International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 02 | February - 2023 Impact Factor: 7.185 ISSN: 2582 3930

Diabetes Prediction using Machine Learning Algorithms

Dr. Chandrashekar D. K[1], Imran Pasha C[2], Roshan Ameen[3], Shashidhar Kundur[4], Suraj Gurjar M[5]

[1]Dr. Chandrashekar D. K, Dept of Computer Science and Engineering, CITech, Bengaluru-560036
[2]Imran Pasha C, Dept of Computer Science and Engineering, CITech, Bengaluru-560036
[3]Roshan Ameen, Dept of Computer Science and Engineering, CITech, Bengaluru-560036
[4]Shashidhar Kundur, Dept of Computer Science and Engineering, CITech, Bengaluru-560036
[5]Suraj Gurjar M, Dept of Computer Science and Engineering, CITech, Bengaluru-560036

Abstract – Diabetics detecting using machine learning is a process that utilizes algorithms to dissect medical data and prognosticate the liability of a patient developing diabetes. The aim of this process done to detect the citing of diabetes early and give applicable treatment to help farther complications. Machine literacy algorithms like Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest (RF), K-Means, K-Nearest Neighbour (KNN) and Naïve Bayes (NB) can be used to analyze a variety of factors, such as patient information, lifestyle and previous medical history, to make predictions about how likely patient developing diabetes. The accuracy of these predictions can be improved through training various algorithms on large number of data and fine-tuning their parameters. By using machine learning for diabetic prediction, healthcare providers can make more informed decisions about patient care and improve patient outcomes.

Key Words: Classification, Datasets, Diabetes, Machine Learning, Prediction,

1. INTRODUCTION

Diabetes is a habitual health condition that affects millions of people worldwide. It's characterized by high situations of sugar (glucose) in the blood, which can lead to serious health complications if not duly managed. Early prediction of diabetes is critical for individuals to take proactive measures to prevent or manage the disease, which can lead to better health outcomes and reduced healthcare costs. In 1980, the number of individuals worldwide affected by diabetes was a staggering 122 million, which rose to a shocking 433 million by 2014. Predictions indicate that this figure will surge to an alarming 640 million by 2040. Tragically, diabetes was directly responsible for over 1.5 million deaths.

Every time, roughly 3 to 5 million cases succumb to death as a result of diabetes. thus, it's an intimidating figure to us. Machine literacy is a fleetly growing field that has the implicit to greatly ameliorate diabetes vaticination. Machine learning (ML) algorithms use large amounts of data to train models that can accurately predict outcomes. In the case of diabetes mellitus prediction, these algorithms can analyze various factors such as medical history, demographic information, and lifestyle habits to identify individuals at risk for the disease. The accuracy of machine learning algorithms in predicting diabetes mellitus has been extensively studied and validated in numerous clinical trials. Many studies have shown that ML algorithms can accurately identify individuals at risk for the disease, even when traditional predictive models, such as logistic regression, fail to do so. This is because machine learning algorithms can process a wider range of factors and find complex relationships between them A plethora of machine learning algorithms exist and are widely employed in diabetes mellitus prediction, including decision trees, random forests, and ANN.

Each algorithm has its own positives and negatives, and the selection of appropriate algorithm is contingent upon the unique requirements of the investigation. For e.g., decision tree is easy to understand, but may not always produce the most accurate results. Artificial neural networks, on the other hand, retain a high degree of complexity and are able of recycling vast quantities of data. However, their output may be challenging to comprehend. This, coupled with their ability to enhance the prediction of diabetes, makes them a valuable tool, machine learning algorithms can also help healthcare providers make more informed treatment decisions. For example, machine learning algorithms can analyze patient data to identify individuals who are likely to respond well to certain treatments and those who are more likely to experience adverse side effects. This information can be used to personalize treatment plans and improve patient outcomes.

The diverse structures of data can be analyzed through the utilization of various Machine Learning (ML) algorithms. Predictive analysis in the healthcare sector is an area of study that utilizes ML algorithms on healthcare datasets for analysis. This particular study focuses on gravid diabetes and employs K-Nearest Neighbor(KNN), Support Vector Machine(SVM), logistic retrogression(LR), and Random Forest(RF) are Machine Learning ways on the Pima Indian Diabetes Database(PIDD) to determine the delicacy of diabetes vaticination. The trial involves assessing different parameters, similar as glucose situations, blood pressure, and BMI, to measure the perfection of the results. In conclusion, machine learning has the potential to revolutionize the way diabetes mellitus is predicted and treated. By leveraging the power of machine learning, healthcare providers and individuals alike can gain valuable insights into the risk of developing diabetes and take proactive measures to prevent or manage the disease. This can lead to better health outcomes and reduced healthcare costs, making machine learning a critical tool in the fight against diabetes mellitus.

2. RELATED WORK

Shahadat Uddin, *et al.* [1] This comprehensive exploration aimed to identify studies that employed multiple supervised ML ways for the diabetes vaticination of a single complaint. The study conducted a hunt on two databases, Scopus and PubMed, using a variety of hunt terms. As a consequence of this, a comprehensive review was conducted which involved the comparison of further than 40 papers related to the application of colorful machine learning algorithms for complaint vaticination. Out of the 17 studies that applied RF, it had the loftiest delicacy in 9 studies, which amounts to 53. SVM came in alternate with the loftiest delicacy in 41 of the studies in which it was estimated.

Shejal Kale, *et al.* [2] The authors of the paper developed a system that enhances the delicacy of early diabetes vaticination by combining the issues from colorful machine learning algorithms. The machine literacy ways incorporated in this system encompass K- Nearest Neighbor(KNN), Logistic Retrogression(LR), Random Forest(RF), Support Vector Machine(SVM), and Decision Tree(DT). The precision of the model was evaluated for each of the algorithms. The results of the diabetes prediction were analyzed to assess the accuracy of the forecasting, the time taken to compute the accuracy, and the correct and incorrect classification of the prediction results. The KNN algorithm was used for diabetes prediction and was classified into "Yes" or "No". The same KNN algorithm was used for the time prediction module. The efficacy of the project was determined by contrasting the testing data with the real data.

Deepti Sisodia, *et al.* [3] In this exploration paper, the Naive Bayes(NB), Decision Tree and SVM algorithms are enforced. The evaluation of the algorithms was carried out using a 10-fold cross-validation approach. The criteria employed to assess the performance of the classifiers were delicacy, F- Measure, recall, perfection, and ROC(Receiver Operating Characteristic).The experiments were performed utilizing the WEKA software. The primary thing of the disquisition was to read the circumstance of diabetes in womanish cases through the use of the WEKA software and the PIDD medical database. The PIDD dataset utilized in the study contains medical information for 768 female patients and comprises of 8 numerical attributes. The values of class '0' and class '1' were interpreted as follows: '0' represented a negative class label for diabetes, while '1' represented a positive class label for diabetes. The conclusion of the study was that the Naive Bayes classification algorithm exhibited the best performance among all the algorithms, with an accuracy of 76.30%. This made it the most effective supervised machine learning method in the experiment.

K. VijiyaKumar, *et al.* [4] The project adopted a two-phased approach. In the first phase, the data was obtained from UCI Repository. In second phase, the raw data underwent pre-processing, which consisted of cleaning, integration, and transformation. The findings indicated that the Random Forest algorithm achieved lesser delicacy compared to the other ways. Throughout the disquisition, the Random Forest algorithm was completely examined and estimated through colorful criteria. The end of this study was to produce a system that can directly prognosticate diabetes at an early stage, by integrating machine literacy styles that enhance the capability to read diabetes with bettered perfection.

Smriti Gupta, *et al.* [5] The PIDD dataset was differentiated into 80 as training sets and 20 as testing sets. The objective of the paper was to accurately predict, based on medical history, whether an individual has diabetes or not (designated as class 0 or 1) The data pre-processing, and the Naive Bayes, Support Vector Machine(SVM), and k-FoldCross-Validation algorithms were employed. The authors of the study determined the perfection of their prognostications by testing different values of k(3, 5, 10) and opting the stylish result with the loftiest delicacy from the performing list.

Krati Sexena *et al.* [6] The author focused on the subject of Diabetes and considered the KNN algorithm was crucial artificial intelligence technique. The level of accuracy indicates how closely the results of the test database match the output data of the various attribute data sets. To evaluate the deviation from the expected outcome by comparing the test data results to the labors of the features in the training dataset. The findings indicate that as the value of k rises, the perfection and misclassification rates will drop. The K- nearest neighbor algorithm is a prominent fashion within artificial intelligence, widely employed for diagnostic applications. Moreover, you can achieve greater accuracy through this method, which proves to be highly efficient when dealing with substantial training data sets.

Mercaldo *et al.* [7] This study utilized six varying classifiers methods. Classifiers: Random Forest, J48, Bayes Net, Multilayer Perceptron, JRip, and Hoe ding Tree. The research also incorporated the Pima Indians dataset. Despite the lack of pre-processing steps mentioned by the authors, they focused on two algorithms, Best First and Greedy Stepwise [7], to identify the differentiating attributes that enhance categorization efficiency. The authors choose attributes age of diabetic function, BMI, PCG and birth. The dataset was subordinated to a10-foldcross-validation. Classifiers were compared grounded on Precision, Recall, and F- measure values [7]. The findings demonstrated that the Hoeding Tree algorithm achieved a perfection value of.75, a recall of.76, and an F- score of.75, which achieved advanced performance when compared to the other ways.

Veena Vijayan V, Anjali C *et.al.* [8] The authors submitted a decision support system that supports the AdaBoost, with Decision Refuse serving as the primary bracket fashion. This study used datasets collected from colorful regions across Kerala, India. The AdaBoost algorithm was paired with colorful base classifiers, including SVM, NB, Decision Refuse, and DT. The proposed system was estimated using the original dataset in the coming step, and eventually, the delicacy of the AdaBoost algorithm when combined with the base classifiers was determined. The results demonstrated that the AB algorithm which includes Decision Refuse as the classification base had the loftiest delicacy in prognosticating diabetes, with a result of 80.7, and a low error rate was recorded.

Kumari Mukesh, Vohra Rajan, Arora Anshul *et.al.* [9] The study found that AdaBoost algorithm achieved highest accuracy in diabetes prediction at 80.7%, and a low error rate was noted. The datasets used in this study were attained from hospitals and comported of 9 attributes and 209 records. Preprocessing was performed on the dataset to define and elect suitable records. The method proposed was implemented using the WEKA software, and a Bayesian network, which employs a probability-

based graphical model, was utilized. The researchers achieved an accuracy rate of 99.51% using the Bayesian network classifier algorithm, resulting in a reduced error rate of 0.48%.

K. Rajesh *et.al.* [10] The author proposed incorporating data mining techniques to classify and analyze medical data related to diabetes and make predictions about a patient's likelihood of having the disease. The goal was to establish an efficient classification relationship through data mining. The authors introduced a system that first performs a feature relevance analysis on the training data, then compares different classification algorithms, selects the best classifier, improves it, and finally evaluates its performance in comparison to the training data. The improved classification algorithm used was the C4.5 algorithm and it achieved a classification rate of 91%.

Muhammad Azeem *et al.* [11] The goal of study to determine if an individual has diabetes or not based on their diagnostic test results. The researchers used 6 discriminating machine learning algorithms in order to make this prediction. This study aimed to determine diagnostic measurements. To accomplish this goal, the authors utilized 6 different ML algorithms including KNN, NB, SVM, DT, LR, and RF. They employed a dataset acquired from the PIDD(National Institutes of Diabetes and Digestive and order conditions) to read diabetic cases and ameliorate clinical opinions. Which contained records for 768 Pima Indian women and had 9 attributes. The Enthought Canopy tool, and SVM and KNN algorithms were found to have the highest accuracy, with a value of 77%.

Md Faisal Faruque *et al.* [12] In this study, 200 cases were named to gather a real- world dataset, including colorful records related to diabetes mellitus. The experimenters also compared the performance of several machine literacy ways, estimated the prognostications made grounded on the applicable threat factors, and aimed to determine the most accurate system. confirmation was performed using several bracket algorithms including SVM, NB, KNN, and the C4.5 DT algorithm. In this study, experimenters attained data on diabetes cases from the Medical Centre Chittagong in Bangladesh, which encompasses information on 200 individualities and their threat factors associated with the complaint. The N-fold Cross Validation fashion was employed, with N equal to 10. The report showed that the C4.5 DT algorithm performed well than other classifiers in prognosticating diabetes mellitus, with a perfection of 72, a recall of 74, and an f- measure of 72, which was advanced than other literacy styles.

Roshan Birjais *et al.* [13] In their article, three ML techniques were used to improve the diagnosis of diabetes: Gradient boosting, Logistic Regression, and Naive Bayes. A variety of evaluation metrics were employed to gauge the effectiveness of the classifiers, including accuracy, specificity, sensitivity, and error rate, in order to give a quantitative assessment of their performance. to analyze the values of the classifiers, variant of metrics was considered such as accuracy, specificity, sensitivity, and error rate. These metrics give a numerical measure of the classifiers' performance. Furthermore, the ROC curve was used to showcase the balance between sensitivity and specificity. This ROC plot is crucial to understand more and less prediction. Additionally, the AUC was used as another method to determine the values of the classifiers. The accurate finding of the Gradient Boosting algorithm was found to be 86% while NB had a accurate margin 77% and LR had an accurate margin of 79%. The ROC and AUC metrics were also taken into consideration to evaluate the performance of these classifiers.

Arvind Aada *et al.* [14] Taking into account the situation described in the opening, we present a model that aims to improve predictions of individuals suffering from diabetes. This model incorporates various classifiers, including DT, KNN and NB, and utilizes dimensionality reduction methods like PCA, to streamline the dataset. After applying PCA on the attributes, the method yielded 6 key characteristics for training the classifiers, including DT, KNN, and NB. Researchers found that the Decision Tree algorithm was more effective than other classifiers and previous studies in predicting diabetes. It achieved an accuracy rate of 94.44%, making it a simple and effective classifier for analytics of diabetes.

Md. Maniruzzaman *et al.* [15] The target of development is to establish a ML related solution for diagnosing diabetes patients. The logistic regression algorithm was utilized to determine the key concepts for the diseases using the approach of odds ratio and technique known as p-value. To improve the accuracy of predictions, four different classifiers were implemented, including NB, DT, Adaboost, and RF. In this research effort, a dataset acquired from the NHANES spanning the years 2009 to 2012 was analyzed. A ML based concept system was devised for diabetic patient prediction using the logistic regression algorithm. The study involved experimenting with four distinct classifiers - NB, DT, Adaboost, and RF - to determine the most accurate method of predicting diabetes. The research protocol was executed through three types of partitioning, K2 protocol, K5 protocol, and K10 protocol, with each protocol repeated 20 times to solidify the results. The accuracy (ACC) and AUC metrics are employed to assess the efficiency of fetched classifiers. With the amalgamation of feature selection and classifier, the K10 protocol attains an ACC of 94.25% and an AUC of 0.95.

O.S. Soliman *et al.* [16] A new mongrel algorithm has been suggested Omar and Eman employed the Modified- flyspeck mass Optimization(MPSO) and Least Places- Support Vector Machine(LS- SVM) algorithms to classify type 2 diabetes. The MPSO algorithm was employed to classify cases into two groups, while the LS- SVM algorithm was used to identify the optimal hyperactive- plan for this division-alive and departed. Therefore, the algorithm is quite sensitive to variations in its parameter values. The Modified-PSO algorithm was utilized as a means of optimizing the parameters for the study. In this research, the algorithm being suggested consists of two distinct stages: classification and optimization of parameters. The data utilized was from PIDD. The alternate phase of the bracket involved two sub-phases- training and testing- exercising the optimized parameters and RBF kernel. In the training phase, the evaluation was carried out using the10-foldcross-validation system, which yielded an delicacy rate of 97.8. The

performance of these algorithms was also compared to that of other algorithms that were run on the same dataset.

K. Rajesh, V *et al*. [17] This study introduces a data mining system designed to categorize diabetes data and diagnose whether a patient has diabetes. The system was trained using the PIDD dataset. During the experiments, the first stage emphasized on feature selection to determine the crucial features for the classification process. The relevance feature analysis ranked the features based on the importance of the class label. The dataset was subjected to various filtering and classification techniques. The study compared the results of 10 different classification methods, including RND TREE, LDA, Naive Bayes, K-NN, ID3, PLS-DA, SVM, C4.5, C-RT, CS and RT Surprisingly, the RND TREE algorithm emerged as the most accurate, with a perfect score of 100%. However, the algorithm was prone to overfitting the data, as the resulting rule-set was quite extensive. An accuracy of approximately 91% was achieved through the use of the C4.5 classification technique, a learning method based on decision tree induction. The healthcare data was analyzed using an advanced version of the ID3 algorithm, developed by Quinlan and known as this technique. The conclusion of the study was that among the ten algorithms tested, the C4.5 algorithm was the most effective for classification and had the highest accuracy.

Bampoe-Addo *et al.* [18] The cutting-edge management system aims to enhance the wellbeing of people living with diabetes by integrating various AI technologies. The system employs a comprehensive approach by dividing the complicated task of diabetes management into smaller and more manageable objectives. These objectives encompass constructing a neural network model using TensorFlow for food categorization, allowing users to determine appropriate meals by submitting an image, utilizing the K-Nearest Neighbor (KNN) algorithm for meal suggestions, creating a chatbot that answers diabetes-related questions using cognitive sciences, monitoring user activity and location, and producing PDF reports of blood sugar readings. A model was trained on images of unique Ghanaian dishes that contain specific nutritional properties crucial for managing diabetes. The result was a highly accurate image categorization system with clearly defined categories and a high level of precision. The food recognition and classification model exceeded all expectations with an outstanding accuracy rate of over 95% in identifying specific calorie levels.

K. Sridar *et al.* [19] In this study a comprehensive system was developed to collect clinical data from the Pima Indians Diabetes Database based on various attributes. The system received real-time inputs from a Glucometer and additional attributes were manually inputted. The diagnosis of diabetes was made using a combination of both Artificial Neural Network (ANN) and Apriori Algorithm (ARM), utilizing all of the collected inputs. The proposed diabetes diagnosis system leverages advanced technologies, including artificial neural networks and association rule mining, to accurately assess a patient's risk level for developing diabetes. Rather than relying solely on manual inputs, the system also integrates real-time data from a Glucometer. Additionally, the management system is designed to be adaptable, providing implementation alternatives in widely used programming languages such as Java and Dotnet, with the ultimate goal of delivering the results through an online platform. The goal of this study is to equip patients with the power of self-assessment for the risk of Diabetes Mellitus using a state-of-the-art integration of Back Propagation and Apriori algorithms. These algorithms were applied to real-time inputs collected from various sources, such as glucometers and manual data inputs, to determine the patient's risk level with an accuracy rate of 91.2%. This innovative approach provides improved reliability and ease of use compared to existing methods, enabling patients to stay ahead of their health without the need for constant doctor visits.

Terry Jacob Matthew *et al.* [20] The aim of this research was to identify the most economical method for forecasting polygenic ailments using non-medical criteria. A comprehensive analysis of supervised learning methods was carried out, with the results indicating a high accuracy rate in disease diagnosis. The results showed that among the algorithms tested, Naive Bayes was the most accurate with a score of 80.37%, followed by REP trees at 78.5%.and Logistic Regression at 77%. The findings demonstrate the potential of these methods for accurately predicting polygenic diseases.

3. METHODOLOGIES

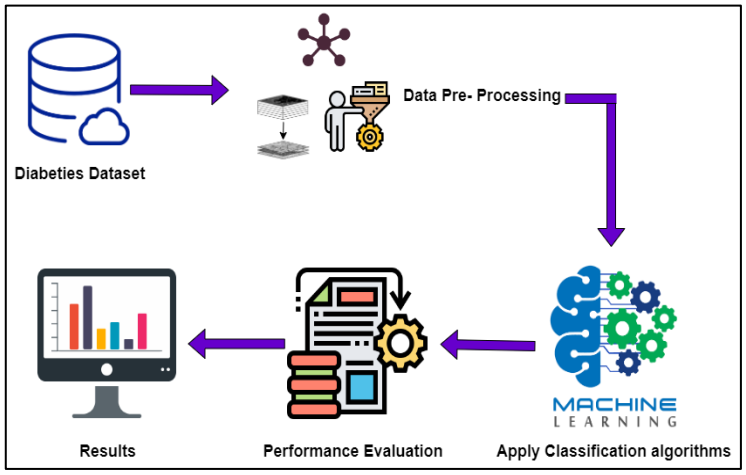

Fig -1: Architecture of Predicting Diabetes

1. Datasets and Attributes:

The study utilized a publicly accessible dataset obtained from the UCI Machine Learning Repository. This dataset is a component of a larger collection owned by the National Institutes of Diabetes and Digestive and Kidney Diseases. Many researchers have employed this dataset for their predictive analytics work. It holds records of 768 female Pima Indian patients, including 9 relevant attributes. In this research, a dataset containing 8 characteristics was employed to forecast the focus of the study was to detect the onset of gestational diabetes in female patients.

By utilizing historical data found in the dataset, including variables such as age, Body Mass Index, blood pressure, and number of pregnancies, the classifiers were instructed to provide a diagnosis of either positive or negative for diabetes. The dataset utilized for the study focuses solely on the demographic of Indian female patients who are at least 21 years of age. The attributes included in the dataset are numerical constant values. The objective of this research is to establish the presence of diabetes in a patient, utilizing a combination of diagnostic evaluations documented for each individual as shown in Table-1.

Table -1: PIMA Dataset Attributes

Attributes	Description
P_Preg	Times of repeated Preg
G_Gluco	Hemoglobin gluco level
B_Pressure	Systolic blod pressure measurement
Thickness_Skin	Measure of thicky nature in skin
I_Insu	Serum insulin contamination
B_bmi	Measurement of body by its height and weight
Diabetes_Pedigree	Thickness of skin located on triceps (measured in millimeters).
A_Age	Age of the patient.
O_Outcome	Outcome in value 1 or 0

2. Data Pre-Processing:

Data Pre-Processing is an important procedure in the data analysis process where raw data is transformed into an organized and structured format to be further analyzed. The goal of preprocessing is to remove errors, inconsistencies, and missing values and make the data suitable for further analysis. The importance of this step cannot be overstated as it can greatly impact the accuracy of the results obtained from the analysis. Data preprocessing is a vital aspect in the data mining process that ensures high quality and dependability of the data. It encompasses cleaning, transforming, and structuring the data to make it appropriate for modeling and analysis. Techniques such as imputation, normalization, and feature scaling are used to correct and remove inaccuracies in the data.

Dealing with missing data and outliers is crucial to obtain accurate results in data analysis. This is especially important in medical datasets, where missing values are more common due to various factors such as incomplete records or data recording errors. Failing to handle these issues can lead to incorrect conclusions and misinterpretation of data. Additionally, selecting the relevant features plays a crucial role in constructing models that are free from correlated variables, biases, and irrelevant noise. Preprocessing of data is a crucial step in artificial intelligence that involves converting raw data into an organized and comprehensible format. Information pre-processing is a key element in the procedure of transforming raw data into a more organized and usable format. It aims to eliminate inconsistencies, inaccuracies, and other problematic features that often occur in real-world data.

The pre-processing stage plays an important role in preparing the data for further analysis and modeling by carrying out activities such as data cleaning, integration, transformation, reduction, and discretization. The data undergoes rigorous inspection in the pre-processing phase was conducted to guarantee error-free and consistent data. This involved verifying for duplicated values, any missing information, and misaligned data types. By cleaning the data in this manner, the classifier is better equipped to uncover hidden patterns and insights within the dataset.

The provision of a well-defined feature vector to the classifier enhances its learning efficiency. Imputation involves filling in missing values with suitable data. In this process, missing entries are replaced with the average of the relevant attributes, ensuring every cell contains a value. Additionally, scaling is performed to standardize the magnitude and units of the dataset's features.
Here are some common data preprocessing steps:

1. Data cleaning: This involves identifying and handling missing or erroneous data, dealing with outliers, and removing irrelevant or redundant data.

2. Data transformation: The transformation of the data into a format that can be easily analyzed by machine learning algorithms. This includes scaling the data to a common range, encoding categorical variables, and creating new features from existing sets.

3. Data reduction: The Size of the data is reduced, without losing important information. This can be achieved through techniques like feature selection and dimensionality reduction.

4. Data integration: This process involves merging data from different resources into a one dataset. This can be grueling when dealing with data that's stored in different formats or has different structures.

5. Data normalization: The process of transforming the data into a standard form so that it can be compared and analyzed. This includes techniques like standardization and normalization.

6. Data discretization: This involves converting continuous data into discrete values. This is often used in machine learning algorithms that require categorical data.

7. Data sampling: The technique useful for selecting a part of the data for data analysis.

Dimensionality reduction is a data preprocessing technique that involves in which features are reduced or variables in a dataset. This is often necessary when dealing with high-dimensional datasets, which can be difficult to analyze and visualize, and can lead to overfitting in machine learning models.

Dimensionality Reduction has two main types:

Feature selection: This includes selecting a subset of the original features that are most relevant to the problem being solved. This can be done using statistical tests, correlation analysis, or other techniques.

Feature extraction: This involves transforming the original features into a low-dimensional space using terminologies like Linear Discriminant Analysis (LDA), t-SNE or Principal Component Analysis (PCA). The new features are usually a combination of the original features that capture the most important information in the data.

Both feature selection and feature extraction can be used to reduce the dimensionality of a dataset, but they have different strengths and weaknesses. Feature selection is faster and easier to interpret, but it may not capture all the important information in the data. Feature extraction can capture more information, but it can be more computationally expensive and harder to interpret.

3. Algorithms:

i. Artificial Neural Networks (ANNs): The concept behind it is inspired by the anatomy and functioning of the human brain. These algorithms are known for their ability to identify patterns in data and make predictions based on these observations. The architecture of an ANN includes interconnected nodes, referred to as artificial neurons, which perform the task of processing information. The connections between neurons are determined by weights, which are numerical values representing the strength of the relationship. To train an ANN, a technique called backpropagation is utilized where the model adjusts its weights based on the error between its predicted outcome and the actual outcome. This process is repeated multiple times until the model can make accurate predictions. Artificial Neural Networks (ANNs) are applied in several domains, including natural language processing, image classification, and speech recognition. It is also used for predictive tasks such as stock market forecasting, weather prediction, and disease diagnosis. However, ANNs have some disadvantages including overfitting, high computational requirements, and low interpretability.

ii. K-Means: Algorithm is a widely recognized unsupervised learning approach used in many fields for grouping and division tasks. The idea behind it is to divide the data into K clusters, each cluster containing similar elements. The procedure for the algorithm is outlined as follows: start with randomly placing K centroids, assign each data point to the nearest centroid, then find the average of all the points in each cluster and relocate the centroids. The algorithm continues until either the centroids stop moving or the specified number of iterations is reached. The aim of K-Means Clustering is to reduce the total sum of squared distances between the data points and the centroids to a minimum., which is known as the objective function. K-Means is a rapid and intuitive technique that can handle large datasets and high dimensional spaces with ease. Despite its ease of use, there are some disadvantages to the algorithm, such as its sensitivity to starting conditions, the requirement of specifying the number of clusters ahead of time, and difficulty in handling non-globular clusters. It is commonly employed in various applications like customer segmentation, image division, abnormality detection and document grouping and also used as a preprocessing step for other machine learning techniques, for instance, reducing dimensions and detecting anomalies.

iii. K-Nearest Neighbors (KNN): The method is a simple machine learning technique that can be utilized for both regression and classification problems. It is based on the idea that the closest data points to a test sample will play the most significant role in determining its class or value. The steps of the algorithm are: 1) Store all the labeled training data, 2) Calculate the distance between the test point and all training points, 3) Choose the K closest training points, where K is a user-defined value, 4) For classification, assign the test point to the most frequent class among the K nearest neighbors and for regression, predict the value of the test point as the mean the K nearest neighbors. Selecting the right value of K is important, as a low value of K can result in the algorithm being too sensitive to noise and outliers while a high value of K can cause the algorithm to not be flexible enough to detect relationships in the data. A commonly used method for choosing the optimal value of K is cross-validation. KNN is simple, fast, and suitable for small datasets and low-dimensional spaces. However, its computational time grows linearly with the size of the training data, making it less scalable than other algorithms. It is commonly used for image recognition, recommendation systems, and anomaly detection.

iv. Naïve Bayes classifier: is a widely used probabilistic classification algorithm that follows Bayes' theorem. As per Bayes' theorem, the likelihood of a class, given a set of features., can be calculated by multiplying the likelihood of the features given the class with the prior probability of the class, divided by the evidence of the features. There are multiple types of Naive Bayes classifiers, such as Gaussian, Multinomial, and Bernoulli. Gaussian is appropriate for continuous features with a normal distribution, Multinomial is suited for features expressed as discrete counts, and Bernoulli is used for features that have binary values. Naive Bayes is highly effective for classification tasks, especially when the number of features is large, and it is particularly useful for text classification like spam filtering and sentiment analysis. The central formula used in Naive Bayes is founded on Bayes' theorem, which can be stated as Equation-1:

$$P(X \mid Y) = \frac{P(Y \mid X) * P(X)}{P(Y)} \qquad \text{Eqn-1}$$

The Naive Bayes classifier can be easily adapted to multiple classes by computing the posterior probability for each class and selecting the class with the highest probability.

v. Random Forest: Random Forest is a highly effective ensemble learning method applied in regression and classification problems. Unlike traditional decision trees, which can easily overfit and produce unreliable results, Random Forest utilizes multiple decision trees to arrive at a more robust and accurate prediction. The algorithm starts by selecting a random sample of the training data and training a

decision tree model on this sample. This process is repeated multiple times, with different samples of the training data being used each time to train a decision tree. The final prediction made by the Random Forest model is the average of the predictions made by all the individual decision trees. Random Forest has several benefits compared to traditional decision tree models. Firstly, it reduces the risk of overfitting, which is a common problem in decision trees. Secondly, it is able to handle large datasets and high dimensional data. Thirdly, it can easily handle both categorical and numerical data types. Fourthly, it is simple to implement and understand, making it a popular choice for many machine learning tasks. Despite its benefits, Random Forest is not without its disadvantages. Firstly, it is computationally intensive, which can be a problem when working with large datasets. Secondly, it can be slow when making predictions, as it needs to average the predictions made by all the individual decision trees. Finally, the results of a Random Forest model can be difficult to interpret, as When the connection between the features and target variable is intricate, Random Forest is a valuable machine learning technique that leverages the advantages of decision trees and the stability of an ensemble model. Ultimately, Random Forest is a formidable algorithm. Although it has its disadvantages, it remains a popular choice for many machine learning tasks due to its simplicity, ability to handle complex data, and overall accuracy and reliability.

vi. Support Vector Machines (SVMs): are a multi-purpose machine learning method that can be applied to both regression and classification problems. They work by finding the optimal decision boundary between data points, which separates them into different classes. The boundary is determined based on the support vectors, which are the data points closest to the boundary. To determine the optimal boundary, SVMs transform the input data into a higher dimensional space where a boundary can be drawn. The boundary is optimized by maximizing the margin, which is the distance between the boundary and the closest data points of each class. SVMs leverage optimization techniques such as gradient descent and quadratic programming to determine the optimal parameters for the decision boundary. This boundary, established using these techniques, is resilient to outliers and noise in the data. SVMs are widely used in image classification, text classification, and bioinformatics. They perform well with high dimensional data, making them a useful tool for datasets with many features. However, SVMs have some limitations that should be considered. SVMs make use of optimization methods like gradient descent and quadratic programming to discover the best parameters for the decision boundary. This boundary, established through these methods, is able to withstand anomalies and disruptive elements in the data, especially for large datasets. Finally, they may not perform well with non-linearly separable data, and in such cases, In contrast, there may be other algorithms that can perform better in certain situations, such as decision trees or neural networks.

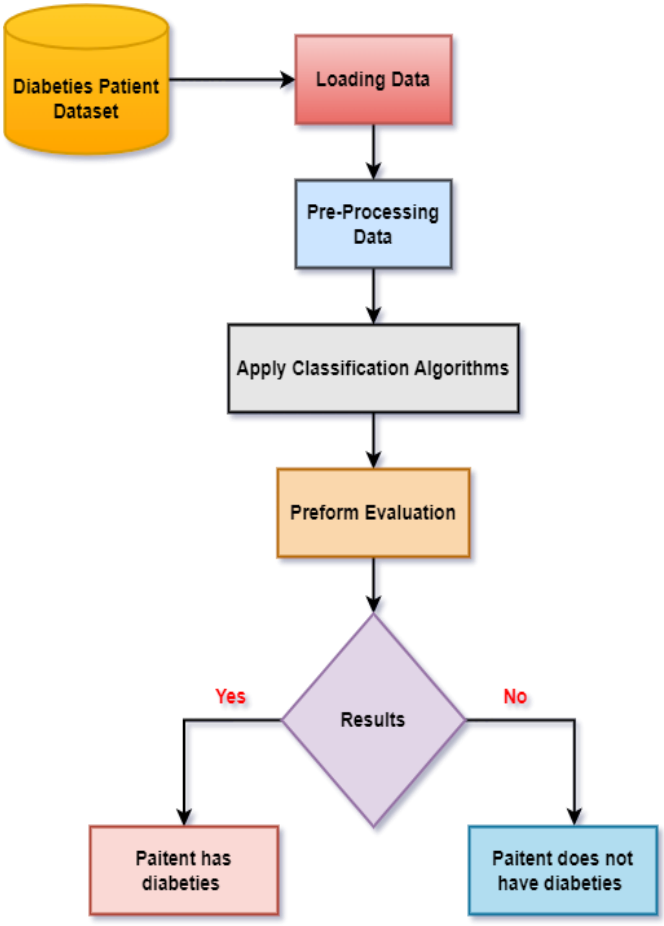

Fig -2: Flow Diagram

4. Performance Evaluation:

1. Evaluation Metrics: In this part of the study, we outline all the metrics utilized in our evaluations. Precision is defined as the proportion of accurate positive predictions compared to the combination of true positive and incorrect positive predictions is referred to as the positive predictive value as shown in Equation-2. This is in accordance with the definition presented in reference [5].

$$Precision = \frac{TP}{TP+FP} \quad \text{Eqn-2}$$

Additionally, recall is calculated using the formula that includes FN as shown in Equation-3.

$$Recall = \frac{TP}{TP+FN} \quad \text{Eqn-3}$$

The F-measure is a composite metric that balances precision and recall to give a more holistic view of the performance of a machine learning model, and is calculated using the following formula given in Equation-4.

$$F - measure = \frac{2*Recall*Precision}{Precision+Recall} \quad \text{Eqn-4}$$

Additionally, the effectiveness of the machine learning techniques can be evaluated using various performance metrics. One of the most commonly used metrics is accuracy, as shown in Equation-5, which is calculated as the fraction of

correctly predicted instances among all instances. The accuracy metric gives us a quick and simple way to understand the overall accuracy of the model.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{Eqn-5}$$

2. Confusion Matrix: The performance of the machine learning classification model was evaluated using a Confusion Matrix, which compares the actual target values to the predictions generated by the model. This matrix provides a comprehensive evaluation of the accuracy of the model, and is depicted in Figure-3. The matrix displays the number of instances in the test data that were correctly and incorrectly classified. The Confusion Matrix helps in analyzing the effectiveness of the machine learning model in categorizing individuals as either having early-stage diabetes or not having it, by comparing the actual results with the predictions generated by the model. The number of classes in the matrix is determined by the number of target classes in the dataset, and the values in the matrix highlight the model's ability to differentiate between these classes.

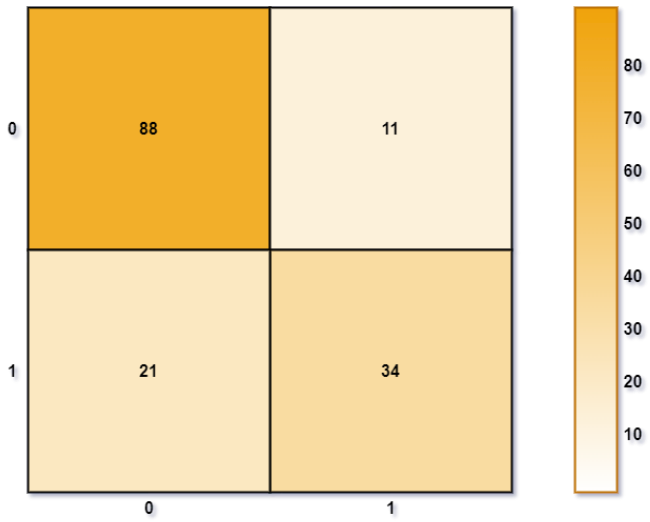

Fig-3: Confusion Matrix

A standard confusion matrix is composed of four sections that symbolize the following:

TP: These is Ture Positive and are instances that were correctly classified as positive.

FP: These are False Positive and are instances that were incorrectly classified as positive, but are actually negative.

TN: These are Ture Negative and are instances that were correctly classified as negative.

FN: These are False Negative and are instances that were incorrectly classified as negative, but are actually positive.

By using these values, different evaluation metrics is calculated, which includes accuracy, precision, recall, and F1-score. These metrics give a clear understanding of the model's performance, its success rate in making accurate predictions, and the kinds of mistakes it is committing.

5. Results:

In this research project, a total of six machine learning algorithms were evaluated. These algorithms included NB, KNN, SVM, K-Means Clustering, ANN, and RF. The experiment was conducted using the PIMA Indian dataset, O the other hand, the data could be split into a larger training set that consists of 70% of the data, and a smaller testing set made up of 30% of the data. The evaluation of the algorithms was based on their accuracy in predicting outcomes and was performed using Enthought Canopy software. The results obtained from the analysis were then used to determine the best-performing algorithm.

Outcome To evaluate the performance of various machine learning methods, we computed the prediction results based on precision, recall, and f-measure. delicacy of algorithms was measured and presented in Figure 4. ANN gives 72 delicacies, SVM gives 79 delicacy, 77 delicacy was achieved by using K-Means and KNN and RF achieved 80 delicacy and Naïve Bayes achieved 82 delicacies. So Naïve Bayes (NB) achieved loftiest delicacy which is 82. From the experimental results attained, it can be concluded that the Naïve Bayes algorithm is applicable for prognosticating the diabetes status of cases.

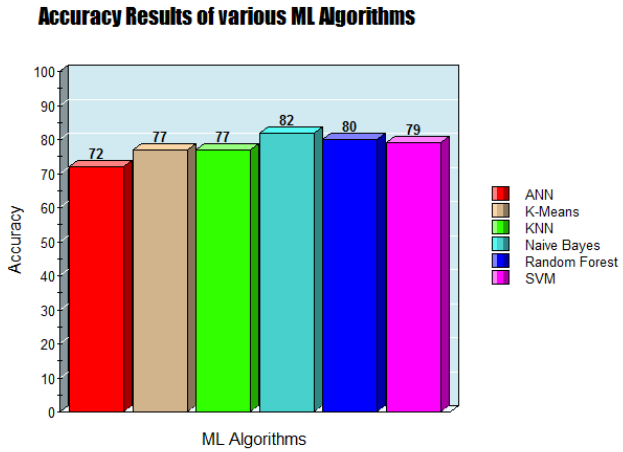

Fig-4: Accuracy of Algorithms

4. CONCLUSION

In our review, we've taken colorful factors related to complaint also we've anatomized the diabetes vaticination using machine literacy (ML) languages. Gleaning perceptivity from factual healthcare data holds the implicit to anticipate the presence of diabetes in individualities. In order to make accurate prognostications about diabetes mellitus, our disquisition employed various machine knowledge ways analogous as Naive Bayes(NB), K- Nearest Neighbor(KNN), Support Vector Machine(SVM), K- Means, Artificial Neural Network(ANN), and Random Forest(RF) on a large- scale dataset of adult populations. The perceptivity gained from these trials have the eventuality to save innumerous lives by enabling healthcare professionals to take early precautionary measures and make informed clinical opinions. still, there's still important work to be done to optimize these models for indeed lesser delicacy. In the future, we plan to continue our

exploration by incorporating fresh styles to fine- tune the parameters of our models, and by testing these models on larger datasets with minimum missing trait values to gain indeed deeper perceptivity and bettered vaticination delicacy.

REFERENCES

1. Shahadat Uddin, Arif Khan, Md Ekramul Hossain and Mohammad Ali Moni "Comparing different supervised machine learning algorithms for disease prediction"
2. Shejal Kale, Priti Rahane, Mansi Ghumare, Snehal Patil "Diabetes Prediction Using Different Machine Learning Approaches"
3. Deepti Sisodiaa, Dilip Singh Sisodiab "Prediction of Diabetes using Classification Algorithms"
4. K. VijiyaKumar, S. Sofia Caroline, I. Nirmala and B. Lavanya "Random Forest Algorithm for the Prediction of Diabetes"
5. Smriti Gupta, Harsh Kumar Verma, and Divyansh Bhardwaj "Classification of Diabetes Using Naïve Bayes and Support Vector Machine as a Technique"
6. Dr. Zuber khan, shaifali singh and Krati Sexena "Diagnosis of Diabetes Mellitus using K- Nearest Neighbor Algorithm"
7. Mercaldo, F.; Nardone, V Santone, A "Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques" Procedia Comput. Sci. 2017, 112, 2519–2528.
8. Veena Vijayan V, Anjali C "Prediction and Diagnosis of Diabetes Mellitus – A Machine Learning Approach. IEEE 2015:6."
9. Kumari Mukesh, Vohra Rajan, Arora Anshul. Prediction of Diabetes Using Bayesian Network. Int. J. Comput. Sci. Inf. Technol. 2014; 5:5
10. K. Rajesh and V. Sangeetha" Application of Data Mining Methods and Techniques for Diabetes Diagnosis," in proceedings of international journal of Engineering and Innovative Technology, vol.2, Issue 3, September 2012
11. Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, Munam Ali Shah "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare"
12. Md. Faisal Faruque, Asaduzzaman and Iqbal H. Sarker "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus"
13. Roshan Birjais, Ashish Kumar Mourya1, Ritu Chauhan2 and Harleen Kaur1 "Prediction and diagnosis of future diabetes risk: a machine learning approach."
14. Arvind Aada, Prof. Sakshi Tiwari "Predicting Diabetes in Medical Datasets Using Machine Learning Techniques"
15. Md. Maniruzzaman, Md. Jahanur Rahman, Benojir Ahammed and Md. Menhazul Abedin "Classification and prediction of diabetes disease using machine learning paradigm
16. O.S. Soliman, E. AboElhamd "Classification of Diabetes Mellitus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine, 2014 arXiv:1405.0549."
17. K. Rajesh, V. Sangeetha "Application of Data Mining Methods and Techniques for Diabetes Diagnosis."
18. Robert A. Sowah, Adelaide A, Bampoe-Addo, Stephen K. Armoo, Firibu K. Saalia, Francis Gatsi and Francis Gatsi "Design and Development of Diabetes Management System Using Machine Learning
19. K. Sridar, Dr. Shanthi "Medical diagnosis system for the diabetes mellitus by using back propagation-Apriori algorithms
20. Terry Jacob Mathew, Elizabeth Sherly "Analysis Supervised Learning Techniques for Cost Effective Disease Prediction using Non-Clinical Parameters" July 05-07,2018.
21. Toshita Sharma and Mannan Shah "A Comprensive review of machine learning techniques on diabetes detection" July 05,07,2018.
22. Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC,978-1-5090-3243-3,2017.
23. Ayush Anand and Divya Shakti," Prediction of Diabetes Based on Personal Lifestyle Indicators", 1st International Conference on Next Generation Computing Technologies, 978-1-4673-6809-4, September 2015.
24. B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 978 1-5386-2745-7,2017.
25. Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing,2015.
26. Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly," Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.
27. P. Suresh Kumar and S. Pranavi "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics",International Conference on Infocom Technologies and Unmanned Systems, 978-1-5386-0514-1, Dec. 18-20, 2017.
28. Mani Butwall and Shraddha Kumar," A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier", International Journal of Computer Applications, Volume 120 - Number 8,2015.
29. K. Rajesh and V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.
30. Humar Kahramanli and Novruz Allahverdi,"Design of a Hybrid System for the Diabetes and Heart Disease", Expert Systems wit Applications: An International Journal, Volume 35 Issue 1-2, July, 2008.
31. B.M. Patil, R.C. Joshi and Durga Toshniwal,"Association Rule for Classification of Type-2 Diabetic Patients", ICMLC '10 Proceedings of the 2010 Second International Conference on Machine Learning and Computing, February 09 - 11, 2010.
32. Dost Muhammad Khan1, Nawaz Mohamudally2, "An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm ", Journal Of Computing, Volume 3, Issue 12, December 2011