Das Gupta Girish, et. al. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 12, Issue 10, October 2022, pp. 73-80

RESEARCH ARTICLE OPEN ACCESS

Diabetes Prediction Using Machine LearningTechniques.

Das Gupta Girish[1], Srivastava Pratima Kumari[2,]Dr.B.B.V. Sailaja[3]
Department of Computer Engineering, Amity University Lucknow Campus India[1]
Department of Zoology,Ch.Sd.St.Theresas College for Women,Eluru,India[2]
Department of Chemistry, AndhraUniversity, Visakhapatnam, India[3]

Abstract: This quantitative research was carried out to demonstrate, "Diabetes Prediction Using Machine Learning Techniques.". Diabetes is a medical condition that emerges once the body's natural glucose levels become excessively high. Diabetes should not be disregarded; if left untreated, it can result in serious complications for an individual, causing deterioration to the eyes, heart rate, kidneys, heart, and other organ systems. If diabetes is addressed early, it is possible to be cured. By employing a range of machine learning approaches and algorithms, we will be doing early diabetes forecasting in a human body or patient for a higher degree of accuracy.By building models using patient records, machine learning approaches offer superior results for prediction.In this research, we will Machine Learning Techniques on a dataset to anticipate diabetes. As compared to similar models, each model's or technique's aka algorithm's accuracy can vary. This research finding suggests that the model can accurately predict diabetes nearby accurately. The results suggest that Logistic Regression algorithm outperformed competing algorithms of Machine Learning in terms of accuracy.

Keywords: Machine Learning, Diabetes, Logistic regression algorithm, Random Forest classifier algorithm, K-Nearest Neighbour(K-NN) algorithm, Support Vector Machine.

Date of Submission: 01-10-2022 Date of Acceptance: 11-10-2022

I. INTRODUCTION

Thegreatest threat to healthy human body could be Diabetes which is not considered as a serious condition util it worsens day by day.Diabetes is a widespread debilitating disease that gravely jeopardises health of the public. Whenever the body fails to produce adequate insulin, diabetes develops. Diabetes is one of the worst human body conditions there seems to be. Obesity, an elevated blood glucose level, and other factors can cause diabetes. It disrupts the activity of the insulin deficiency, which enables individual's body to exhibit an irregular metabolism and raises blood sugar levels. The hallmark of diabetes is elevated blood glucose levels, which are spurred on by ineffective insulin production or its attenuated biological effects.Type 1 diabetes (T1D) and type 2 diabetes are the two subtypes of diabetes. The average age of type 1 diabetes patients is under 30 years old. High blood sugar levels, extreme thirst, and recurrent urination are the commonly diagnosed signs. Patients with this kind of diabetes necessitate insulin treatment because oral pills alone might not be able to adequately treat it. Middle-aged and older individuals are more inclined to develop type 2 diabetes, which would be typically associated with the emergence of overweight, high blood pressure, atherosclerosis, coronary artery disease, and some other maladies.The smoother it is would be to control, if earlier the diagnosis is reached. As the daily physical examination data goes, people may use machine learning to produce a preliminary diagnosis of diabetes mellitus, and it can be used as a reference by doctors as well for further diagnosis.

II. LITERATURE SURVEY

S. Selvakumar et al. [1] addressed concerning diabetes problems. Data mining techniques are utilised to forecast whether a person will get diabetes or not. The algorithms K-Nearest Neighbor, Multilayer Perception, and Binary Logistic Regression are categorised. Multilayer Perception, K-Nearest Neighbor, and Binary Logistic Regression all have accuracy levels of 0.69, 0.71, and 0.80 respectively. The accuracy of K-Nearest Neighbor is higher than that of Binary Logistic Regression and Multilayer Perception.The goal of this work, as stated by B. Tamilvanan et al. in [2], is to forecast diabetes more accurately. The accuracy rates of the three classification algorithms—Naive Bayes, Random Forest, and NB-Tree—are compared. application of the Weka tool. As a consequence, Naive Bayes has the highest accuracy rate (76.3%) and lowest error rate (23.7%), making it the top predictive model.According to Rahul Joshi et al. in [3], employing machine

learning approaches to forecast medical datasets at an early stage is safe for human life. to evaluate the diabetes dataset for Pima Indians. The algorithms that are being used are KNN, Naive Bayes, Random Forest, and J48. When we combine several methodologies and approaches, the ensemble approach yields the best results. Additionally known as a hybrid model. Compared to only one, this offers the best performance and accuracy.In his research, Yunsheng[4] developed a novel method for employing the KNN algorithm that involved deleting outliers/OOBs (out of bag) using DISKR (reduce the size of the training set for K-nearest neighbour). Additionally, the storage space was kept to a minimum in this investigation. As a result, the space complexity is reduced and more effective when some parameters or situations were eliminated. This improved the researchers' accuracy.Monika and Pooja [5] have talked about the most contemporary advancements in medical science research as well as historical data retrieval methods. Additionally, we have clarified the language and learning methods used in data mining and machine learning.S M Hasan Mahmud et al.in[6] prediction 's of diabetes has found the five most significant machine learning classification algorithms for predicting diabetes were examined in this article. Methods of 10-fold cross validation were used to see how well the categorization approaches performed. The analysis's findings demonstrate that Naive Bayes, which obtained an F1 value of 0.74, outperformed the other classifiers in terms of performance.This work by Ayan Mir et al. [7] concentrated on diabetes prediction. Diabetes databases for Pima Indians are utilised. The Weka interface uses the Naive Bayes, SVM, Random Forest, and Simple CART algorithms for categorization. As a consequence, SVM offers better accuracy than the competition.In this work, DeepikaVermaet al. [11] use two illness datasets. This dataset from the UCI machine learning repository includes data on breast cancer and diabetes. WEKA, a useful classification tool, was used in this study.

III. PROPOSED ALGORITHM

1.LogisticRegression:

Why use Logistic Regression for this study? When utilising logistic regression algorithm, it might be advantageous to forecast the likelihood of an occurrence. It helps with probability calculations between any two classes. Logistic regression, in short, predicts whether or not a person would be diagnosed with diabetes based on dataset provided.The idea of machine learning existed before analyst DR Cox introduced logistic regression in 1958. It is a method of machine learning that is used in characterization tasks (to predict an output after ingesting training data from the dataset). Logistic regression employs a condition, much like linear regression does, but the result is unaltered even while the condition serves as a driver for other regression models.

There are two outcomes that are expected to result from the autonomous components. The overall work process is as follows:(1) Adding data from the dataset is step one. (2) Dividing the data into variables to train on and test against.(3) By using the classifier to anticipate the previous sentence, the additional stages of what to setup during the procedure were previously supplied. is a very strong texture classification algorithm.

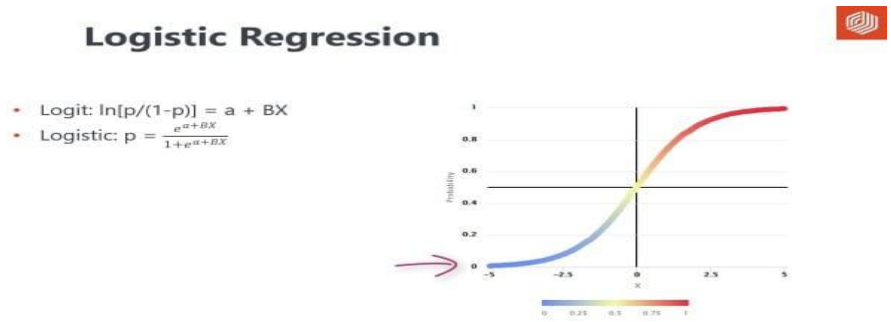

Fig-1:LogisticRegression algo predictive analysis.

2.Gaussian Naïve bayes Classifier:

A statistical classifying technique named Gaussian Naive Bayes is based on the implementation of the Bayes theorem alongside stringent exogenous inputs.The method combines elements into its model that are independent of one another, thus the term "Nave." Any changes to the value of one parameter of the algorithm have no direct effect on the classifiers assume that the value. The Naive Bayes algorithm's primary aspect is that it provides an extremely simple yet useful tactic.

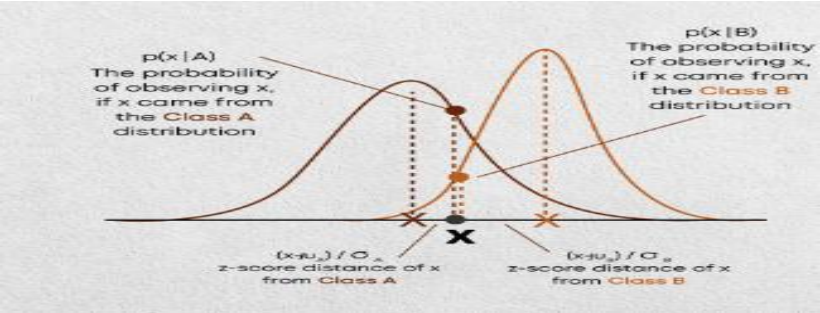

Fig-2:Gaussian Naïve bayes Classifier algorithm boosting analysis.

3.Support Vector Machine (SVC):

Why use SVM for this study? A sort of machine learning technique that may be used to evaluate and categorise data is called an SVM classifier, or support vector machine classifier. An approach for supervised machine learning called a support vector machine may be utilised for both classification and regression problems. In order to discover the hyperplane that optimises the separation between the two classes, the support vector machine classification explores seeking data.If input can be split linearly, the high degree of precision afforded by SVM would be available (Hard Margin). Loosening the margin is all that is needed to account for predictive performance when data cannot be differentiated linearly (Soft Margin).

Focusing on depiction and deriving rules from data is done using a directed machine learning algorithm called support vector machines (SVM). In design affirmation issues, it excels brilliantly. It is advisable to apply this computation when there are several components and visuals. In an SVM model, where n is the number of factors, each feature is viewed as the value of an element in the n-dimensional space, and each piece of input is represented as a centre in the space.

This is how it goes:

(1) The first lines or constraints that are checked are those that correctly request the preparation dataset.

(2) Then, it chooses the specific point where, amongst lines or cut-off points from the nearest point or element, has the best partition.

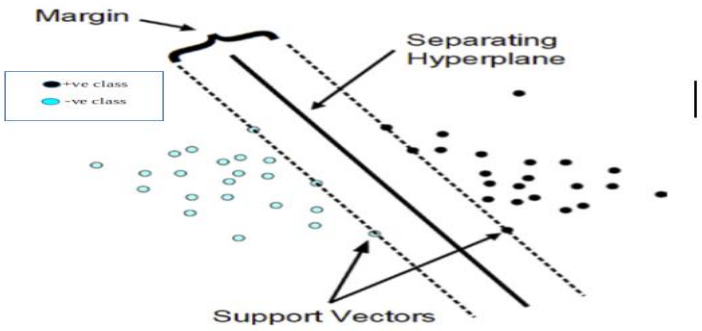

Fig-3:Support Vector machinealgo predictive analysis.

4.Random Forest Classifier:

Why use Random Forest Classifier for this study?The aforementioned classification algorithm helps determine whether or not an individual is vulnerable of Diabetes by splitting the information into several trees. Additionally, it lowers variance, improving accuracy for exact diagnosis of Diabetes prediction throughout research. Furthermore, it lessens the problem of overfitting and unbalanced datasets in decision trees.

As its name implies, the random forest algorithm is a directed ML method that combines a number of single-choice braids. Every single tree in the Random Woods emits a class hypothesis, and our model's estimate or forecast is based on the class that receives the most votes.The application of this technique to the two generated datasets has also resulted in the projection of the accuracy score. The class that receives the most votes becomes our model's forecast or prediction.

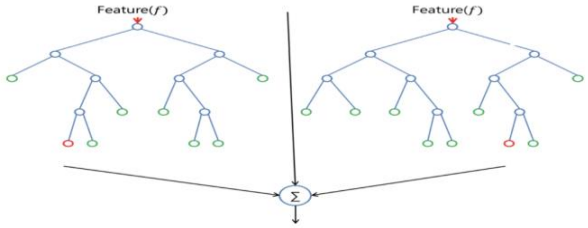

Fig-4:Random Forest Classifieralgo predictive analysis.

5.Decision Tree Classifier:

Why use Decision Tree Classifier for this study?Decision Tree Classifiers are typically the methodology of choice for predictive modelling since they are extremely easy to learn and massively successful. Numerous Tree-based models may indeed be taught all through procedure since the tissue in this study is segregated into categories. It is simple to use and has a strong resemblance to a flow chart used in human-like strategic decision-making processes.

A decision tree is a supervised machine learning algorithm that may be used to categorise or predicts the outcome variable based on the responses to prior comments. Since this model is supervised learning in composition, it is developed and evaluated using data sets that incorporate the requisite categorization. The decision tree pathways may also be depicted using these "if-then" rules. A decision tree model is created using a hierarchy of branches. Each path from the root node through internal nodes to the leaf node reflects a grading determination rule. The decision tree classifier creates the classification model by building a decision tree (Pang-Ning et al., 2006). Each branch that descends from a tree node offers a test.

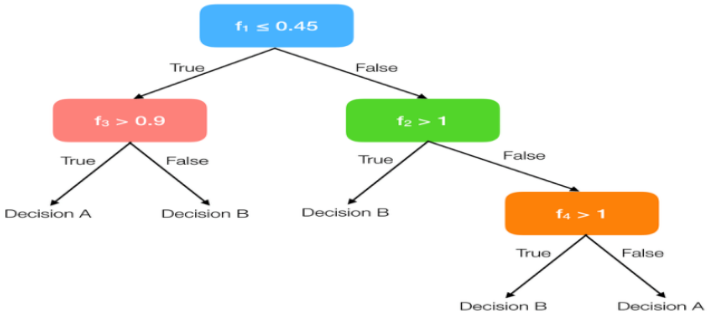

Fig-5:Decision Tree Classifier algo predictive analysis.

IV. EXPERIMENTAL SETUP AND RESULTS

Machine learning, which encompasses the concept of generating predictions about additional info dependent on enormous databases of prior data, is the project's driving principle. Making judgments based on understanding of a particular issue or phenomena depends on building models from impressions known as training data. The idea is illustrated by a flowchart, where the training component is portrayed by the characterized by strong of the channel, and the assessment or evaluation section by the flat portion of the stream. Out of the several models it comprises, the model of supervised learning is the subject of this study. The goal of monitoring outputs during the information preparation process, per the supervised learning, is to identify accuracy.Unsupervised learning, on the contrary, does not establish a correlation between the input and the result. The major goal of supervised machine learning is to uncover the underlying organisation or flow of data in order to reliably compare input and output. In this project, I'll be employing both supervised and unsupervised machine pedagogical strategies to predict the results and analyse the accuracy of each algorithm. and the ability or model that appropriately links all inputs and outputs provided into the device which can be then used.The key objective of supervised machine learning is to reveal the fundamental organisation or flow of data in order to accurately compare input and output. In this project, I'll be utilising both supervised and unsupervised machine learning approaches to predict the results and assess the accuracy of each algorithm.

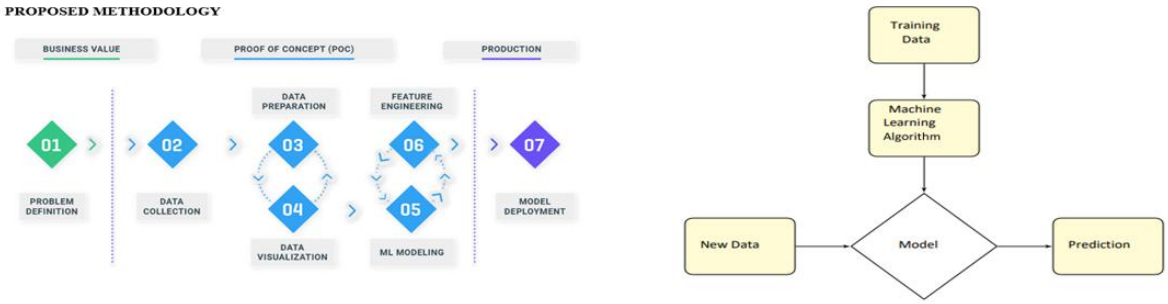

Fig.6(a),6(b): Clearly laid out Machine Learning prediction and analysis procedure.

The following are the four primary supervised machine learning methods I'll be using for the dataset:

- Logistic regression
- Support Vector Machine (SVM)
- Random Forest Classifier.
- Gaussian NB algorithm
- Decision Tree Classifier.

To obtain the anticipated result and make the intended forecast, the following actions must be taken:

Using training and test variables, data is obtained, filtered, split into input and output, and tested. The data will then be normalised (scaled), a classifier or regressor will be performed, the model will be equipped, its output will be predicted, and the accuracy score will be computed eventually through the end. The overall process would involve taking the data, filtering it, and then dividing it into input and output. The central feature of this accuracy prediction would instead involve training and testing the test parameters, normalising the data, and then running a classifier. Ultimately, the model would indeed be fitted, and the accuracy score would be determined.

Fig-7: Machine Learning process involved in accuracy prediction.

Result for Diabetes Detection:

Algorithm used **Random Forest Classifier:**

Random Forest Classifier

```
In [55]:
1 from sklearn.ensemble import RandomForestClassifier
2 ranfor = RandomForestClassifier()
3 ranfor.fit(X_train, Y_train)
4 Y_pred_ranfor = ranfor.predict(X_test)
5 from sklearn.metrics import accuracy_score
6 accuracy_ranfor = accuracy_score(Y_test, Y_pred_ranfor)
7 print("Random Forest: " + str(accuracy_ranfor * 100))

Random Forest: 77.27272727272727
```

Fig.8: Random Forest Classifier matrix with accuracy score of 77.27%

Algorithm used: **Support Vector Machine:**

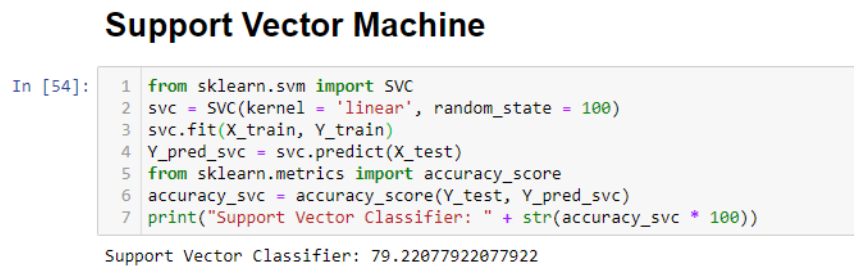

Fig.9: Support Vector Machine matrix with accuracy score of 79.22%

Algorithm used: **Decision Tree Classifier:**

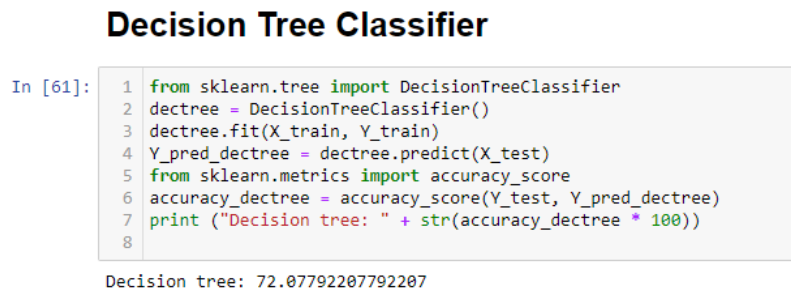

Fig.10: Decision Tree Classifier confusion matrix with accuracy score of 72.07%

Algorithm used:**Gaussian Naïve bayes Classifier:**

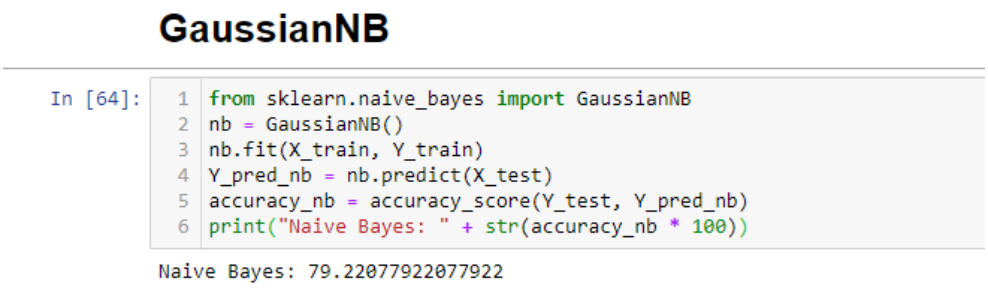

Fig.11:Gaussian Naïve bayes classifier confusionmatrix with accuracy score of 79%

As the aforementioned algorithms have already been applied to the input data, let's try incorporating some additional algorithms and working with them to create a standard, usable algorithm that could truly cross the levels and assist us in obtaining the desired accuracy for the crime prediction of a given set of data, as I did.

Algorithm used: **Logistic Regression:**

Logistic Regression

```
In [39]:
1 from sklearn.linear_model import LogisticRegression
2 logreg = LogisticRegression(random_state = 42)
3 logreg.fit(X_train, Y_train)
4 Y_pred_logreg=logreg.predict(X_test)
5 from sklearn.metrics import accuracy_score
6 accuracy_logreg = accuracy_score(Y_test, Y_pred_logreg)
7 print("Logistic Regression: " + str(accuracy_logreg * 100))

Logistic Regression: 79.87012987012987
```

Fig.12: Logistic Regression confusion matrix with accuracy score of 79.87%.

Result for Diabetes Prediction and Accuracy Score:

We can draw a conclusion by evaluating the precision score's accuracy using the Matrix network found in the measurements module from SK Learn, working with the Diabetes dataset in this project, and using machine learning calculations, specifically Logistic Regression, Random Forest Classifier, Support Vector Machine, Decision Tree Classifier, and Gaussian Nave Bayes. On referencing the percentage bar chart below a clear picture of the accuracy score can be withdrawn.

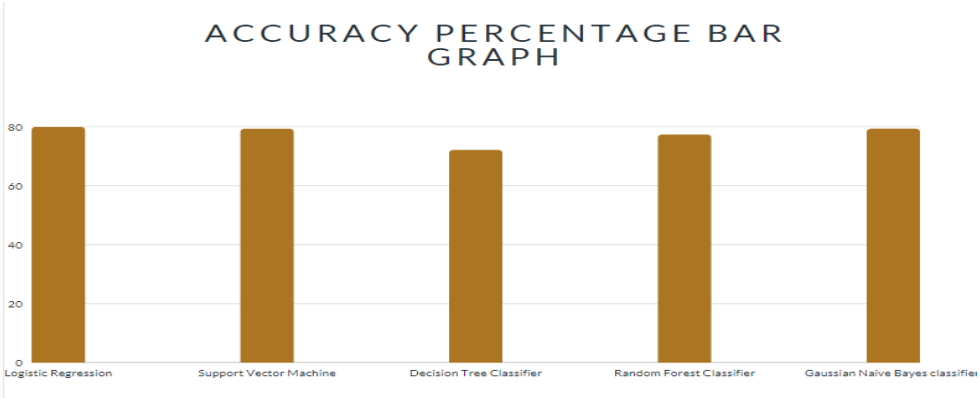

Fig13: Accuracy percentage of algorithms deployed.

Per the research, the preferred approach or algorithm for said prognosis of Diabetes is logistic regression classifier algorithm because of the presented input dataset and its high proficiency of accuracy in terms of metrics.

V. RESULT AND DISCUSSION

This project's primary goal was to build and implement techniques for predicting diabetes using machine learning, and to assess the effectiveness of such approaches. By developing a methodology that might be deployed as an aid in diagnosing the patients who have the greatest risk of developing diabetes, this research provides a minor contribution to the previously managerial aspects of diabetes diagnosis. This research accomplishes this by the review and meta - analysis of patients' health records and the examination of several significant considerations i.e attributes, including the patient's blood glucose level, body mass index, etc. Depending on the pertinent medical information that is gathered from a dataset or the data that would be supplied by the user, the concept anticipates the advent of diabetes in a patient.The user's complete set of essential patient information is provided into the online Web application, and this input is then communicated to the trained model so it can evaluate whether the individual possesses diabetics or otherwise. The suggested strategy makes use of a variety of classification and ensemble learning techniques, including classifiers from SVM, Random Forest, Decision Tree, Logistic Regression, and Naïve Bayes. The experimental results can help medical professionals make early predictions and decisions to treat diabetes and save a patient's life.It has become easy to uncover links and examples among diverse pieces of information thanks to machine learning innovation, which is essential to successfully complete this job. Employing the idea of machine learning, we have created a model using prepared informative index that have undergone information cleansing and modification. In addition to what has already been stated, the beneficiary will use supervised adaptation if there is a smaller

amount of information that is clean and named identically; however, if there is no named information that is surprisingly present in any way, the beneficiary will use unsupervised learning.NLP (Natural Language Processing), which is essentially MACHINE LEARNING (Artificial intelligence) and is on the brink of total of evolving into the generation of a considerable part of advanced technologies, is another technique for machine learning that is deeply an integral component in furthermore to the aforementioned. Deep learning, reinforcement learning, and neural networks are other examples. Augmented intelligence frameworks are now pioneering new anti-toxins that are intended to cure specific ailments as they are given datasets to assess and learn from. It is extremely likely that in the future it will be feasible to include AI into medical services much more, allowing it to evaluate patients and treat them for the particular diseases.

REFERENCES

[1]. Rahul Joshi and Minyechil Alehegn, "Analysis and prediction of diabetes diseases using machine learning algorithm": Ensemble approach, International Research Journal of Engineering and Technology Volume: 04 Issue: 10 | Oct - 2017.

[2]. VeenaVijayan V. And Anjali C, Prediction and Diagnosis of Diabetes Mellitus, "A Machine Learning Approach" ,2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)| 10- 12 December 2015 | Trivandrum.

[3]. Santhanam, T. and Padmavathi, M.S., 2015. Application of K-means and genetic algorithms for dimension reduction by integrating SVM for a diabetes diagnosis .Procedia Computer Science, 47, pp.76-83.(2015).

[4]. Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: an Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.

[5]. 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.

[6]. 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.

[7]. Sisodia, D. and Sisodia, DS, 2018. Prediction of diabetes using classification algorithms .Procedia computer science, 132, pp.1578-1585.(2018) .

[8]. 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.

[9]. S.Selvakumar, K.Senthamarai Kannan and S.GothaiNachiyar.2017. Prediction of Diabetes Diagnosis Using Classification Based Data Mining Techniques from International Journal of Statistics and Systems, ISSN 0973-2675 ,pp. 183-188.

[10]. Y. A. Christobel and C. Sivaprakasam.2013.New Classwise K Nearest Neighbor (Cknn) Method For The Classification Of Diabetes Dataset, Int. J. Eng. Adv. Technol., vol. 2, pp. 396–400.