Cracking the Sugar Code: Predicting Diabetes in Patients

# Introduction

Diabetes, a chronic medical condition characterized by elevated blood sugar levels, has emerged as a global health crisis in recent years. Its prevalence has been steadily increasing, with millions of individuals affected worldwide. What is particularly concerning is that diabetes often remains undiagnosed until it reaches an advanced stage, leading to severe health complications and an increased burden on healthcare systems. This undermines the critical importance of early diabetes detection, as it can significantly improve the quality of life for those at risk and prevent or delay the onset of serious complications. This essay will explore multiple applications in which data mining can help address this global issue by giving doctors tools to help identify patients at risk and start preventative and early intervention treatments. By utilizing Explanatory Analysis, Linear Regression, Association Rule Mining, Naïve Bayes, Support Vector Machine (SVM), Random Forest Decision Tree (RFDT), and Clustering algorithms; We hope to identify the mains factors behind the cause diabetes as well as the demographic the disease mainly affects.

# Data Description

The data set was obtained from Kaggle and is a collection of medical and demographic data from patients, along with their diabetes status.

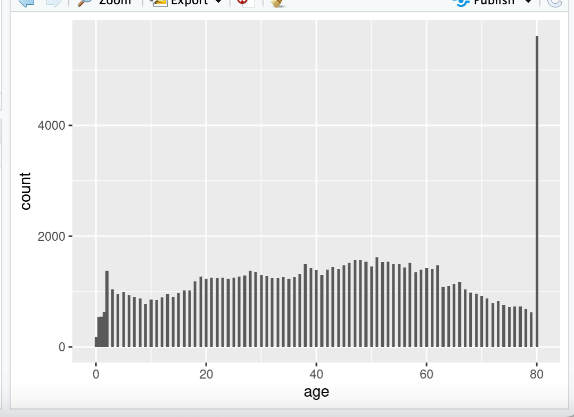
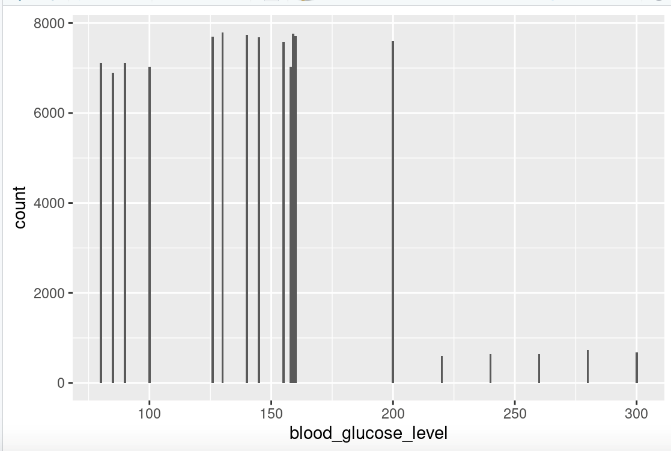
The variables are:

* Gender – categoric variable (Male, Female, or Other).
* Age – numeric variable representing age in years (ranges from 0-80).
* Hypertension – binary variable (1 or 0) on whether the patient has hypertension or not.
* Heart\_disease – binary variable (1 or 0) on whether the patient suffers from heart disease or not.
* Smoking\_history – categoric variable representing patients’ smoking status (no info, never, current, not current, former, ever)
* BMI – numeric value which is the measure of body fat based on weight and height.
* HbA1c\_level – numeric variable which is the measure of a person's average blood sugar level over the past 2-3 months.
* Blood\_glucose\_level – numeric variable representing the amount of glucose in the bloodstream at a given time.
* Diabetes – binary variable (1 or 0) representing whether the patient has diabetes or not.

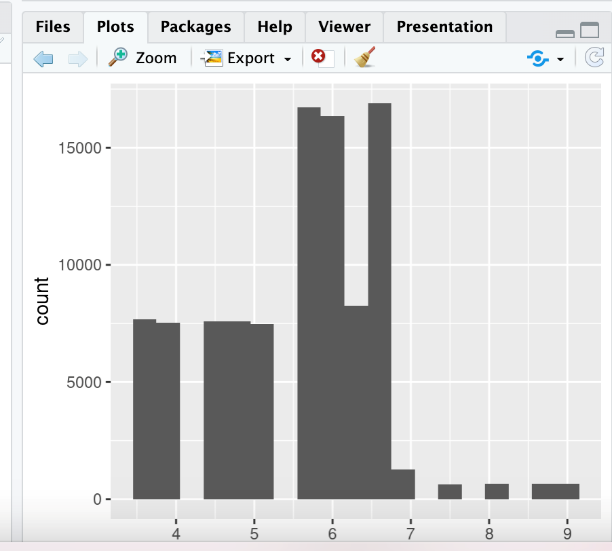
# Methodology

## Explanatory Analysis

For the explanatory analysis, the preparation needed was to download and load the data from Kaggle into R Studio. After the data is loaded into R, I then computed histograms of each variable from the dataset; I found that the ones with the most numerical data were blood glucose level, age, and HbA1c level. Seen below are histograms of these 3 variables (first is blood glucose level, second is age, and third is HbA1c level). The first histogram (*figure 1A*) of blood glucose level shows how scattered and high of a count the level was for patients with slightly elevated levels. Similarly, you can see how low the count was for patients with very high levels; these higher levels are seen in middle aged people; 40 to 60 years old. The second histogram (*figure 1B*) for age simply shows how many people have been tested in each age group. The third histogram (*figure 1C*) shows the HbA1c levels; it shows that the highest levels were between 5.5 and 7.



**FIGURE 1A**  **FIGURE 1B**

 **FIGURE 1C**

In an article “Correlation between Estimated Average Glucose Levels Calculated from HbA1c Values and Random Blood Glucose Levels in a Cohort of Subjects”, summarizes the correlation between age, blood glucose level, and the HbA1c when being diagnosed/undiagnosed with diabetes. For this study (from the article) a group was selected from patient reporting to the laboratory for HbA1c estimation. The simple random sampling technique was used to obtain laboratory records of both sexes in the age range of 12 to 90 years presenting as outpatients. Pregnant females and patients diagnosed with renal disorder were excluded from the study. In one of the tables in the article, it showed the distribution of baseline characteristics of the study subject. The results of that table were very similar to ours from our dataset; ours had a significantly larger sample of people. In another table, they showed the correlation between HbA1c with random blood sugar and the average blood glucose with random blood sugar; the charts were almost identical. In their results, there were very few limitations on the study, and they did not categorize patients into type 1 or type 2 diabetes.

Association Rule Mining

Before running the apriori algorithm, the needed packages were loaded in, and the data was prepared. Since the data set was already clean, the only steps taken were changing the data type to nominal for ARM method and discretizing of the age column. Discretization is a process of taking continuous data attribute values and grouping them into intervals and associating with each interval some specific data value. The age values were separated into their respective intervals, "child", "teens", "twenties", "thirties", "forties", "fifties", "old". *Figure 2A* shows the script and the rules returned from the algorithm.

A computer screen shot of a computer screen

Description automatically generated

**Figure 2A: R Script for Data Preprocessing and Apriori Algorithm set up**

A screenshot of a computer

Description automatically generated

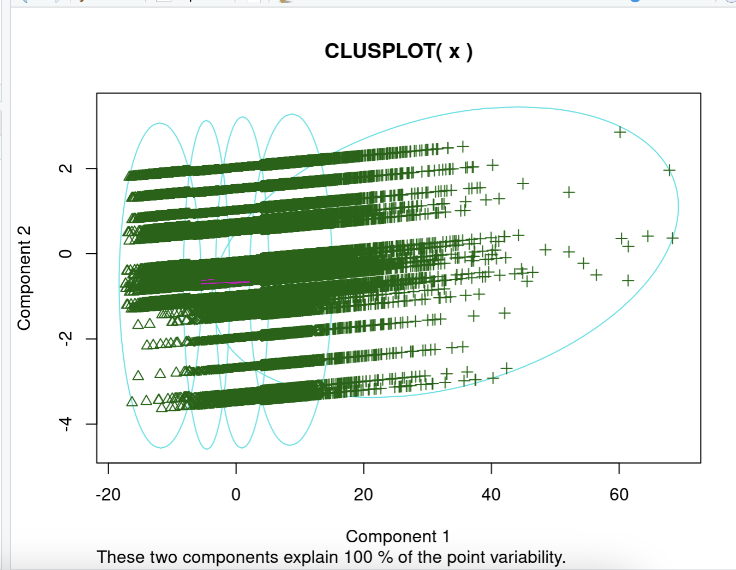
**Figure 2B: List of Rules Based on Diabetes Data set**

Looking at the list in *figure 2B*, it is evident the rules predominantly involve high blood glucose and HbA1c levels, hinting that these may be the main factors behind a person getting the chronic disease. Also, it seems hypertension and heart disease may not be significant when it comes to obtaining diabetes as the rules, they’re involved in show the binary attributes are equal to 0, meaning those conditions are not present. Lastly, old age (individuals 60 and up) was also involved in many of the top rules which gives us an idea the demographic that is mostly affected by diabetes.

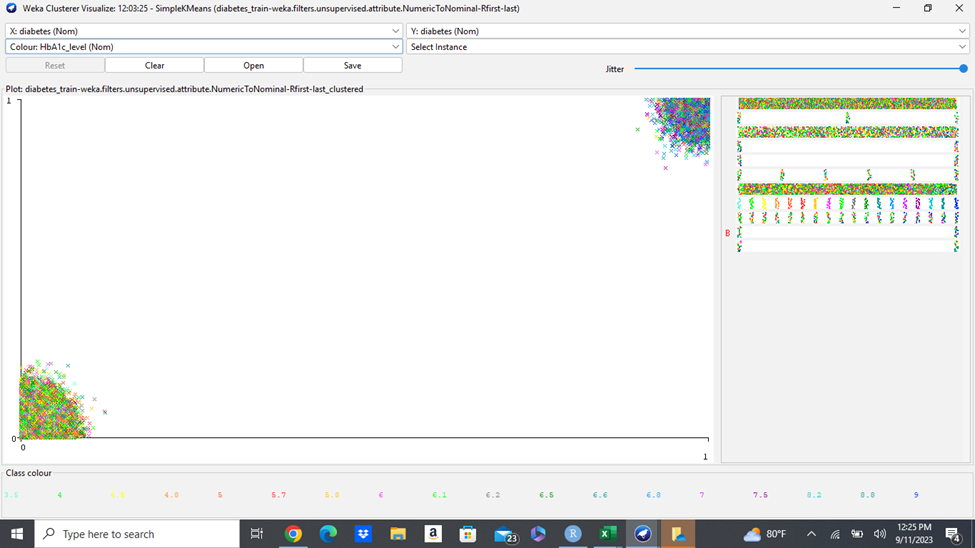
“Association rule for classification of type -2 diabetic patients” by B. M. Patil, R. C. Joshi, and Durga Toshniwal discusses the use of association rule mining to reveal relationships between measured fields such as plasma-glucose levels, BMI, age, etc. and a diabetes diagnosis. The Apriori algorithm is used on a data set that was taken from UCI machine learning repository which contains a total of 768 instances and 8 numeric attributes. For preprocessing, all numeric attributes are converted to nominal attributes to ensure proper results and applied equal interval binning with approximate values based on medical expert advice. Looking at the rules returned, ones containing high plasma-glucose levels and/or high BMI levels, older age, and later stage pregnancies were associated with a positive diabetes diagnosis at a decent confidence level. In their rules where the RHS was a negative diagnosis, medium to low BMI, plasma-glucose, early pregnancy, younger age, and normal Diastolic blood pressure were common. The authors were able to determine that high plasma-glucose and BMI levels as well as older age are heavy factors in a positive diagnosis. I was able to come to the same conclusion as high blood-glucose and BMI levels and older age returned as rules with high confidence in my own study. One difference, however, is the incorporation of pregnancy and more numeric measurements while my dataset included more categorical aspects such as hypertension and heart disease.

Clustering

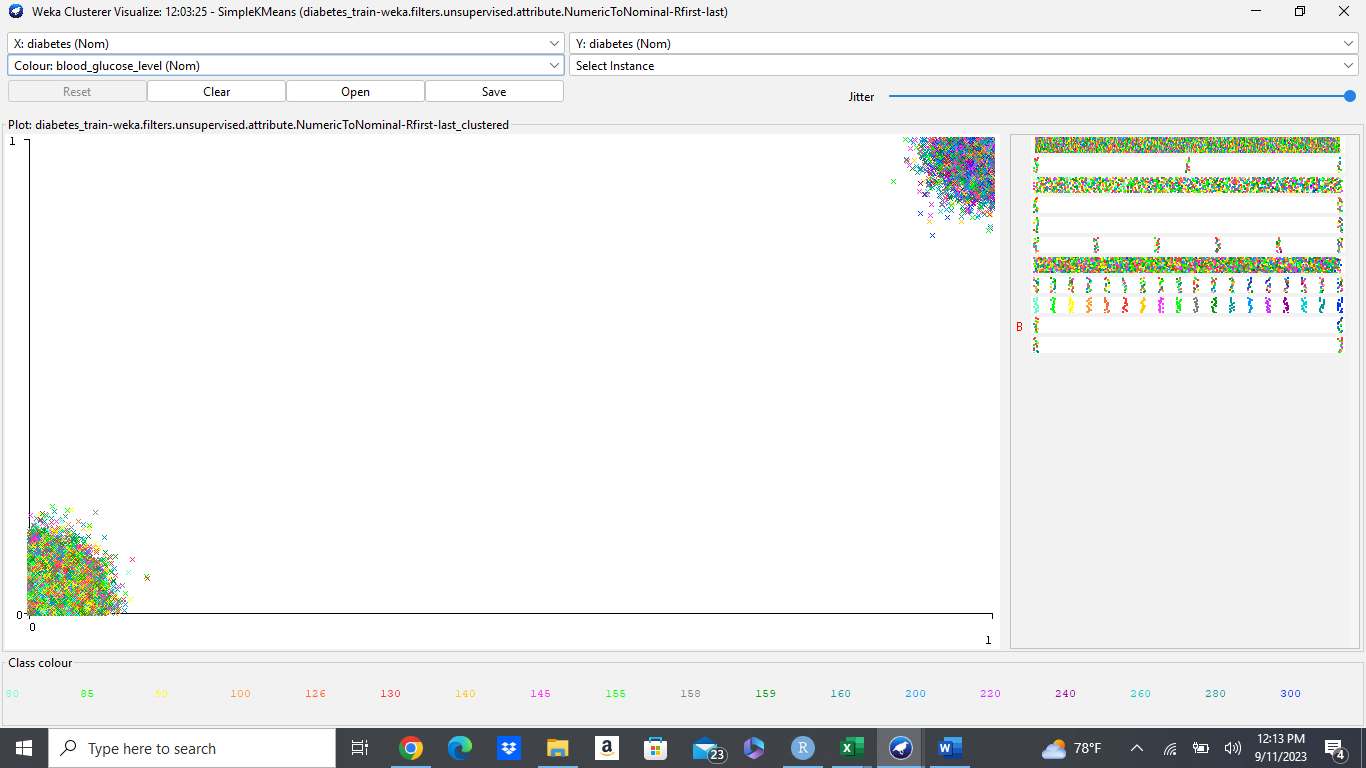
For clustering analysis, I loaded the data into R-Studio and WEKA. Below will show two different clusters: one in R and the other in WEKA. The top cluster (*figure 3A*), plotted in R, shows an overview of all variables from our dataset; it is a little strewn. The second cluster (*figure 3B)*, plotted in WEKA, shows the Hba1c levels. While it shows to be more compact, it is easier to see what color correlates with each variable of the dataset. These models show how different the results are when using two different programs. The WEKA clusters shown below are the only ones out of the 9 attributes to display distinct differences color-wise between the two clusters. The others that did not such as gender, smoking\_history, hypertension, and heart\_disease showed no distinct differences which in turn means there's no considerable differences demographically. Looking at the visual clusters in *figure 3B*, cluster 0 (no diabetes) is dominated by lowers HbA1c levels while cluster 1 (diabetes) is the opposite, showing that patients with diabetes usually measure for high HbA1c levels. It is the same case in *figure 3C* as cluster 1 is mostly occupied by high blood glucose levels while cluster 0 contains data of low blood glucose levels. *Figure 3D* shows our two clusters color sorted by age. Cluster 0 contains an array of colors, telling us the age range is wide and probably contains the data patients of all ages within it while cluster 1 is mostly of red, orange, and yellow hue which represents patients of older age. These clusters tell us the consensus of people who have diabetes display traits of high HbA1c and blood glucose levels as well as older age.



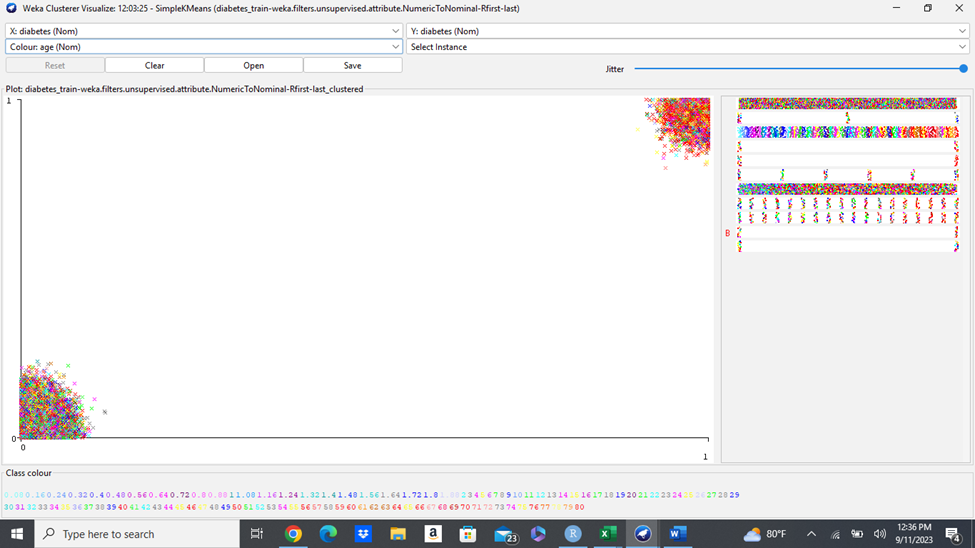
**FIGURE 3A**



**FIGURE 3B**



**Figure 3C**



**Figure 3D**

## Linear Regression

For the linear regression, the only preparation was to create dummy variables for our categorical attributes. As shown below in *figure 3*, the model turned out not to be a great fit. The significance of F is exactly 0 which is below the alpha level of 0.05, meaning there is statistical significance. However, the R Square value is approximately 0.35. This tells us that only 35% of the data can be explained by the model and for that reason we’ve decided to disregard any of the findings usings this model as it looks like linear regression is not good in this situation.

A screenshot of a computer

Description automatically generated

**Figure 4: Linear Regression Results Using Data set**

In the article “Predicting Diabetes with Machine Learning — Part I” by Federico Trotta, the author discusses his attempt at using a simple linear regression model to predict a diabetes diagnosis. The data set used contained measurements such as age, gender, BMI, blood pressure, glucose levels, etc. In the produced model, the author mentions it did not have a good fit as the R2 was close to 0.5, “This is telling us that this type of model is not suitable for this type of problem.” It is also noted that the mean absolute error is well above 0 which is also not a good sign as the metric should be as close to 0 as possible. In my attempt at doing the same thing, I was met with similar results as I got an even worse fitting model with an R2 of about 0.33. As a result, it is best to go ahead and try other analysis methods on the data set to seek more conclusive answers.

## Naïve Bayes

The next analysis used to model the data set was naïve Bayes. Before the model was created, the data set had to first be converted into an ARFF file so that Weka could read it properly. Binary variables such as hypertension, heart\_disease, and diabetes were converted from numeric to nominal. Next, the model was created multiple times, toggling parameters like kernel estimator and supervised discretization to see what returned the best model. After doing so, it was determined that the default parameters worked best. Looking at the 3-fold cross validation, we were able to get a very good classifier model with 94% accuracy. Comparing the most significant mean differences, the mean age of individuals with diabetes was about 60.9 while individuals who didn’t have diabetes was around 40.1, an almost 21-year difference. As for the HbA1c, patients who tested negative for diabetes had a level of 5.4 out of 10, a safe level. Patients who tested positive for diabetes had a mean of 6.9, well into the danger zone. The other significant mean difference was blood glucose levels where the classes 0 (no diabetes) and 1 (diabetes) had means of 133.3 and 195.7 respectively. Considering the slight difference of means and totals of the other attributes, it is clear these findings concur with that of the other analysis where age, blood glucose level, and HbA1c level are the apparent main causes behind diabetes.

A screenshot of a computer

Description automatically generated                      A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Figure 5: Naïve Bayes Classifier Accompanied with 3-Fold Cross Validation**

PhD. Pragya Paudyal also explored using naïve Bayes on a diabetes data set to predict whether patients have diabetes and reported the findings in the article “Classification of Diabetes using Naive Bayes in Python.” The data set Dr. Paudyal used consisted of 9 columns, “Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age and Outcome”. Using python, Dr. Paudyal was able to produce a classifying model with an accuracy of 78%. Compared to this model, our naïve Bayes came out much better. This is most likley a result of our data set being much bigger as it has 100,000 observations compared to Paudyal’s data set only having 768 observations.

## SVM

SVM is a supervised machine learning algorithm which strives to find the optimal hyperplane to classify the data in a multi-dimensional (2D/3D/etc.) space, making this extremely well suited for datasets with a vast number of data points. The optimal hyperplane maximizes the distance between the hyperplane and the datapoints, increasing the classification accuracy. The kernel of the model can change to transform boundaries to different dimensional spaces to help identify nonlinear boundaries. Due to the complexity of the algorithm, a smaller subset of data of 3500 training records to 1500 testing records. To determine which SVM model to use, the approach was to first see which kernel is better suited for this data set using a normalization filter, cost of 1, and 3-fold cross validation. The kernels used were linear, binomial, trinomial and radial basis function (RBF) and were analyzed for accuracy and time to create the model. The results were as follows (Time – Accuracy):

* Linear: 11.09 seconds – 96.9%
* Binomial: 58.43 seconds – 95.4%
* Trinomial: 106.69 seconds – 94.3%
* RBF: 110.47 seconds – 91.1%

Based on these results, a linear kernel was selected to move on to the next tuning step, cost adjustment. The cost was adjusted to .75, .5, and .25. All these values returned the same accuracy of 97.1% but decreased the run time of the model to a range of 9.78 – 13.66 seconds respectively. .75 was selected for the increased efficiency in model creation. The final tuning step was to see if normalization, standardization, or neither best impacted model performance. When standardizing the data set, the accuracy decreased to 92.6% and the time to create increased to 28.13 seconds. No normalization also resulted in a decrease of accuracy to 96.9% and an increase in implementation time to 11.02 seconds. This ended with the selection of the normalized linear model with a cost of .75 as the best fit model for the data set. The classifier errors can be seen in *figure 6* below.

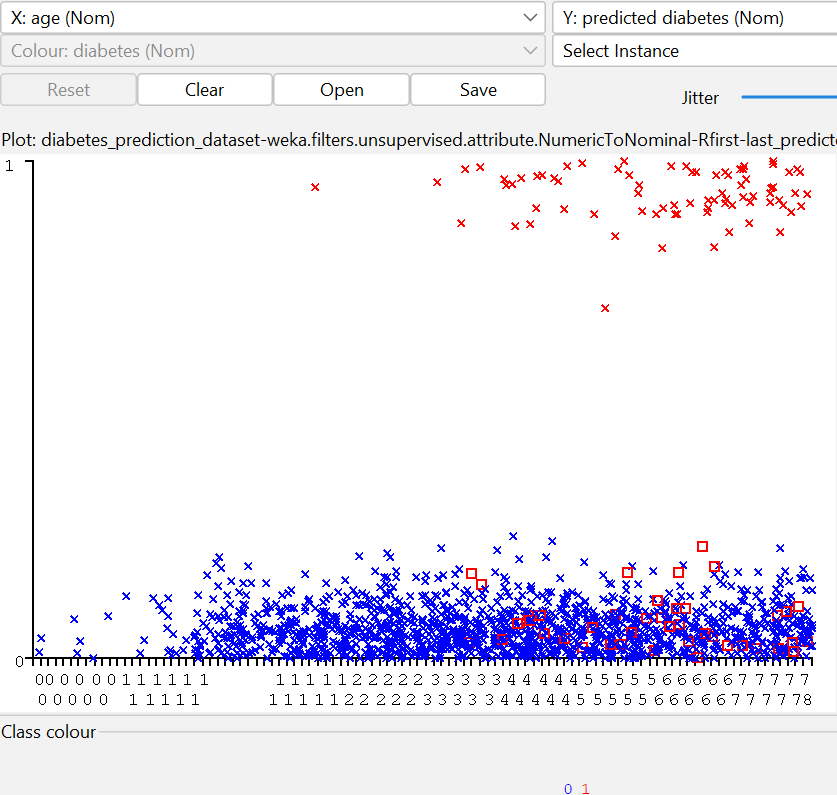


Figure 6 - SVM Classifier Errors

SVM application in the medical field, especially in diabetes prediction has been performed many times. In a recent study conducted in 2022 by Srikar Sistla of Viswa Mahavidyalaya University, created SVM models based on the following attributes:

* Pregnancies
* Glucose, blood pressure
* Skin Thickness
* Insulin
* BMI
* Diabetes Pedigree Function
* Age
* Outcome

The models were able to give promising results of 77.5% and 80.5% accuracy (Sistla) for their training and test sets respectively. This is significantly less than our results, but this could be due to data set size and/or our data set covers more indicative attributes such as HbA1c / blood glucose levels and smoking history.

## RFDT

RFDT is a collection of decision trees that are independently on random subsets of the data and features (variables). Each tree then provides a prediction on the classification of the data until they all come to a consensus (majority rules). Each tree provides input of variable importance to create the final model for classification. Tuning consists of modifying the number of trees in the model. It is important to note that the more trees used runs the risk of overfitting. Since this algorithm demands similar computing power as the SVM, the same subset of data was used for testing RFDT. The RFDT model was tested against 5, 10, 25, 50, 100, 250 trees and the results were as follows (time – accuracy):

* 5: .12 seconds – 93.7%
* 10: .13 seconds – 93.5%
* 25: .34 seconds – 92.8%
* 50: .6 seconds – 93.6%
* 100: 1.34 seconds – 92.9%
* 250: 2.7 seconds – 92.7%

Based on these results, the RFDT model with 5 trees was selected as the best performing. The model has the highest accuracy, the least creation time, and the least likely to have overfitting. The classifier error results can be seen in *figure 7 below.*

A screenshot of a computer

Description automatically generated

Figure 7 - RFDT Classifier Errors

Random Forest is yet another algorithm that has been heavily investigated in the medical field to help doctors detect many different ailments in their patients, more specifically, RFDT has shown promise in diabetes detection as well. A recent study in 2021 was conducted by Umair Butt et al and published with the Journal of Healthcare Engineering diving into multiple different algorithms for diabetes detection, one of them being RFDT. The attributes covered in this study were:

* Pregnancies
* Glucose
* Blood Pressure
* Skin Thickness
* Insulin
* BMI
* Diabetes Pedigree
* Function
* Age
* Outcome

RFDT performed relatively well for their dataset boasting a 77.4% accuracy, 75.7% recall, and 76.9% precision. These results are significantly less than ours but could be explained by data set size, differing attributes, and different tuning methods. Ultimately, both results are extremely promising and further exploration of utilizing RFDT for diabetes detection would be advised.

# Conclusion

To conclude, the global diabetes crisis highlights the urgent need for early detection and intervention to mitigate its severe health and economic impacts. This paper has explored how data mining techniques, including Linear Regression, Exploratory Analysis, Association Rule Mining, Naïve Bayes, Support Vector Machine (SVM), Random Forest Decision Tree (RFDT), and Clustering algorithms, can be applied. Using these analytical methods, we were able to pinpoint high HbA1c and blood glucose levels as having an integral part in obtaining diabetes. These can be caused by poor dietary practices or deficiencies. The methods carried out also pinpointed age as a vital contributor. The results provided hints that people of older age are the ones most affected by the illness. By using these methods, healthcare professionals and researchers can develop tools to identify individuals at risk of diabetes, enabling timely interventions and personalized care plans. These data-driven approaches offer the potential to significantly improve the quality of life for those susceptible to diabetes and to reduce the strain on healthcare systems worldwide. The fusion of healthcare expertise with data mining capabilities presents a promising avenue for combating the diabetes epidemic by enhancing early detection and intervention strategies, ultimately leading to healthier lives and more sustainable healthcare systems.

# References

1. Butt, U. M., Letchmunan, S., Ali, M., Hassan, F. H., Baqir, A., & Sherazi, H. H. (2021). Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications. *Journal of Healthcare Engineering*, *2021*, 1–17. https://doi.org/10.1155/2021/9930985
2. B. M. Patil, R. C. Joshi and D. Toshniwal, (2010). Association Rule for Classification of Type-2 Diabetic Patients. S*econd International Conference on Machine Learning and Computing*, 2010, 330-334, doi: 10.1109/ICMLC.2010.67.
3. Paudyal, P. (2020, February 4). *Classification of diabetes using naive bayes in python*. Medium. <https://medium.com/@pragya_paudyal/classification-of-diabetes-using-naive-bayes-in-python-44385b279277>
4. [Pinky Garg](https://pubmed.ncbi.nlm.nih.gov/?term=Garg%20P%5BAuthor%5D),1 [Karthikeyan Pethusamy](https://pubmed.ncbi.nlm.nih.gov/?term=Pethusamy%20K%5BAuthor%5D),2 and [Rajiv Ranjan](https://pubmed.ncbi.nlm.nih.gov/?term=Ranjan%20R%5BAuthor%5D)1 Correlation between Estimated Average Glucose Levels Calculated from HbA1c Values and Random Blood Glucose Levels in a Cohort of Subjects <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10264116/>
5. Sistla, S. (2022). Predicting Diabetes u sing SVM Implemented by Machine Learning. *International Journal of Soft Computing and Engineering*, *12*(2), 16–18. https://doi.org/10.35940/ijsce.b3557.0512222
6. Trotta, F. (2023, June 11). *Predicting diabetes with machine learning - part I*. Medium. <https://towardsdatascience.com/predicting-diabetes-with-machine-learning-part-i-f151cb764aee>