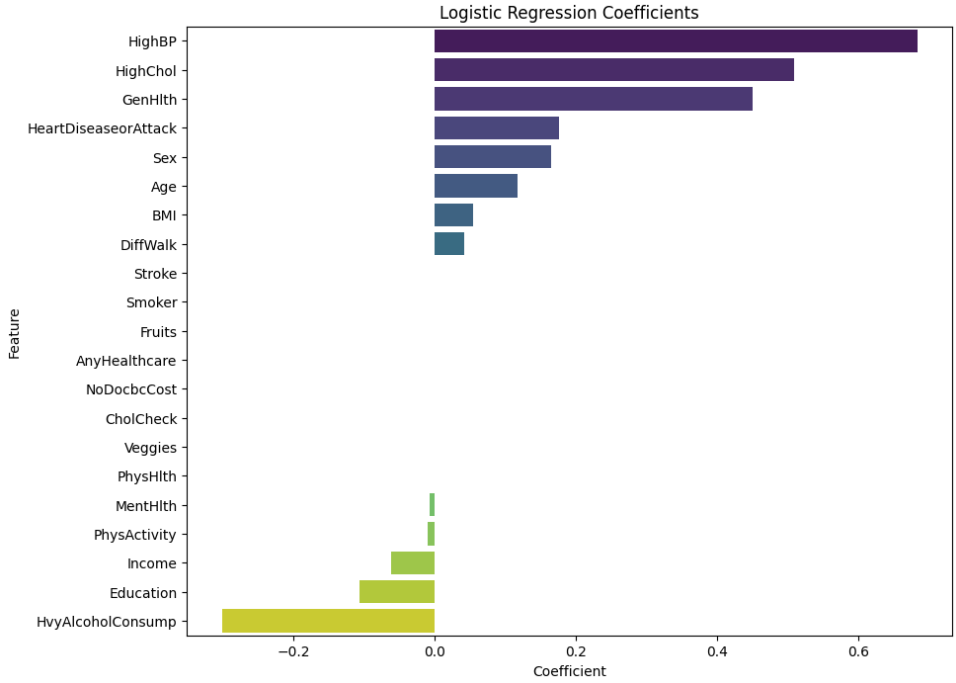


Figure 33: Bar plot showing the coefficients of features in the logistic regression model for predicting diabetes risk.

6.3 Feature Subset Evaluation

By focusing on key risk factors such as HighBP, HighChol, GenHlth, Age, and BMI, we observed significant improvements in prediction accuracy compared to using the entire set of features. This suggests that a subset of risk factors can effectively capture patterns associated with diabetes risk, enhancing model interpretability and efficiency.

6.4 Short Form Development

Our investigation successfully identified a streamlined set of questions from the BRFSS dataset, including HighBP, HighChol, GenHlth, Age, and BMI, which demonstrated significant predictive power in assessing diabetes risk. This approach simplifies screening processes, facilitates early detection, and enables tailored interventions for individuals at elevated risk of diabetes.

7 Conclusion

In this study, we investigated the predictive capabilities of survey questions from the Behavioral Risk Factor Surveillance System (BRFSS) dataset in identifying individuals at risk of diabetes. Through the application of machine learning techniques, we aimed to address several research questions about diabetes prediction, risk factor identification, feature subset evaluation, and short-form development.

Our findings demonstrate that survey questions from the BRFSS dataset can provide valuable insights into diabetes prediction. By leveraging logistic regression modeling and hyperparameter tuning, we achieved a prediction accuracy of approximately 71.00%, with a recall score of 76.14%. These results indicate that our model can effectively distinguish between individuals with and without diabetes, thereby aiding in early detection and intervention.

Furthermore, our analysis identified several key risk factors associated with diabetes risk, including High Blood Pressure (HighBP), High Cholesterol (HighChol), General Health Status (GenHlth), Age, and BMI. These risk factors serve as crucial indicators for assessing an individual's likelihood of developing diabetes and can inform targeted preventive measures and healthcare interventions.

Moreover, feature subset evaluation revealed that a streamlined set of risk factors, such as HighBP, HighChol, GenHlth, Age, and BMI, significantly improve prediction accuracy compared to using the entire feature set. This underscores the importance of identifying and focusing on the most informative predictors to enhance model performance and interpretability.

As a next step, further research could explore the development of a short form of survey questions derived from the BRFSS dataset, leveraging feature selection techniques to efficiently capture the most influential risk factors for diabetes. Additionally, incorporating additional data sources, such as genetic and lifestyle factors, could enhance the predictive power of the model and provide a more comprehensive understanding of diabetes risk.

8 Referrals

- Best techniques and metrics for Imbalanced Dataset
- ML — Underfitting and Overfitting
- Random Forest Simple Explanation
- Bayes Classifier and Naive Bayes
- How to Create Decision Trees for Business Rules Analysis
- Stratified Random sampling – An Overview
- Logistic Regression Explained with Examples
- Data source: Kaggle
- Data Visualization
- Relationship Testing (Chi-Square)
- Implementation of Chi-Square
- Course on Coursera
- Contingency Table
- Imbalances
- Best techniques and metrics for Imbalanced Dataset
- Tour of Evaluation Metrics for Imbalanced Classification
- Outlier detection methods!
- How to Choose the Right Machine Learning Algorithm: A Pragmatic Approach
- Traditional and emerging risk factors that explain the increased risk of adverse events in patients with HF and associated diabetes
- How data leakage affects machine learning models in practice
- What is Information Leakage?
- An Intro to Hyper-parameter Optimization using Grid Search and Random Search

9 Appendix: Python Notebook

An optional Python notebook containing the code used for this analysis is available upon request.

Figure 34: Python Notebook