A Comparative Analysis of Machine-Learning Algorithms to Build A Predictive Model for Diabetes Disease

OLATUNDE OLUKEMI VICTORIA[1], ADESOMOJU FISAYO ADEYEMO[2]
[1] *Department of Software Engineering, School of Computing, Federal University of Technology, Akure, Nigeria.*
[2] *Department of Computer Science, School of Computing, Federal University of Technology, Akure, Nigeria.*

Abstract- ***Machine Learning is concerned with the development of algorithms and techniques that allows the computers to learn and gain intelligence based on the past experience. It is a branch of Artificial Intelligence (AI) closely related to statistics. By learning it means that the system is able to identify and understand the input data, so that it can make decisions and predictions based on the data. In this paper, a web based comparative analysis of various machine learning algorithms (Decision Tree, Support Vector Machine, K-Nearest Neighbor, Logistic Regression) was design in order to recognize accurate model for detecting diabetes disease.***

Indexed Terms- ***Machine Learning, Diabetes diseases, Decision Tree, Support Vector Machine, K-Nearest Neighbor, Logistic Regression.***

I. INTRODUCTION

Diabetes mellitus (DM) or just Diabetes is one of the most common chronic diagnosed disease that occurs as result of the pancreas not producing enough insulin or caused by the increase level of the sugar (glucose) in the blood when the body cannot effectively use the insulin it produces and also a major public health challenge worldwide with more than 370 million people diagnosed with the disease in 2012 projected to reach 439 million in 2030, it is the third leading cause of death following diseases of heart and cancer (Lal, 2016; Mirzajani and Salimi, 2018; Alam, *et al*., 2019; Warke *et al., 2019*). Machine learning (ML) algorithms are characterized by the ability to learn over time without being explicitly programmed, Diabetes diagnosis have been carried out using different machine learning algorithms to predict the disease onset for an improved treatment. In this study, we used decision tree, random forest and neural network to predict diabetes mellitus. The dataset consisting of 9 attributes gotten from First Mercy Hospital physical examination data in Akure, Nigeria was used in the study.

The IDF Diabetes Atlas Ninth edition 2019 provides the latest figures, information and projections on diabetes worldwide. In 2019, Approximately 463 million adults (20-79 years) were living with diabetes; by 2045 this will rise to 700 million, the proportion of people with type 2 diabetes is increasing in most countries.79% of adults with diabetes were living in low- and middle-income countries**,** 1 in 5 of the people who are above 65 years old have diabetes, 1 in 2 (232 million) people with diabetes were undiagnosed. Diabetes caused 4.2 million deaths. 10% of total global expenditure spent on diabetes. More than 1.1 million children and adolescents are living with type 1 diabetes, more than 20 million live births (1 in 6 live births) are affected by diabetes during pregnancy, 374 million people are at increased risk of developing type 2 diabetes

Diabetes causes a lot of damages to different parts of the human body some of which are: eyes, kidney, heart, and nerves. According to the Centers for Disease Control and Prevention (CDCP) information were given for the duration of 9 ensuing years that is between 2001 and 2009 type II diabetes increased 23% in the United States (US). There are different countries, organization, and different health sectors worry about this chronic disease control and prevent before the person death. (Minyechil *et al.* 2017)

The machine learning algorithms can be roughly categorized into three types namely supervised learning, unsupervised learning and semi-supervised learning. The supervised learning algorithms are used when human expertise does not exist (navigating on Mars), humans are unable to explain their expertise (speech recognition). Solution changes in time series (routing on a computer function) and to solution needs to be adapted to particular cases (user biometrics). The supervised learning algorithms are classified into different types such as probability-based, function-based, rule-based, tree-based, instance-based, etc. The unsupervised learning is the descriptive type learning. This learning is used to describe or summarize the data. The examples of the unsupervised learning algorithms are clustering, association rule mining, etc. The semi-supervised learning is the combination of supervised and unsupervised. This study presents a diabetes prediction system to diagnosis the diabetics. Moreover, the supervised learning algorithm is used to train the diabetes predication system for diagnosing diabetes. The accuracy of this prediction system is improved using pre-processing technique. (Antony *et al.* 2017)

II. MATERIALS

A lot of systems have been developed using various machine learning techniques: Vijayakumar, *et al*. (2019) developed a Machine Learning system that could predict diabetes over big data from healthcare communities. The objective was to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy. Pre-processing, noise removal and clustering were done before the classification, SVM (Support Vector Machine) classification techniques was used for the classification of diabetic and non-diabetic data for earlier detection of the diabetes disease. Aminul, *et al*. (2017) developed a system that could predict the onset of diabetes using machine learning techniques (SVM, Naive Bayes, Logistic Regression). Their objective was to detect diabetic patient's onset from the outcomes generated by machine learning classification algorithms that is been used. Datasets were collected among the Pima Indian female population, the prediction of outcome, the patient was classified into one of the two categories (tested positive and tested negative).

Aakansha, *et al*. (2017) developed a system that could predict diabetes using supervised learning towards better healthcare for women. The aim was to detect diabetes risk using PIMA Indian Diabetes Data-set. Minyechil, *et al*. (2017) with the aim was to predict diabetes disease and compare the algorithms used in diagnosis with the intent of picking the best employed naïve Bayes, K-nearest neighbor and random forest machine-learning algorithms in the prediction using dataset obtained from UCI repository PIDD. Basharat, *et al*. (2018) developed a prediction expert system using Random forest, Decision tree, Random Stump, and Random tree learning algorithm for diagnosis of diabetic disease. (Tejas, *et al*., 2018 and Rabina *et al.* 2016) developed diabetes diagnostic systems to predict diabetes/juvenile diabetes using machine learning techniques.

III. METHODS

The architecture presented in figure 3.0 consist of a 503 dataset, gotten from a medical institution and serve as an interface where users and administrator can access diagnostic prediction based on the reqiured information and view their result.

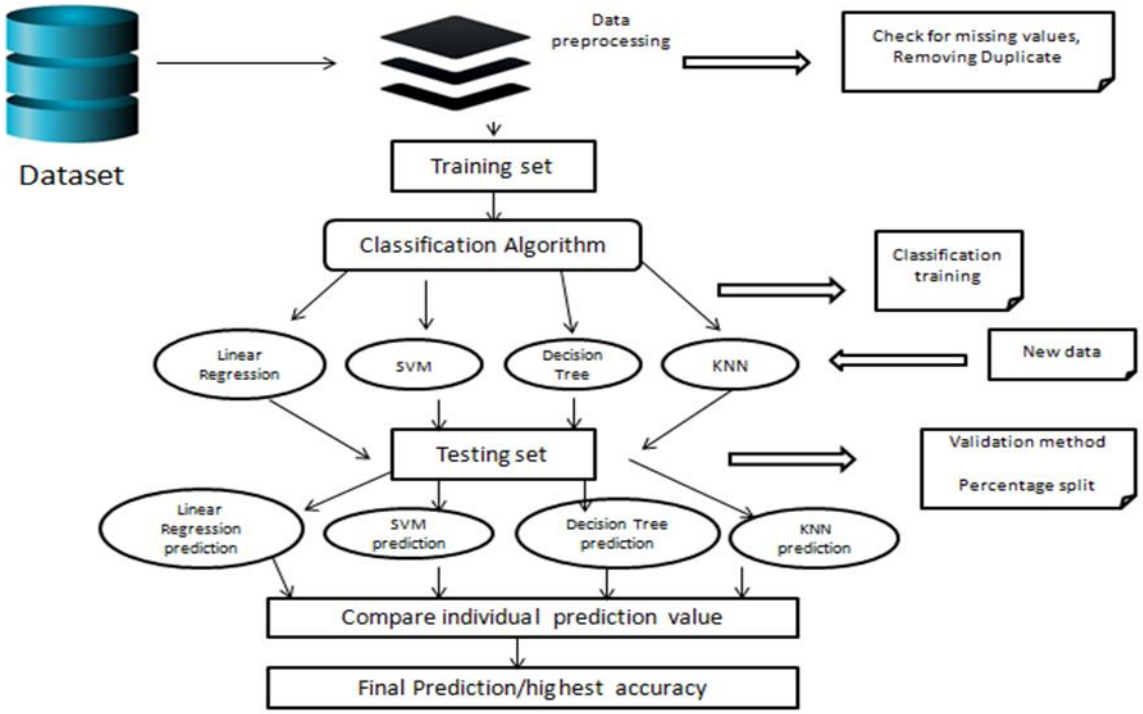


Figure. 3.0 Architecture for the Predictive Model For Diabetes Disease
(Adapted from Minyechil, 2017)

In the data preprocessing stage, dataset are screened or checked for missing values, that is, all missing values in the dataset are identified and replaced with a "Null". The trainning set consist of some percentage of data that were initially gathered, and these data were used to train the system, using various algorithms which are Decision tree, Linear regression, Support vector machine and KNN. The testing set also consist of some percentage of the dataset that were initially gathered, and this were used to test the sytem, still using the same classification algorithms that were used to train the system. Individual predicting accuracy is displayed and the algorithm with the highest accuracy is selected as the best choice, hence considered as the algorithm to be finally used to predict any other case.

3.1 Implementation

At the implementation stage, several component of the evaluation system were integrated together, this involves the preparation of the resources including equipment and personnel with the testing of the system.

The design of the system was done using: Python, Jupyter Notebook and Microsoft Excel spreadsheet. Python was used to write the code for the entire system. Jupyter Notebook was the IDE, which is Integrated Development Environment, which was used to write the machine learning codes. Microsoft Excel Spreadsheet was use to analyze, structure the data and clean the dataset used.

3.1.1 Data Selection

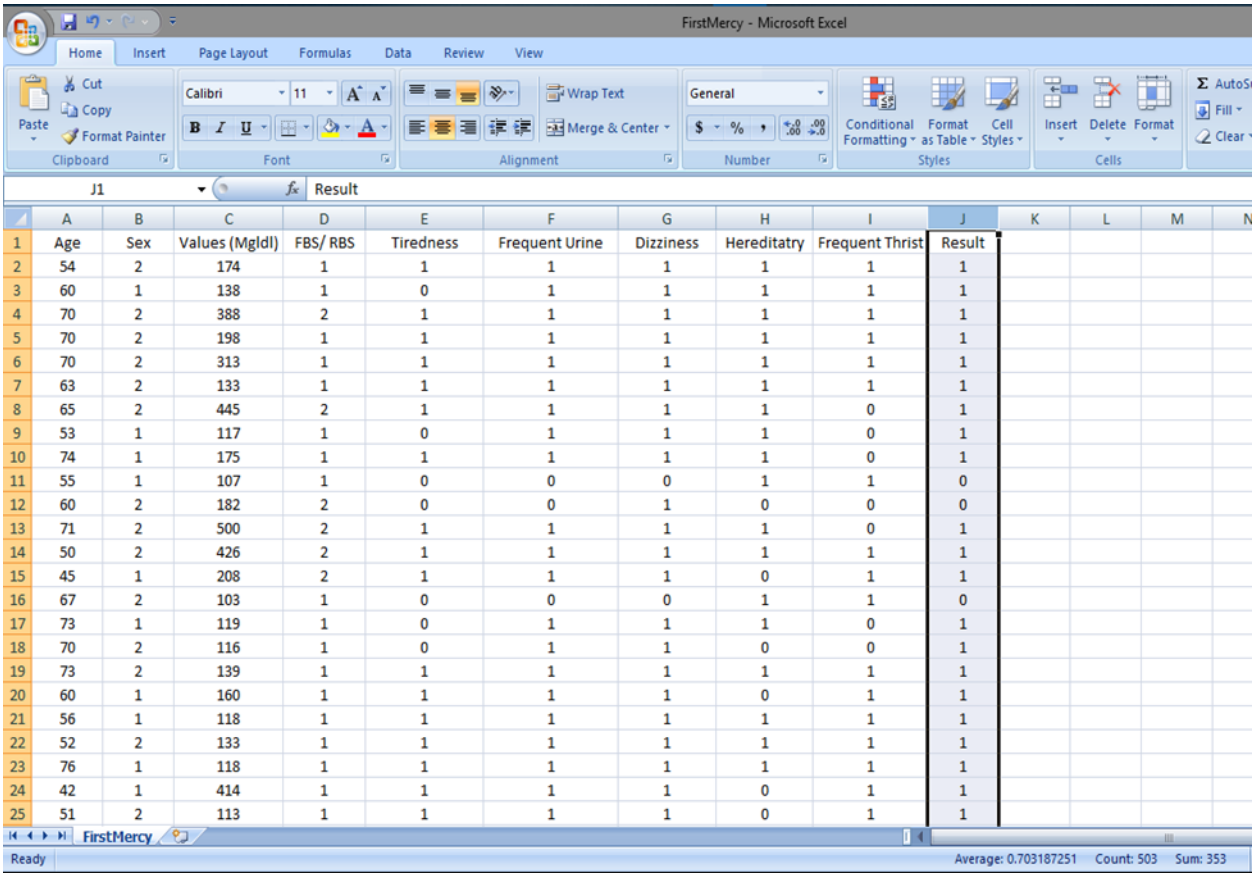
The excel spreadsheet interface contains the entire database, about 503 patient with 9 attributes and these attributes are numerical as shown in *Table 3.1* and *Table 3.2* respectively. (Age, Sex, Blood-Sugar Level, FBS/RBS, Tiredness, Frequent Urine, Frequent, Thirst, Dizziness and Hereditary) and the result. As shown in *figure 3.1.*

Table 3.1 Attribute of dataset

S/N	Name of attributes	Type
1	Age	Numeric
2	Sex	Numeric
3	Values (Mgldl)	Numeric
4	FBS/RBS	Numeric
5	Tiredness	Numeric
6	Frequent urine	Numeric
7	Dizziness	Numeric
8	Frequent Thirst	Numeric
9	Hereditary	Numeric

Table 3.2 Analysis of data

Data	Value
Male	1
Female	2
FBS	1
RBS	0
Yes	1
No	0



	A	B	C	D	E	F	G	H	I	J
1	Age	Sex	Values (Mgldl)	FBS/ RBS	Tiredness	Frequent Urine	Dizziness	Hereditatry	Frequent Thrist	Result
2	54	2	174	1	1	1	1	1	1	1
3	60	1	138	1	0	1	1	1	1	1
4	70	2	388	2	1	1	1	1	1	1
5	70	2	198	1	1	1	1	1	1	1
6	70	2	313	1	1	1	1	1	1	1
7	63	2	133	1	1	1	1	1	1	1
8	65	2	445	2	1	1	1	1	0	1
9	53	1	117	1	0	1	1	1	0	1
10	74	1	175	1	1	1	1	1	0	1
11	55	1	107	1	0	0	0	1	1	0
12	60	2	182	2	0	0	1	0	0	0
13	71	2	500	2	1	1	1	1	0	1
14	50	2	426	2	1	1	1	1	1	1
15	45	1	208	2	1	1	1	0	1	1
16	67	2	103	1	0	0	0	1	1	0
17	73	1	119	1	0	1	1	1	0	1
18	70	2	116	1	0	1	1	0	0	1
19	73	2	139	1	1	1	1	1	1	1
20	60	1	160	1	1	1	1	0	1	1
21	56	1	118	1	1	1	1	1	1	1
22	52	2	133	1	1	1	1	1	1	1
23	76	1	118	1	1	1	1	1	1	1
24	42	1	414	1	1	1	1	0	1	1
25	51	2	113	1	1	1	1	0	1	1

Figure 3.1: Excel Spreadsheet

3.1.2 Data processing

A statistical analysis was carried out with a bar chart showing the number of patient with diabetic and the number of people without diabetes, the blue bar and orange bar with the result zero represent the numbers of people without diabetes while those with the result one represent the numbers of those with diabetes. As shown in *figure 3.2*. It was proofed from *figure 3.3* that

the data are clean, that is no missing values were found.

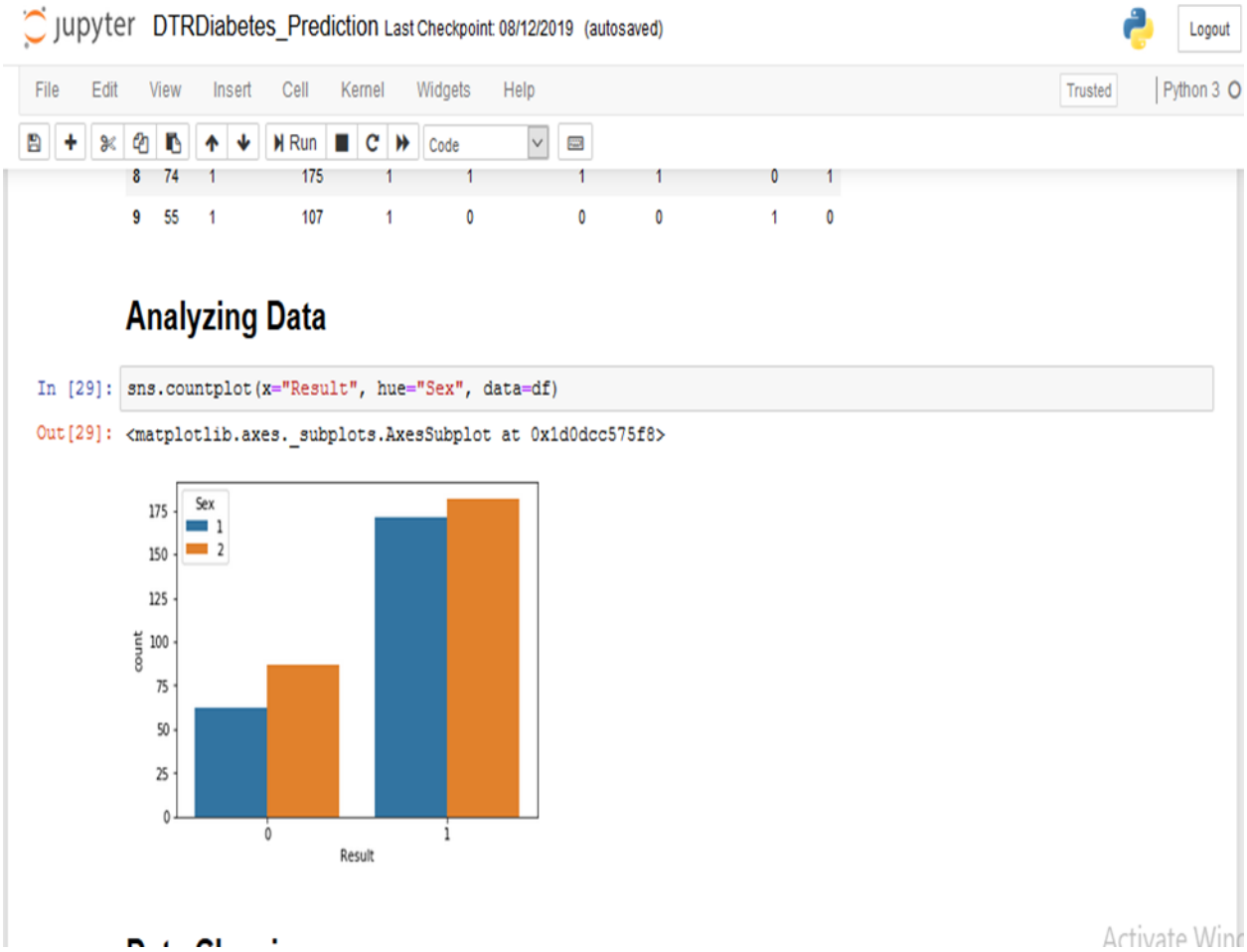
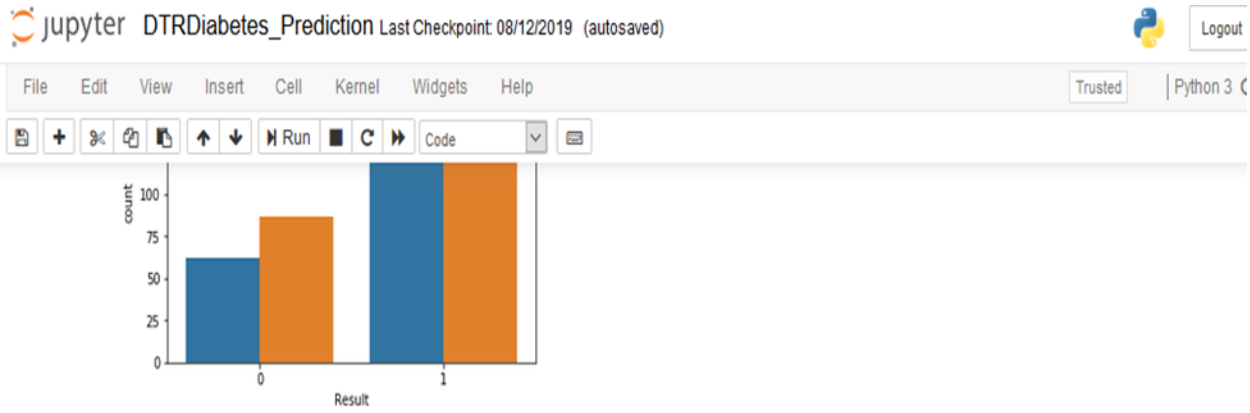


Figure 3.2 Data Analysis



Data Cleaning

```
In [30]: df.isnull().sum()

Out[30]: Age                0
         Sex                0
         Values (Mgldl)     0
         FBS/ RBS           0
         Tiredness          0
         Frequent Urine     0
         Dizziness          0
         Frequent Thrist    0
         Result             0
         dtype: int64
```

Figure 3.3 Data Cleaning

IV. THEORY/CALCULATION

4.1 Decision Tree

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning. A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

4.2 Support Vector Machine

A support vector machine (SVM) is a machine learning algorithm that constructs a hyper plane or set of hyper planes in a high dimensional space, which can be used for classification. It is a group of supervised learning methods and good separation can be achieved by the hyper plane that has the largest distance to the nearest training data points of any class. Sometimes it happen that sets are not linearly separable. Kernel function improving the SVM and solve dimensional and over fitting problem.

In computing the (soft-margin), SVM classifier is use to minimizing expression of the form, as shown in equation 1. (*wikipedia, 2019*).

$$\frac{1}{n}\sum_{i=1}^{n}\max([0,1-y_i(w\,.\,x_i-b)]\;+\;\lambda||\,w\,||^2) \quad \text{----- (1)}$$

4.3 K-Nearest Neighbour (K-NN)

k-Nearest neighbour is a simple algorithm but yields very good results. It is a lazy, non-parametric and instance based learning algorithm. This algorithm can

be used in both classification and regression problems. In classification, k-NN is applied to find out the class, to which new unlabeled object belongs. For this, a 'k' is decided (where k is number of neighbours to be considered) which is generally odd and the distance between the data points that are nearest to the objects is calculated by the ways like Euclidean's distance, Hamming distance, Manhattan distance or Minkowski distance.

The *k*-nearest neighbour classifier can be viewed as assigning the *k* nearest neighbours a weight 1/*k* and all others *0* weight. This can be generalised to weight nearest neighbour classifiers. That is, where the *i*th nearest neighbour is assigned a weight ω_{ni} , with $\sum_{i=1}^{n} \omega_{ni} = 1$. An analogous result on the strong consistency of weighted nearest neighbor classifiers also holds.

Let C_n^{wnn} denote the weighted nearest classifier with weights $\{w_{ni}\}_{i=1}^{n}$. Subject to regularity conditions] on the class distributions the excess risk has the following asymptotic expansion as shown in equation 2 and 3 (*Wikipedia*).

$$R_R(C_n^{wnn}) - R_R(C_n^{Bayes}) = (B_1 s_n^2 + B_2 t_n^2)\{1 + o(1)\} \quad \text{----- } 2$$

for constants B_1 and B_2 where $s_n^2 = \sum_{i=1}^{n} w_{ni}^2$ and $t_n^2 = n^{-2/d} \sum_{i=1}^{n} w_{ni} \{ i^{1+\frac{2}{d}} - (i-1)^{1+\frac{2}{d}} \}$. The optimal weighting scheme $\{w_{ni}^*\}_{i=1}^{n}$, that balances the two terms in the display above, is given as follows: set $k^* = Bn^{4/(d+4)}$.

$$w_{ni}^* = \frac{1}{k^*}\left[1 + \frac{d}{2} - \frac{d}{2k^{*2/d}}\{ i^{1+\frac{2}{d}} - (i-1)^{1+\frac{2}{d}} \}\right] for\ i = 1,2,\dots,k^* \quad \text{------ } 3$$

$$w_{ni}^* = 0\ for\ i = k^* + 1, \dots, n.$$

4.4 Linear Regression

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables. In statistics, linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.

Given a data set $\{ y_i, x_{i1}, \dots, x_{ip} \}_{i=1}^{n}$ of *n* statistical units, a linear regression model assumes that the relationship between the dependent variable *y* and the *p*-vector of regressors **x** is linear. This relationship is modeled through a *disturbance term* or *error variable* ε an unobserved random variable that adds "noise" to the linear relationship between the dependent variable and regressors. Thus the model takes the form, as shown in equation 4 and 5. (*Wikipedia*)

$$y_i = \beta_o + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \epsilon_i = x_i^T \beta + \epsilon_i, \quad \text{---- } 4$$

$$i = 1, \dots, n,$$

$$y = x\beta + \epsilon \quad \text{------} \quad 5$$

Where,

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad x = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

V. RESULT/DISCUSSION

Detection of Diabetes Mellitus in its early stages is the key for treatment. This work has detailed machine learning approach to predicting diabetes diseases. This project was developed as a web based comparative machine learning model to predict diabetes diseases for patient.

The system is driven by a particular Machine learning algorithm, proven to be the best among all other machine learning (SVM, KNN, Linear regression, Decision Tree) techniques. This system can be accessed and used by medical practitioners, patients, and other individual that might want to know their health status as regards diabetes.

The result indicated that linear regression and Decision Tree assured an accuracy of 100%, Support Vector Machine (96.6%) and KNN (92.7%).

CONCLUSION

Machine learning has the great ability to revolutionize the diabetes prediction with the help of availability of large amount of genetic diabetes dataset. Detection of diabetes in its early stages is the key for treatment.

There is no cure for diabetics but early detection can reduce the long-term complications and reduce the cost. Millions of people in the world have diabetes without knowing, the ability to predict diabetes early plays an important role for the patient's appropriate treatment strategy. However, the correct prediction accuracy of current machine learning algorithms is often low. Linear Regression and Decision tree performed the best among all. It predicted whether an individual was diabetes positive or not.

REFERENCES

[1] Aakansha, R., Sakshi, G., & Simran, C. (2017). Detecting and Predicting Diabetes Using Supervised Learning: An Approach Towards Better Healthcare For Women. International Journal of Advanced Research in Computer Science, Volume 8, No. 5. Retrieved from pdfs.semanticscholar.org/ec92/6200d2b424d8d5260954c91ec0b777ff255a.pdf

[2] Alam TM, Iqba MA, Ali Y, Wahab A, Ijaz S,… and Abbas Z (2019). A model for early prediction of diabetes. Informatics in Medicine Unlocked 16 () 100204 retrieved from https://doi.org/10.1016/j.imu.2019.100204

[3] Aminu, I., & Nusrat, J. (2017). Prediction of Onset Diabetes Using Machine Learning Techniques. International Journal of Computer Applications, Volume 180 – No.5. Retrieved from pdfs.semanticscholar.org/2c3a/6609a76762e3d40bdd90d8c07fa714d611fa.pdf

[4] American Diabetes Association (2013).Hyperosmolar hyperglycemic nonketotic syndrome (HHNS). Retrieved from www.diabetes.org/diabetes.

[5] American Diabetes Association. (2016).Diagnosing diabetes and learning about prediabetes. Retrieved from www.diabetes.org/diabetes-basics/diagnosis/

[6] American Diabetes Association. (2018).Diabetes symptoms. Retrieved from www.diabetes.org/diabetes-basics/symptoms/

[7] Antony, GS., Jebamalar, L., & Shanawaz B. (2017). Diabetes Prediction Using Medical Data. 2018-2020 Diabetes Research Institute Foundation. Volume 10, Number 1. Retrieved from researchgate.net/publication/316432650_Diabetes_Prediction_Using_Medical_Data.

[8] Basharat, N., Arshad, A., Muhammad, AH. & Muhammad, A. (2018). Prediction Techniques for Diagnosis of Diabetic Disease. International Journal of Computer Science and Network Security, Volume 18, No.8. Retrieved from paper.ijcsns.org/07_book/201808/20180818.pdf

[9] Centers for Disease Control and Prevention (2017). New CDC report: More than 100 million Americans have diabetes or pre-diabetes [Press release]. Retrieved from www.cdc.gov/media/releases/2017/p0718-diabetes-report.html

[10] Diabetes facts and figures (2019). IDF Diabetes Atlas 9th Edition 2019. Retrieved from www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html

[11] Erika, F.B. (2019, May). Diabetes Mellitus Disorder. Retrieved from www.msdmanuals.com/home/hormonal-and-metabolic-disorders/diabetes-mellitus-dm-and-disorders-of-blood-sugar-metabolism/diabetes-mellitus-dm.

[12] Harleen, K., & Kumari V. (2018). Predictive modeling and analytics for diabetes using machine learning approach. Retrieved from www.sciencedirect.com/science/article/pii/S2210832718303650X

[13] Health Jade Team (2018) .What is type 2 diabetes and how do I prevent it? Retrieved from www.healthjade.net/what-is-type-2-diabetes-and-how-do-i-prevent-it/

[14] Lal, BS (2016). Public Health Environment and Social Issues in India Edition: Diabetes: Causes,

Symptoms and Treatments (pp 55 – 67). Serial Publication

[15] Mirzajani, SS and Salimi, S (2018). Prediction and Diagnosis of Diabetes by Using Data Mining Techniques. Avicenna Journal of Medical Biochemistry 6(1):3-7 doi:10.15171/ajmb.2018.02

[16] Minyechil, A., & Rahul, J. (2017). Analysis and Prediction of Diabetes Diseases Using Machine Learning Algorithm: Ensemble Approach. Volume: 04, Issue 10. Retrieved from www.irjet.net/archives/V4/i10/IRJET-V4I1077.pdf

[17] Pretty, M., Rahul, J., & Minyechil, A. (2018). Analysis and Prediction of Diabetes Mellitus using Machine Learning Algorithm. International Journal of Pure and Applied Mathematics, Volume 118, No. 9. Retrieved from acadpubl.eu/jsi/2018-118-7-9/articles/9/87.pdf

[18] Rabina, & Anshu, C., (2016). Diabetes Prediction by Supervised and Unsupervised Learning with Feature Selection. International Journal of Advance Research, Ideas and Innovations in Technology. Volume 2, Issue 5. Retrieved from www.ijariit.com/manuscripts/v2i5/V2I5-1136.pdf

[19] Tejas, N.J. & Pramila, M.C. (2018). Diabetes Prediction Using Machine Learning Techniques. Volume 8, Issue 1. Retrieved from www.ijera.com/papers/Vol8_issue1/Part-2/C0801020913.pdf

[20] Vijayakumar, Kavin P.A., Manivel S., Karthikeyan L. (2019). Diabetes Prediction by Machine Learning Over Big Data from Healthcare Communities. Volume 06, Issue 04. Retrieved from www.irjet.net/archives/V6/i4/IRJET-V6I4918.pdf

[21] Vinodini, S.V. & Saravanan, S. (2018). Data mining techniques using e-health information for diabetes disease prediction. International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE), Volume 25 Issue 7. Retrieved from www.ijetcse.com/wp-content/plugins/ijetcse/file/upload/docx/213Data-mining-techniques-using-e-health-information-for-diabetes-disease-prediction-pdf.pdf

[22] Warke M, Kumar V, Tarale S, Galgat P and Chaudhari DJ (2019). Diabetes Diagnosis using Machine Learning Algorithms. International Research Journal of Engineering and Technology (IRJET) 6(3): 1470 – 1476 retrieved from https://www.irjet.net/archives/V6/i3/IRJET-V6I3277.pdf

[23] Yamini, C., Robert, L., Gloria, L., Luebering, J., & Grace G. (2019). Diabetes mellitus. Retrieved from www.britannica.com/science/diabetes-mellitus/Insulin-therapies.