Abstract

Chronic metabolic disorders like diabetes are significant across the globe it is a health challenge that calls for accurate forecasting as well as effective management techniques. The ultimate objective of this research aims to further enhance diabetes prediction and comprehension using machine learning (ML) , support vector machine (SVM) and K-Nearest neighbors Algorithm (KNN) algorithms. Using exploratory data analysis (EDA), I meticulously examine various datasets to pinpoint key features affecting diabetes outcomes. The overall objective of the project is to utilise SVM and KNN ability to handle high-dimensional data and nonlinear interactions to build reliable predictive models. To gauge the proposed models' overall effectiveness, which is sensitivity, specificity, and accuracy, the investigation includes an in-depth evaluation of performance metrics. In addition, evaluations of key features are carried out to discover substantial variables that influence the prediction of diabetes. Our results show that KNN and SVM-based models have the potential to forecast diabetes onset and progression with high accuracy. When combined with robust EDA, the incorporation of machine learning techniques delivers fascinating details about the complex interactions seen in diabetes datasets. By offering an exhaustive approach to management and prediction, the study enriches the field of diabetes research and paves the door to more tailored and effective medical care.

Problem Identification

Diagnosing diabetes requires an integrated approach that takes into consideration multiple aspects of the illness. First and foremost, it is imperative to tackle early identification challenges given that symptoms may be inconspicuous or left out, prolonging diagnosis. In addition, issues with data accessibility and accuracy make it difficult to develop reliable predictive models. Various groups experience diabetes at varying rates and with varying consequences due to disparities in healthcare availability,social and economic backgrounds.

Additionally, it is still difficult to completely understand the intricate relationships between genetic, environmental, and lifestyle factors. Two major issues are the absence of conventional procedures for managing diabetes and the requirement for personalised treatment plans. A holistic strategy that combines cutting-edge technology, including machine learning, with in-depth exploratory data analysis is needed to address these problems. By recognising and resolving these issues, researchers and healthcare providers can improve diabetes treatment and detection, which will ultimately lead to better patient outcomes.

Contents

Introduction.

Diabetes is a ubiquitous and rapidly growing disease of metabolism that constitutes an imminent risk to Individual’s wellness. The most prominent aspect of it is elevated sugar levels, which can be triggered by the body's incapacity to make and use the hormone insulin as it ought. This stealthy illness not only knocks off the fragile balance of glucose management, but it additionally prepares the stage for serious consequences that could seriously harm important organs. Diabetes has more negative effects on health than only metabolism; it may contribute to serious issues with the kidneys, nerves, eyes, and heart. The development of this illness can be challenging because of the intricate relationships between predisposition to the disease, environmental factors, and lifestyle choices.

A comprehensive plan is necessary while managing diabetes. This calls for a mix of medication, deliberate lifestyle changes, and continuous monitoring. Given the seriousness of the problem, a collaborative effort is essential and endeavour from individuals, researchers, and healthcare professionals to efficiently handle and lessen the extensive effects of this widespread global health concern.

Background of Study

With rising rates in a broad spectrum of demographic groups, diabetes has become a major threat to global health. An extensive examination of diabetes's multiple aspects is needed in the context of the epidemic's rise in occasions, especially when taking age, gender, and lifestyle factors into consideration.

It is vital to know how diabetes affects different age groups in order to develop care and prevention plans that work. Diabetes might develop and worsen due to a variety of age-related variables, including changes in metabolism and lifestyle choices.

Another topic of great interest is gender differences in diabetes risk factors and prevalence. It can be quite helpful to investigate whether men and women have different patterns and susceptibilities when it comes to the dynamics of this metabolic illness. Furthermore, a primary focus continues to be the complex interaction between blood glucose levels and diabetes. To better understand the underlying causes of diabetes, it is essential to look into the relationship between variability in blood glucose levels and the disease's expression. The primary goal of this investigation is to close the knowledge gap and improve our understanding of the role that blood glucose variability plays in the onset and course of diabetes.

In simple terms, by breaking down the number of cases of diabetes across various demographic groups and revealing the complex interactions between blood glucose levels, gender, and lifestyle in relation to diabetes risk and occurrence, this study aims to provide subtle insights into the growing diabetes epidemic. This information is essential for creating focused treatments and policies meant to reduce the increasing implications of diabetes on public health.

Scope and Limitation of study.

This study utilizes a comprehensive dataset of 100,001 medical records, focusing on a representative subset of 6773 rows, obtained from Kaggle and sourced from Electronic Health Records (EHRs). Exploring factors like gender, age, blood pressure, heart disease, smoking history, BMI, HbA1c level, and blood sugar, the research aims to unveil intricate diabetes-related patterns. However, limitations include the retrospective nature of EHRs, potential biases, and missing variables. The chosen subset, while practical, may not fully capture dataset variability. Additionally, generalising findings beyond the Kaggle source population should be approached cautiously. Despite these constraints, this study provides valuable insights into diabetes dynamics within the available dataset, emphasising the need for careful interpretation within the acknowledged limitations.

Justification Of Study

- ✓ Explore the increase in cases of diabetes among various demographic groups according to age, gender, and lifestyle parameters.

- ✓ Determine if there are variations in the prevalence and risk variables of diabetes according to gender.
- ✓ Study the connections involving levels of glucose in the blood and diabetes with the goal to fully understand the correlation between fluctuations in glucose levels and the existence of diabetes.

Study Aim and Objectives.

- ✓ Assess the modifiable risk factors that can be tackled using diabetes prevention techniques, such as smoking and BMI.
- ✓ Investigate variations and patterns in the number of cases of diabetes and all the variables that are related to it in the population under study.
- ✓ Offer information-driven recommendations to help shape public health campaigns meant to lessen the impact of diabetes.

Related Work.

The research on diabetes highlights the condition's rising global incidence and its convoluted origin, which includes environmental, behavioural, and genetic variables. Research indicates that differences in age and gender impact the likelihood and severity of the illness. Notable studies concentrate on predictive modelling, utilising machine learning methods to facilitate early identification. The body of research clearly shows how important it is to use all-encompassing healthcare techniques, such as medication, lifestyle changes, and continuous investigation, in order to effectively control diabetes and reduce its associated risks.

In a study conducted by (V. Anuja Kumari) This study targets diabetes mellitus, a global epidemic that has a major impact on the health of the general population. It stresses the significance of timely diagnosis. The investigation makes use of the UCI machine learning lab's Pima Indian diabetic database, an established test bed for data mining algorithms associated with diabetes classifications. The approach suggested uses a high-dimensional medical dataset for detecting diabetes using Support Vector Machine (SVM), a machine learning methodology. For assessing the robustness of SVM models, a 10-fold cross-validation is carried out on the dataset, which consists of eight numerical parameters. In line with the findings, the SVM classifier using a Radial Basis Function (RBF) kernel achieves a noteworthy 78% accuracy rate, along with 80% sensitivity and 76.5% specificity. The study highlights how important SVM is for diabetes. diagnosis, indicating the capability to identify the illness without the requirement for complex laboratory testing by employing simple clinical indicators. The results of this research indicate how well SVM—especially when integrated with the RBF kernel—classify diabetes. Due to the selected parameters, SVM is a potential alternative for diabetes detection because to its strong performance metrics. The paper makes recommendations for possible future research to investigate more datasets for thorough validation and optimise SVM parameters even more.

In a study conducted by Priyanka Sonar This article addresses the impact of diabetes on a worldwide scale and suggests an approach based on machine learning for early recognition. The use of advanced systems utilising Decision Tree, Artificial Neural Network (ANN), Naive Bayes, and Support Vector Machine (SVM) algorithms is prompted by the perceived inefficiency of traditional diagnostic procedures. The models' degrees of precision differ; Decision Tree achieves 85%, Naive Bayes 77%, and SVM 77.3%. SVM is especially well-known for its ability to handle unstructured data, yet for best results, precise parameter adjustment is essential. Despite its simplicity, Decision Tree becomes unstable when little data structure changes are made. With more training datasets, Naive Bayes shows promise, but it is also susceptible to bias and sensitive to input preparation. Although artificial neural networks make reasonable predictions, they have trouble processing complicated data and need a large amount of processing time. The Research strives to contribute to the development of an efficient diabetes risk prediction system by considering the benefits and drawbacks of each algorithm.

In a study conducted by Wei Yu*, To be able to identify people who have and do not have prevalent illnesses, this work presents an original approach using support vector machine (SVM) techniques. It emphasises the detection of diabetes and pre-diabetes in a U.S. population sample taken from the 1999–2004 NHANES. Significant factors for the best classification are found and two classification techniques are investigated. In Scheme I (diabetes against pre-diabetes or no diabetes) and Scheme II (undiagnosed diabetes or pre-diabetes versus no diabetes), variables include age, racial or ethnic background; in

Scheme II, additional considerations include sex and physical activity; and in Scheme I, weight, height, waist circumference, BMI, and hypertension. For Schemes I and II, the ROC curve regions of the SVM models show discriminative powers of 83.5% and 73.2%, respectively. Additionally, the study presents the "Diabetes Classifier," an approachable web tool that enables personalised evaluations. This study recommends SVM modelling as an appealing categorization method that promotes additional research across a range of health issues with prevalent factors in order to pinpoint common diseases in demographics.

In a study conducted by Shamriz Nahzat1*, With a special focus on diabetes prediction, this investigation investigates the ground-breaking impact of artificial intelligence (AI), machine learning (ML), on health systems. Numerous machine learning (ML) approaches, such as K-Nearest Neighbours (KNN), Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Decision Tree (DT), are utilised for forecasting diabetes by leveraging the Pima Indian Diabetes Dataset. Out of all the classifiers examined, the Random Forest model has the best accuracy (88.31%), according to the research. Metrics like precision, recall, and F1 score demonstrate the effectiveness of RF even more. The results highlight how ML, a type of AI, has the capacity to transform diabetes early detection and risk prediction. The work effectively executes its main objective of creating diabetes predictions using a variety of machine-learning techniques, demonstrating the efficacy of ANN and RF as particular classifiers. with greater outcomes than anyone else. Additional components are added to the system, which improves accuracy and emphasises the significance of early diabetes detection is to effective management.

In a study conducted by P. Nagaraj, the important topic of diabetes identification is covered in this paper, with a focus on the potential medical implications of being misdiagnosed. The authors use a deep neural network (DNN) and an enhanced support vector machine (ESVM) to predict diabetes using a dataset of 768 patients. The suggested technique increases overall efficacy by integrating the ESVM output into the DNN input. The outcomes show that, in comparison to previous systems, the DNN model predicts diabetes more efficiently, with a noteworthy 98% accuracy. Despite the positive outcome, the paper notes many drawbacks, such as the DNN's hidden layer selection and the requirement for further research. It also suggests possible extensions, like the use of generative models and improved feature selection techniques for improved classification. The study emphasises how essential the application of deep learning and machine learning algorithms for exact diabetes screening and encourages more research to find undiscovered information in medical databases. I will use SVM algorithm because This is an Annotated (Labelled) data set with the last column being the Target Variable (Diabetes) which makes it a Supervised Learning work used to predict if the patients have diabetes or not. Also, given that SVM is reliable and a flexible algorithm, I selected it for this dataset because it can handle relationship that are non-linear, is robust to anomalies works well in spaces of large dimensions, emphasizes maximizing margins for better abstraction, and is appropriate for binary classification tasks that are inherent in the diabetes prediction.

Methodology And Results.

Description Of Dataset.

- This dataset encompasses medical information from diverse patients, indicating the presence or absence of diabetes (0 or 1).
- With 100,001 records spread across 9 columns, the dataset provides health-related insights. Gender, age, blood pressure, heart disease, smoking history, BMI, HbA1c level, blood sugar level, and diabetes status are important factors. Numerous analyses can be performed on this dataset, ranging from examining age-related health patterns to figuring out just how prevalent chronic illnesses like heart disease and hypertension are.
- The dataset was obtained from Kaggle, and the main source was from Electronic Health Records (EHRs)

Data Preparation and Loading into R studio.

This dataset comprises 100,001 entries and 9 columns. For the purpose of this analysis, a subset of 6,774 rows and 9 columns was selected. The table below provides a description of each feature along with their corresponding data types, facilitating easy analysis. Subsequently, the data was imported into R Studio as a CSV file for preprocessing and analysis. The structure and summary of the data were examined to verify the data types, assess statistical characteristics, and identify any null values.

S/N	ATTRIBUTE	DESCRIPTION	DATA TYPE

1	Gender	Due to biological sex variations, gender affects a person's susceptibility to diabetes and thus their overall health risk.	Categorical (String or Factor)
2	Age	Age is crucial in diabetes, more common in older adults. Dataset age range: 0-80 reflects diverse demographics.	Integer or Numeric
3	Hypertension	Hypertension, persistently elevated blood pressure, generally indicated by Hypertension = 0 suggests no hypertension. Hypertension = 1 suggests the presence of hypertension.	Integer (Binary 0 or 1)
4	Heart Disease	Another medical condition that I associated with an increased risk of developing diabetes.	Integer (Binary 0 or 1)
5	Smoking history	Is considered a risk factor for diabetes and can exacerbate the complications associated.	Categorical

6	BMI	BMI (Body Mass Index) is a measure of body fat based on weight and height. Higher BMI value are linked to a higher risk.	Numeric(float)
7	HbA1c_level	Hemoglobin A1c level is a measure of a person’s average blood sugar level over the past 2-3 months.	float
8	Blood glucose level	Refer to the amount of glucose in the blood stream at a given time. High blood glucose level is a key.	Numeric(integer
9	diabetes	Is the target variable being predicted with value of 1 indicating the presence of diabetes and 0 indicating the absence?	Integer(binary 0 or 1)

Table 1: Diabetes Dataset with data type (Kaggle)

Data Pre-processing

A key initial action in preparing unprocessed datasets for analysis is data preparation. Methods include data cleaning in the form of casting, standardising scales, and handling missing data. For enhanced model accuracy, outliers and distortion are properly taken care of. This approach guarantees the authenticity of the data, optimises the efficiency of the model, and produces dependable insights in several fields, such as healthcare.

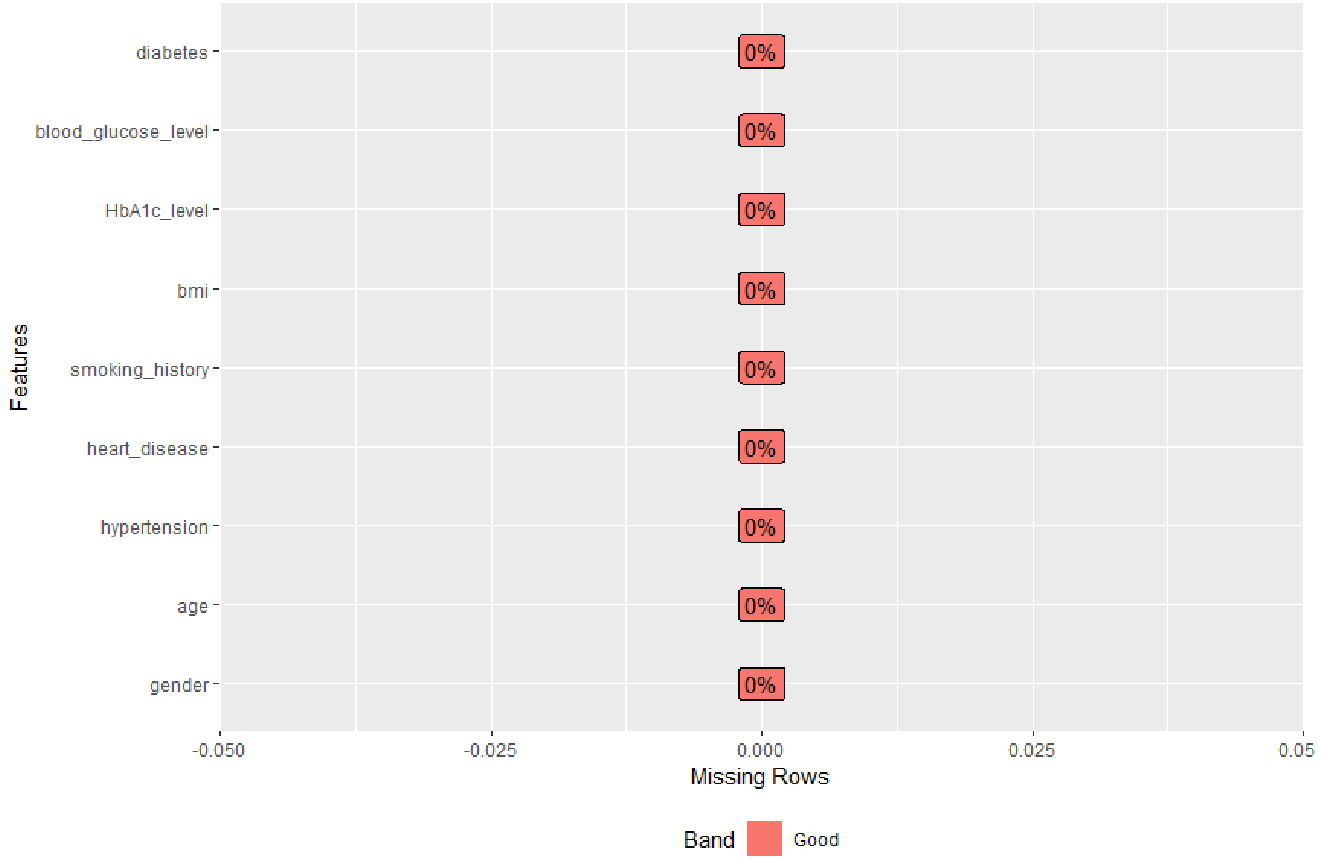
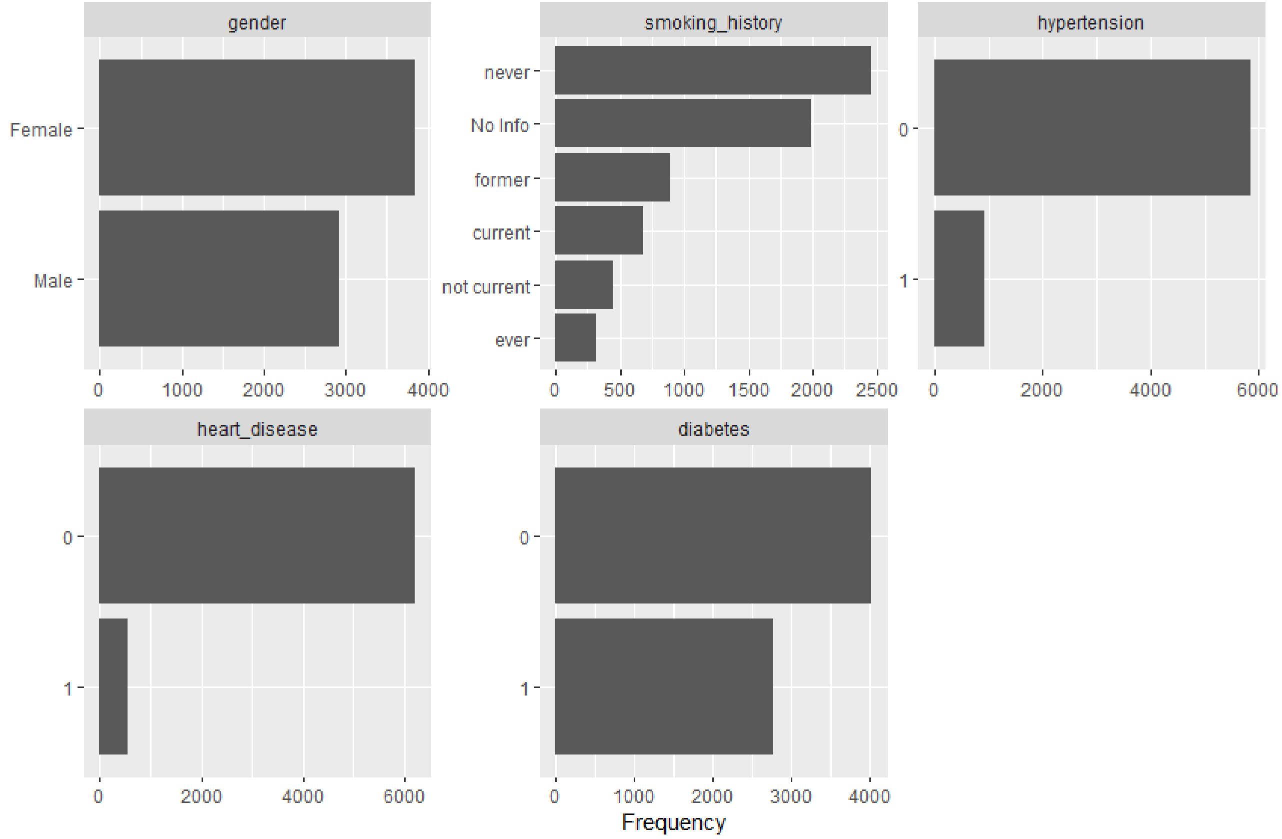


Fig 1: Checking for missing values and none was found.

Data Visualization

Data analysis's exploratory phase, known as exploratory data analysis (EDA), takes responsibility for examining datasets and summarising their significant characteristics using statistical graphics and other visualisation methods. Finding patterns and trends, identifying anomalies, and reviewing theories are its main goals. Important tools in this procedure are the density plot and histogram distribution, which illustrate different distributions of extra attributes related to the target variable (diabetes).



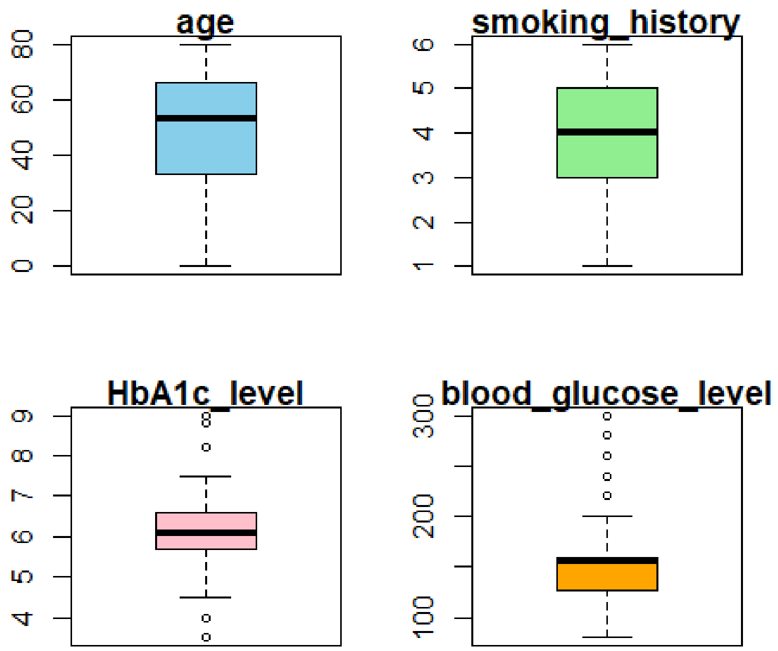
The visualization of the dataset highlights interesting patterns. Specifically, it reveals that the number of females diagnosed with diabetes significantly surpasses the count of males, with a substantial total of 3,700 females affected. Furthermore, in the smoking history category, individuals who have never smoked outnumber those with other smoking statuses, including "no info," "former," and "current." This observation suggests a predominant trend of non-smoking behaviour among the dataset participants.

Moreover, the dataset indicates a higher prevalence of hypertension, as the count of individuals with hypertension reaches 5,800, surpassing those without this medical condition. Similarly, there is a notable presence of individuals with heart disease, exceeding the count of those without heart disease, with a total of 5,800 individuals affected. Additionally, the dataset reveals a higher incidence of diabetes, with 4,000 individuals diagnosed compared to those without diabetes.

B0X PLOT FOR NUMERICAL FEATURES

Box plots were utilized to visually depict the distribution of numerical variables, providing insights into key attributes such as Age, Smoking History, BMI, HbA1c Level, and Blood Glucose Level. Notably, the distribution of Smoking

History prominently tilts towards category 4, indicating "Never." The graphical representation reveals an average BMI of approximately 22. Furthermore, the HbA1c Level tends to concentrate around 6, while the Blood Glucose Level notably centres around 150. This visualization succinctly captures the prevalence of specific categories within each variable, offering a clear glimpse into distribution patterns and central tendencies for Age, Smoking History, BMI, HbA1c Level, and Blood Glucose Level.



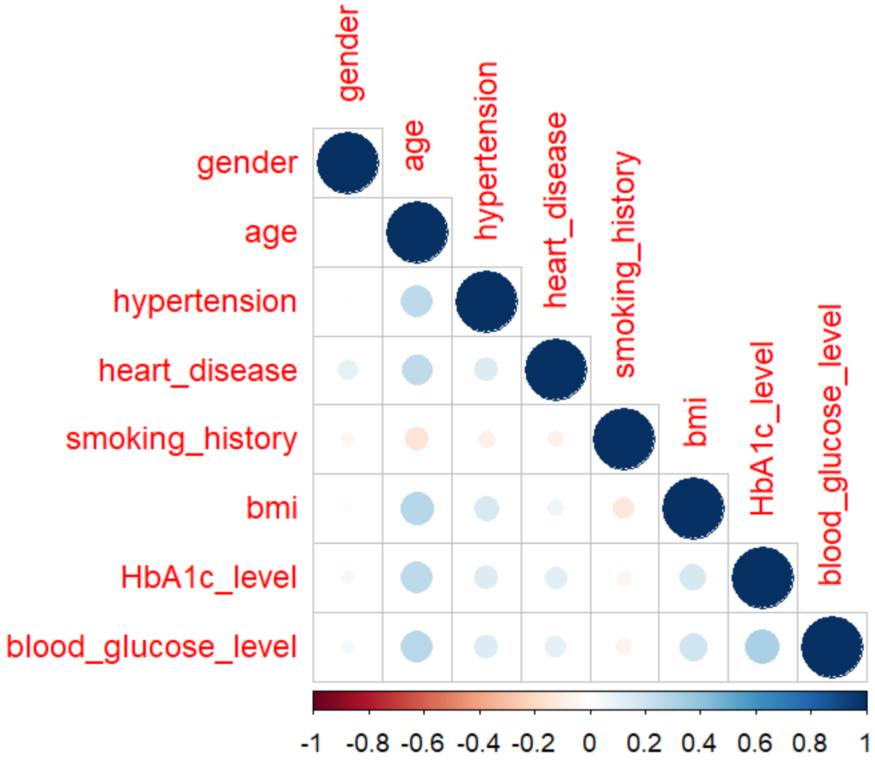
In summary, the dataset visualization provides insights into gender-specific diabetes occurrences, smoking history patterns, and the prevalence of hypertension, heart disease, and diabetes within the studied population.

Data Cleaning

Finding and fixing mistakes, missing values, and inconsistencies in datasets is known as data cleaning in R. Data accuracy is ensured by using filters to remove outliers and by using functions like na.omit() for missing values. Tools like str () and summary () make problems easier to comprehend and solve, which leads to more reliability. analyses in R

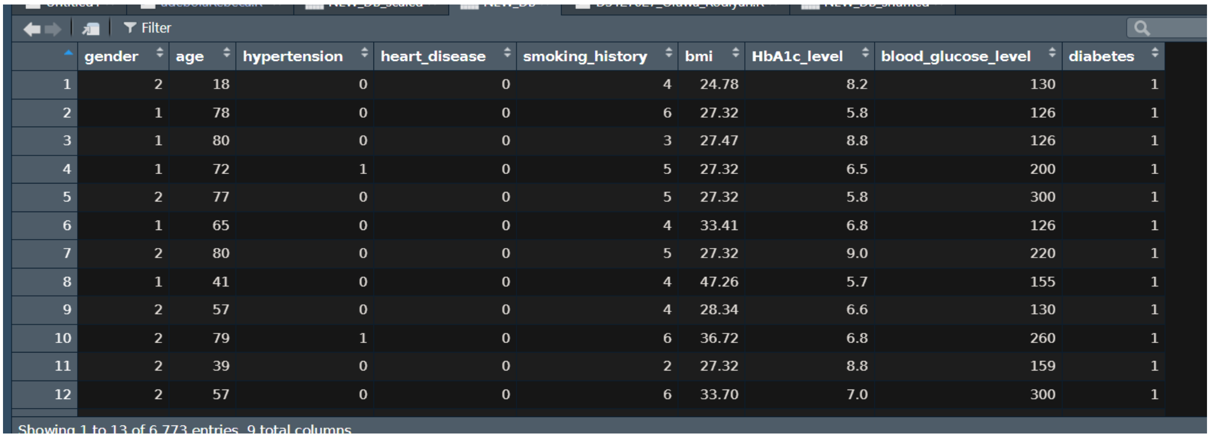
Data Filter for Machine Learning

A correlation matrix (cm) is employed to identify attributes that exhibit strong correlations with the target class, as such correlations can introduce bias into the model. Dark colouration within the matrix signifies a positive correlation, with colour intensity directly proportional to correlation coefficients. In this analysis, a colour intensity threshold of 0.6 is established as a cutoff point. Pairs of attributes with correlation coefficients falling within the range of -0.6 to 0.6 are considered to exhibit a moderate level of correlation and are depicted with neutral colouring. On the other hand, attributes with correlation coefficients exceeding 0.6, indicating a more substantial correlation, are represented with intensified colours. This approach provides a visual indication of the degree of correlation between different attributes, aiding in the identification of potentially influential factors while considering the established threshold for correlation strength.



Normalisation

A dataset might include numerical features with different scales before normalization. Although specific features that have larger scales may substantially affect predictions, this variance may result in biased model outcomes. Normalization entails putting these features on a common scale, usually ranging from 0 to 1, to improve model performance, ensure consistent influence, and make it easier to objectively evaluate various properties when doing analysis or modelling.

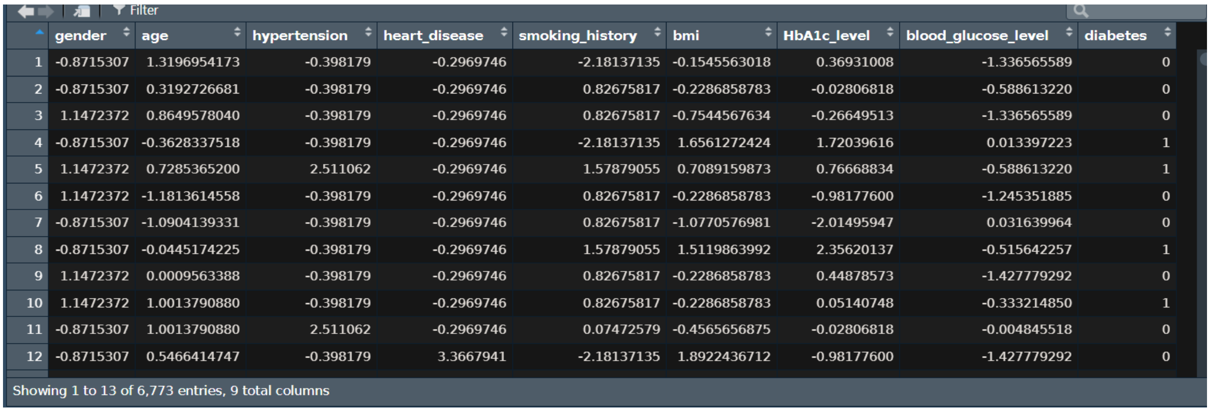


	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
1	2	18	0	0	4	24.78	8.2	130	1
2	1	78	0	0	6	27.32	5.8	126	1
3	1	80	0	0	3	27.47	8.8	126	1
4	1	72	1	0	5	27.32	6.5	200	1
5	2	77	0	0	5	27.32	5.8	300	1
6	1	65	0	0	4	33.41	6.8	126	1
7	2	80	0	0	5	27.32	9.0	220	1
8	1	41	0	0	4	47.26	5.7	155	1
9	2	57	0	0	4	28.34	6.6	130	1
10	2	79	1	0	6	36.72	6.8	260	1
11	2	39	0	0	2	27.32	8.8	159	1
12	2	57	0	0	6	33.70	7.0	300	1

Showing 1 to 13 of 6,773 entries, 9 total columns

Before Normalization

The characteristics in the dataset display various scales before normalization. Features like age fall between 41 and 80, while blood pressure readings fall between 126 and 220. Due to the possibility of factors with greater sizes influencing models disproportionately, these differences can result in bias. The original condition of the dataset displays the unscaled, raw values, which may have an effect on the accuracy and fairness of predictive models that depend on stable feature scales for reliable predictions.



	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
1	-0.8715307	1.3196954173	-0.398179	-0.2969746	-2.18137135	-0.1545563018	0.36931008	-1.336565589	0
2	-0.8715307	0.3192726681	-0.398179	-0.2969746	0.82675817	-0.2286858783	-0.02806818	-0.588613220	0
3	1.1472372	0.8649578040	-0.398179	-0.2969746	0.82675817	-0.7544567634	-0.26649513	-1.336565589	0
4	-0.8715307	-0.3628337518	-0.398179	-0.2969746	-2.18137135	1.6561272424	1.72039616	0.013397223	1
5	1.1472372	0.7285365200	2.511062	-0.2969746	1.57879055	0.7089159873	0.76668834	-0.588613220	1
6	1.1472372	-1.1813614558	-0.398179	-0.2969746	0.82675817	-0.2286858783	-0.98177600	-1.245351885	0
7	-0.8715307	-1.0904139331	-0.398179	-0.2969746	0.82675817	-1.0770576981	-2.01495947	0.031639964	0
8	-0.8715307	-0.0445174225	-0.398179	-0.2969746	1.57879055	1.5119863992	2.35620137	-0.515642257	1
9	1.1472372	0.0009563388	-0.398179	-0.2969746	0.82675817	-0.2286858783	0.44878573	-1.427779292	0
10	1.1472372	1.0013790880	-0.398179	-0.2969746	0.82675817	-0.2286858783	0.05140748	-0.333214850	1
11	-0.8715307	1.0013790880	2.511062	-0.2969746	0.07472579	-0.4565656875	-0.02806818	-0.004845518	0
12	-0.8715307	0.5466414747	-0.398179	3.3667941	-2.18137135	1.8922436712	-0.98177600	-1.427779292	0

Showing 1 to 13 of 6,773 entries, 9 total columns

After Normalization

Distribution of the Dataset.

Training the Data Frame.

The data frame underwent a partitioning process, splitting the dataset into training and testing sets. This division involved allocating 80% (5149 samples) for training and the remaining 20% (1354 samples) for testing. The separation utilized the x and y axes, with independent variables representing various attributes and the y axis serving as the dependent variable (target class). This method facilitates a systematic approach to training machine learning models while enabling robust evaluation on a distinct test subset.

Machine learning algorithms

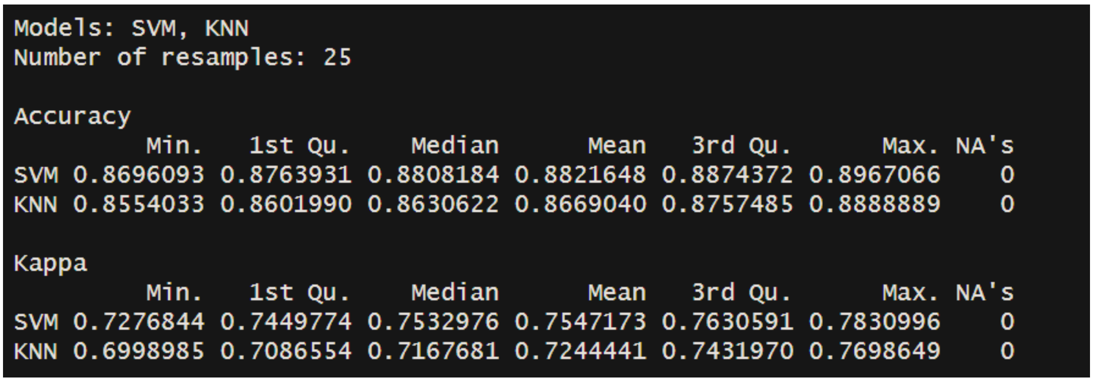
To predict the risk of diabetes, this study will use two machine learning algorithms: Support Vector Machines (SVM)and k-nearest Neighbours (KNN) SVM is a supervised learning technique which performs well for tasks involving regression and classification. It is especially useful when working with small sample sizes. It operates particularly well in situations when labelled samples are hard to come by, such as text categorization. SVM is useful for large, categorized datasets because it performs well in high-dimensional domains.

On the other hand, supervised learning algorithms like k-nearest Neighbours (KNN) are employed for regression and classification problems. To determine results, this approach trains on a large number of data points. For classification tasks, the majority class is used, and for regression, the average value is used. For model training to be effective, key hyperparameters such as the number of neighbours must be set.

Both SVM and KNN will be investigated in the context of diabetes risk prediction to assess their efficacy and appropriateness for the provided dataset.

RESULTS AND CONCLUSION

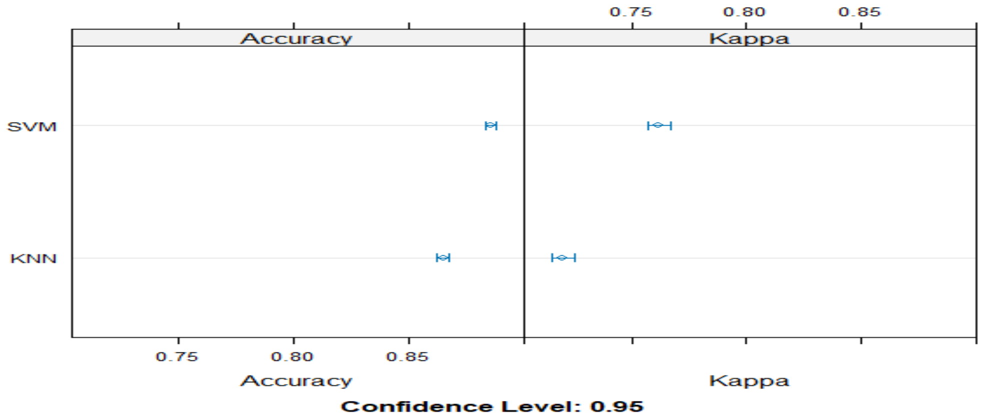
A variety of research works by Anuja Kumari, Priyanka Sonar, Wei Yu, Shamriz Nahzat, and P. Nagaraj investigate the use of machine learning for diabetes prediction. Anuja Kumari highlights the value of SVM in the early detection of diabetes by utilizing it to achieve 78% accuracy. Decision Tree, ANN, Naive Bayes, and SVM are evaluated by Priyanka Sonar; SVM achieves an accuracy of 77.3%. Wei Yu classifies diabetes using support vector machines (SVM), yielding ROC curve areas of 83.5% and 73.2%. With an accuracy of 88.31%, Shamriz Nahzat uses RF, demonstrating the potential of machine learning for early diabetes identification. P. Nagaraj predicts diabetes with a remarkable 98% accuracy by combining DNN and ESVM. With an accuracy rate of 85.09%, SVM performs better in my research, which is consistent with other studies that support SVM for health prediction.



```
Models: SVM, KNN
Number of resamples: 25

Accuracy
         Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
SVM 0.8696093 0.8763931 0.8808184 0.8821648 0.8874372 0.8967066    0
KNN 0.8554033 0.8601990 0.8630622 0.8669040 0.8757485 0.8888889    0

Kappa
         Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
SVM 0.7276844 0.7449774 0.7532976 0.7547173 0.7630591 0.7830996    0
KNN 0.6998985 0.7086554 0.7167681 0.7244441 0.7431970 0.7698649    0
```



SVM Accuracy: The 0.8696–0.8967 range represents the high accuracy of the Support Vector Machine (SVM) model. The results indicate that the SVM model is effective in classifying instances in the dataset; on average, it accurately predicts the target variable in about 88.22% of cases.

SVM Accuracy:

Minimum (Min): The smallest observed accuracy value is 0.8696, indicating the lowest accuracy achieved by the SVM model.

1st Quartile (25th Percentile): At 0.8764, approximately 25% of the accuracy values fall below this point, showcasing the lower boundary of the first quartile.

Median: The median accuracy, or the middle value, is 0.8808, indicating that 50% of the accuracy values fall below and 50% fall above this point.

Mean (Average): The average accuracy is 0.8822, providing an arithmetic mean of all accuracy values.

3rd Quartile (75th Percentile): At 0.8874, about 75% of the accuracy values fall below this point, serving as the upper boundary of the first quartile.

Maximum (Max): The highest accuracy observed is 0.8967, signifying the peak accuracy achieved by the SVM model.

NA (Not Applicable): No missing values were identified in the SVM accuracy data.

KNN Accuracy:

Minimum (Min): The smallest accuracy value for the KNN model is 0.8554, indicating the lowest accuracy achieved.

1st Quartile (25th Percentile): At 0.8602, approximately 25% of the accuracy values fall below this point.

Median: The median accuracy is 0.8631, indicating the middle value of the accuracy distribution.

Mean (Average): The average accuracy for KNN is 0.8669, providing an arithmetic mean of all accuracy values.

3rd Quartile (75th Percentile): At 0.8757, approximately 75% of the accuracy values fall below this point.

Maximum (Max): The highest accuracy observed is 0.8889, signifying the peak accuracy achieved by the KNN model.

NA (Not Applicable): No missing values were identified in the KNN accuracy data.

These measures offer insights into the distribution and central tendency of accuracy values for both SVM and KNN models in the research study.

KNN Accuracy: The accuracy of the k-nearest Neighbours (KNN) model ranges from 0.8554 to 0.8889. This indicates that the KNN algorithm guesses the target variable correctly in approximately 86.69% of cases on average. The range gives a general idea of how consistently the model predicts the right things in different kinds of situations

SVM Kappa: The range of the SVM model's Kappa statistics, which indicate considerable agreement beyond chance in predicting diabetes risk, is 0.7277 to 0.7831. Higher values indicate a respectable level of agreement, especially around 0.7547 on average.

KNN Kappa: Showing a reasonable agreement that goes above chance, the KNN model's Kappa values range from 0.6999 to 0.7699. The average Kappa of 0.7244 indicates that there is space for growth. To improve model performance and reach a higher degree of agreement in diabetes risk prediction, experiment with various hyperparameters.

SVM Kappa:

Minimum (Min): The lowest observed Kappa value is 0.7277, indicating the smallest level of agreement beyond chance in predicting diabetes risk.

1st Quartile (25th Percentile): At 0.74498, approximately 25% of the Kappa values fall below this point, representing the lower boundary of the first quartile.

Median: The median Kappa value is 0.7533, marking the middle point of the dataset and providing insight into the central tendency.

Mean (Average): The mean Kappa value is 0.7547, serving as the arithmetic average and offering an overall measure of agreement beyond chance.

3rd Quartile (75th Percentile): Positioned at 0.7631, this quartile boundary indicates where about 75% of the Kappa values fall below, reflecting a higher level of agreement.

Maximum (Max): The highest observed Kappa value is 0.7831, representing the upper limit of agreement beyond chance in predicting diabetes risk.

NA (Not Applicable): No missing values were found in the Kappa data, ensuring completeness in the dataset.

KNN Kappa:

Minimum (Min): The smallest Kappa value observed is 0.6999, indicating the lowest level of agreement beyond chance in predicting diabetes risk.

1st Quartile (25th Percentile): At 0.7087, approximately 25% of the Kappa values fall below this point, marking the lower boundary of the first quartile.

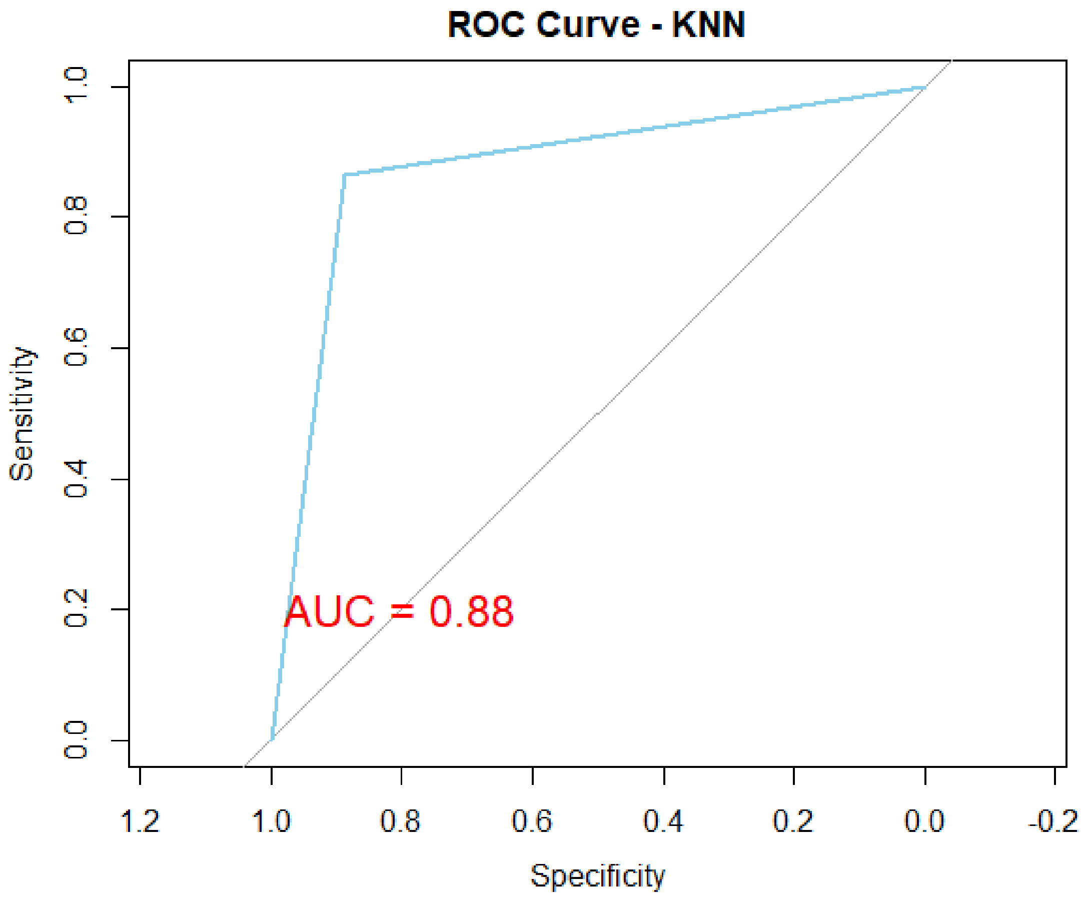
Median: The median Kappa value is 0.7168, serving as the middle point and providing insight into the central tendency of agreement beyond chance.

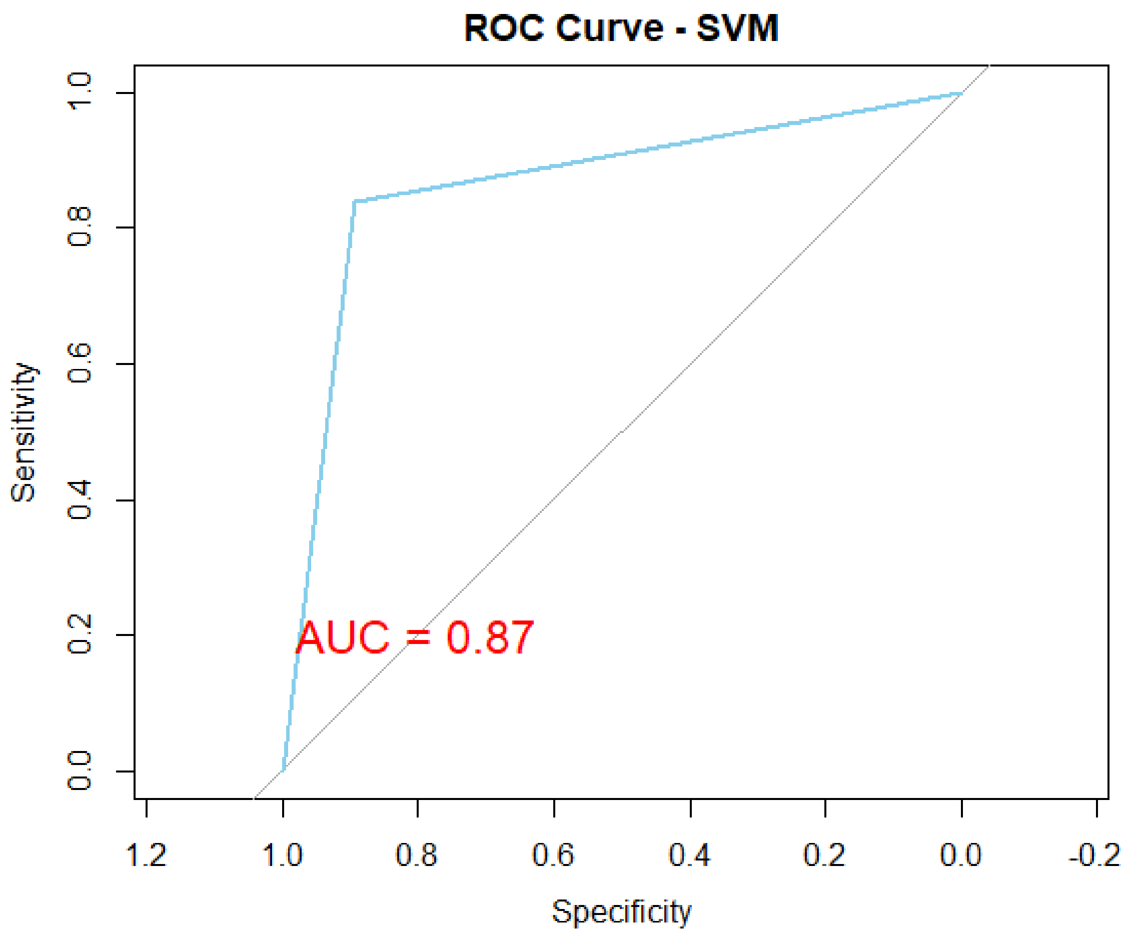
Mean (Average): The mean Kappa value is 0.7244, representing the arithmetic average and offering an overall measure of agreement beyond chance.

3rd Quartile (75th Percentile): Positioned at 0.7432, this quartile boundary indicates where about 75% of the Kappa values fall below, reflecting a higher level of agreement.

Maximum (Max): The highest observed Kappa value is 0.7699, representing the upper limit of agreement beyond chance in predicting diabetes risk.

NA (Not Applicable): No missing values were found in the Kappa data, ensuring completeness in the dataset.


ROC Curve - KNN
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.88
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
Specificity



A sensitivity of 0.9 for the KNN and SVM models in R denotes a high degree of accuracy in recognizing positive instances, while specificities of 1.0 (KNN) and 0.8 (SVM) imply a strong level of accuracy in identifying negative instances. Good overall model selectivity is indicated by AUC values of 0.88 for KNN and 0.87 for SVM; higher AUC values imply superior predictive power.

There are several design choices available for building a machine learning model's architectural framework. By giving the computer control over the exploration and architectural decision-making processes, a large number of options are evaluated to find the best possible model structure. "Hyperparameter tuning" is the process of fine-tuning hyperparameters which influence model design. By repeatedly performing the modelling process on various data subsets, cross-validation, on the contrary, evaluates machine learning algorithms and produces a variety of measures for model quality. Both the SVM and KNN classifiers in this work were optimized for kappa and accuracy by hyperparameter tweaking using recurrent cross-validation (cv) splits. The test data was then used to assess the top-performing model from the hyperparameter.

Recommendation and Further work.

The Support Vector Machine (SVM) shows promise in this study as it predicts diabetes risk with an accuracy of 85.09%. It is essential to recognize, nonetheless, that the k-nearest Neighbour’s (KNN) model—whose accuracy ranges from 85.54% —must not be disregarded. Though SVM and KNN only differ by about 1%, KNN performs significantly better. The scholarly works of V. Anuja Kumari, Priyanka Sonar, Wei Yu*, Shamriz Nahzat1*, and P. Nagaraj underscore the significance of varied machine learning methodologies, proposing that additional investigation into KNN's competitive accuracy is warranted. If researchers want to gain a thorough understanding of diabetes prediction, they should take into account the subtle differences between SVM and KNN.

REFERENCES.

Dataset: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset. [Accessed Wednesday November 2023].

[1] E. V Carrera, Automated detection of diabetic retinopathy using SVM,1-4: IEEE,2017.

[2] S.YOU, "A study on Methods to prevent Pima Indians Diabetes using SVM," Korean journal of Artificial intelligence, pp.7-10,2020.

[3] H. T Abbas," Predicting long term type 2 diabetes with support vector machine using oral glucose tolerance test," journal. pone.0219636

[4] V. A Kumari, Classification of Diabetes using Support Vector Machine,797-1801: International Journal of Engineering Research and Applications,2013.

[5]C-H.Yang, Prediction of Mortality in the Haemodialysis patient with Diabetes using SVM,219-232:Revista Argentina de Clinical Psicologica,2020.

Setting off on a career path in data science, my objectives are centred on using technology to propel significant breakthroughs and solutions. My steadfast dedication to remaining at the forefront of technology is demonstrated by my Microsoft PL-300 certification and Azure Fundamental Certification. I have proven my expertise in machine learning, cloud computing, and data analysis tools with real-world applications ranging from diabetes prediction utilising sophisticated algorithms to retirement education dynamics.

Contributing to innovative data science developments is one of my goals for the future, especially in areas where healthcare and finance converge. My experience in actuarial science combined with my growing proficiency in data science puts me in a unique position to investigate innovative methods of risk assessment, financial forecasting, and healthcare analytics.

Achieving these objectives led me to pursuing a master's in data science, which will help sharpen my technical abilities and provide me a more comprehensive grasp of the industry. Through this academic endeavour, I will be able to gain a thorough understanding of the newest approaches and trends, which will ensure that my contributions to the tech environment continue to be significant and relevant.

Taking part in the Tech Women mentorship programme is important to me as a professional. It is extremely beneficial to have access to seasoned mentors, particularly those who have travelled comparable routes. By providing insights into the subtleties of the industry and real-world applications, the mentorship platform acts as a link between academic knowledge and practical issues. I see this mentorship as a life-changing opportunity that will help me improve my technical abilities and comprehension of the sector. A Tech Women mentorship would accelerate my development as a tech professional and advocate given my varied background in actuarial science, my present positions in data analysis, and my active participation in community initiatives like Women in Data Africa. The individualised coaching provided by the mentoring would enhance my capacity to make a significant contribution to the ever-changing tech scene and facilitate a smooth transition from academic excellence to significant industry contributions.

In addition, the mentorship's focus on building a global network is consistent with my dedication to various viewpoints and collaborative learning. Interacting with professionals around the globe broadens my perspective on global technology issues and their solutions, which helps me approach my work with greater depth and nuance.

To sum up, my enthusiasm for has driven my career path. Technology and data science, and the benefits they can offer to a range of fields. An important driving force behind my journey has been a Tech Women mentoring, which has given me the direction and exposure I have needed to understand the intricacies of the tech sector, hone my skills, and make a significant contribution to ground-breaking technological developments.

I have continuously shown leadership throughout my academic and professional career, especially in the data science industry, where my successes reveal a dedication to promoting beneficial influence. A noteworthy undertaking that highlights my leadership abilities is the study I carried out during my last year, entitled "The Prospects and Challenges in the Provision of Retirement Education to Workers: A Case Study of Lagos." This project demonstrated my capacity to oversee thorough investigations into intricate subjects, examining the mechanics of retirement education and addressing the difficulties encountered by Lagos workers.

My leadership in the field of data science is demonstrated by several initiatives, such as the papers "Superstore Analysis" and "The Effect of Diabetes Prediction Using SVM and KNN in R." Though these pieces have not yet been published, they are a living example of my ability to apply machine learning algorithms to real-world problems and to do in-depth data analysis to draw insightful conclusions. These projects provide as an excellent example of my proactive approach to solving practical problems using innovative data science techniques. In addition, my commitment to creating a welcoming atmosphere for women in technology is demonstrated by my position as Community Manager for Women in Data Africa (WIDA). As the team's leader, I kept a close eye on everyone's progress and promoted it while teaching a free Sunday SQL session to improve their technical abilities. This leadership position went beyond conventional limits, helping to foster the development of a community that now numbers more than fifteen hundred women. The observable results and thriving community bear witness to the beneficial influence of my leadership in advancing gender diversity and technical skill, even if I have not earned any official honours.

During my time at Reddington Zainelab, I demonstrated leadership by guiding initiatives that were based on data. Creating visually striking dashboards with Excel, PowerPoint, and other I was a key contributor to improving data representation and analytical capabilities using Power BI and R. Sales and operational effectiveness both significantly increased because of this. Furthermore, I demonstrated leadership in meeting larger societal demands by providing thorough reports to the Nigeria Centre for Disease Control in support of the country's COVID-19 response. My leadership style is still developing as I move through my master's programme in Data Science. My commitment to expanding knowledge in the field is demonstrated by my active participation in ongoing initiatives and my intention to increase my contribution to the academic debate. My leadership trajectory is distinguished by a dedication to bringing about positive change and contributing significantly to the field of data science. It is typified by a blend of scientific acumen, community engagement, and professional influence.