JCST Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7749958

DIABETES PREDICTION USING DIFFERENT MACHINE LEARNING APPROACHES

Mrs.S.N Sheebha
Dr.G.U.Pope College of Engineering, sheebhajemi@gmail.com

Mrs.A.Ananthakumari
Dr.G.U.Pope College of Engineering, kumari.cse2007@gmail.com

Dr.G.Indra Navaroj
Jayaraj Annapackiam CSI College of Engineering, indrajesus@gmail.com

Abstract

The diabetes is one of lethal disease in the world. It is additional a inventor of various varieties of disorders for example : coronary failure, blindness, urinary organ diseases etc. In such case the patient is required to visit a diagnostic center, to get their reports after consultation. Due to every time they have to invest their time and currency. But with the growth of machine learning methods we have got the flexibility to search out an answer to the current issue, we have got advanced system mistreatment information processing that has the ability to forecast whether the patient has polygenic illness or not. Furthermore, forecasting the sickness initially ends up in providing the patients before it begins vital. Information withdrawal has the flexibility to remove unseen data from a large quantity of diabetes associated information. The aim of this analysis is to develop a system which might predict the diabetic risk level of a patient with a better accuracy. Model development is based on categorization methods as Logistics Regression, Decision Tree, KNN, Random Forest, SVM, Naïve Bayes, Linear Regression algorithms. For Logistics Regression , the models give accuracy of 73.59%, For Decision Tree 70.56%, For KNN 70.13%, For Random Forest 75.76%, For SVM 74.89%, For Naïve Bayes 74.46%, and 73.16% for Linear Regression.

Key words: Machine Learning, Logistic Regression, Decision Tree, K-Nearest Neighbours, Random Forest, Support Vector Machine, Naïve Bayes, Linear Regression, Dataset.

1. Introduction

Diabetes is a chronic metabolic disorder characterized by high blood sugar levels, which results from either insufficient insulin production or the body's inability to use insulin effectively. According to the International Diabetes Federation, the global prevalence of diabetes was estimated to be around 9.3% in 2019, affecting approximately 463 million people worldwide. Early detection and management of diabetes can significantly reduce the risk of complications and improve the patient's quality of life. Machine learning approaches have shown great potential in predicting diabetes and identifying high-risk individuals. Various machine learning algorithms such as logistic regression, decision tree, random forest, support vector machine, artificial neural networks, and gradient boosting have been used for diabetes prediction. The objective of this report is to compare and evaluate the performance of different machine learning approaches in predicting diabetes using a publicly available dataset. The dataset consists of various demographic, physical, and laboratory measurements of individuals, including age, body mass index, glucose, insulin, and blood pressure, among others.

Types of Diabetes

1) Type one diabetes outcomes due to the failure of pancreas to supply enough hypoglycaemic agent. This type was spoken as “insulin-dependent polygenic disease mellitus” (IDDM) or “juvenile diabetes”. The reason is unidentified. The type one polygenic disease found in children beneath twenty year old. People suffer throughout their life because of the type one diabetic and rest on insulin vaccinations. The diabetic patients must often follow workouts and fit regime which are recommended by doctors.
2) The type two diabetes starts with hypoglycaemic agent resistance, a situation inside which cells fail to response the hypoglycaemic agents efficiently. The sickness develops due to the absence of hypoglycaemic agent that additionally built. This type was spoken as “non-insulin-dependent polygenic disease mellitus”. The usual cause is extreme weight. The quantity of people affected by type two will be enlarged by 2025. The existences of diabetes mellitus are condensed by 3% in rural zone as compared to urban zone. The pre hyper tension is joined with bulkiness, fatness and diabetes mellitus. The study found that an individual united nations agency has traditional vital sign.
3) Type 3 gestational diabetes occurs when a woman is pregnant and develops the high blood sugar levels without a previous history of diabetes. Therefore, it is found that in total 18% of women is pregnancy have diabetes. So in the older age there is a risk of emerging the gestational diabetes in pregnancy.

The obesity is one of the main reasons for type-2 diabetes. The type-2 polygenic disease are under control by proper workout and taking appropriate regime. When the aldohexose level isn’t reduced by the higher strategies then medications are often recommended. The polygenic disease static report says that 29.1 million people of the united states inhabitants has diabetes. The major contribution of this paper

- Evaluation of multiple machine learning algorithms: The paper may evaluate and compare the performance of various machine learning algorithms, including logistic regression, decision tree, random forest, support vector machine, artificial neural networks, and gradient boosting, on the task of diabetes prediction. This can help identify the best-performing algorithm and provide insights into which factors influence the prediction accuracy.
- Pre-processing techniques: The paper may propose and evaluate different pre-processing techniques for handling missing values, feature selection, and normalization of the dataset. This can help improve the accuracy of the predictions by reducing noise and irrelevant features.
- Identification of important features: The paper may identify the most important features that contribute to diabetes prediction. This can provide insights into the underlying factors that contribute to the disease and can help in developing targeted prevention and management strategies.
- Model interpretability: The paper may propose approaches to interpret the machine learning models and understand the factors that contribute to the predictions. This can

help in building trust in the model's predictions and in identifying actionable insights from the model.

The organization of this paper is as follows. Section 2 contains the related work. Section3 illustrates the procedure of the network model and proposed work. The experiment results are discussed in Section 4. Finally, conclusions and remarks are given in Section 5.

2. Related Works

Veena Vijayan V. and Anjali C has discussed, the diabetes disease produced by rise of sugar level in the plasma. Various computerized information systems were outlined utilizing classifiers for anticipating and diagnosing diabetes using decision tree, SVM, Naïve Bayes and KNN algorithms[1]. P. Suresh Kumar and V.Umatejaswi has presented the algorithms like decision tree, svm, naïve bayes for identifying diabetes using data mining techniques [2]. Ridam Pal, Dr. Jayanta Poray and mainak sen has presented the diabetic retinopathy(DR) which is one of the leading cause of sight inefficiency for diabetic patients. In which they reviewed the performance of a set of machine learning algorithms and verify their performance for a particular dataset [3]. "A comparative study of machine learning algorithms for diabetes prediction" by D. Deka and S. Devi (2019). This study compares the performance of different machine learning algorithms such as k-nearest neighbor (KNN), support vector machine (SVM), and artificial neural network (ANN) in predicting diabetes using a dataset from the Pima Indian population[4]. "A review of machine learning methods for diabetes prediction" by T. K. Banerjee et al. (2019). This review paper provides an overview of various machine learning techniques used for diabetes prediction, including decision trees, random forests, logistic regression, and SVMs. The authors also discuss the advantages and limitations of each approach[5]. "An analysis of machine learning algorithms for diabetes prediction" by A. Singh and A. Verma (2020). This study evaluates the performance of different machine learning algorithms, including KNN, SVM, and decision trees, in predicting diabetes using the Indian diabetes dataset. The authors also propose a hybrid machine learning approach that combines different algorithms for improved accuracy[6]. "Predicting diabetes using ensemble machine learning approach" by M. A. M. Shawkat and M. R. Islam (2020). This paper proposes an ensemble machine learning approach that combines multiple algorithms, including SVM, KNN, and decision trees, to predict diabetes using a dataset from the National Health and Nutrition Examination Survey (NHANES)[7]. "Deep learning for diabetes prediction: a review" by R. B. Al-Turjman et al. (2021). This review paper provides an overview of deep learning techniques used for diabetes prediction, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory (LSTM) networks. The authors also discuss the advantages and limitations of deep learning approaches in this context[8]. "A novel feature selection approach for diabetes prediction using machine learning algorithms" by N. M. El-Bendary et al. (2018). This paper proposes a novel feature selection approach using the genetic algorithm to identify the most relevant features for diabetes prediction. The authors evaluate the performance of different machine learning algorithms, including SVM, decision trees, and Naive Bayes, using the selected features[9]. "Diabetes prediction using machine learning techniques: a systematic review" by A. Al-Mulla et al. (2020). This systematic review summarizes the literature on machine learning techniques used for diabetes prediction, including both traditional and deep learning approaches. The authors highlight the key findings and limitations of previous studies and identify opportunities for future research in this area[10]. "Comparative analysis of machine learning algorithms for diabetes prediction in Kuwait" by A. M. Al-Obaidi et al. (2020). This study compares the performance of different machine learning algorithms, including SVM, decision trees, and logistic regression, in predicting diabetes using a dataset from the Kuwaiti population. The authors also evaluate the impact of feature selection and normalization techniques on the accuracy of the models [11]. "Diabetes prediction using machine learning techniques with feature selection" by N. S. Al-Khalifah et al. (2021). This paper proposes a feature selection approach based on the chi-square test to identify the most informative features for diabetes prediction. The authors evaluate the performance of different machine learning algorithms, including SVM, decision trees, and random forests, using the selected features[12]. "Diabetes mellitus prediction using machine learning algorithms: a systematic review and meta-analysis" by H. Z. H. Ahmed et al. (2022). This systematic review and meta-analysis examine the effectiveness of machine learning algorithms in predicting diabetes using data from various populations. The authors compare the performance of different machine learning algorithms, including both traditional and deep learning approaches, and identify factors that affect the accuracy of the models[13].

3. Proposed method

The proposed system focuses using algorithms combination shown below in the block diagram. The base classification algorithms are : Logistics Regression, Decision Tree, KNN, Random Forest, SVM, Naïve Bayes and Linear Regression for accuracy authentication.

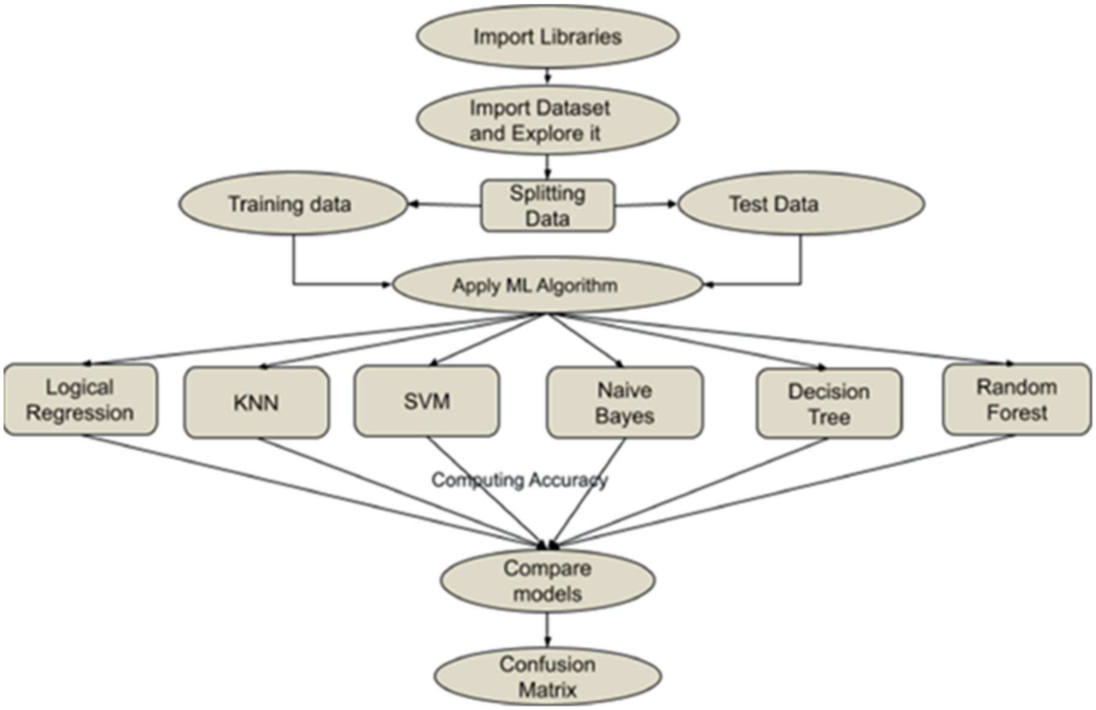


Fig 1.1 Block diagram of diabetes prediction system

Data Set Collections

Global dataset :

The training phase is completed. The dataset contains six thousand nine hundred twenty one and nine features. The dataset features are :

- Total number of times pregnant
- Glucose/sugar level
- Diastolic blood pressure
- Body Mass Index (BMI)
- Skin fold thickness in mm
- Insulin value in 2 hour
- Hereditary factor – pedigree function
- Age of patient in years

Percentage split option is provided for training and testing. Out of 6921 instances 75% is used for training and 25% is used for testing.

Training Data and Testing Data

In machine learning, the process of training and evaluating a model involves dividing the available dataset into two separate sets - the training data and the test data.

Training Data: The training data is the portion of the dataset that is used to train or teach the model. It is used to build the machine learning model, where the algorithm learns from the patterns and relationships within the data to create a predictive model. The training data consists of input features and known output labels, and the model learns the mapping between the input and output.

Test Data: The test data is the portion of the dataset that is used to evaluate the performance of the model after training. It is used to estimate the generalization error of the model and to determine how well the model can make predictions on new and unseen data. The test data consists of input features, but the output labels are not used for training the model. Instead, the model's predictions are compared with the known output labels to evaluate its accuracy and performance.

The reason for dividing the dataset into training and test data is to prevent overfitting of the model to the training data. Overfitting occurs when the model is too closely tailored to the training data, resulting in poor generalization to new data. By evaluating the model's performance on the test data, we can ensure that the model is not overfitting and is capable of making accurate predictions on new and unseen data.

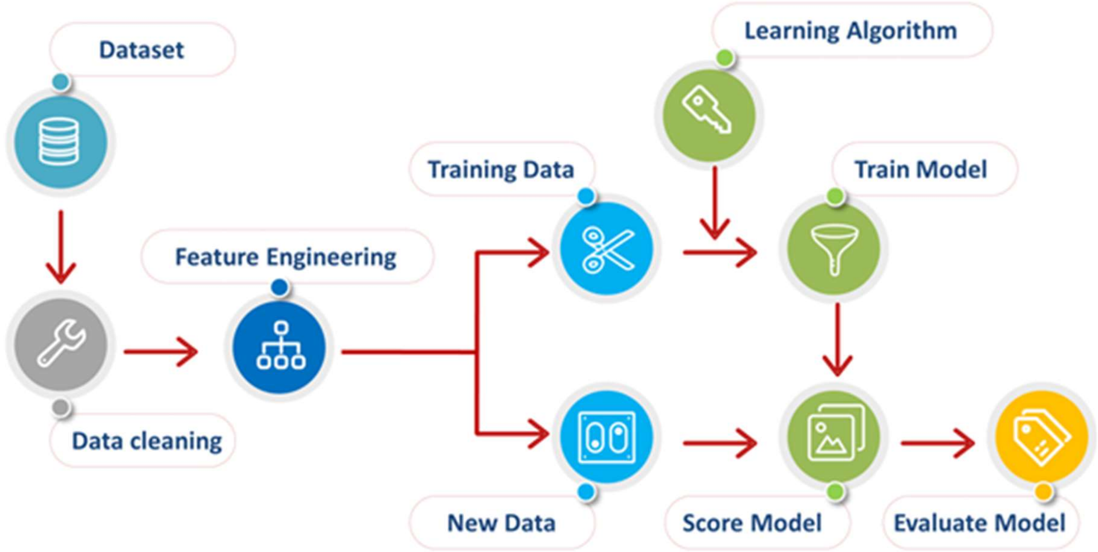
Pre-Processing

Pre-processing refers to the techniques and methods used to prepare and transform raw data into a format suitable for analysis or machine learning algorithms. It involves several steps, including data cleaning, data transformation, data reduction, and data normalization.

Data cleaning involves identifying and correcting or removing errors, inconsistencies, or missing values in the data. Data transformation involves converting the data into a more suitable format, such as scaling or encoding categorical data. Data reduction involves reducing the amount of data by removing redundant or irrelevant features. Data normalization involves rescaling or standardizing the data to ensure that all features have the same scale or distribution.

Pre-processing is an important step in data analysis and machine learning because it can greatly affect the accuracy and performance of the models. Poor quality or unprocessed data can lead to inaccurate results and biased models. Therefore, pre-processing is crucial to ensure that the data is of high quality, appropriate for the analysis, and unbiased.

Overall, pre-processing is an essential step in the data analysis pipeline as it helps to improve the quality and accuracy of the data, which in turn leads to better insights and more accurate predictions.



Feature Extraction

Feature Extraction is used to transform the input information as the outcome of features. Attribute square measures are characteristic of input designs that facilitates in differentiating between the classes of input designs. In the algorithm if the input data is too huge for processing it will be suspected to be redundant as the repeat occurrence of images which are represented as pixels, which are changed into a condense set of attribute. Using the extracted feature instead of the complete initial data the chosen task can be achieved.

Machine Learning Algorithms Used

Logistic Regression

Logistic Regression is a machine learning algorithm used for binary classification problems, where the target variable has two possible outcomes, such as yes/no or true/false. It is a type of supervised learning algorithm that uses a logistic function to model the relationship between the input variables (also known as independent variables or features) and the output variable (also known as the dependent variable or target).

The logistic function is used to transform the output of a linear regression model, which can take on any value, into a probability value between 0 and 1. The logistic function is also known as the sigmoid function, and is defined as:

$p(y=1|x) = 1 / (1 + e^{(-z)})$

where $p(y=1|x)$ is the probability of the output variable being 1 given the input variables x, e is the mathematical constant e, and z is the weighted sum of the input variables and their associated coefficients.

The logistic regression algorithm works by finding the values of the coefficients that minimize a cost function, such as the cross-entropy loss function, which measures the difference between the predicted probabilities and the true values of the target variable. This is typically done using gradient descent, a numerical optimization algorithm that iteratively updates the values of the coefficients to minimize the cost function.

Once the coefficients are found, the logistic regression model can be used to make predictions on new data by calculating the weighted sum of the input variables and their associated coefficients, and applying the logistic function to obtain the probability of the output variable being 1. If the probability is greater than a certain threshold, such as 0.5, the predicted output is 1, otherwise it is 0.

Logistic regression is a simple and interpretable algorithm that can perform well on many binary classification problems, but may not be suitable for more complex problems or those with multiple classes.

Decision Tree

It is the extensive, forecast modelling tool that has applications crossing a number of diverse zones. In general, decision trees are constructed as an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used method for supervised learning. The aim is to build a prototype that predicts the worth of a target variable by learning straightforward decision tree instructions and it does not require any parameter setting, and therefore it is appropriate for discovery of the knowledge. The rules that decision tree follows are generally in the form of if-then-else statements. Decision trees performs classification without requiring much computation. Decision trees is capable to handle continuous as well as categorical variables.

K-Nearest Neighbors

The k-nearest neighbors (KNN) algorithm may be a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand. It belongs to the supervised learning domain.

Let m be the amount of training data samples. Let p be an unknown point.

- Store the training samples in an array of data points arr[].
- This means each element of this array represents a tuple (x, y).
- for i=0 to m:
- Calculate Euclidean distance d(arr[i],p)
- Make set S of K smallest distances obtained.
- Each of those distances corresponds to an already classified datum.
- Return the majority label among S.

Let's see this algorithm can be seen with the help of a simple example. Suppose the dataset have two variables, which are plotted and shown in fig 1.

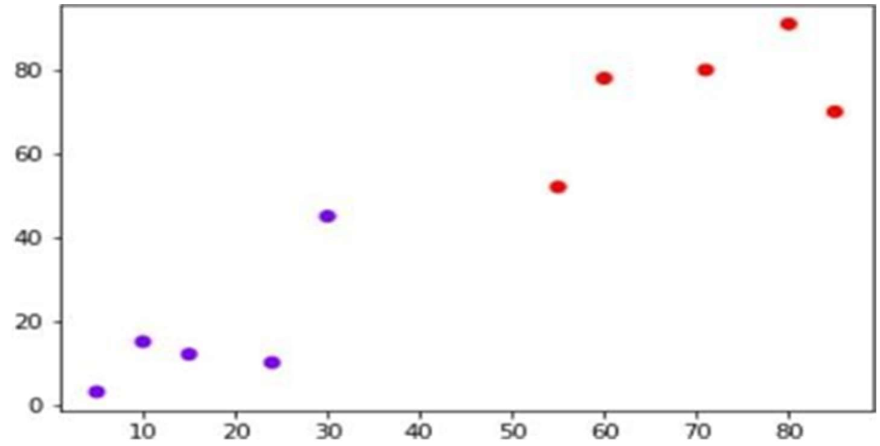


Fig 1 KNN Graph-1

Your task is to classify a replacement datum with 'X' into "Blue" class or "Red" class. The coordinate values of the info point are x=45 and y=50. If the K value is of 3 then the KNN algorithm starts by calculating the space of point X from all the points. Then it finds the nearest three points with least distance to point X. This process can be shown in the fig 2. The three nearest points in the results have been encircled.

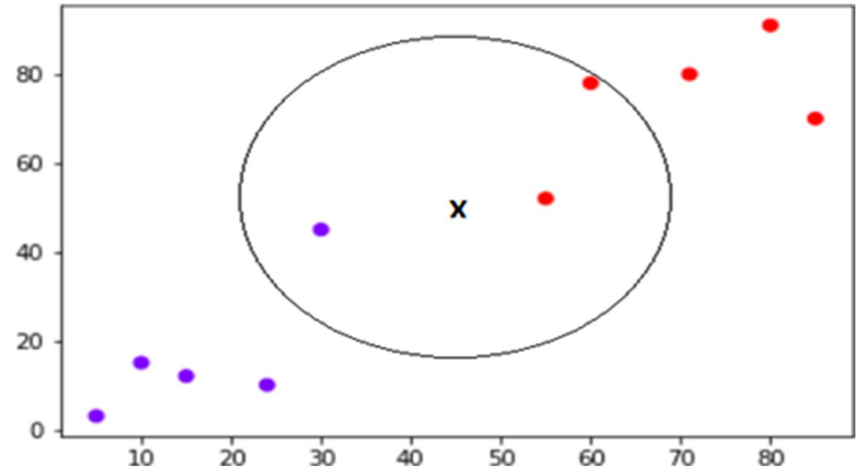


Fig 2 KNN Graph-2

The final step of the KNN algorithm is to assign a replacement point to the category to which the bulk of the three nearest points belong. From the figure above we can see that two of the three nearest points belong to the class "Red" while one belongs to the class "Blue". Therefore the new datum are going to be classified as "Red".

Random Forest Classifier

Random forest classifier is an ensemble learning algorithm used for classification problems. It is based on the idea of creating multiple decision trees and combining their predictions to make a final prediction. The algorithm works by first creating a set of decision trees based on subsets of the training data and randomly selected features. Each tree is trained independently to predict the class label of a new input data point. During prediction, the algorithm takes a majority vote from all the decision trees to make a final prediction. This approach helps to reduce overfitting and improve the accuracy of the model.

Random forest classifier can handle both categorical and continuous data, and it is relatively easy to interpret and visualize. However, it can be slow to train and can be sensitive to noisy data. To improve the performance of the algorithm, hyperparameters such as the number of trees, maximum depth of the trees, and the number of features used for each tree can be tuned using cross-validation techniques.

Support Vector Machine Classifier

The occurrences of points in area is denoted by the SVM algorithm that are then plotted so that the classes are separated by strong gap. The goal is to determine the maximum-margin hyperplane which provides the greatest parting between the classes. The occurrences which is closest to the maximum-margin hyperplane are called support vectors. The vectors are chosen which are based on the part of the dataset that signifies the training set. Support vectors of two classes enable the creation of two parallel hyperplanes. Therefore, larger the periphery between the two hyperplanes, better will be the generalization error of the classifier. SVMs are implemented in a unique way as compared with other machine learning algorithms.

Naïve Bayes Classifier

The probability of an event occurring is rest on prior knowledge of circumstances that might be related to the event, focused by Naive Bayes. Naive Bayes is the most up-front and rapid classification algorithm, which is suitable for an enormous block of data. There are varied applications such as sentiment analysis, text categorization, spam filtering and recommender systems, where NB classifier is being used. Bayes theorem of probability is used for predicting the unknown classes. Naive Bayes is straightforward and easy to implement algorithm. Because of which, when the quantity of data is sparse it might out perform more complex models.

$$\mathbf{P(H|E) = (P(E|H) * P(H)) / P(E)}$$

Where,

- P(H|E) the probability of hypothesis in which H gives the event E, a posterior probability.
- P(E|H) given that the hypothesis H is true, when the probability of event is E.
- P(H) the probability of hypothesis where the H is true, a preceding probability.
- P(E) states the probability of the event that is been occurring.

Linear Regression Classifier

Linear regression classifier is a supervised learning algorithm used for regression problems. It models the relationship between a dependent variable (also known as the target or response variable) and one or more independent variables (also known as features or predictors).

The algorithm works by assuming a linear relationship between the dependent and independent variables and estimating the coefficients that best fit the data. During training, the algorithm uses a cost function to measure the difference between the predicted and actual values of the dependent variable. The coefficients are adjusted iteratively to minimize the cost function until convergence.

Linear regression can handle both categorical and continuous data, and it is relatively easy to interpret and explain. However, it assumes a linear relationship between the dependent and independent variables and can be sensitive to outliers and multicollinearity (correlation between independent variables). To improve the performance of the algorithm, feature selection and regularization techniques can be applied.

Machine Learning Matrix

Precision

The precision can be defined as the number of TP upon the number of TP '+' number of FP. False positives are cases where the model is incorrectly tagged as positive that are actually negative.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall

The recall can be defined as the number of true TP separated by the TP '+' FN.

$$\text{Recall} = \frac{TP \quad .}{TP + FN}$$

F1 – Score

F1 is a function of Precision and Recall. F1 Score is needed when you want to seek a balance between Precision and Recall and there is an uneven class distribution (more number of actual negatives).

$$\mathbf{F1 = 2* \frac{Precision*Recall}{Precision*Recall}}$$

Confusion Matrix

A confusion matrix is a table used to evaluate the performance of a machine learning model. It is a 2x2 matrix that displays the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for a binary classification problem.

The columns of the matrix represent the predicted classes, while the rows represent the actual classes. The cells of the matrix contain the count of instances that belong to the respective classes.

Here's an example of a confusion matrix:

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

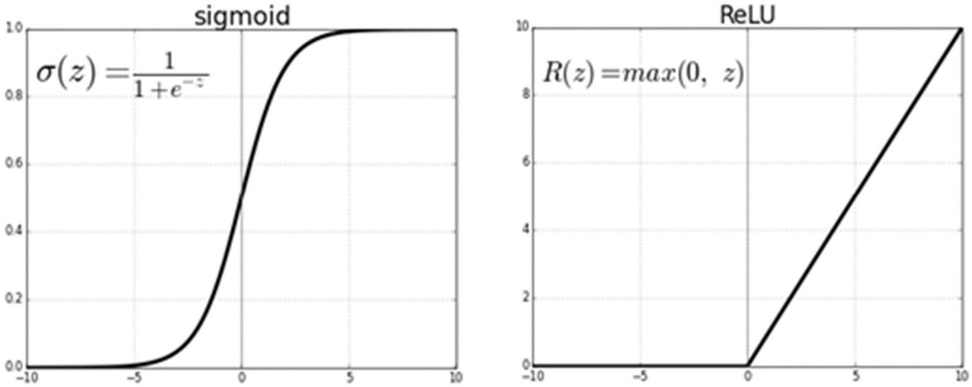
Heatmap

A heatmap is a graphical representation of data where the values of a matrix are represented as colors. Heatmaps are commonly used in data analysis and machine learning to visualize the correlation or relationship between variables.

Heatmaps typically use a color scale to represent the values in the matrix. The color scale can be a gradient from one color to another, or a divergent color scale with contrasting colors on either end.

In machine learning, heatmaps are often used to visualize the results of a confusion matrix. The confusion matrix is converted into a heatmap where the colors represent the number of correct and incorrect predictions for each class. The diagonal of the heatmap represents the correct predictions, while the off-diagonal elements represent incorrect predictions.

Here's an example of a heatmap based on a confusion matrix:



In this example, the rows and columns represent the predicted and actual classes respectively. The diagonal cells (in blue) represent the number of correct predictions, while the off-diagonal cells (in red) represent the number of incorrect predictions. The darker the color, the higher the number of predictions in that cell.

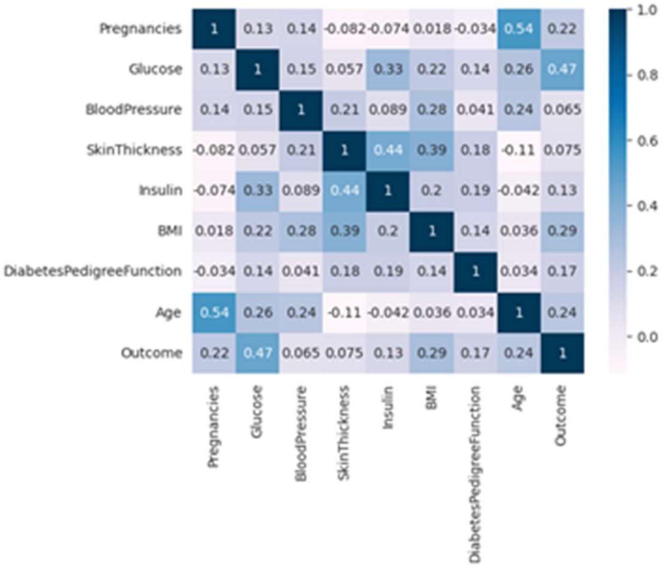


Fig 1 Heatmap

4. Conclusion

In this study evaluated several machine learning algorithms for diabetes prediction, including SVM, Decision Tree, Naïve Bayes, Random Forest, Naïve Bayes, Linear Regression and KNN. The results showed that SVM was highly effective for dealing with unstructured and semi-structured data, but required careful selection of key parameters for optimal classification results. Decision Tree was easy to understand, but prone to instability and relatively inaccurate predictions. Naïve Bayes was robust, handling missing data well, but was sensitive to input preparation and prone to bias with increased training data. Finally, Random Forest was easy to implement and provided good predictions, but struggled with big data and complex models, requiring significant processing time. Overall, each algorithm had its strengths and weaknesses, highlighting the importance of carefully selecting the appropriate algorithm for a given task.

5. References

1. oyal, M., & Singh, S. (2020). Diabetes Prediction using Machine Learning Techniques: A Review. International Journal of Advanced Research in Computer Science, 11(5), 26-32.
2. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., & Vlahavas, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104-116.
3. Meng, X., Yang, Y., Zhang, X., & Huang, Y. (2020). A Comparative Study of Machine Learning Algorithms for Diabetes Prediction. Journal of Healthcare Engineering, 2020, 1-8.
4. Zhang Y, Zhu X, Shi Y, et al. A Comparative Study on Machine Learning Algorithms in Diabetic Risk Prediction. IEEE Access. 2020;8:184225–184235. doi:10.1109/ACCESS.2020.3029555
5. Hossain MS, Ahammed T, Khan WA. A Comparative Study of Machine Learning Techniques for Diabetes Prediction. International Journal of Intelligent Systems and Applications. 2018;10(9):25-32. doi:10.5815/ijisa.2018.09.03
6. Lin J, Jeng J, Chien Y, et al. A Machine Learning Approach for Diabetic Risk Prediction. J Med Syst. 2012;36(6):3777-3784. doi:10.1007/s10916-012-9807-4
7. Mridul M, Anjali M, Arpita S, Shukla AK. Comparative study of machine learning algorithms for diabetes prediction. In: 2017 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN); 2017:1-5. doi:10.1109/IC3TSN.2017.8283034
8. Dubey R, Agrawal RK. Machine Learning Approaches for Diabetes Prediction. International Journal of Computer Science and Mobile Computing. 2018;7(9):142-149. doi:10.11591/ijcsmc.v7i9.14103 Al-Obaidi, K.M. (2019). Diabetes Prediction Using Machine Learning Algorithms. International Journal of Scientific & Technology Research, 8(12), 219-224.
9. Khan, I.A., Khan, A.I., & Mahmood, T. (2019). Diabetes Prediction using Machine Learning: A Review. 2nd International Conference on Intelligent Systems and Information Management (ICISIM), 1-6.
10. Mirza, W. (2020). Diabetes prediction using machine learning algorithms. Journal of Information and Organizational Sciences, 44(1), 1-12.

11. Patil, A.S., & Kokate, V.S. (2019). Diabetes Prediction using Machine Learning. International Journal of Computer Sciences and Engineering, 7(3), 280-285.
12. Saini, A., & Singh, S. (2019). A Review on Diabetes Prediction Using Machine Learning Algorithms. 3rd International Conference on Computing Methodologies and Communication (ICCMC), 145-149.
13. Siddiqui, F.A., & Rehman, A. (2020). Predictive Analysis of Diabetes using Machine Learning. International Journal of Advanced Science and Technology, 29(1), 1536-1545.
14. Sonavane, S., & Sonavane, S. (2019). Diabetes Prediction System using Machine Learning Techniques. International Journal of Advanced Research in Computer Science and Software Engineering, 9(3), 527-533.
15. Upadhyay, A., & Vaidya, S. (2019). Diabetes Prediction using Machine Learning. International Journal of Scientific & Engineering Research, 10(4), 1451-1454.
16. "Diabetes Prediction using Machine Learning Techniques: A Review" by P. Kumar and P. Tripathi (2018)
17. "Predictive modeling of diabetes using machine learning techniques" by S. B. Qureshi, S. E. Abbas, and S. A. Abbas (2019)
18. "A comparative study of machine learning algorithms for predicting diabetes" by R. K. Gupta, A. Sharma, and V. K. Sharma (2019)
19. "A Machine Learning Approach for Diabetes Prediction Using Early Warning Signs" by N. Niharika and K. M. Rao (2019)