DATA SCIENCE FINAL PROJECT

FALL 2023

DIABETES PREDICTION

PREPARED BY:
MARIAM AMIR
YAHIA QOULQUELA
MOHAMED ALY

INTRODUCTION

Key Objectives:

1. Comprehensive Descriptive Analysis: Uncover patterns and relationships within the dataset, revealing insights into the factors associated with diabetes.
    - Development of Predictive Models:
    - Construct a baseline logistic regression model.
    - Explore two additional machine learning models, carefully selected for their suitability in addressing classification tasks.
2. Rigorous Model Comparison: Evaluate performance metrics to identify the most accurate and reliable model for diabetes prediction.
3. Best Model Application: Utilize the selected model to predict diabetes risk in new patients, empowering healthcare professionals with a valuable decision-making tool.

Dataset

- diabetes_prediction_dataset.csv
- Features: Age, gender, BMI, hypertension, heart disease, smoking history, HbA1c, blood glucose level
- Target variable: Diabetes status (positive or negative)
- Data source: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

INTRODUCTION

Approach:

- Exploratory Data Analysis (EDA):Visualize data distributions.
  - Identify correlations and potential patterns.
  - Address missing values and anomalies.

Model Development:

  - Implement logistic regression as a baseline.
  - Experiment with one machine learning model.
  - Fine-tune hyperparameters for optimal performance.

Model Evaluation:

  - Assess accuracy, precision, recall.
  - Model Selection and Prediction:Select the best-performing model based on evaluation results.
  - Utilize the chosen model to predict diabetes risk in new patient cases.

Project Significance:

- Improved Healthcare Outcomes: Facilitate early identification and intervention for those at risk, potentially reducing diabetes incidence and its associated complications.
- Personalized Treatment Plans: Inform tailored prevention and management strategies based on individual risk profiles.
- Research Advancement: Contribute to a deeper understanding of diabetes risk factors and the effectiveness of machine learning in healthcare applications.

DATA DESCRIPTION

.Data Source and Population:

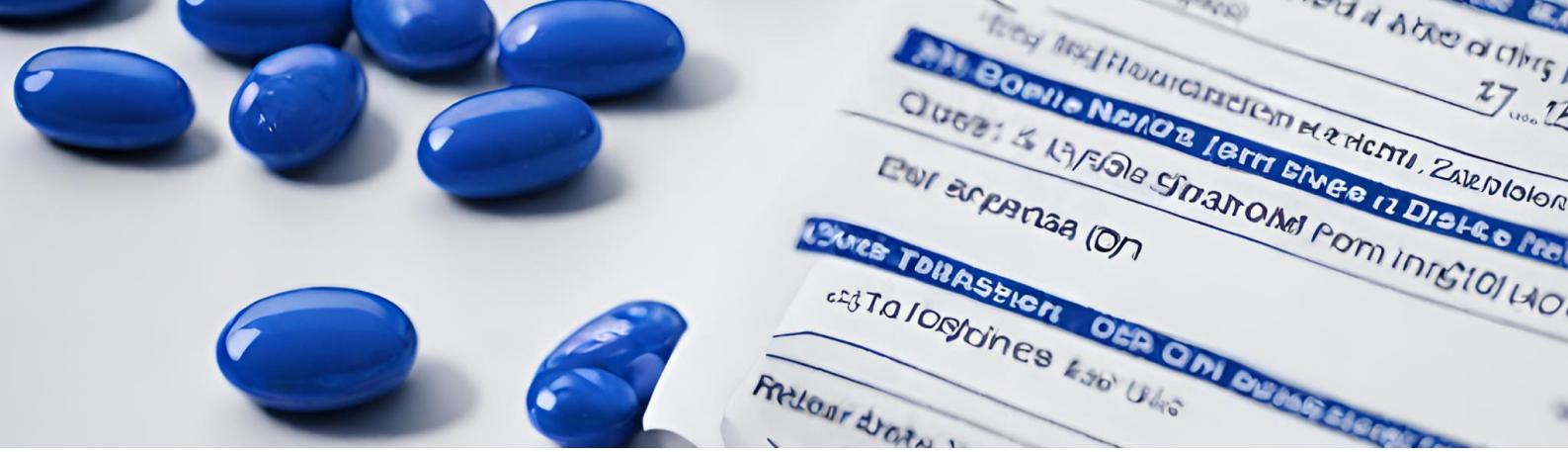
This dataset was obtained from Kaggle, based on data aggregated from Electronic Health Records (EHRs) of multiple healthcare providers. While the target population remains unspecified, EHRs typically represent patients actively seeking medical care and might not perfectly reflect the broader population.

Variables and Explanations:

- Gender: Categorical variable (male, female, other) with potential hormonal and metabolic differences influencing susceptibility.
- Age: Continuous variable (range 0-80) indicating increased risk with advancing age.
- Hypertension: Binary variable (0/1) representing presence or absence of high blood pressure, a known risk factor.
- Heart disease: Binary variable (0/1) highlighting existing comorbidities that heighten diabetes risk.
- Smoking history: Categorical variable (not current, former, no info, current, never, ever) highlighting a modifiable risk factor with potential adverse effects.
- BMI: Continuous variable (range 10.16-71.55) reflecting body fat percentage, with higher values increasing diabetes risk.
- HbA1c level: Continuous variable (indicating average blood sugar over 2-3 months), with levels above 6.5% typically suggesting diabetes.
- Blood glucose level: Continuous variable (reflecting glucose concentration in the bloodstream at a given time), with elevated levels suggesting potential diabetes.
- Diabetes: Binary variable (1/0) serving as the target variable for prediction, signifying the presence or absence of diabetes.

Collection Methodology:

The data was gathered through surveys, medical records, and laboratory tests administered to patients with or at risk of developing diabetes. This diverse data collection approach helps capture various aspects of health and lifestyle, potentially enriching the analysis.



DESCRIPTIVE ANALYSIS

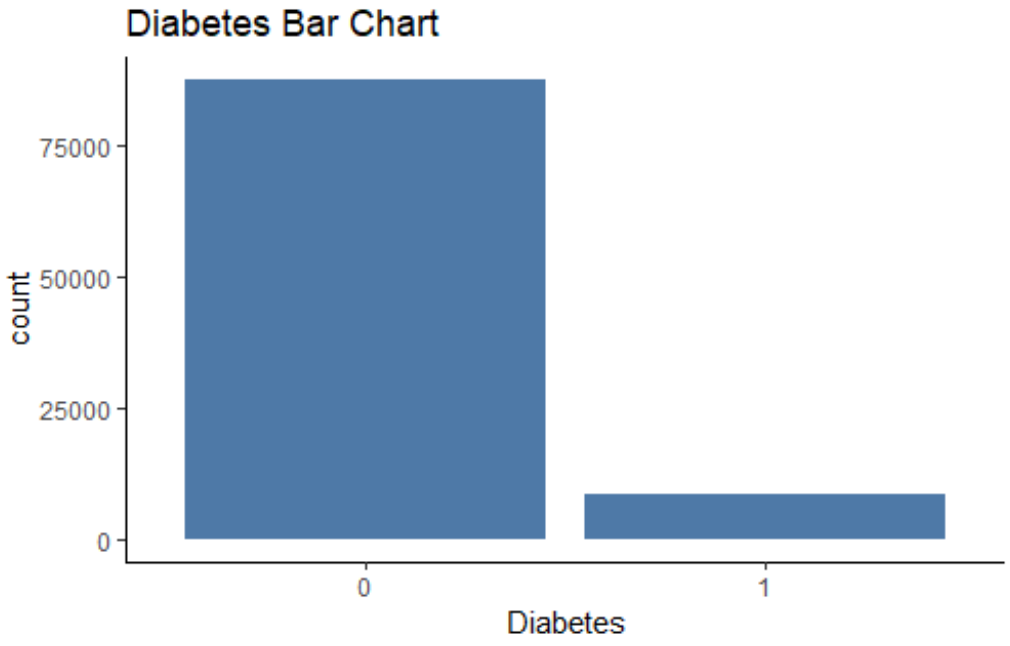
UNIVARITAE ANALYSIS

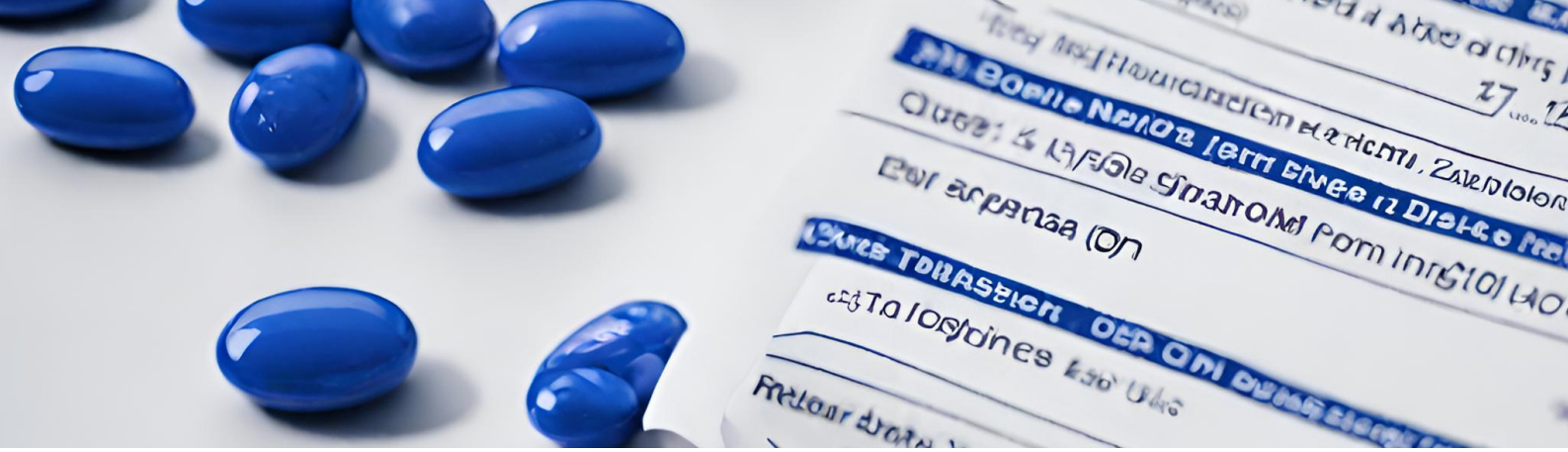
RESPONSE VARIABLE: CATEGORICAL

DIABETES

The response variable under investigation in our analysis is diabetes, categorized as 1 for its presence and 0 for its absence. Among the total population studied, comprising **96,128** individuals, a notable majority of **87,646** individuals are classified as not having diabetes (labeled as 0), indicating the absence of this condition. Conversely, **8,482** individuals are identified as having diabetes (labeled as 1), signifying the presence of this health concern within a subset of the surveyed group. Understanding the prevalence of diabetes within this population is pivotal for identifying risk factors, establishing preventive measures, and tailoring interventions.

Diabetes	Frequency
0	87646
1	8482





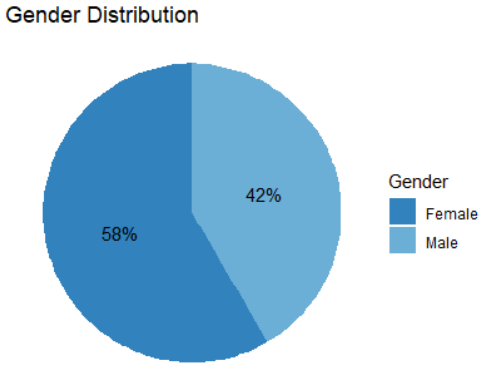
DESCRIPTIVE ANALYSIS

UNIVARITAE ANALYSIS

CATEGORICAL GLIMPSE

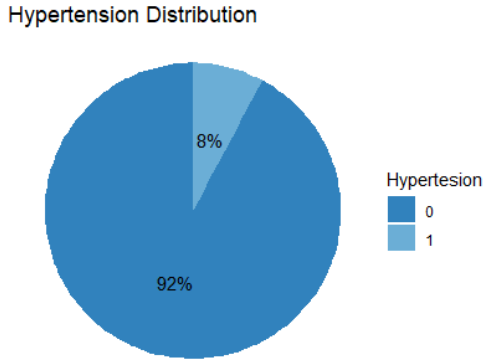
GENDER

The data collected illustrates the gender distribution within our sample population. Among the total respondents, **56,161** identified as female, representing approximately **58.4%** of the population, while **39,967** identified as male, accounting for approximately **41.6%** of the population. This clear gender breakdown underscores a higher representation of females compared to males in our dataset, highlighting a notable difference in gender distribution within the surveyed group.



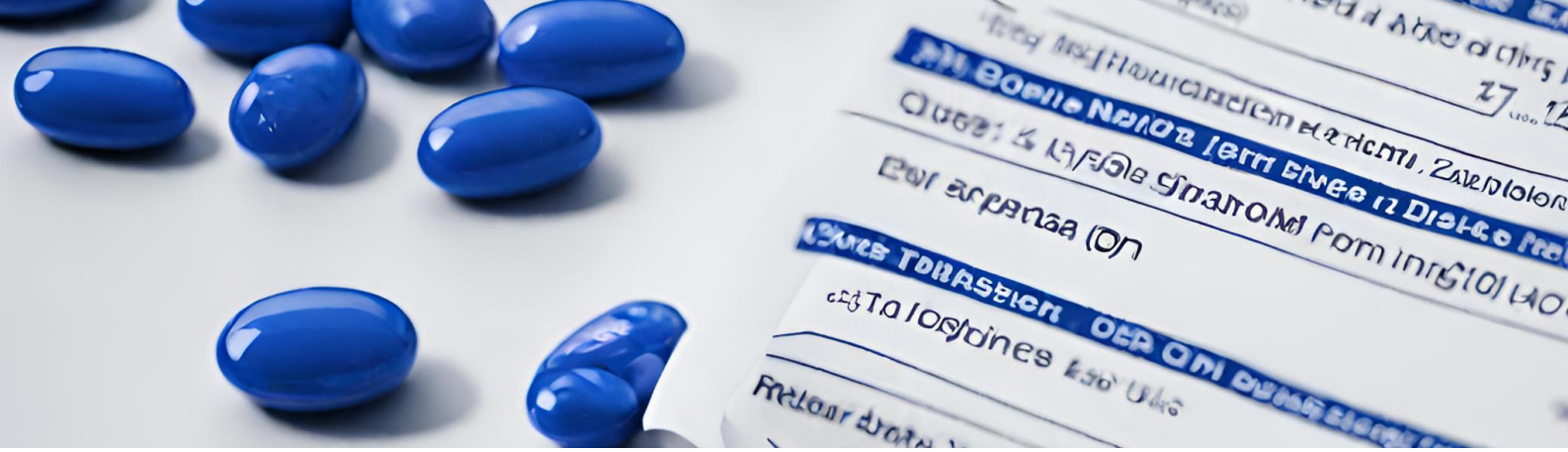
Gender	Frequency	Percentage
Female	56161	58.42314
Male	39967	41.57686

HYPERTENSION



The descriptive analysis of hypertension within our studied cohort reveals compelling insights. Among the total participants, comprising **88,667** individuals, the absence of hypertension, denoted by the value **0**, is predominant, encompassing the majority of our sample. Specifically, 88,667 individuals do not exhibit signs of hypertension. Conversely, **7,461** individuals, a smaller yet notable subset of the population, have been identified with hypertension, marked by the value **1**. This data underscores a substantial discrepancy, with a larger proportion showing no indications of elevated blood pressure compared to those diagnosed with hypertension within our surveyed group.

Hypertesion	Frequency
0	88667
1	7461



DESCRIPTIVE ANALYSIS

UNIVARITAE ANALYSIS

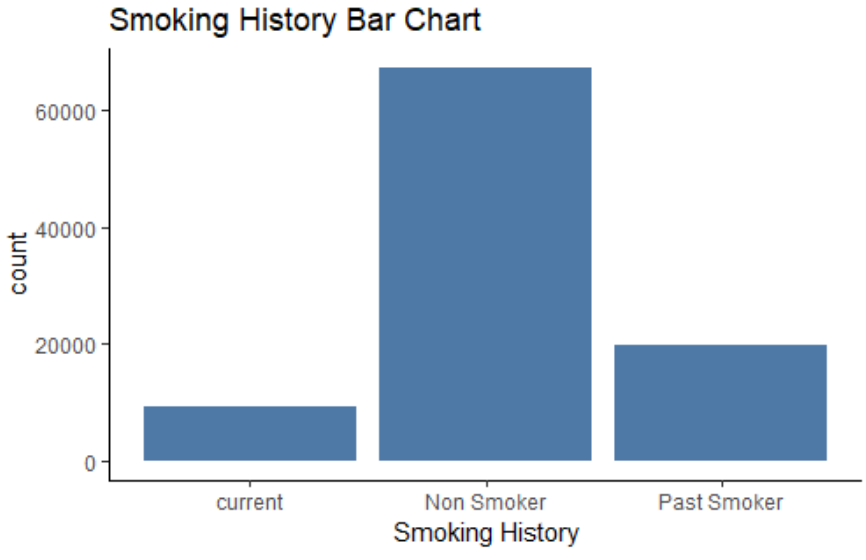
CATEGORICAL GLIMPSE

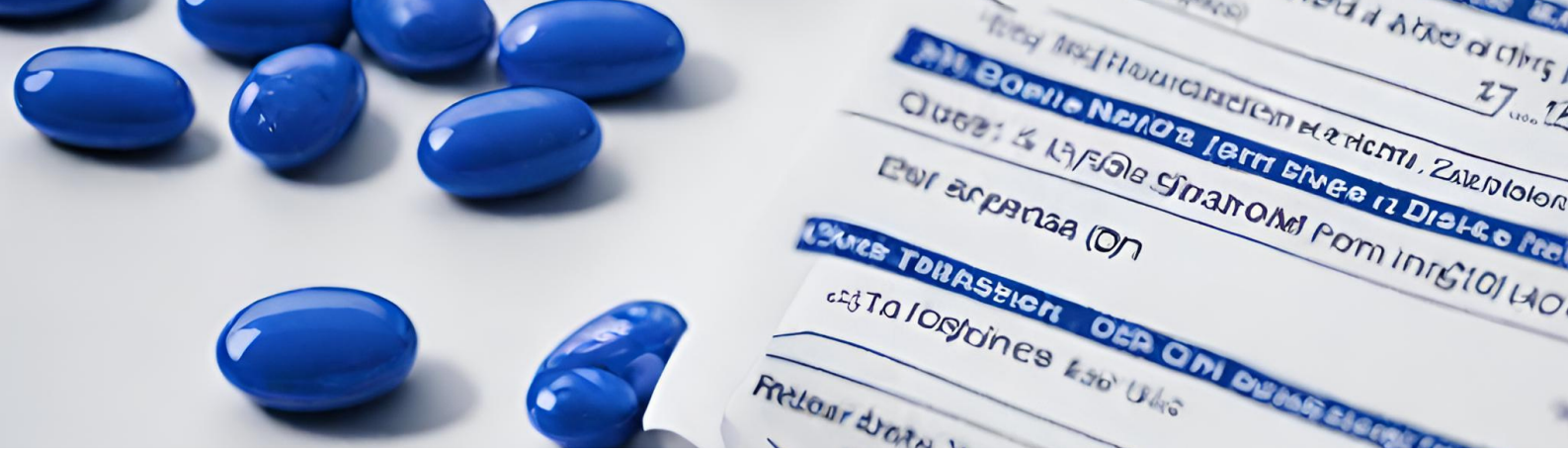
SMOKING HISTORY

Smoking History	Frequency
current	9197
Non Smoker	67276
Past Smoker	19655

The smoking history data initially had different descriptions like "never," "No Info," "ever," "former," and "not current." To make it easier to understand, these descriptions were grouped together.
The group "never" and "No Info" became "Non-Smoker," representing those who never smoked or had unknown smoking data. Similarly, "ever," "former," and "not current" became "Past Smoker," including people who smoked before, regardless of if they currently smoke.

The bar chart showcases the distribution of individuals based on their smoking history across three categories: "Non-Smoker," "Current Smoker," and "Past Smoker." Among these categories, the highest frequency count is observed within the Non-Smoker group, totaling **67,276** individuals, signifying the largest segment in this dataset. Following this, the Past Smoker category denotes a substantial count of **19,655** individuals, representing those who have previously smoked but are not current smokers. Conversely, the Current Smoker group illustrates the smallest count among the three categories, comprising **9,197** individuals. This depiction emphasizes a notable prevalence of Non-Smokers within the surveyed population, a significant representation of individuals with a history of smoking, and a comparatively smaller count of current smokers in this dataset.





DESCRIPTIVE ANALYSIS

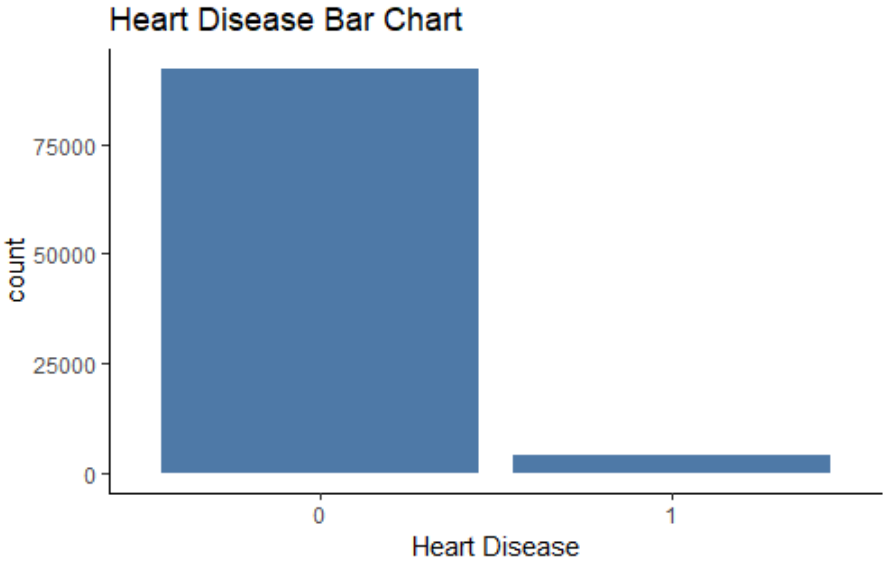
UNIVARITAE ANALYSIS

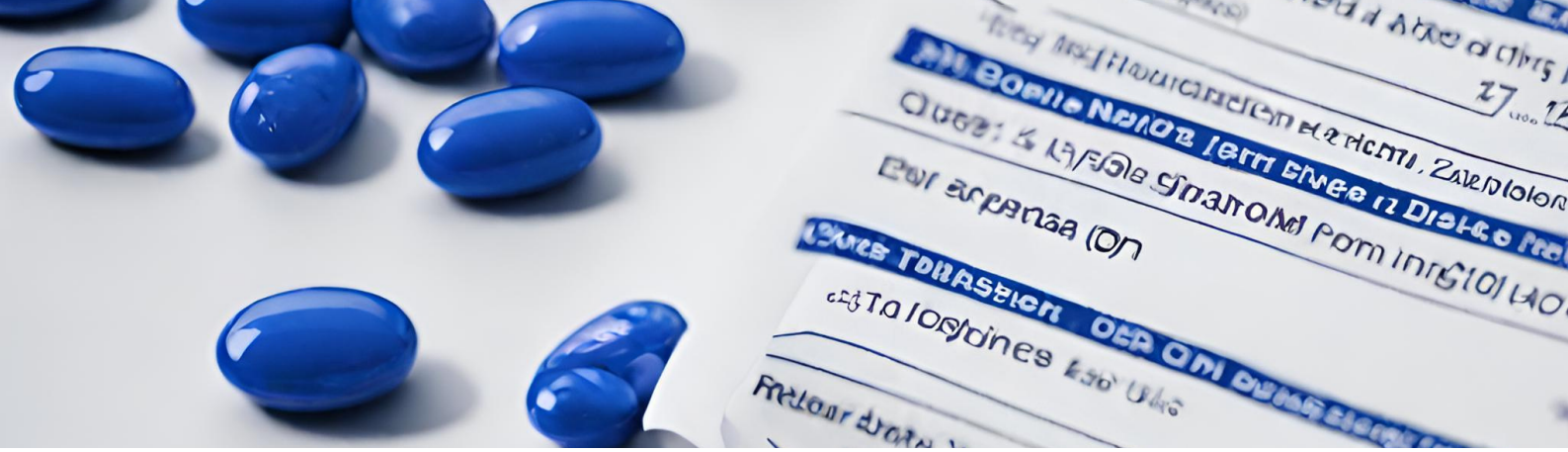
CATEGORICAL GLIMPSE

HEART DISEASE

Heart Disease	Frequency
0	92205
1	3923

The Bar Chart illustrates the distribution of heart disease within our surveyed population. Among the observed categories, the absence of heart disease ('0') emerges as the predominant condition, with a substantial frequency count of **92,205** individuals. In contrast, the presence of heart disease ('1') records a notably lower frequency count of **3,923** individuals. This discrepancy underscores a substantial prevalence of individuals without heart disease compared to those diagnosed with this condition within our studied group.





DESCRIPTIVE ANALYSIS

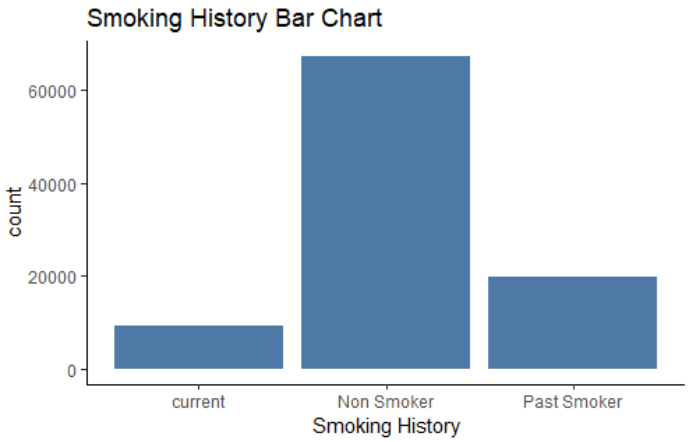
UNIVARITAE ANALYSIS

CATEGORICAL GLIMPSE

SMOKING HISTORY

Smoking History	Frequency
current	9197
Non Smoker	67276
Past Smoker	19655

The bar chart showcases the distribution of individuals based on their smoking history across three categories: "Non-Smoker," "Current Smoker," and "Past Smoker." Among these categories, the highest frequency count is observed within the Non-Smoker group, totaling 67,276 individuals, signifying the largest segment in this dataset. Following this, the Past Smoker category denotes a substantial count of 19,655 individuals, representing those who have previously smoked but are not current smokers. Conversely, the Current Smoker group illustrates the smallest count among the three categories, comprising 9,197 individuals. This depiction emphasizes a notable prevalence of Non-Smokers within the surveyed population, a significant representation of individuals with a history of smoking, and a comparatively smaller count of current smokers in this dataset.



DESCRIPTIVE ANALYSIS

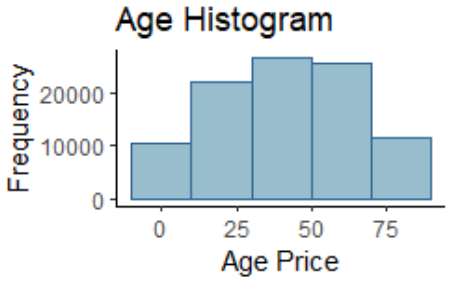
UNIVARITAE ANALYSIS

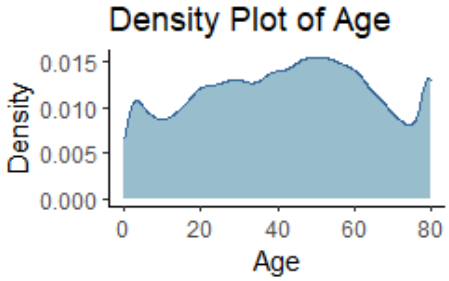
QUANTITATIVE VARIABLES

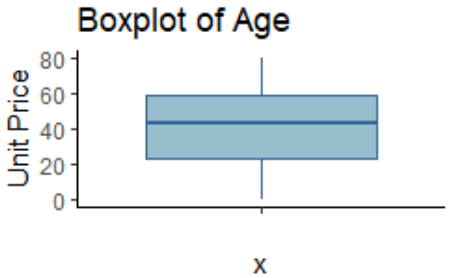
AGE

Min	Max	Mean
0.08	80.00	41.80
1st Quartile	**Median**	**3rd Quartile**
24.00	43.00	59.00

The analysis of the 'age' variable reveals compelling insights into the dataset's age distribution. The histogram demonstrates a positively skewed distribution, with ages predominantly clustered between approximately **24 and 59** years. Notably, there's a notable concentration around the median age of **43**, suggesting a relatively higher occurrence within this age group. The boxplot further emphasizes this concentration, showcasing the spread of ages and highlighting the presence of outliers beyond the minimum and maximum values. Additionally, the density plot reaffirms the prevalent age range of **24 to 59**, depicting a prominent peak around the median age of **43**. Overall, these visualizations and summary statistics collectively portray a skewed yet concentrated distribution of ages within the dataset, offering valuable insights into the age demographics under study.







DESCRIPTIVE ANALYSIS

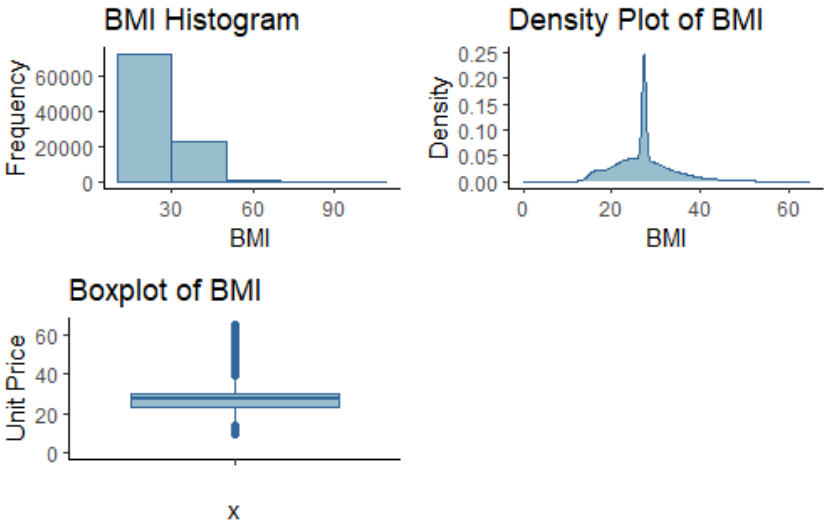
UNIVARITAE ANALYSIS

QUANTITATIVE VARIABLES

BMI

Min	Max	Mean
10.01	95.69	27.32
1st Quartile	**Median**	**3rd Quartile**
23.40	27.32	29.86

The analysis of Body Mass Index (BMI) data reveals a distribution concentrated around the median of **27.32,** spanning a range from **10.01** to **95.69**. The histogram illustrates a relatively symmetrical distribution with the majority of BMI values clustered between **23.40** and **29.86,** indicating average to moderately high BMI levels. However, the boxplot showcases outliers beyond the upper whisker, notably extreme values exceeding the upper quartile. These outliers, representing exceptionally high BMI levels, contribute to an elongated upper whisker and are visibly highlighted in the density plot. Their presence, while rare, emphasizes the potential impact of extreme observations on the dataset's overall BMI distribution, warranting further scrutiny for potential data irregularities or genuine extreme weight instances



DESCRIPTIVE ANALYSIS

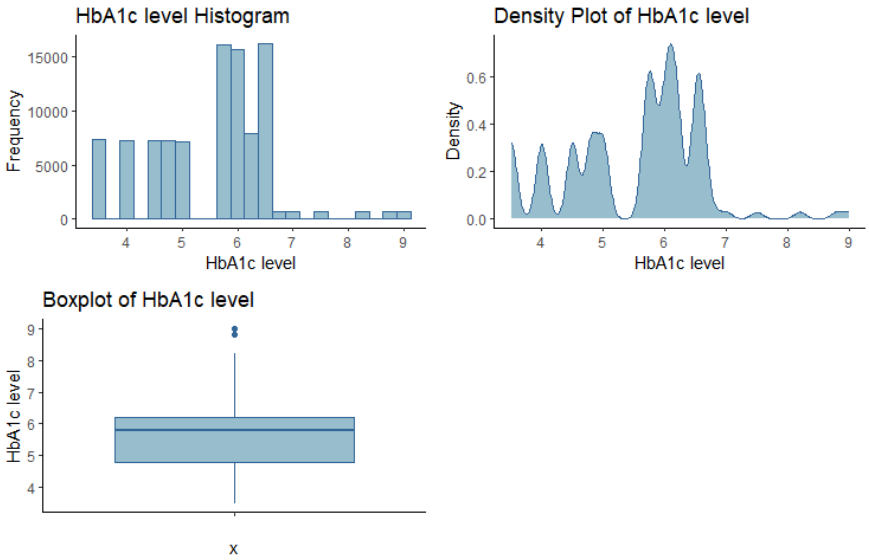
UNIVARITAE ANALYSIS

QUANTITATIVE VARIABLES

HBA1C LEVEL

Min	Max	Mean
3.500	9.000	5.533
1st Quartile	**Median**	**3rd Quartile**
4.800	5.800	6.200

The analysis of Hemoglobin A1c (HbA1c) levels, reflecting average blood sugar over recent months, reveals insightful findings. The dataset spans from **3.500 to 9.000**, with a median of **5.800**, indicating a central tendency around this measure. The mean of **5.533** suggests a slight skew towards lower values. Notably, while the median and mean values fall below the clinical threshold of **6.5%** associated with a higher risk of diabetes, the third quartile of **6.200** approaches this critical threshold closely. This implies a notable segment of individuals within or nearing the diabetic range. The majority of individuals exhibit HbA1c levels below the diabetic threshold, but the proximity of the upper quartile to the threshold underscores the importance of continued monitoring and potential intervention for those at risk or nearing diabetic levels.



DESCRIPTIVE ANALYSIS

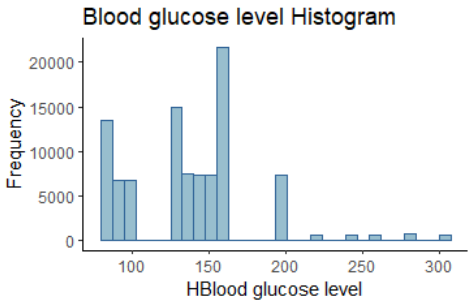
UNIVARITAE ANALYSIS

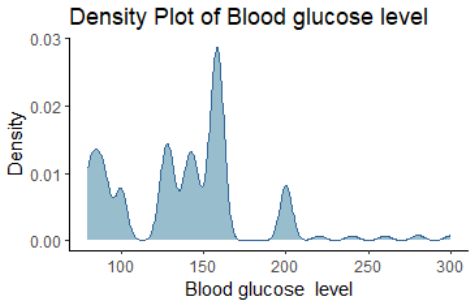
QUANTITATIVE VARIABLES

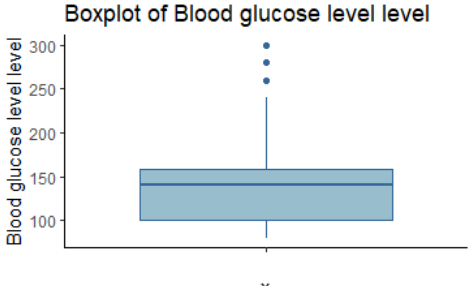
BLOOD GLUCOSE LEVEL

Min	Max	Mean
80.0	300.0	138.2
1st Quartile	**Median**	**3rd Quartile**
100.0	140.0	159.0

The analysis of blood glucose levels provides insightful findings into the dataset. Ranging from **80.0** to **300.0**, these levels exhibit a diverse spectrum. The median blood glucose level of **140.0** indicates the central tendency, while the mean of **138.2** suggests a slight skew towards lower values. Elevated blood glucose often signals diabetes, with diagnostic thresholds typically set at 126 mg/dL for fasting glucose or **200** mg/dL for random glucose levels. The dataset's median and mean values, hovering around **140.0** and **138.2**, respectively, align closely with these thresholds, indicating a significant portion of individuals potentially at risk for elevated blood sugar levels associated with diabetes. This underscores the importance of further clinical evaluation or monitoring for individuals within this range to ascertain their diabetes status and consider appropriate interventions.







RANDOM FOREST

.Introduction to Our Approach in Random Forest Development

At the core of our modeling approach lies the Random Forest algorithm, a robust ensemble learning technique renowned for its versatility and predictive power. Our objective was to leverage this algorithm to craft predictive models capable of accurately classifying observations in a dataset containing variables relevant to our domain.

Understanding Random Forest

Random Forest operates by constructing multiple decision trees, each trained on different subsets of the data and features. This ensemble approach aggregates the predictions from individual trees to generate more accurate and stable results, effectively mitigating overfitting and enhancing predictive performance.

Balancing Classes Through Upsampling

We noticed that our dataset had unequal numbers of examples for different categories. To make things fair for our model, we added more examples of the smaller group. But here's the thing: we only did this when teaching the model, not when we tested it.

Adding More Examples Only in Training

We made sure to add extra examples only when teaching the model. This way, when we tested the model's ability, we used the original dataset without these extra examples. That's because we wanted to see how well the model could handle new, unseen examples.

Why We Did This

By teaching the model on a more balanced set, it got better at understanding both sides of the data. But when we tested it, we made sure it wasn't influenced by these extra examples. This helped us see if the model could work well in the real world, not just on the training data.

Making Things Fair for Better Predictions

Our goal was to give the model a fair chance to learn from all the data without changing the original dataset. This way, it could make better predictions without favoring one side too much.

RANDOM FOREST

Methodology Overview

1. **Data Sampling:** We began by sampling a subset of our training dataset to expedite model training and validation while retaining the dataset's representative nature.
2. **Initial Model Building**: Initially, we constructed Random Forest models with varying tree counts, evaluating their performance through out-of-bag (OOB) error rates and confusion matrices.
3. **Optimization Strategies:** Further exploration involved optimizing the model by varying the number of variables considered at each split (mtry). This allowed us to identify the configuration that minimizes the model's error rate.

Guiding Principles

1. **Model Flexibility:** Harnessing the power of ensemble learning to create models robust against overfitting and capable of handling diverse datasets.
2. **Addressing Class Imbalance:** Employing upsampling techniques to rectify class imbalance, promoting equitable learning and better representation within our models.
3. **Iterative Exploration:** Systematically exploring model parameters to unearth configurations yielding optimal predictive performance.
4. **Objective Validation**: Relying on OOB error rates, confusion matrices, and visualization techniques to objectively evaluate model performance and make informed decisions.

Conclusion

By integrating an upsampling technique to mitigate class imbalance and leveraging the strengths of Random Forest, our aim is to develop models that offer accurate predictions while maintaining robustness and generalizability across various scenarios.

RANDOM FOREST

Random Forest Model Evaluation

Sampled Data and Model Training

We randomly sampled a subset of the training dataset (train_set2), comprising 10,000 observations, for training the Random Forest models.

Model 1: 500 Trees

- Number of Trees: 500
- Variables Tried at Each Split: 2
- OOB Error Rate: 7.65%

	0	1	Class Error
0	4581	410	0.08214787
1	355	4654	0.07087243

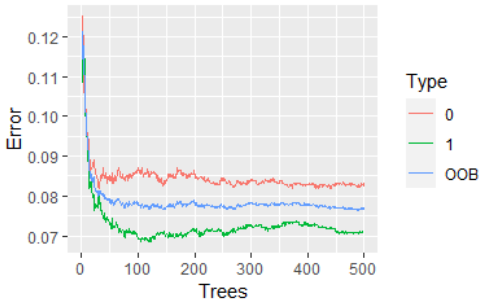
Analysis:

The initial model with 500 trees exhibited an out-of-bag (OOB) error rate of 7.65%. The confusion matrix indicates a misclassification rate for both classes, with Class 0 showing a slightly higher error rate compared to Class 1.

Random Forest Model Evaluation

Model 1: 500 Trees

- Number of Trees: 1000
- Variables Tried at Each Split: 2
- OOB Error Rate: 7.69%



	0	1	Class Error
0	4571	420	0.08415147
1	349	4660	0.06967459

Analysis:

Increasing the number of trees to 1000 led to a marginal rise in the OOB error rate to 7.69%. However, the confusion matrix suggests a similar trend in misclassification rates for both classes compared to the previous model.

Both models performed similarly in terms of error rates and misclassification across the two classes. Increasing the number of trees from 500 to 1000 did not significantly improve the model's performance.

RANDOM FOREST

Random Forest Model Evaluation

Model 3: Optimal mtry Selection

- Optimal mtry: 5
- OOB Error Rate with Optimal mtry: 6.95%

	0	1	Class Error
0	4568	423	0.08475255
1	272	4737	0.05430226

Analysis:

Through exploration, the optimal mtry value was found to be 5, resulting in an improved OOB error rate of 6.95%. The confusion matrix indicates a substantial decrease in the misclassification rate for Class 1 compared to the previous models.

Conclusion:

- The models with different tree counts and mtry values were evaluated.
- Increasing the number of trees from 500 to 1000 did not significantly improve model performance.
- Optimal mtry selection led to a decrease in the overall error rate, especially in correctly classifying observations in Class 1.
- Model 3 with optimal mtry of 5 exhibited the lowest error rate among the models evaluated.

LOGISTIC REGRESSION MODEL

.Introduction to Our Approach in Random Forest Development

We constructed a logistic regression model to predict the likelihood of diabetes based on a dataset with various health-related predictors. The model aimed to estimate the probability of an individual having diabetes, considering factors such as gender, age, medical history (hypertension, heart disease), lifestyle choices (smoking history), and biomarkers (BMI, HbA1c level, blood glucose level).

Model Coefficients and Interpretations:

The coefficients obtained from the model highlight the influence of each predictor on the log-odds of having diabetes. For instance:

- Age: With each one-unit increase in age, the log-odds of having diabetes increase by 0.051, holding other variables constant.
- HbA1c Level: Higher HbA1c levels significantly increase the log-odds of having diabetes (odds ratio = 8.77).

Balancing Classes Through Upsampling

We noticed that our dataset had unequal numbers of examples for different categories. To make things fair for our model, we added more examples of the smaller group. But here's the thing: we only did this when teaching the model, not when we tested it.

Model Performance and Assessment:

McFadden's R^2: The model shows a pseudo-R^2 value of approximately 0.645, signifying a good fit.

LOGISTIC REGRESSION MODEL

.Predictive Performance Metrics:

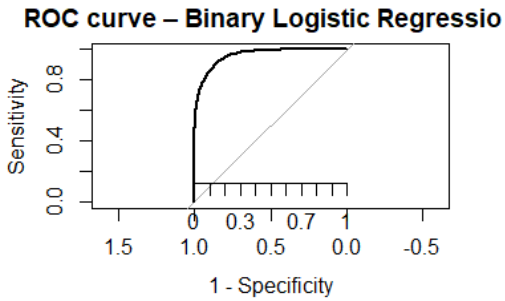
- .**Confusion Matrix:** Using an optimal cutoff, the confusion matrix displays the classification performance of the model.

Prediction	Reference 0	1
0	22813	3549
1	245	2232

- **Accuracy:** The model achieved an accuracy of 86.84% on the test set.
- **Sensitivity & Specificity: Sensitivity** (True Positive Rate) is high (98.94%), while Specificity (True Negative Rate) is comparatively lower (38.61%).

Diagnostic Metrics:

- **AUC-ROC Curve:**



- **Variance Inflation Factor (VIF)**: VIF values suggest no significant multicollinearity issues among the predictor variables.

Conclusion:

The logistic regression model presents a valuable framework for predicting diabetes based on various health-related features. While it exhibits high sensitivity, further refinement is needed to reduce false positive rates and enhance specificity for more reliable predictions.

Recommendations and Further Steps:

- The model demonstrates good predictive power but has higher false positives and lower specificity. This could be an area for further improvement.
- We could try in the future to get different models such as stepwise and backward Models and compare them to the model we developed

COMPARATIVE ANALYSIS

Random Forest: Shows a lower overall error rate (6.95%) but higher class errors for "0" predictions compared to Logistic Regression.
L**ogistic Regression**: Demonstrates a good fit (McFadden's R^2 of 0.6447), indicating a strong association between predictors and the target variable.
Conclusion:

- **Random Forest** performs well with an overall lower error rate, although it shows a bit higher class error for one category.
- **Logistic Regression** offers strong predictive power, indicating a good fit of the predictors to the target variable.

Both models showcase strengths. The choice between them may depend on the specific emphasis required on accuracy or interpretability. For instance, if precision in certain classes is critical, Random Forest might be more suitable. Conversely, if understanding the contribution of individual predictors is vital, Logistic Regression provides more interpretability.

Comparing their performances using confusion matrices

- **Logistic Regression** seems more conservative in predicting the positive class, having fewer false positives but also fewer true positives.
- **Random Forest,** on the other hand, predicts more instances of the positive class but tends to have more false positives.

Deciding between these models would rely on the importance of minimizing specific types of errors. If minimizing false positives is critical, the Logistic Regression model may be preferred. If maximizing true positives is more crucial, then the Random Forest model might be the choice, even though it comes with a higher rate of false positives.