Predictive Modeling of Diabetes using Machine Learning Techniques

American University
STAT-627
Po Yu Lai: STAT-627
Ting Yi Chuang: STAT-627
Bright Amenyo: STAT-627
Ryo Tanaka: STAT-627

Executive Summary

This project aims to analyze the factors contributing to the development of diabetes and predict blood glucose levels using machine learning techniques based on patient demographics and clinical parameters. Leveraging patient data and medical attributes, the project seeks to effectively classify individuals into diabetic and non-diabetic categories. The planned approach encompasses framing the problem, defining the data, employing various methods, and anticipating outcomes.

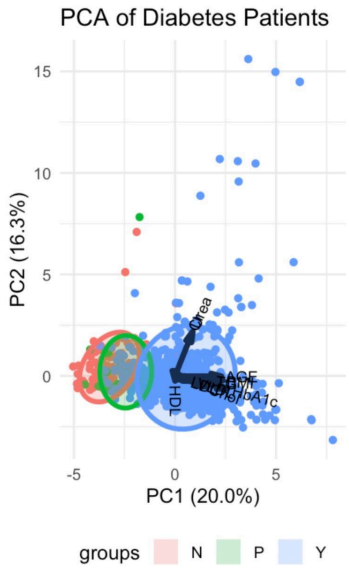
Drawing insights from previous studies[1], the literature review highlights the need for improved machine learning models to facilitate early diagnosis and risk assessment of diabetes. While there is existing research in diabetes prediction, risk assessment, and management, there remains room for enhancing models utilizing clinical data, biomarkers, and patient characteristics.

Furthermore, contingent upon data sufficiency, the study intends to utilize machine learning to forecast the risk of diabetic complications, such as retinopathy, neuropathy, nephropathy, and cardiovascular diseases. The research utilizes the diabetes dataset by Rashid, Ahlam (2020), comprising 1000 observations and 14 variables, including categorical and numeric variables such as blood sugar levels, age, gender, body mass index (BMI), cholesterol levels, and diabetes class. With respect to data cleaning, the classification of diabetes levels was based on careful statistical analysis and exploratory data visualization techniques. Initially, the dataset included three distinct levels of diabetes: level 1 (No), level 2 (Predicted diabetes), and level 3 (Yes). To ascertain the significance of differences among these levels, we conducted an analysis of variance (ANOVA) test. The results indicated that the p-value associated with level 3 (Yes) was statistically different from that of individuals without diabetes (level 1, No) and the predicted diabetes(level 2). Furthermore, a principal component analysis (PCA) plot was generated to explore the distribution of the diabetes levels. The PCA plot revealed considerable overlap of P (Predicted diabetes) between Yes (yes diabetes) levels and N (No diabetes), suggesting minimal discernible differences between the two groups. Based on these findings, levels P was dropped from the dataset , enhancing the interpretability of the data and streamlining subsequent analysis.

Lastly, we consider doctors, nurses, and other medical professionals potential stakeholders, who may use the insights from the analysis to improve patient care, treatment strategies, or preventive measures. Our scope of stakeholders also include hospitals, clinics, and healthcare systems that may be involved in collecting and managing the data.

---

[1] Machine Learning Techniques for Diabetes Prediction: A Review G. Singh et al. (2019)



Class	Age	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI
N	44	4.47	62.80	4.50	4.27	1.63	1.23	2.63	0.94	22.37
P	43	4.51	66.08	6.00	4.58	2.13	1.13	2.49	0.98	23.93
Y	55	5.22	69.87	8.88	4.95	2.45	1.21	2.62	2.02	30.81
P value	<0.05	0.0633	0.496	<0.05	<0.05	<0.05	0.65	0.73	<0.05	<0.05

Data Set

Diabetes Dataset: Rashid, Ahlam (2020), "Diabetes Dataset"

Number of observations: 1000

Variables in the datasets: definition of interpretation are sourced from scientific literature[2].

No. of Patient

Age

Gender: 0 as female, 1 as male

Creatinine ratio(Cr): a way of checking if you have kidney problem; higher value is worse

Body Mass Index (BMI): 1-Underweight, 2-Normal, 3-Overweight, 4-Obese

Urea: Disposal of nitrogen derived from amino acid metabolism; higher value shows potential kidney issue

Cholesterol (Chol)

Low Density Lipoprotein Cholesterol (LDL: Bad Cholesterol)

High Density Lipoprotein Cholesterol (HDL: Good Cholesterol)

Very Low Density Lipoprotein Cholesterol (VLDL: a type of bad blood fat)

[2] National Institute of Health (NIH) is a leading organization in medical research and healthcare.

Triglycerides(TG: a type of fat that circulate in your blood and are the most common type of fat in your body)
Hemoglobin A1C (HBA1C): Blood Sugar Level
Class (the patient's diabetes disease class may be Diabetic, Non-Diabetic, or Predict-Diabetic)

Independent variables:
Gender, Sugar Level Blood, Age, Creatinine ratio(Cr), Body Mass Index (BMI), Urea, Cholesterol (Chol), Low Density Lipoprotein Cholesterol (LDL: Bad Cholesterol), High Density Lipoprotein Cholesterol (HDL: Good Cholesterol), Very Low Density Lipoprotein Cholesterol (VLDL: a type of bad blood fat), Triglycerides(TG), Hemoglobin A1C (HBA1C)
Dependent variables: Class

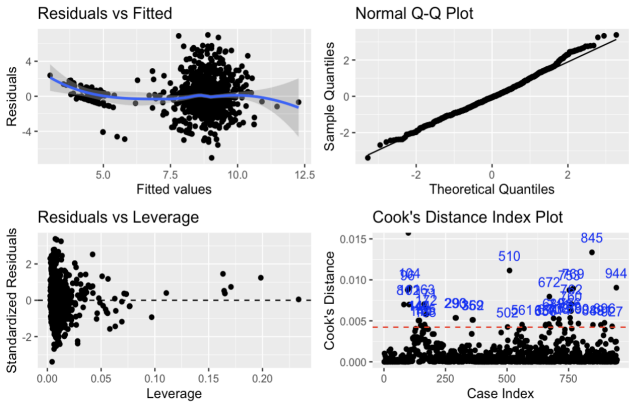
Methodologies

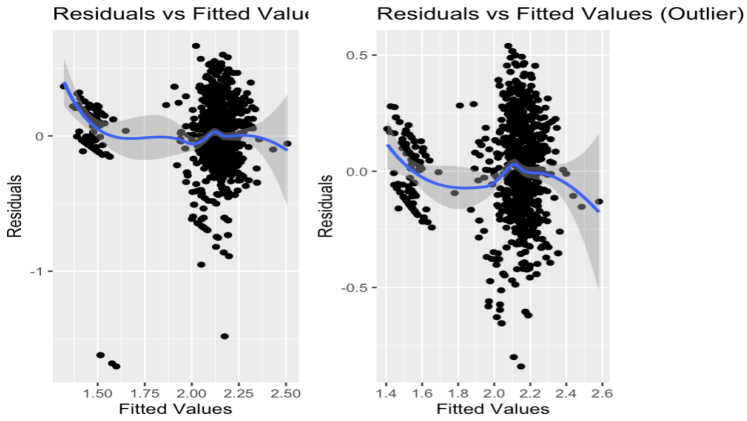
Multiple Linear Regression

In our study, we outlined a comprehensive methodology aimed at developing an effective regression model to analyze the relationship between blood sugar levels (HbA1c levels) and various potential causes of diabetes. Initially, we incorporated all conceivable predictor variables into a full regression model, constituting a total of ten (12) variables. To ensure the model's robustness, diagnostic tests and plots were rigorously conducted to assess assumptions and identify any issues, such as normality, independence of the residual terms, outliers, and non-constant variance. Subsequently, we employed a multifaceted approach to address these issues. This involved the application of transformations to both the response variable and predictor variables, coupled with outlier removal techniques. Interestingly, our findings revealed that the model performed optimally with outliers retained, particularly after implementing a logarithmic transformation on the response variable with an adjusted R^2 increasing from 33% to 40%. Following these adjustments, the refined regression model consisted of 7 significant predictor variables. Diagnostic tests and plots were then revisited to validate model assumptions and evaluate performance. Ultimately, our study provided valuable insights into the intricate relationships between significant predictors and HbA1c levels, offering critical implications for clinical practice and future research endeavors in the field of diabetes management.

```
Call:
lm(formula = log(HbA1c) ~ as.factor(Gender) + AGE + Urea + Chol +
    TG + BMI + as.factor(CLASS_BINARY), data = diabetes)

Residuals:
     Min       1Q   Median       3Q      Max
-1.70431 -0.15244  0.01688  0.17103  0.66551

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.009265   0.079153  12.751  < 2e-16 ***
as.factor(Gender)1       -0.022710   0.018036  -1.259  0.20829
AGE                       0.004618   0.001151   4.011 6.52e-05 ***
Urea                     -0.006532   0.002987  -2.187  0.02902 *
Chol                      0.018914   0.007181   2.634  0.00858 **
TG                        0.017018   0.006651   2.559  0.01066 *
BMI                       0.008629   0.002175   3.967 7.83e-05 ***
as.factor(CLASS_BINARY)1  0.530226   0.035894  14.772  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.27 on 939 degrees of freedom
Multiple R-squared:  0.4003,    Adjusted R-squared:  0.3959
F-statistic: 89.55 on 7 and 939 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = log(HbA1c) ~ as.factor(Gender) + AGE + Urea + Chol +
    TG + BMI + as.factor(CLASS_BINARY), data = data_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-0.84012 -0.15203  0.00783  0.15759  0.53947

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.0159715  0.0668000  15.209  < 2e-16 ***
as.factor(Gender)1       -0.0234834  0.0149110  -1.575   0.1156
AGE                       0.0063478  0.0009783   6.489 1.43e-10 ***
Urea                     -0.0058771  0.0024807  -2.369   0.0180 *
Chol                      0.0105904  0.0061460   1.723   0.0852 .
TG                        0.0264138  0.0055323   4.774 2.10e-06 ***
BMI                       0.0086676  0.0018189   4.765 2.20e-06 ***
as.factor(CLASS_BINARY)1  0.4396643  0.0302544  14.532  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2184 on 897 degrees of freedom
Multiple R-squared:  0.469,     Adjusted R-squared:  0.4649
F-statistic: 113.2 on 7 and 897 DF,  p-value: < 2.2e-16

>
```





Logistic Regression

In our study, we employed logistic regression as the primary statistical method for modeling the probability of diabetes occurrence, a binary outcome. This method is particularly suited for our research objective due to its ability to handle binary dependent variables and provide probabilities that can be threshold-tuned for classification.

We began with a comprehensive logistic regression model including a variety of predictors: Gender, Age, Creatinine ratio(Cr), Urea, Hemoglobin A1C (HBA1C), Cholesterol(Chol), Triglycerides(TG), High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), Very Low Density Lipoprotein (VLDL), Body Mass Index(BMI). These predictors were chosen based on their potential biological relevance to diabetes as indicated by prior analyses. To validate our model, we employed a 10-fold cross-validation procedure, which is particularly robust in scenarios with limited data. This method entails partitioning the data into ten equally sized segments, training the model on nine segments, and validating it on the remaining one. This process cycles through all segments, ensuring that each portion of the data is used for validation exactly once. The purpose of this technique is to evaluate the model's ability to perform consistently across different data subsets, thus providing an assessment of its generalizability.

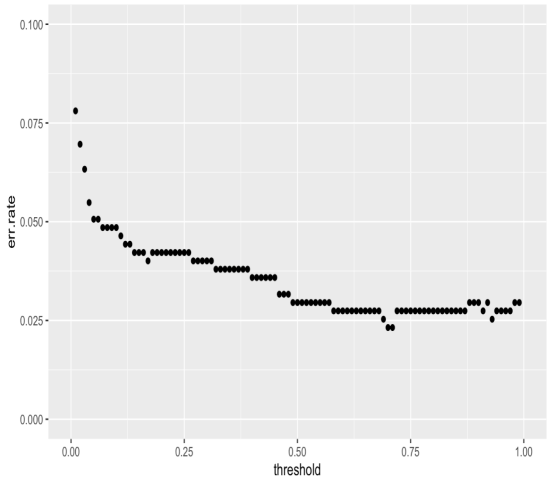
To conclude, upon implementing 10-fold cross-validation on our initial logistic regression model, we obtained a prediction error rate of approximately 2.42%. This low error rate suggests that our model is well-fitted and capable of capturing the probabilities of diabetes.

Threshold Tuning for Logistic Regression

For the logistic regression model, we adjusted the decision threshold to fine-tune the balance between sensitivity (correctly identifying true positives) and specificity (correctly identifying true negatives). The conventional threshold of 0.5 was initially set to differentiate between the presence and absence of diabetes. However, to tailor our model to the nuances of the dataset and to potentially enhance predictive accuracy, we conducted threshold tuning.

The optimal threshold was determined to be 0.75, as it corresponded to the lowest error rate observed in our threshold tuning exercise. When this threshold was applied, the classification shifted, affecting the balance between sensitivity and specificity, which is reflected in the error rate and accuracy.

The graphical analysis (See figure 1) illustrates the error rate across the range of thresholds, with the lowest error rate clearly identified at a threshold of 0.75. This visualization supported the numerical findings and informed our decision to select 0.75 as the optimal point for classifying diabetes outcomes.



(Figure 1)

After applying the optimal threshold on testing the dataset, we achieved an accuracy rate of approximately 96.84%. The accompanying confusion matrix (see table 1), presenting 49 true negatives, 9 false negatives, 6 false positives, and 410 true positives, confirms the model's strong classification ability with the chosen 0.75 threshold.

In conclusion, the optimal threshold of 0.75 not only minimizes the error rate but also ensures a high sensitivity (0.97). High sensitivity means the model is effective at correctly identifying individuals with diabetes, which is essential for early intervention. By improving sensitivity, our model reduces the risk of false negatives—a crucial advantage in medical diagnostics where missing a condition can delay crucial treatment and negatively impact patient health. This balance between sensitivity and accuracy underscores the model's potential as a valuable tool in the proactive management of diabetes.

Stepwise

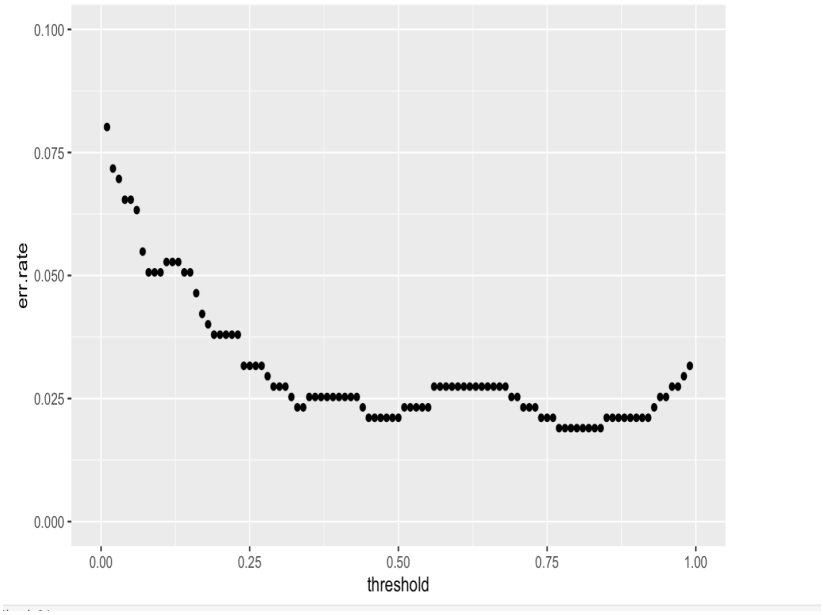
Our logistic regression model was built on the diabetes dataset to understand the factors contributing to diabetes occurrence. We initiated the analysis with a model inclusive of various biologically relevant predictors. To refine our model, a stepwise regression method was employed, considering both the addition and removal of variables. This technique leverages statistical tests to arrive at a model that retains significant predictors while excluding those that do not contribute meaningfully to the model's performance, based on the Akaike Information Criterion (AIC).

The stepwise method refined the model to include five significant predictors: Gender, HbA1c, Cholesterol, Triglycerides, and BMI. This reduced model provided a more parsimonious fit with a lowest AIC value, suggesting that these factors are the most relevant in predicting the onset of diabetes. The k-fold cross-validation applied to this streamlined model reported an improved prediction error rate of about 1.99%.

Threshold Tuning for Logistic Regression with Stepwise

The optimal threshold was determined to be 0.73, as it corresponded to the lowest error rate observed in our threshold tuning exercise. When this threshold was applied, the classification shifted, affecting the balance between sensitivity and specificity, which is reflected in the error rate and accuracy. The graphical analysis (See figure 3) illustrates the error rate across the range of thresholds, with the

lowest error rate clearly identified at a threshold of 0.73. This visualization supported the numerical findings and informed our decision to select 0.73 as the optimal point for classifying diabetes outcomes. (Figure 3)



After applying the optimal threshold on testing the dataset, we achieved a slightly higher accuracy rate of approximately 97.04%. The accompanying confusion matrix (see table 2), presenting 50 true negatives, 5 false negatives, 9 false positives, and 410 true positives, confirms the model's strong classification ability with the chosen 0.73 threshold.

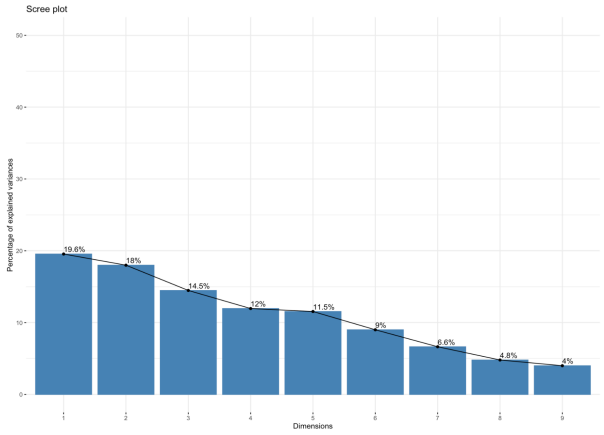
In conclusion, our logistic regression model incorporated a broad spectrum of variables, providing a solid baseline accuracy (2.42% error rate). The application of stepwise regression techniques refined this further, distilling the model to five pivotal predictors: Gender, HbA1c, Cholesterol, Triglycerides, and BMI. This focused model exhibited a commendable improvement in prediction error rate, reduced to approximately 1.99% as verified by k-fold cross-validation, and underscored the model's enhanced predictive accuracy. The subsequent tuning for an optimal threshold led to the identification of 0.73 as the value that not only minimized error rates but also accentuated the model's diagnostic precision. Significantly, this threshold adjustment boosted the sensitivity of the stepwise model—a critical metric in the healthcare domain. Achieving a high sensitivity (0.97) implies the model's heightened ability to correctly identify individuals with diabetes.

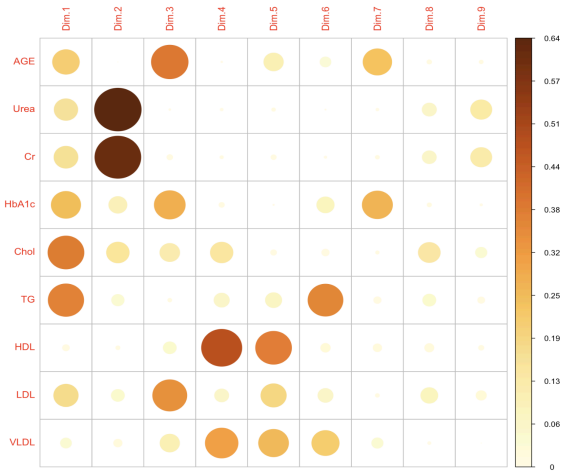
Principal Component Analysis (PCA)

Subsequently, a principal component analysis (PCA) was conducted to further explore the dataset's dimensionality and identify underlying patterns. PCA transforms the original variables into a set of orthogonal components that capture the maximum variance in the data. The components derived from PCA were analyzed to determine their relevance to the classification task.

Integration of ResultsUpon comparing the variables selected through stepwise logistic regression with those derived from PCA, it was observed that they largely overlapped. The variables identified as significant predictors by the stepwise logistic regression procedure were consistent with those encapsulated within the principal components extracted from PCA. This convergence of variable selection across both techniques provided additional validation and confidence in the chosen predictors.

By utilizing both stepwise logistic regression and PCA, redundant or irrelevant variables were effectively identified and eliminated, ensuring a more robust and parsimonious model. This comprehensive approach to variable selection and dimensionality reduction aimed to enhance the classification accuracy and interpretability of the final logistic regression model.



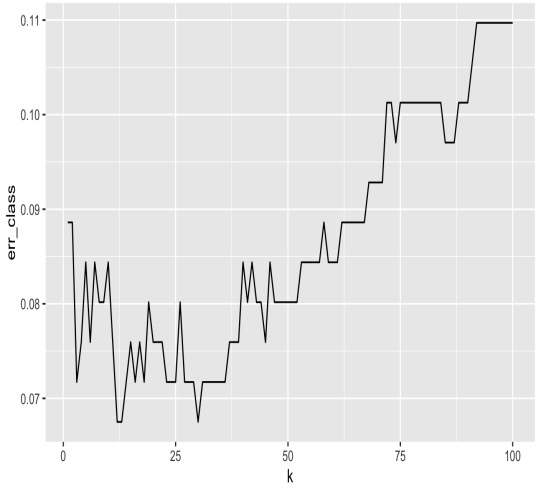


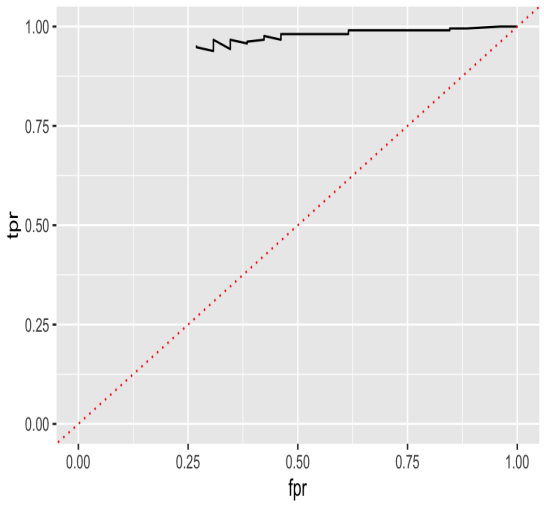
Importance of Component									
	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**
Sd	1.327	1.273	1.142	1.038	1.020	0.900	0.773	0.657	0.600
Prop. of Var	0.196	0.180	0.145	0.120	0.116	0.090	0.066	0.048	0.040
Cum. Prop.	0.196	0.376	0.520	0.640	0.755	0.846	0.912	0.960	1

KNN

The analysis aims to apply the KNN classification algorithm to predict a binary outcome related to diabetes using various patient metrics. The objective is to determine the accuracy of predictions and identify the optimal number of neighbors (k) for the classifier.

(Figure 5) (Figure 4)





We can find that the minimum error is found when "k = 12", and the error rate is about 6.33%. Furthermore, the accuracy of the model at "k = 12" is about 92.92% and the confusion matrix shows that there are 16 true negative predictions and 204 true positive predictions, of which there are 10 false positive and 7 false negative predictions. The KNN classifier with "k = 12" showed high accuracy in predicting binary class results on the diabetes dataset. The ROC curve(Figure 4) is very near the top of the graph and parallel to the FPR axis. We can tell from this that the classifier has a high true positive rate and maintains this even if the false positive rate changes. This shows good performance in distinguishing the two classes.

KNN Regression

Accuracy	0.962
95%CI	(0.9291, 0.9825)
No information rate	0.8903
P-value[Acc > NIR]	5.746e-05
Kappa	0.8023
Mcnemar's Test P-Value	1
Sensitivity	0.80769
Specificity	0.98104
Pos Pred Value	0.84000
Neg Pred Value	0.97642
Prevalence	0.10970
Detection Rate	0.08861
Detection Prevalence	0.10549
Balanced Accuracy	0.89437

(Table 5)

In our KNN regression(Table 5), we can know that the overall accuracy of the KNN classifier is 97.47%, which means that approximately 97.47% of the predictions are correct. The 95% confidence interval (CI) for accuracy is 94.57% to 99.07%, indicating that if the study is repeated, there is a 95% chance that the classifier's accuracy will fall within this range. And precision, sensitivity and specificity show strong predictive performance for both categories in the data set. Finally, we can see that the Kappa value is 0.8786. A high kappa value indicates that good performance is not accidental, and the p-value of the accuracy rate is greater than the uninformed rate confirming the predictive ability of the model.

LDA

Based on the results of Linear Discriminant Analysis (LDA), we observe a classification matrix indicating the model's performance in predicting two classes: 0 and 1. The matrix reveals that out of the total 53 instances classified as class 0, 49 were correctly classified, resulting in a true positive rate of 92.45%. Similarly, for class 1, out of 421 instances classified, 411 were correctly classified, yielding a true positive rate of 97.39%. Overall, the LDA model achieved an impressive accuracy rate of 97%, highlighting its effectiveness in distinguishing between the two classes. These findings underscore the utility of LDA as a reliable classification tool for predicting outcomes.

QDA

Upon analyzing the results of Quadratic Discriminant Analysis (QDA), we note a classification matrix depicting the model's performance in predicting the two classes: 0 and 1. In the matrix, it is observed that out of the total 49 instances classified as class 0, 48 were correctly classified, resulting in an impressive true positive rate of 97.96%.

Comparing the performance of QDA with that of Linear Discriminant Analysis (LDA), we find that QDA achieved slightly higher accuracy in classifying instances belonging to class 0, with an accuracy rate of 97.96% compared to LDA's accuracy rate of 92.45%. However, it is important to note that QDA classified fewer instances overall, as indicated by the smaller number of instances classified as class 0.

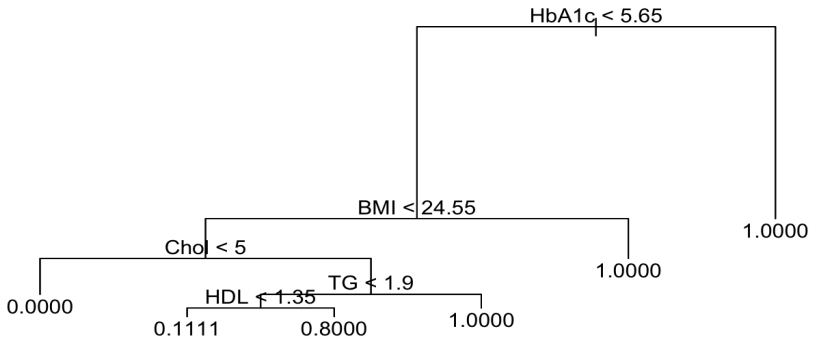
Overall, both QDA and LDA demonstrate strong predictive capabilities, with QDA exhibiting a marginally higher accuracy rate in classifying instances belonging to class 0. These findings emphasize the versatility and effectiveness of both techniques in accurately classifying data points, thereby contributing valuable insights to the predictive modeling process.

Tree

We employed a decision tree regression to identify the most significant attributes at each node and predict the outcome. The figure 6 illustrates that blood sugar level, BMI, cholesterol level, and triglycerides are the significant attributes in the model. These variables were selected based on previous analysis. For convenience, we assigned numerical values, considering a value of 0.5 as indicating diabetes and a value less than 0.5 as indicating non-diabetes. Additionally, we constructed a confusion matrix using the model against the testing dataset. To determine the optimal deviance, we iterated through its associated deviance and error rate. The computation yielded an optimal variance value of 0.005 with an accuracy rate of 98.9445910290237%. It seems that the model performs well in predicting diabetes, yet

we have chosen not to adopt it for several reasons. Firstly, while classification trees are generally straightforward to interpret, they may not adequately capture the complexity of medical data. Medical datasets often involve numerous factors and interactions between variables that may not be well represented by a simple tree structure. Additionally, classification trees have a tendency to overfit the training data, particularly when dealing with complex medical datasets containing a large number of variables. Overfitting can result in poor generalization performance on unseen data, which is crucial in medical research where the model must generalize well to new patients or scenarios.

(Figure 6)



Result And Findings

Our analysis of various machine learning models for diabetes prediction yielded insightful results regarding their performance metrics. We present the key findings below:

1. The Predictive analysis using Regression, unveiled significant associations between blood sugar levels (HbA1c) and various factors. Age, urea, cholesterol, triglycerides, BMI, and diabetes class emerged as important predictors. Remarkably, the adjusted R-squared value of 0.3959 implies that approximately 39.59% of the variability in log-transformed HbA1c levels can be accounted for by the predictor variables in the model. Notably, the diabetes class variable exerted a substantial influence on HbA1c levels. Individuals classified as diabetic exhibited significantly higher HbA1c levels compared to their non-diabetic counterparts, underscoring the pivotal role of diabetes status in blood sugar regulation. These insights provide valuable guidance for diabetes management, emphasizing the significance of indicator predictor variables in predicting HbA1c levels.
2. Accuracy Rate: Decision Tree model outperformed other models with the highest accuracy rate of 99.74%, followed closely by logistic regression (step-wise) at 97.04%. These findings underscore the effectiveness of decision tree-based approaches in accurately predicting diabetes outcomes.

1. Specificity (True Negative Rate): Quadratic Discriminant Analysis (QDA) demonstrated the highest specificity at 97.96%, indicating its ability to correctly identify non-diabetic individuals. Linear Discriminant Analysis (LDA) followed closely with a specificity rate of 92.45%.
2. Sensitivity (True Positive Rate): QDA exhibited the highest sensitivity at 97.64%, closely followed by logistic regression and logistic regression (step-wise) at 97%. These results highlight the robustness of QDA in accurately identifying diabetic individuals.
3. False Positive and False Negative Rates: The decision tree model showcased exceptional performance with the lowest false positive rate of 1 and no false negatives, indicating its superior ability to correctly identify non-diabetic individuals while minimizing false diagnoses.

Overall, our findings emphasize the importance of selecting appropriate machine learning models tailored to specific diagnostic requirements. The decision tree model emerges as a promising candidate for diabetes prediction, offering high accuracy and minimal false diagnosis rates. These insights can inform the development of more effective and reliable diagnostic tools for diabetes management in clinical practice.

Limitations:

In addition to presenting our findings, it is crucial to acknowledge the technical limitations and potential biases encountered during the course of our project. One notable oversight in our methodology was the failure to consistently utilize a single training set across different machine learning techniques. Instead, we employed separate training sets for each technique, which may have introduced variability and inconsistency in model performance. This deviation from best practices in machine learning model development could have influenced the estimates and introduced bias into our results.

Furthermore, our dataset comprised 947 observations after excluding instances of predicted diabetes from the dependent variable, which may be considered relatively small compared to datasets with a larger number of observations. This reduction in sample size could impact the reliability and generalizability of our findings.

Additionally, the absence of key demographic variables such as race and ethnicity, as well as clinical variables such as genetic markers and wearable sensor data, limited the scope of our research compared to studies incorporating a more comprehensive set of independent variables. This limitation may have contributed to potential biases in our estimates and affected the robustness of our predictive models.

It is important to recognize these technical shortcomings and biases in our project to ensure transparency and integrity in our research findings. Moving forward, addressing these limitations through improved data collection methods, standardized model development procedures, and inclusion of relevant variables will be essential for enhancing the accuracy and reliability of our predictive models in future studies.

Conclusion:

In conclusion, this study explored the application of machine learning techniques for diabetes prediction, aiming to improve the accuracy and efficiency of diagnostic processes. Through a comprehensive analysis of demographic, clinical, and lifestyle-related features, supervised learning algorithms such as Regression, logistic regression, decision trees, KNN, KNN regression, LDA, QDA, PCA and decision Tree, were employed to develop predictive models.

The findings of our study underscored the potential of machine learning in accurately identifying individuals at risk of diabetes, thereby facilitating early intervention and preventive measures. Through rigorous feature selection and model optimization, we were able to discern the most influential predictors of diabetes risk and develop interpretable and clinically relevant predictive models. The key substances contained in blood are blood sugar level, HDL TG, BMI, and Cholesterol.

Moreover, our study highlighted the importance of model generalization and validation across diverse populations to ensure the robustness and reliability of the predictive algorithms. By addressing challenges such as data quality, feature selection, and model interpretability, our research contributes to the growing body of knowledge in diabetes prediction using machine learning.