# Introduction

Diabetes mellitus is chronic illness that affects millions of people globally and has big influence on healthcare systems and quality of life. It is characterized by high blood glucose levels and if left untreated can lead to several problems. For management and treatment to be effective early diagnosis and prediction are essential. The goal of this study is to identify people who are at high risk of diabetes by using predictive modelling and data analysis approaches. We attempt to construct dependable predictive model by utilizing dataset that includes variety of health indicators including blood pressure body mass index and glucose levels among others. This approach aims to serve as proactive health management tool by not only helping medical practitioners identify patients at an early stage but also educating patients about their health concerns through an interactive Shiny application.

# Data Overview

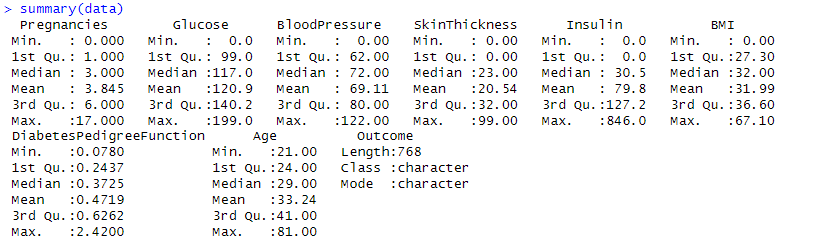
The primary dataset for this research includes records from 768 people collecting various health variables thought to affect onset of diabetes. The number of pregnancies body mass index (BMI) diabetes pedigree function, age, triceps skinfold thickness, diastolic blood pressure, plasma glucose concentration and 2 hour serum insulin are all included in each record. The dataset offers a binary outcome variable that may be used to determine if patient has diabetes or not. It was originally shown as 'Y' for positive instances and 'N' for negative cases but it has since been numerically recoded for analytical purposes. Notably zero values in physiological measurements which are considered missing data because they are biologically implausible needed to be handled by preprocessing data. After that various imputation techniques were used to impute missing values to guarantee reliability and integrity of ensuing analyses. This meticulously selected dataset provides basis for feature selection, exploratory data analysis and predictive modeling all of which seek to identify patterns and variables that are important for diabetes risk.

# Data Preprocessing

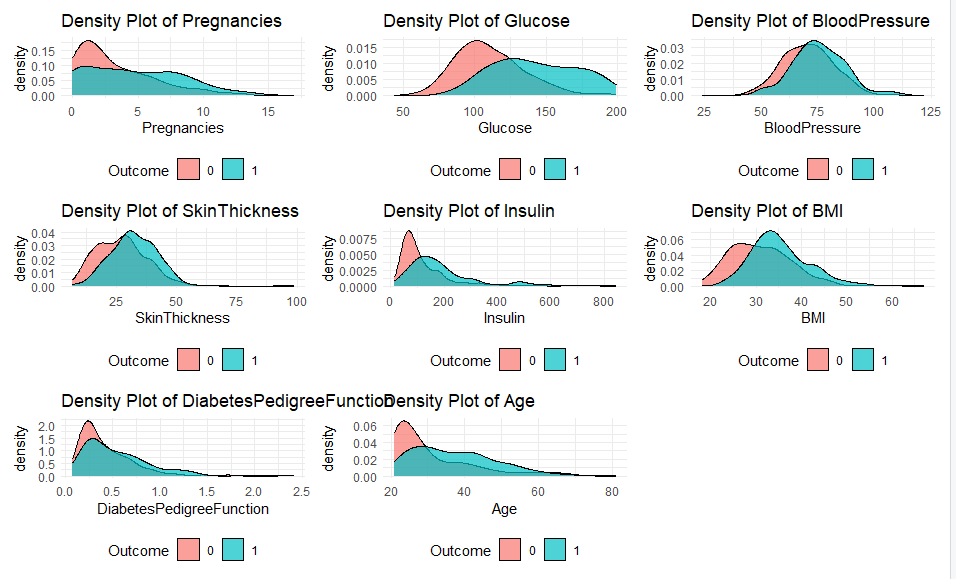
To guarantee caliber and precision of analysis dataset underwent thorough cleaning procedure throughout data preprocessing stage. Initial analyses of several important variables including BMI and glucose, showed zero values which were deemed missing data since they were biologically impossible. NA placeholders were used in place of these zero values to address this. Next five complete datasets with predictive mean matching imputed missing values were produced using R mice package for multiple imputation. The binary outcome variable 'Outcome' was then transformed from category labels 'Y' and 'N' to a binary format that was numerical where 1 denoted existence of diabetes and 0 its absence. The application of several statistical modelling techniques was made easier by this conversion. These preparatory actions were essential in producing a trustworthy and analyzable dataset which prepared ground for solid exploratory data analysis and modelling procedure.

# Exploratory Data Analysis (EDA)

### Summary Statistics

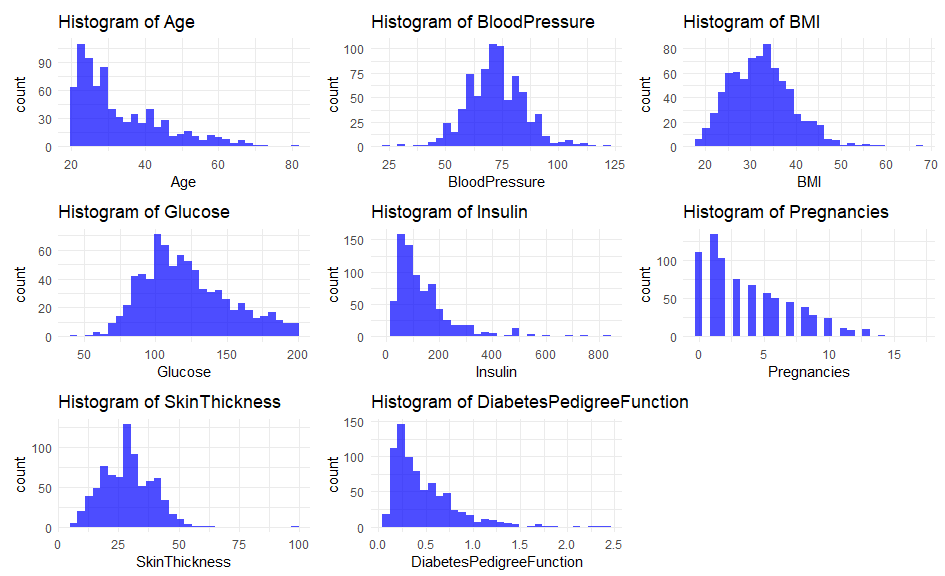


### Density Plots for all variables



The image displays series of density plots comparing the distribution of various healthrelated variables for two groups of individuals differentiated by an 'Outcome' of 0 (no diabetes) and 1 (diabetes). Each plot shows how measurements for Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction and Age are distributed across these two groups. The plots aim to identify differences in these distributions which could be related to presence or absence of diabetes. For instance individuals with diabetes might show higher glucose levels as indicated by right-skewed distribution for 'Outcome' 1.

### Histogram Plot



The image shows collection of histograms representing frequency distribution of various health-related measurements: Age, Blood Pressure, Body Mass Index (BMI), Glucose, Insulin, Skin Thickness, Diabetes Pedigree Function and Number of Pregnancies. Each histogram displays number of individuals (count) within specific ranges (bins) for each variable. These graphs are useful for understanding distribution and central tendencies of each health metric within given population such as commonality of certain age ranges or BMI values. They're essential for initial data exploration and can indicate presence of outliers or unusual values in dataset.

### Correlation Matrix Heat Map

Isn’t this supposed to include the outcome so we can see which variables have the strongest correlation with it?

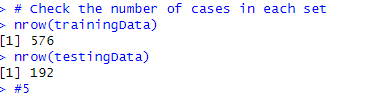
A diagram of a health care system

Description automatically generated with medium confidence

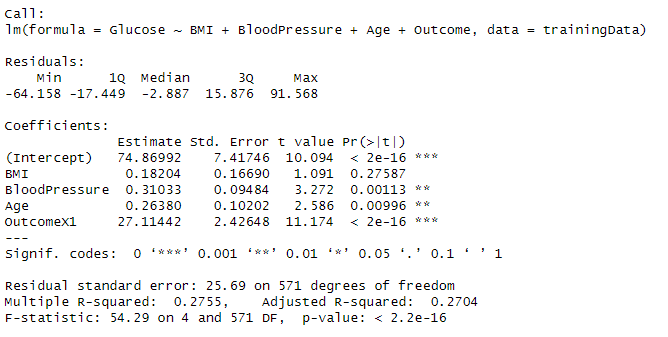
On the basis of color strength, we conclude the variables which correlated most. Understanding interdependencies between variables is possible with help of the correlation matrix heat map. Darker blue hues indicate stronger correlations whereas lighter blue hues show direction and intensity of associations. Notably factors like BMI and glucose had more robust positive associations with diabetes outcome confirming their significance as model predictors. In meantime majority of other variable pairings do not exhibit dark shades indicating lack of strong linear correlations. This is important to know to prevent multicollinearity during modelling process. Using factors that may be more important in forecasting beginning of diabetes this heat map helps with feature selection.

# Model Building and Evaluation

### Training and Test Split



### Linear Model



There is substantial correlation between glucose levels and existence of diabetes according to linear regression study using glucose as response variable and diabetes as binary outcome. The outcome's coefficient is positive (estimate = 30.762) meaning that people with diabetes typically have glucose levels that are around 30.76 units higher than those without disease. With a p-value of less than 2e-16, this difference is highly significant and supports documented association between higher glucose levels and diabetes. This discovery is visually supported by plot which clearly distinguishes between two groups' glucose levels those without diabetes are concentrated at lower values while those with diabetes are distributed over larger range of higher values.

A screenshot of a graph

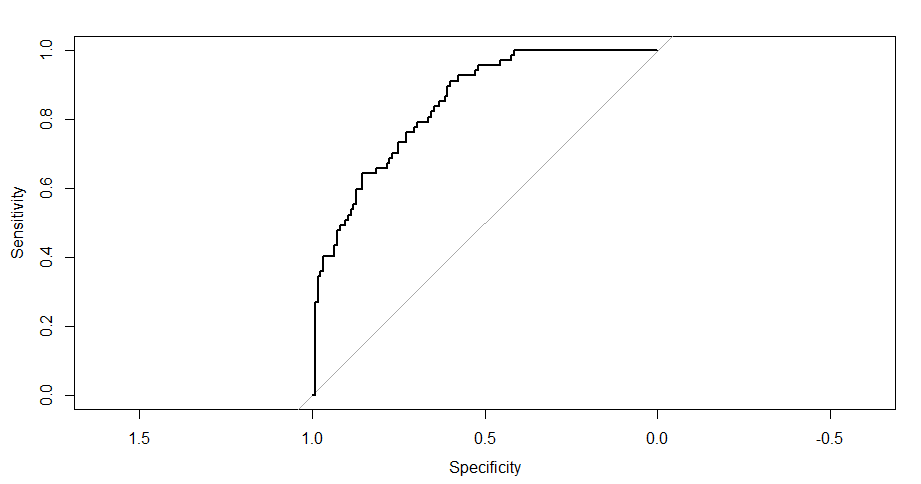
Description automatically generated

While glucose is significant factor additional variables not included in this model may also contribute to variability in glucose levels linked with diabetes outcomes. The model explains roughly 23.81% of variance in glucose levels (Multiple R-squared: 0.2381).

### Logistic Regression Model

A screenshot of a computer

Description automatically generated



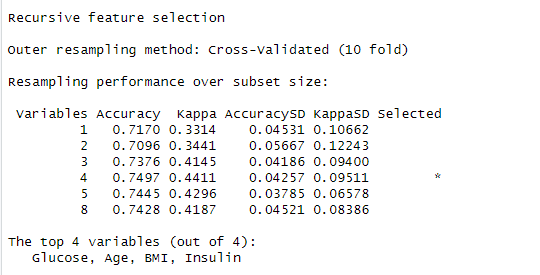
The results of logistic regression and ROC curve offer important new information on variables affecting diabetes risk and diagnostic performance of model. The regression coefficients show that probability of getting diabetes is highly influenced by BMI, glucose, DiabetesPedigreeFunction and age with glucose having most impact (Estimate = 0.0376942, p < 0.001). These results support clinical knowledge that age, body mass index, genetic variables and elevated glucose levels are significant predictors of diabetes.

A number and text on a white background

Description automatically generated

The confusion matrix is used to calculate model's performance metrics which show that accuracy is 77.60%, the sensitivity (also called the true positive rate) is 69.35%, and the specificity (also called the true negative rate) is 81.54%. This indicates that 81.54% of the cases without diabetes and 69.35% of cases with diabetes were accurately diagnosed by model.

# Feature Selection and Model Optimization



Employing Recursive Feature Elimination (RFE) method from caret package to select most predictive features for a binary classification model predicting diabetes outcome. RFE iteratively builds models with fewer predictors and cross-validates performance to find best subset of features which here seems to be Glucose, Age, BMI, and Insulin. The chosen features are used to train a regularized logistic regression model using glmnet method which is beneficial in preventing overfitting and improving model generalizability.

The model’s performance is then evaluated using ROC analysis plot that illustrates diagnostic ability of a binary classifier system as its discrimination threshold is varied provided by pROC package. The AUC (Area Under the ROC Curve) is calculated which quantifies overall ability of the model to correctly classify individuals with and without diabetes. This step is crucial for understanding how well model might perform when making predictions on new unseen data. The entire process is encapsulated within a reproducible framework which is important for consistent data science practices.

Age, BMI, insulin, and glucose were found to be most important variables predictive of diabetes by use of recursive feature selection. These findings are in line with established risk factors for disease. The model performed better since just the most important predictors were included thanks to this feature selection method.

A white background with black text

Description automatically generated

201 patients have diabetes and 375 do not indicating moderate imbalance in outcome variable's balance. One class may inherently dominate other in medical datasets this imbalance was controlled during modelling procedure.

**During Model Training:** Imbalance handling comes into play during model training. The trainControl function specifies use of repeated cross validation and includes classProbs=TRUE, which suggests that probabilities are being estimated for each class. This can be useful if you later decide to adjust decision thresholds. The summaryFunction=twoClassSummary indicates that evaluation metrics beyond simple accuracy are employed. These metrics (like sensitivity, specificity, and AUC) are better suited for datasets with class imbalance as they give a more nuanced view of model performance.



Overall model's accuracy and AUC score's validation show that it has potential to be useful tool for predicting diabetes risk, which could help with early diagnosis and improved treatment of illness. The overall model performance increase to 0.83 form 0.77.We have found the best model accuracy with the help of cross-validation that is actually k-fold cross validation. And we have used the regularization technique to increase the accuracy of model.



A graph of a curve

Description automatically generated

The final model's ROC curve, which has an AUC of 0.8349 and a dashed diagonal line that represents a no-discrimination classifier, demonstrates the model's superior performance. This indicates that the model can effectively distinguish between those who have diabetes and those who do not. The curve indicates a solid balance between sensitivity and specificity across various thresholds because it remains considerably above the line of no discrimination.

**Black Stepped Line:** This represents the ROC curve itself. Each point on this line reflects sensitivity/specificity pair corresponding to particular decision threshold. Sensitivity (true positive rate) is on the y-axis, and 1-specificity (false positive rate) is on x-axis. The curve starts at bottom left (representing a threshold that classifies all instances as the negative class) and ends at top right (representing threshold that classifies all instances as positive class).

**Grey Diagonal Line:** This line represents 'no-skill' classifier, which is model that provides no discriminative ability between positive and negative classes. It's where model's probability of correctly identifying true positives is equal to probability of incorrectly identifying negatives as positives. Essentially it's line a model would follow if it was guessing randomly.

**Red Dashed Line:** This usually represents line of reference to help visualize where ROC curve stands in relation to perfect performance. In a typical ROC space, this line might be drawn from top left to the bottom right representing trade-off between sensitivity and specificity (although in diagram it seems incorrectly drawn extending into nonsensical negative specificity values). A perfect model would have curve that goes to top left corner,meaning 100% sensitivity (no false negatives) and 100% specificity (no false positives).

The closer ROC curve is to the top left corner higher overall accuracy of the test. The Area Under Curve (AUC) can be calculated for this ROC curve to quantify the model’s performance and an AUC of 1.0 represents a perfect model, while an AUC of 0.5 represents a model with no discriminative ability.

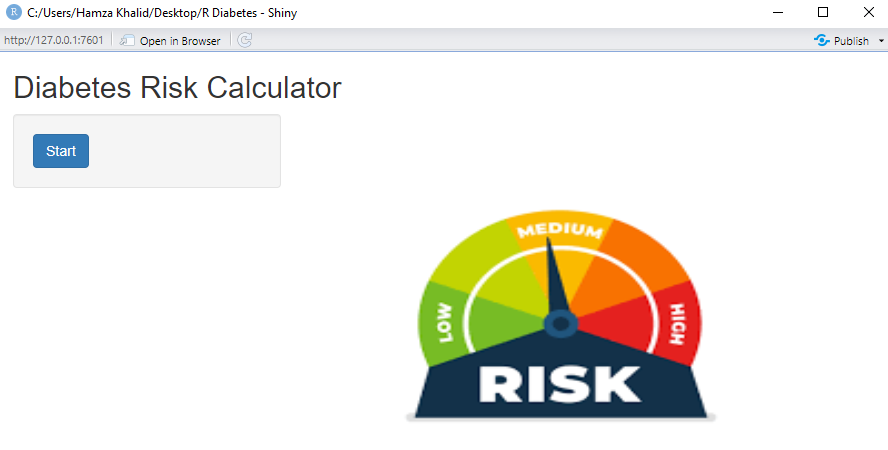
# 

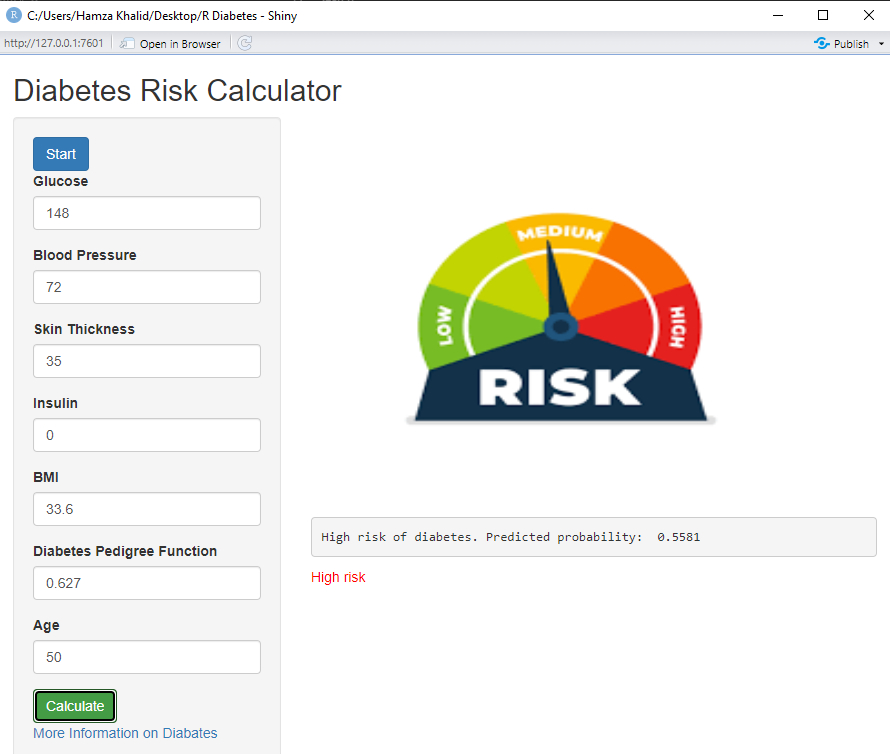
# Shiny Application Development

The diabetes risk calculator displayed in screenshots is a Shiny application that exemplifies an intuitive user interface that turns predictive model into a useful utility. Users are first shown "Start" button which when clicked displays form asking for vital health information including Age, Diabetes Pedigree Function, Skin Thickness, Insulin, Blood Pressure and BMI. Most likely most important variables found during the feature selection stage of the model building are source of these inputs.

The program employs predictive model to estimate user's chance of developing diabetes when they enter their health characteristics and click 'Calculate' button. This risk is then displayed as a probability score. The application reads given example's projected probability of 0.5581 as indicating a high risk of diabetes.

This application demonstrates how statistical models can be practically applied in real-world settings. The Shiny app gives people ability to understand their health and maybe take preventative action by allowing them to enter their health data and receive instant risk assessments. Patients and healthcare providers alike can utilize this concrete result of data science effort to determine their risk of diabetes.





# Conclusions

After thorough investigation of diabetes dataset with statistical and machine learning techniques reliable prediction model that can determine a person's probability of developing diabetes was produced. An AUC of 0.8349 indicates that this model which was created by meticulously preprocessing data selecting important features and validating predicted accuracy has shown good performance characteristics. The Shiny program which offers user-friendly interface for real-time diabetes risk assessment illustrates practical implementation of this research.

After thorough investigation of diabetes dataset with statistical and machine learning techniques reliable prediction model that can determine a person's probability of developing diabetes was produced. An AUC of 0.8349 indicates that this model which was created by meticulously preprocessing data, selecting important features and validating predicted accuracy has shown good performance characteristics. The Shiny program which offers a user-friendly interface for real-time diabetes risk assessment illustrates practical implementation of this research.

Notwithstanding model's success it's critical to recognize study's inherent limits. Although significant model's predictive capacity is not absolute and there may be room for further predictors or sophisticated modelling strategies to be explored in order to further improve accuracy. The results of this study will have a big impact on public health especially in field of preventive medicine. Early risk identification makes it possible to target interventions more effectively and closely monitor those who are at risk which may postpone or even prevent onset of diabetes.

As a result this project serves as testament to importance of data analysis in healthcare field. It provides promising tool for diabetes risk prediction and opens door for further research involving integration of lifestyle and genetic factors for a more comprehensive approach to diabetes risk assessment investigation of novel modelling techniques and use of more diverse datasets.