Computers in Biology and Medicine

Analysis and Future Risk Prediction of Diabetes Mellitus

--Manuscript Draft--

Manuscript Number:	
Article Type:	Full Length Article
Keywords:	SVM; Logistic Regression; Perceptron; Decision Tree; k-Nearest Neighbor; Multilayer Perceptron; Naïve Bayes; Random Forest
Corresponding Author:	Praffulla Kumar Dubey INDIA
First Author:	Praffulla Kumar Dubey
Order of Authors:	Praffulla Kumar Dubey
	Udbhav Naryani
	Medhavi Malik
Abstract:	The number of patients suffering from diabetes in India are increasing rapidly. One of the major reasons behind this can be that most of the population does not take regular check-ups for diabetes. Hence, they are unable to take proper precautions and have a risk of getting diabetes in the future. The current methodologies followed by the doctors are manual and can be time consuming. The objective of this paper is to predict is the diagnosed patient is diabetic or the diagnosed patient is not diabetic by applying various deep and machine learning algorithms like Perceptron, Logistic regression, Multilayer Perceptron, k-Nearest Neighbor, Naive Bayes, Decision Tree, Random Forest, and Support Vector Machines on Pima Indian diabetes dataset and improve the accuracy of model.
Suggested Reviewers:	
Opposed Reviewers:	

Dear Editor,

We wish to communicate the research paper entitled "**Analysis and Future Risk Prediction of Diabetes Mellitus**". With the submission, authors confirm no conflict of interest of internet and also confirm the manuscript has not been published. '**Declaration of interest: none**'.

Thank you for your time and consideration.

Sincerely,

Praffulla Kumar Dubey

Dear Editor,

We wish to communicate the research paper entitled "**Analysis and Future Risk Prediction of Diabetes Mellitus**". With the submission, authors confirm no conflict of interest of internet and also confirm the manuscript has not been published. Thank you for your time and consideration.

Any further correspondence can be addressed to praffulladubey@gmail.com

Sincerely,

Praffulla Kumar Dubey

Analysis and Future Risk Prediction of Diabetes Mellitus

Praffulla Kumar Dubey
Computer Science and Engineering
SRM Institute of Science and Technology, India
praffullakrdubey@gmail.com

Udbhav Naryani
Computer Science and Engineering
SRM Institute of Science and Technology, India
udbhav.naryani@gmail.com

Medhavi Malik
Computer Science and Engineering
SRM Institute of Science and Technology, India
medhavimalik28@gmail.com

ABSTRACT

The number of patients suffering from diabetes in India are increasing rapidly. One of the major reasons behind this can be that most of the population does not take regular check-ups for diabetes. Hence, they are unable to take proper precautions and have a risk of getting diabetes in the future. The current methodologies followed by the doctors are manual and can be time consuming. The objective of this paper is to predict is the diagnosed patient is diabetic or the diagnosed patient is not diabetic by applying various deep and machine learning algorithms like Perceptron, Logistic regression, Multilayer Perceptron, k-Nearest Neighbor, Naive Bayes, Decision Tree, Random Forest, and Support Vector Machines on Pima Indian diabetes dataset and improve the accuracy of model.

KEYWORDS: SVM, Logistic Regression, Perceptron, Decision Tree, k-Nearest Neighbor, Multilayer Perceptron, Naïve Bayes, Random Forest.

I. INTRODUCTION

Presently in the world, there are various chronic diseases that are distributed throughout the developed as well as developing countries. Diabetes mellitus, which is these days called as diabetes is such one of these chronic diseases and it is responsible for cutting human life at an early age [1]. It is a group of metabolic disease which marks high blood sugar levels over a prolonged period [2]. It has been observed that diabetes disease is on a rapid increase in Asian country like India. Various health sectors are working to predict these chronic diseases in future hence saving the human life. Exercise plays an important role in recovering diabetes and if a diabetic person eats healthy food and does exercise regularly then there are chances of recovery for them. Various human body parts like eye, heart, nerves and kidney can be damaged by diabetes disease [12]. Currently, in order to diagnose diabetes, the doctors take patients sample of blood and measure the sugar concentration that is present in their blood. This methodology followed by the doctors is time consuming and some other features are like blood pressure levels, insulin levels, age of the patient and body mass index (BMI). There are chances for a person to be diabetic in future if an ancestor of the patient shows a presence of diabetes [2]. Diabetes of two types, Type 1 and Type 2 diabetes. At present there are no intrusive techniques to predict whether the patient has type one diabetes. Type two diabetes does not depends on the insulin levels and it has been observed that it is more common in Pima Indians than it is in any other population on earth. Type 2 diabetes can be cured if it is detected at an early age.

II. LITERATURE SURVEY

S. Barik et al., [3] in their paper “Analysis of Prediction Accuracy of Diabetes Using Classifier and Hybrid Machine Learning Techniques” used XGBoost and Random Forest algorithm and they did not perform data cleaning on the dataset. They achieved the highest accuracy of 74.1 % with XGBoost algorithm. H. Balaji et al., [2] in their paper “Optimal predictive analytics of pima diabetics using deep learning” used only Recurrent Neural Networks for training and testing their models. A. Mohebbi et al., [4] in their paper “A deep learning approach to adherence detection for type 2 diabetics” used Convolutional Neural Networks and the dataset used by the authors was very small. Y. N. and SriPreethaa [5] in their paper “Diabetes prediction in

healthcare systems using machine learning algorithms on hadoop clusters" used only Random Forest Algorithm on Hadoop Cluster and the authors discussed the information gain methods for feature selection but they did not mention the pre-processing steps used. F. Mercaldo et al., [6] in their paper "Diabetes mellitus affected patients classification and diagnosis through machine learning techniques" used Hoeffding Tree, JRip, BayesNet, Random Forest. The authors used WEKA tool for analysis and prediction but they did not mention the pre-processing steps and the highest accuracy achieved by the authors was 76 % using Hoeffding Tree. S. Lekha and Suchetha M. [7] in their paper "Real-time non-invasive detection and classification of diabetes using modified convolution neural network" used one dimensional modified Convolution Neural Network on the dataset. The dataset used by them was very small in size. J Pradeep and Saminathan [8] in their paper "Performance analysis of classifier models to predict diabetes mellitus" used J48, KNN and Random Forest algorithm for prediction and they got 73.8 % accuracy with J48 algorithm. The authors did not mention the data pre-processing techniques and methodology used by them. A. Iyer et al., [9] in their paper "Diagnosis of diabetes using classification mining techniques" used only two algorithms i.e., Decision Tree and Naive Bayes algorithms and the authors got an accuracy of 79.5% from Naive Bayes classifier. Pre-processing and data-set transformation was done using WEKA tool. Q. Zou et al., [10] in their paper "Predicting diabetes mellitus with machine learning techniques." used Decision Tree, Neural Networks and Random Forest and algorithms. They achieved highest accuracy of 76% with Neural Networks. MATLAB tool was used for model creation by the authors. S Nanda et al., [11] in their paper "Prediction of gestational diabetes mellitus by maternal factors and biomarkers at 11 to 13 weeks" used only Logistic Regression for prediction and the accuracy of their model was not mentioned.

III. ALGORITHMS USED

In this paper, we propose to use different supervised ML algorithms like Logistic Regression, Support Vector Machine, Random forest, Naive Bayes, Decision Tree and k-NN and DL algorithms like Perceptron and Multilayer Perceptron for diabetes prediction. The different algorithms that we have used in our research work are described below.

A. Supervised learning

Machine learning is a trending domain of computer science and it can also refer to automated detection of patterns in data that are meaningful. There are many far reaching applications like predicting various outcomes. [14]. The task of generating a function that is used to map the given inputs to the expected outputs is called as supervised learning. It defines a function using the training data-set which contains various types of training sample sets. It is also usually used in problems like classification because our objective is to make the machine learn and become a classifier [15]. Supervised learning technique is used for analyzing the training data of the dataset and create a deduced function for mapping the new data objects and using this makes the learning algorithm classify hidden objects uniquely. These algorithms mainly deal with classification algorithms which includes the below mentioned algorithms:

a. Support Vector Machine

Support Vector Machine is a recent ML algorithm also termed as SVM [16]. SVM algorithms are very closely related to deep learning neural networks algorithm "multilayer perceptron". SVM converts the actual data into a data-set with higher dimensions using non-linear mapping. SVM can be used for classification as well as numerical prediction. SVM has a margin on either sides of a hyperplane and this hyperplane is used to separate data into two different groups. The expected generalization error can be minimized by increasing the margin and increasing the distance between instances on either sides and the hyperplane.

b. K-nearest neighbor

K-nearest neighbor or KNN provides a large set of rules which can be used to divide the data into groups. This algorithm makes an assumption that similar things exist close to each other. It finds distance between different data items when plotted on a graph. It finds K samples close to an element. Then the most frequent label is assigned to the entire group. As the dataset is small KNN could be well suited to solve our problem [8].

c. Naive Bayes classifier

Naïve Bayes algorithm is a sequential probabilistic algorithm which performs prediction and classification by using execution and estimation using the Bayes Theorem [9]. In order to find relationships and patterns between diabetes and medical records of a person the problem of quantity of data arises. Data available

to solve this problem is limited. Naïve Bayes algorithm could be well suited because it requires less data for training and it can perform well with numeric data.

d. Random Forest Classifier

Random Forest Algorithm uses bagging method to train multiple decision trees. It makes up a model of many different decision trees as a single model in order to maximize the overall result. It is usually used for both classifier based and regression based problems. It adds randomness to our model while forming the trees. Its goal is not to find the best feature when we split a node but finding the most important feature from randomly generated feature sets. This diversity results in formation of a good model [24].

e. Decision Tree

Decision Tree is a flowchart structured like a tree. It is used for predicting and classifying data values by representing them as nodes of a tree. The root node of this tree is used to distinguish and separate instances with different attribute values [9]. The final leaf node represents the class label assigned. The goal is to predict the value or class of a given set of input variable with the help of decision rules formed from training data.

f. Logistic Regression

Logistic Regression is a type of supervised ML algorithm that basically uses the logistic function in order to perform binary classification of the dependent variables. The result is computed based on the response received from every variable of the "odds ratio" of the event of interest observed. It uses an equation which is used to find the conditional probability of an outcome on specified predictor values [24].

B. Deep Learning

Recently, various machine learning models developed are being used for forecasting the results or in a simpler term these models are being used for prediction of results. These models have become very famous in the medical diagnosis field also and medical diagnosis can be termed as one of the main problems in medical applications. There has been a significant research for improving the accuracy of different neural networks for medical datasets as neural networks improve the medical diagnosis models [17]. There are many neural networks like Recurrent Neural Networks, Convolution Neural Networks and Artificial Neural Networks. Artificial Neural Networks also termed as ANN contain various layers of nodes with computing data and these nodes are capable of processing information. These ANN can also find non linearities that are not originating as inputs and they are able to form clusters, association, approximate functions and forecasting, and helps to perform multi factorial analysis to make complex patterns, where we have less prior information [18]. Today, artificial neural networks (ANN) are being preferred over other models because the non-linearity that they posses allow the data to be fitted more accurately, as ANN are noise-insensitive hence this leads to an accurate prediction irrespective of the measurement errors, ANN helps in fast processing and are hardware failure tolerant because of the presence of high parallelism, in the response to the changing environment ANN's learning and the adaptability permits the models to update the model's internal structure with response to the change in the environment [19]. Frank Rosenblatt in 1957 introduced Perceptron and a perceptron learning rule based on the original MCP neuron was proposed by him. A perceptron can be defined as a neural network that is capable of doing some computations to detect various features or business intelligence in the data that has been provided as an input whereas a multilayer perceptron also termed as MLP is a type of feedforward ANN and it utilizes backpropagation for training which is a supervised learning technique.

a. Perceptron

The perceptron is defined as an algorithm for binary classifiers of supervised learning algorithms where a binary classifier is a function which is used for deciding whether an input provided by the user is part of a specific group or not. It is a form of a linear classification algorithm where a prediction is made on a linear predictor function by combining feature vectors with different set of weights [21]. Perceptron is one of the simplest and first artificial neural networks. It is similar to logistic regression algorithm. It can learn separation feature space for two class classification tasks, although unlike logistic regression, it uses stochastic gradient descent optimization algorithm for learning purpose and perceptron does not predict calibrated probabilities [22].

a.1. Perceptron Algorithm

In the model as shown in figure 1, there are binary inputs (I_1, I_2, …., I_N) and same N is the total amount of weights (W1, W2, …, WN) provided to the model. All the inputs are now multiplied and then summed up together. This sum is called as pre-activation function and denoted as "P".

$$P = \sum_{i=1}^{n} Weight_i I_i = Weight^T I$$

Now we apply the activation function and this produces an activation which is termed as "A" and this activation function can also be termed as "step function".

$$\varrho\,(y) = 1, \text{ when } y > 0$$
$$\varrho\,(y) = 0, \text{ when } y <= 0$$

The model generates an output 1 if the input value is either 0 or greater whereas, our output value is 0 if we have an input value is lesser than 0.

$$A = \varrho\,(Weight^T I)$$

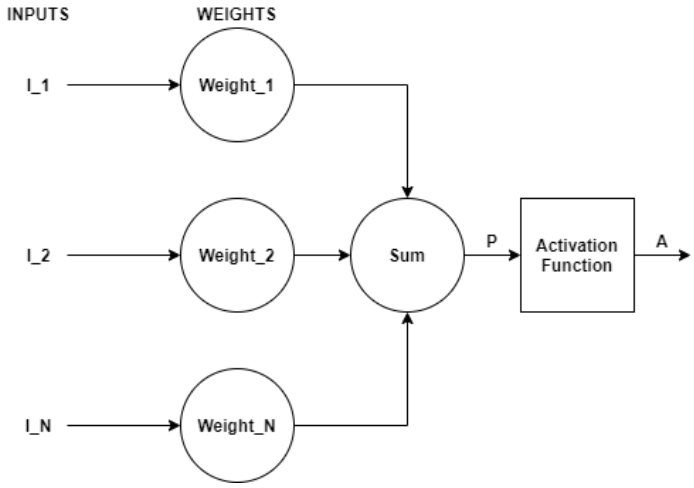

Fig 1. Perceptron

b. Multilayer Perceptron

The multilayer perceptron or MLP algorithm is the most used neural network based algorithm. MLP is considered as a type of a feedforward based artificial neural network [23]. Generally, the signals transmitted from the network are unidirectional i.e., from input to output. As there is no loop in multilayer perceptron, hence output does not affect the neuron and this architecture of multilayer perceptron is termed as feed-forward architecture. There are layers present in multilayer perceptron that do not have a direct connection to the environment and these layers are called as hidden layers. The input layer is the layer that provides input to the model. The non-linear activation function provides all the strength to multilayer perceptron. We can use any of the non-linear functions except for polynomial function and most commonly used functions are single pole sigmoid and bipolar sigmoid function [20].

b.1. Multilayer Perceptron Algorithm

In the model as shown in figure 2, feed forward type of network is implemented. The main objective of feed forward network is to provide approximation for any function T (). Let for a classifier q = T * (I) that is mapping input I to output q, then multilayer perceptron finds the approximation by mapping q = T (I; θ) and it learns the best parameters. Multilayer perceptron consists of many functions together. A MLP network with 4 layers will have the following equation:

$$T\,(I) = T\,(4)\,(T\,(3)\,(T\,(2)\,(T\,(1)\,(I))))$$

and each layer is represented as:

$$q = A\,(Weight\ I\ P + Bias)$$

where,

A is activation function
Weight is weights in the layers
I is the input in form of vectors
Bias is the bias vector

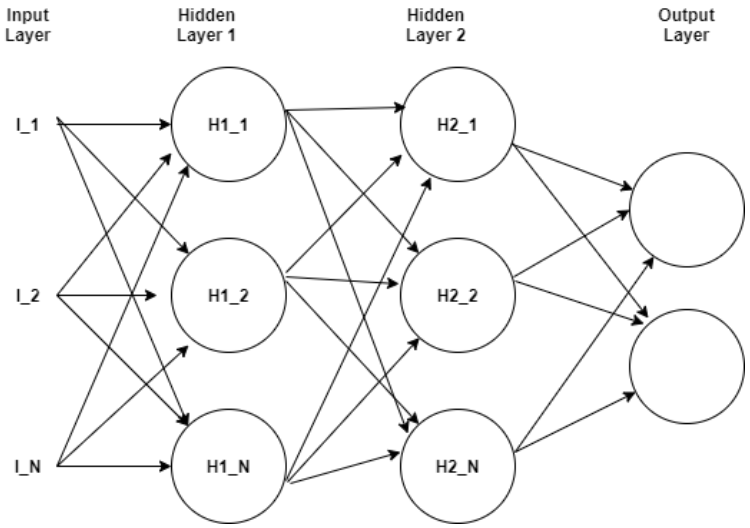

Fig 2. Multilayer Perceptron

IV. METHODOLOGY

The technique of gathering, processing and analyzes data i.e., combine programming, mathematics and statistics domain to extract insights and knowledge from the data is termed as Data Science. This technique is considered to be an important technique for a project as it improves the performance prediction of machine learning algorithms used i.e., help in predicting the outcome by providing input to the machine [13]. The methodology followed is explained below:

- **Data Import:** It is the step where we collect the dataset. The collected dataset is then split into train data and test data for the prediction model. The dataset can be collected from various resources like online or data stored in various databases. It is important to select the correct dataset because our prediction model is trained using the dataset and if use incorrect model then we would not receive the required results from our trained machine learning model.
- **Data Cleaning:** Data cleaning plays an important role as the data is collected from various resources and hence the data in coarse in nature. This coarse data is termed as "dirty data". The dirty data may contain some values that might not be required for the model training purpose hence that should be handled and we handle unrequired data in the data cleaning step. Also, the dataset may contain null values in it. It becomes important to handle null values because we can not use the dataset with null values. We can replace the null values with either mean, median or mode values.
- **Data Visualization:** In this technique we visualize the data present in the dataset by plotting various graphs. This technique makes analysis of data easy and helps us to find co-relation among various attributes present in the dataset.
- **Data Analysis:** Analyze the results of various graphs and gain knowledge about the different attributes of the dataset.
- **Feature Scaling:** As the dataset contains the data which is of different units, hence we need to scale the data present in the dataset into data of similar units and for this we use standard scaler to scale all the values. The accuracy of the prediction model can be increased to a great extent using feature scaling.
- **Select Machine Learning Algorithm:** Select the best Machine/Deep learning algorithm for creation of prediction model.
- **Prediction Model:** Create the prediction model for predicting diabetes.
- **Design User Interface:** Develop a Graphical User Interface (GUI) for the prediction model to be easily used by users.

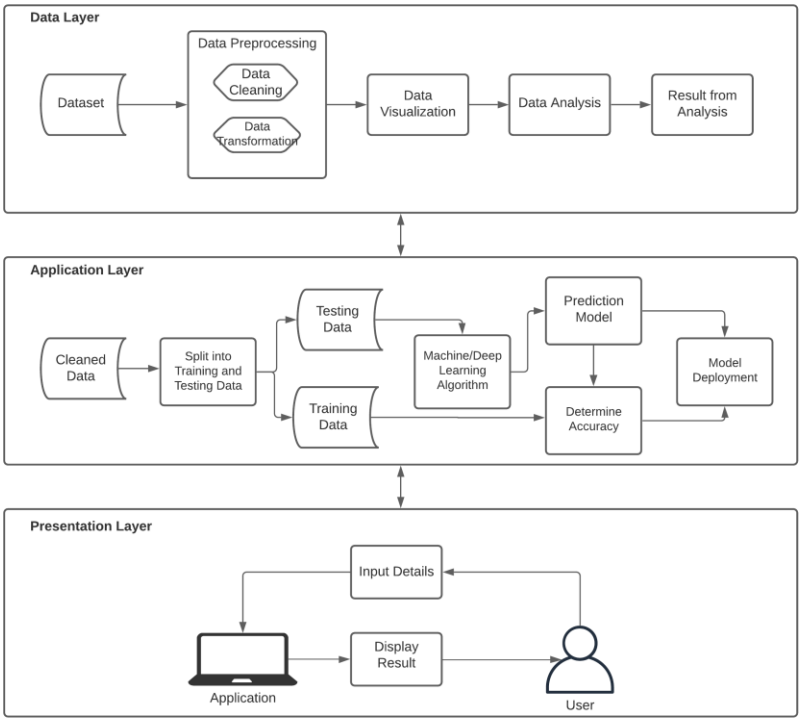

Fig 3. Methodology

V. PROPOSED ALGORITHM

Step i: Collect Diabetes Dataset
Step ii: Perform Preprocessing on the data
1) Drop useless columns
2) Fill in missing values with class median.
Step iii: Split the data into Training set and Testing set.
Step iv: Create ensemble model
Level 0) Stack KNN, SVM and Decision Tree in Base Layer.
Level 1) Make Logistic Regression Model in Meta Layer from the outcomes of base layer.
Step v: Find Predictive Outcome using Testing set on the model.
Ensemble_Model.fit (train_x_data, train_y_data)
Predictive_Outcome = model.predict (test_x_data)
Step vi: Calculate accuracy achieved by the model.

VI. RESULT

In order to analyze the propinquity betwixt the non-identical attributes present in the collected dataset, we plotted a heatmap as shown in figure 4. From the Heatmap we can see that glucose and insulin are highly correlated with diabetes. We can also see that BMI and skin thickness have a high correlation with each other. There is a correlation between age and pregnancies. Insulin and glucose are also strongly correlated with each other.

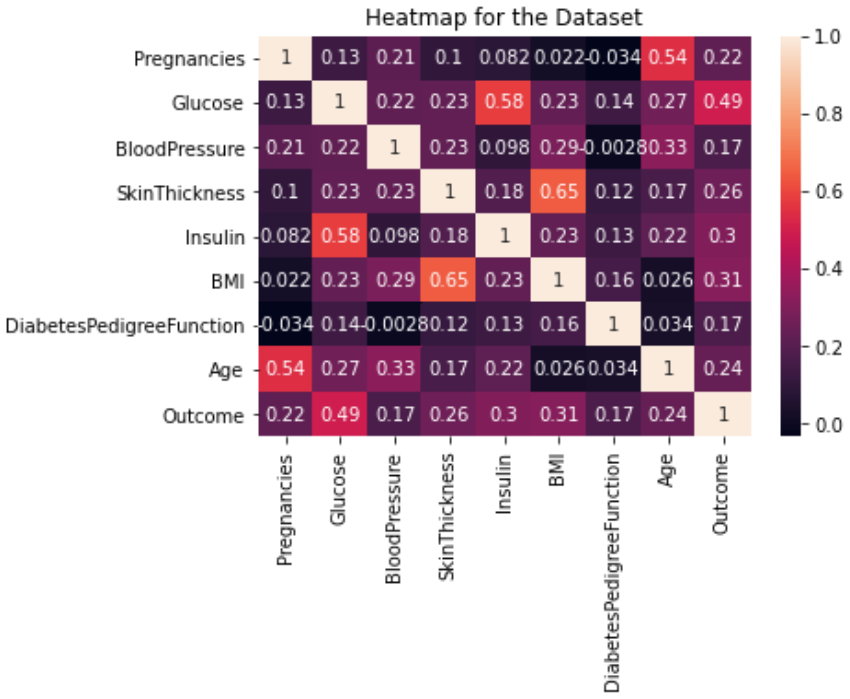

Fig 4. Heatmap for the Dataset

In this research work, we have implemented different machine learning and deep learning algorithms as mentioned earlier in the paper. Upon testing our trained model, accuracy of Perceptron came out to be 67% which is the lowest accuracy achieved. Naïve Bayes gave an accuracy of 77.2%. An accuracy of 83.7% was obtained by using Random Forest. Multilayer Perceptron with scaled input for training was 83.9% accurate. 78.5% accuracy was obtained by using logistic regression algorithm, we used logistic regression in the meta layer (level 1) of our ensemble model. Top three accuracies were achieved 88.3% by k-NN, 87.0% by SVM and 85.7% by Decision Tree, we used these three algorithms in the base layer (level 0). The output from these layers was fed to the meta layer to train our ensemble hybrid model. We achieved an accuracy of 90.6% with our ensemble model.

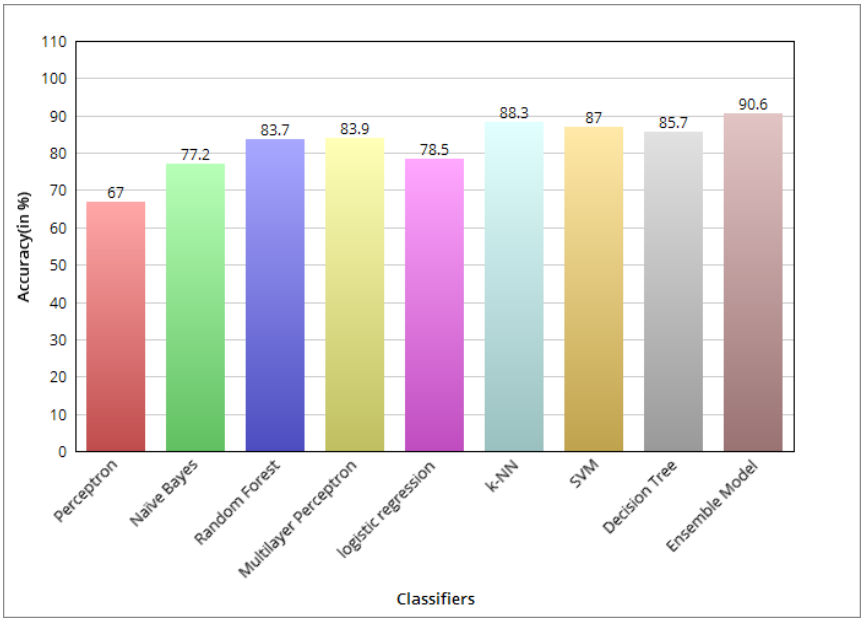

Fig 5. Accuracies of different classifiers

VII. FUTURE SCOPE AND CONCLUSION

Only restriction faced during this project included access to the finite data in the dataset. The next step to improve the accuracy of model created for prediction is to apply the various techniques mentioned in this research paper using different or dataset larger than the dataset used. Also, we can use attributes like daily food habits and exercise habits of persons for training the model as these attributes can affect whether a person has a risk of being diabetic or not. If Deep Neural network (Keras, Thaeno and Tensorflow) and unsupervised learning algorithms are implemented on a larger dataset, then the accuracy of the prediction model can be improved to a great extent. We can develop similar prediction models for various diseases like heart dieses, cancer, brain tumors, asthma and so on.

VIII. REFERENCES

[1] Alehegn, Minyechil & Joshi, Rahul & Mulay, Preeti. (2018). Analysis and prediction of diabetes mellitus using machine learning algorithm. International Journal of Pure and Applied Mathematics. 118. 871-878.

[2] Balaji, H. & Iyenger, N Ch Sriman Narayana & Caytiles, Ronnie. (2017). Optimal Predictive analytics of Pima Diabetics using Deep Learning. International Journal of Database Theory and Application. 10. 47-62. 10.14257/ijdta.2017.10.9.05.

[3] S. Barik, S. Mohanty, and D. Singh. Analysis of Prediction Accuracy of Diabetes Using Classifier and hybrid Machine Learning Techniques, pages 399-409. 01 2021.

[4] Mohebbi A, Aradottir TB, Johansen AR, Bengtsson H, Fraccaro M, Morup M. A deep learning approach to adherence detection for type 2 diabetics. Annu Int Conf IEEE Eng Med Biol Soc. 2017 Jul;2017:2896-2899. doi: 10.1109/EMBC.2017.8037462. PMID: 29060503.

[5] Y. N. and K. SriPreethaa. Diabetes prediction in healthcare systems using machine learning algorithms on hadoop cluster. Cluster Computing, 22, 01 2019.

[6] F. Mercaldo, V. Nardone, and A. Santone. Diabetes mellitus affected patients classification and diagnosis through machine learning techniques. Procedia Computer Science, 112:2519-2528, 12 2017.

[7] S. Lekha and S. M. Real-time non-invasive detection and classification of diabetes using modified convolution neural networks. IEEE Journal of Biomedical and Health Informatics, 22(5):1630-1636, 2018.

[8] J. p. Kandhasamy and S. Balamurali. Performance analysis of classifier models to predict diabetes mellitus. Procedia Computer Science, 47:45–51, 12 2015.

[9] A. Iyer, s. Jeyalatha, and R. Sumbaly. Diagnosis of diabetes using classification mining techniques. International Journal of Data Mining Knowledge Management Process, 5:1–14, 02 2015.

[10] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang. Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics, 9:515, 2018.

[11] S. Nanda, M. Savvidou, A. Syngelaki, R. Akolekar, and K. Nicolaides. Prediction of gestational diabetes mellitus by maternal factors and biomarkers at 11 to 13 weeks. Prenatal Diagnosis, 31(2):135 – 141, Feb. 2011.

[12] http://www.who.ont/mediacenter/factsheets/fs312/en/

[13] https://medium.com/@praffullakrdubey/understanding-data-science-pipeline-and-its-general-terms-5898c6927707

[14] Akinsola, J E T. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. International Journal of Computer Trends and Technology (IJCTT). 48. 128 - 138. 10.14445/22312803/IJCTT-V48P126.

[15] Taiwo Oladipupo Ayodele (February 1st 2010). Types of Machine Learning Algorithms, New Advances in Machine Learning, Yagang Zhang, IntechOpen, DOI: 10.5772/9385. Available from: https://www.intechopen.com/books/new-advances-in-machine-learning/types-of-machine-learning-algorithms

[16] Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. (2nd ed.). Springer Verlag. Pp. 1 – 20. Retrieved from website: https://www.andrew.cmu.edu/user/kk3n/simplicity/vapnik2000.pdf

[17] A. Rana, A. Singh Rawat, A. Bijalwan and H. Bahuguna, "Application of Multi Layer (Perceptron) Artificial Neural Network in the Diagnosis System: A Systematic Review," 2018 International Conference on Research in Intelligent and Computing in Engineering (RICE), San Salvador, El Salvador, 2018, pp. 1-6, doi: 10.1109/RICE.2018.8509069.

[18] Leonardo Maria Reyneri "Implementation Issues of Neuro-Fuzzy Hardware: Going Toward HW/SW Codesign" IEEE TRANSACTIONS ON NEURAL NETWORKS, JANUARY 2003, VOL. 14, NO. 1, , pp(176-194)

[19] Cheol-Taek Kim and Ju-Jang Lee "Training Two-Layered Feed forward Networks With Variable Projection Method" IEEE Transactions on Neural Networks, Vol. 19, No. 2, February 2008.

[20] Marius, Popescu & Balas, Valentina & Perescu-Popescu, Liliana & Mastorakis, Nikos. (2009). Multilayer perceptron and neural networks. WSEAS Transactions on Circuits and Systems. 8.

[21] https://en.wikipedia.org/wiki/Perceptron

[22] https://machinelearningmastery.com/perceptron-algorithm-for-classification-in-python/

[23] https://en.wikipedia.org/wiki/Multilayer_perceptron

[24] Naveen Kishore G, V.Rajesh, A.Vamsi Akki Reddy, K.Sumedh, T.Rajesh Sai Reddy. (2020). "Prediction Of Diabetes Using Machine Learning Classification Algorithms", INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 9, ISSUE 01, JANUARY 2020 ISSN 2277-8616