IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125141

By Using Machine Learning Algorithms we can predict and classify the diabetes mellitus

K. Harsha vardhan[1], Mohd Maaz muntajib [1], S. Sai Kiran[1], Dr. G. Shyama Chandra Prasad[2]

Department of Information Technology, Matrusri Engineering College Hyderabad[1]

Faculty of Information Technology Matrusri Engineering College Hyderabad[2]

Abstract- The many advancements in the healthcare and technological infrastructure has developed in bioscience which led to an incredible production of healthcare of critical and sensitive data. By using various data techniques and many patterns for identify and early onset detection and prevent many fatal diseases.

Diabetes mellitus is a most occurring life threatening disease because it also cause to other l diseases, i.e., heart, kidney, liver and nerves system damage. Here, few machine learning algorithms based approaches has been proposed for the categorization, for early stage prediction and identification of diabetes.

For diabetes classification, we used six different classifiers they are i.e., Logistic Regression, KNN Classifier, Naïve. Bayes classifier, SVM classifier, Decision Trees, random forest classifier. We will be using a prediction of diabetes model for better classification of diabetes which consists few different factors required for diabetes along with basic factors based as Glucose, Body Mass Index, Age, Insulin, etc.

Training and testing will be done to get the possible accurate results on the data set considered.

Index Terms- Logistic Regression , KNN Classifier, naïve bayes classifier, SVM classifier, decision tree classifier, random forest classifier.

INTRODUCTION

"Diabetes mellitus" is Most commonly referred as diabetes which is a constantly recurring disease related with abnormal higher levels of the sugar glucose in blood. Diabetes is due to the two mechanisms which is Inadequate production of insulin (that is made by the pancreas and lowers blood glucose), or not enough sensitivity of cells to the action of insulin. Diabetes mellitus has may develop as a secondary condition connect to another disease, equally pancreatic disease, a genetic syndrome, such as myotonic dystrophy, similarly glucocorticoid, which causes by the hyperglycemia.

The Suffering of hyperglycemia causes intensify thirst, increase starvation and regular urine are the symptoms of diabetes. Many issues will occur if diabetes remains untreated. The early identify is the only remedy to beware from the complications. Machine learning algorithms gain the strength due to the capacity of managing a huge amount of data to merge the data from many different sources and accommodating the background details.

By applying machine learning and data mining methods in Diabetes Mellitus, research is a key approach to make use of large volumes of available diabetes related data for extracting knowledge. The severe social impact of the specific disease renders DM one of the main priorities in medical science research, which inevitably produces high amounts of data. Data mining represents a significant advance in the type of analytical tools. It has been shown that the benefits of introducing data mining into medical analysis are to increase diagnostic accuracy and minimise costs to save human resources.

This study motive is to compare the performance of the most effective machine learning techniques, used to prognosticate diabetes diseases. In this works will be using diabetes dataset that collected from the dataset from the hospital Frankfurt. The dataset contains information about 2000 patients and their corresponding nine unique attributes. Algorithms used for datasets to predict diabetes are Naïve Bayes (NB), Random Forest (RF), K-Near Neighbour Classification algorithms (KNN), SVM, Logistic Regression.

However, according to the obtained outcome it is observed that the proposed model with SVM is achieved an excellent result of accuracy during the comparison with a rest classification algorithm that using in the proposed system.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125141

LITERATURE SURVEY

1 In Pisapia et AI used image analysis and ML for the prophecy of Hydrocephalus. They used the cerebral ventriculomegaly and draw out 77.5 figuring attributes. ML algorithms SVM classfier was put in on the ventricular profile of 25 children. The question is who needs shunts and who does not need? The outcome is acquired and compared. Consequences tells that every 6 out of 8 children need shunts with 76 percent sensitivity and 94 percent specificity.

2 In this the authors was predicated the diabetes types, complications and what are the treatments can be provide to the sufferers. Predictive analysis algorithm and Hadoop map reduce was used for the prediction and the treatment . Larger data sets is gathered from different labs, clinics, EHR and PHR the process is done in Hadoop and distributed among the different porters according to the graphical places.

3 In the bygone work of Cherradi B., et. al. to predict with or without type 2 diabetes mellitus patients, researchers used and evaluated four ML algorithms they are Artificial CNN, K Nearest Neighbours, Decision Tree classifier, and Deep Neural Network. The first was retrieved from the Germany Frankfurt Hospital while the second is a well-known dataset of Pima Indian, which contains the same feature composed of risk factors, mixed data, and some clinical data. The proposed model achieved the best accuracy rate with KNN 97.53% and Deep NN 96.35%.

4 In a healthcare forecast method is based on Independence Bayes algorithms which is hand overed, The proposed system which discovers and hidden data related to different diseases are extracted from the database. The system lets users to allocate their health related troubles and then using Naive Bayes we find the exact problem.

5 Simi et al. range over the significance of initial finding of females infertility in their testing work use 28 variables and 9 classes of female infertility, Results identified that Random Forest technique outperformed other techniques and provide 88% accuracy.

6 In the previous work of Maniruzzaman M., et. al. built a diabetic patient prediction algorithm based on machine learning (ML). Logistic Regression (LR) is used to classify risk factors for diabetes disease using p-values and odds ratios (OR). To forecast diabetic patients, they used four classifiers: Naive Bayes, decision tree , Ada boost (AB), and Random Forest (RF). These protocols were also followed and replicated in 20 trials by three groups of partition protocols (K2, K5, and K10). The accuracy (ACC) and region under the curve (AUC) of these classifiers are used to assess their performance (AUC) The ACC of the ML-based method as a whole is 90.62 %. The K10 protocol has a 94.25 % ACC and 0.95 AUC thanks to a mix of LR-based feature collection and RF-based classifier.

7 These studies have mainly focused on obtaining acceptance accuracy rates to diagnose and detect diabetic patients using different classification methods. Unlike these approaches, the focus of the proposed model is to design a diabetes classification model based on 5 advanced ML Algorithms to achieving higher accuracy from them. These images are often sourced from publicly available datasets.

SOFTWARE REQUIREMENT SPECIFICATIONS:

HARDWARE REQUIREMENTS

- Machine with Windows or Linux platform
- RAM 4 GB or above
- GPU for training model.
- CPU: 2 GHz or faster.
- Architecture: 32-bit or 64-bit

SOFTWARE REQUIREMENTS

- Python IDE
- Google Colab
- Jupyter Notebook

Python Libraries:

- NumPy
- Matplotlib
- Sklearn

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

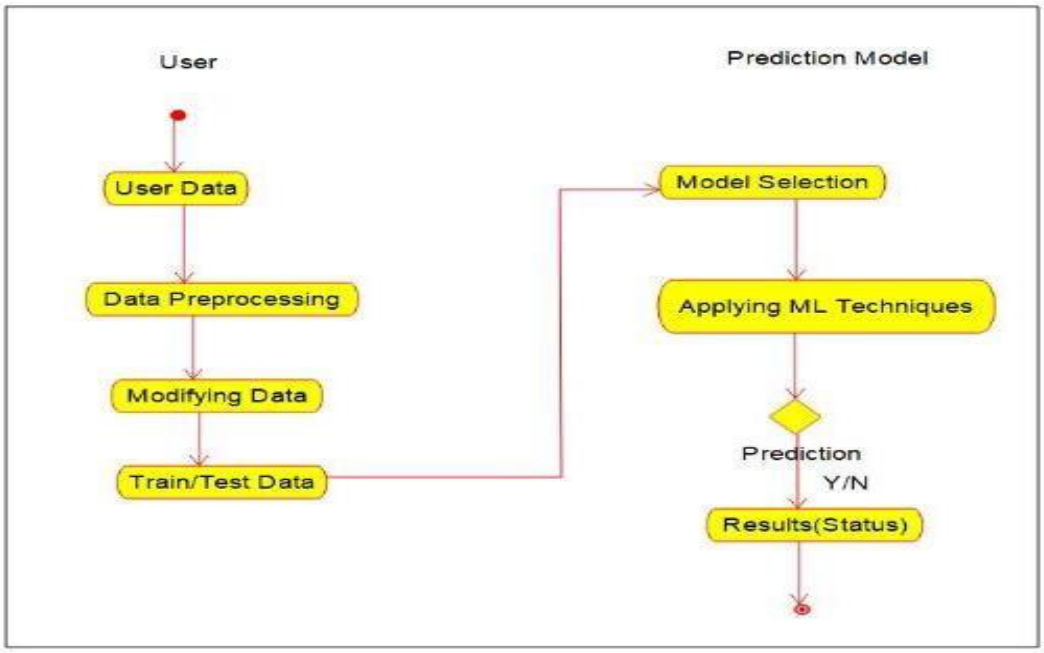
DOI: 10.17148/IJARCCE.2023.125141

A. Data Set:

We have taken the dataset from Kaggle which consists of 768 rows and 9 columns or fields namely: Glucose, BP, heart attack, Functions, lifespan . Outcome: A Sample of 10 rows is shown in above table.

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1

B. Data Preparation



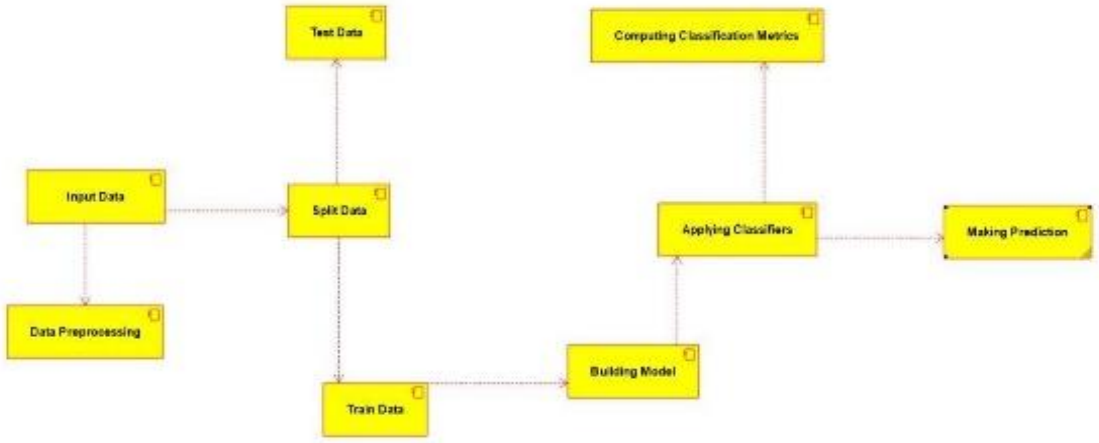
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

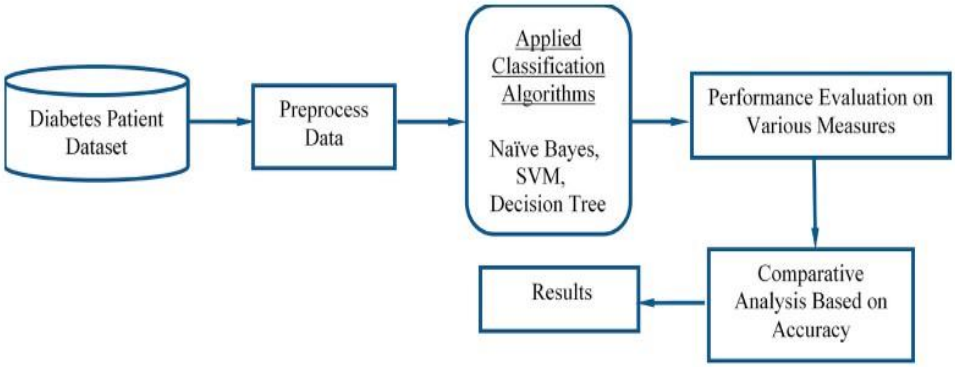
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125141

C. Pre-processing:



D. Proposed Model:



LR Classifier:

- LR is one of the most used ML algorithms, which comes under the Supervised Learning(SL) technique. It is used for forecast the categroized dependent variable using a given set of unanimous variables.

o LR forecast the output of a categorize dependent variable. Therefore, the outcome must be a categorize or discreted value. It can be either Yes/No, 0/ 1 and true/False, etc. but instead of giving the correct value as0 or 1, it gives the probabilistic values which lie between 0 and 1.

o LR classifier is much same to the Linear Regression behalf of that how they were using it. The LR is used for solving Regression problems, whereas the Logistic regression is used for solving the classification of problems.

o In LR , instead of fitting a regression line, we can set a "S" shape logistic function in it allow forecast two maximum values 0 or 1.

o The curve from the logistic function indicates the likelihood of something such as that the cells are cancerous or not and a mouse is obese or not based on its weight, etc. o Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. o Logistic Regression can be used to classify the observations using different types of data and can be easily determine the most effective variables used for the classification. The below image will shows the logistic function.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125141

Sigmoid Function:

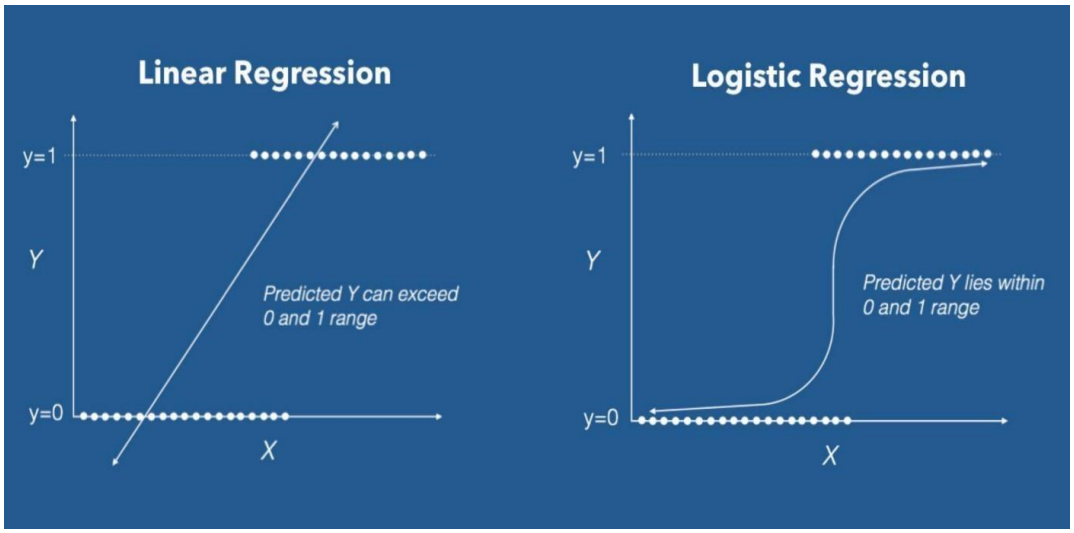
o Sigmoid function is a mathematical function that used to map the predicted values of probabilities.

o This maps real value into another value within a range of (0 to 1).

o The value of the logistic regression must be between (0 and 1), which cannot go beyond this limit, so it forms a curve like the "S" form. The S form curve is called the logistic function.

o In LR, we use the concept of the threshold value, which defines the probability of either(0 or 1). Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

The following graph shows the regression curves.



PROPOSED ALGORITHM AND ANALYSIS

Table 1. The calculations using confusion matrix:

We can perform different types of calculations for this model, such as the model accuracy, using this matrix. The calculations are given below:

o Classification Accuracy: It is the most main parameters to describe the accuracy of the classification model. It provides how often the model predicts the accurate outcome. It can be calculated as the ratio of the number for exact predictions done by the classifier to all number of forecast done by the classifiers.

o The formula is

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+FN+TN}$$

o Error rate: It is also known as misclassification rate, and it shows how often the model gives the incorrect forecasts. The error rate of given value can be calculate as the number of wrong predictions to all number of the predictions that made by the classifier.

- The formula is

$$\text{Error rate}=\frac{FP+FN}{TP+FP+FN+TN}$$

- Precision : It shows the number of outputs given correctly by the method or out of all practical classes that have forecast exactly by the model and how many of them were correct .

- This will be calculated by using the formulae below:

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125141

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

- Recall : It is defined as the out of total positive classes and how our method is predicted correctly. The memory should be as huge as possible.

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

CONCLUSION

- One of the most important real-world medical problems is to identifying of diabetes at its early stage.
- During the work, six ML algorithms are used have studied and accessed on various factors.
- This Experiments are done on Pima Indian Diabetes Dataset .This Experiment shows and determines the adequacy of the design system with an achieved accuracy using the algorithms.
- On our Project we happened to get the More Accuracy for Support Vector Machine for the taken Dataset when compared with all other five Algorithms. It can be used to predict whether a person is suffering from gestational diabetes or not.
- We can predict whether a person is suffering from diabetes early with the help of some attributes like Glucose level, Insulin, BMI, Age etc.
- In future, with the help of ML algorithms we can used to forecast or diagnose many other different diseases.
- This task can be expanded and upgraded for the automatically work of diabetes analysis by using Ensemble ML algorithms.

ACKNOWLEDGMENT

- The satisfaction of completing this project would be incomplete without mentioning our gratitude towards all the people who have supported us. Constant guidance and encouragement have been instrumental in the completion of this project.
- First and foremost, we thank the Chairman, Principal, Vice Principal for availing infrastructural facilities to complete the mini project in time.

REFERENCES

[1] P. A. Chiarelli, J. S. Hauptman, and S. R. Browd, “Machine Learning and the Prediction of Hydrocephalus,” JAMA Pediatr., vol. 172, no. 2, p. 116, Feb. 2018.
[2] N. M. S. kumar, T. Eswari, P. Sampath, and S. Lavanya, “Predictive Methodology for Diabetic Data Analysis in Big Data,” Procedia Comput. Sci., vol. 50, pp. 203–208, Jan. 2015.
[3] Cherradi B.,et., 2014. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing 1, 763– 770. doi:10.1007/978-3-319-11933-5.
[4] International Journal of Advanced Computer and Mathematical Sciences. Bi Publication-BioIT Journals, 2010
[5] M. S. Simi, K. S. Nayaki, M. Parameswaran, and S. Sivadasan, “Exploring female infertility using predictive analytic,” in 2017 IEEE Global Humanitarian Technology Conference (GHTC), 2017, pp. 1–6.
[6] Maniruzzaman, M., Rahman, M. J., Ahammed, B., & Abedin, M. M. (2020).
[7] Han, J., Rodriguez, J.C., Beheshti, M., 2008. Discovering decision tree-based diabetes prediction model, in: International Conference on Advanced Software Engineering and Its Applications, Springer. pp. 99–109.
[8] https://medium.com/analytics-vidhya/na%C3%AFve-bayes-algorithmhttps://medium.com/analytics-vidhya/na%C3%AFve-bayes-algorithm-5bf31e9032a25bf31e9032a2
[9] https://www.javatpoint.com/machine-learning-random-forest-algorithm
[10] https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

AUTHORS:

First Author – K. Harsha Vardhan , B.E, Matrusri Engineering college , hv222666@gmail.com
Second Author – Mohd Maaz Muntajib, B.E, Matrusri engineering college, muntajib110@gmail.com
Third Author – Swarna Sai Kiran, B.E, Matrusri Engineering college, swarnasaikiran113@gmail.com.
Correspondence Author – Dr. G. Shayama Chandra Prasad, 9849244986.