"MACHINE LEARNING ON DIABETES MANAGEMENT: EMPLOYABILITY OF ADVANCED LOGISTIC REGRESSION AND PREDICTIVE ANALYSIS IN EARLY DETECTION OF DIABETES"

Sidharth Grover

ABSTRACT

In general sugar fluctuation expansion in the blood is termed as Diabetics. Various diagnosis methods are already carried out to address this issue in real life. Still, it has a research gap to furthermore improve the performance. To receive higher performance two concepts are introduced in this paper. They are Preprocessing and Predictive Analysis. For the concept of Preprocessing several algorithms are used. They are Logistic Regression, Decision Tree Classifier, Linear Discriminant Analysis, KNN Classifier, GNB and SVM. For predictive analysis, Advanced Support Vector Machine is used. While compared with the earlier method our proposed methods provide high accuracy score, high confusion matrix, high classification reports as well as high average & total prediction values.

Keyworks: *Logistic Regression, Decision Tree Classifier, Linear Discriminant Analysis, KNN Classifier, GNB and SVM.*

INTRODUCTION

Nowadays, in rising kingdoms such as India, Diabetic Mellitus (DM) has to turn out to be a large physical condition exposure. To recognize disease as well as its related menace, Big-data analytics know how to be used at early stages, hence; preserve to make available medical care. DM has been developed into an immense physical condition risk in India based on today's situations. Diabetic Mellitus (DM) is categorizing as a Non-Communicable Diseases (NCD), as well as lots of people are suffering from it. DM is classified into 3 types namely, Type 1, Type 2 and gestational diabetes. The reason owed to patient's body's lack of ability to generate insulin as well as presently, it entail inject insulin to a person in Type 1.

The above-mentioned type is referred to as Insulin - Dependent Diabetes Mellitus (IDDM). The body cells are not capable to use insulin appropriately owed to the cause in Type 2 DM, at the same time we can say Type 2 DM as Non-Insulin-Dependent Diabetes Mellitus (NIDDM). Finally, owing to the growth of high blood sugar in heavy with child devoid of preceding analysis of DM can be caused by gestational diabetes [1].

Data mining technique is the most popular one which is used to predict and analyze the diabetes mellitus. There are many algorithms to predict diabetes and also some of them giving awareness by using mobile applications. In most of the research, they used only a few algorithms which are using frequently and fully focused on accuracy only. The following algorithms are used to enhance the accuracy as K-Nearest Neighbor (KNN), hybrid Adaptive Neuro-Fuzzy Inference System (ANFIS), Bayesian algorithm, J48, Random forest and so on. We are using ML algorithm for the diabetes prediction types which are widespread. In this method, we can able to provide the assurance level of early diagnosis of a patient's risk level.

To transaction in the midst of extroverted dataset, the diabetes estimation organization chips have gone next to a modest dataset generally at the elevated section of the papers. To implement the organization, we have to focus on retreat the therapeutic test, as it may be needed by the extent of a medicinal test. For the prediction of diabetes, the factors or else the feature has been taken in the organization. The requirement of a therapeutic testing system will be overcome by our prospect scheme determination acquire a turn next to a larger dataset [3].

The collection of a database for a person who is having diabetes is done by using data mining as well as ML techniques. In future, there are many different algorithms have taken to enhance and improve the competence of the organization. As well as functioning on with various algorithms, we can easily tackle diabetes. Also, we can get the missing databases by using an algorithm such as Hadoop-Map Reduce environment. Diabetes has been categorized into two types namely, Type 1 and Type 2. In Type 1, initially, they will analyze the database for diabetic patients. Hence, after analyzing the database, we can find the relevant and irrelevant features for the diagnosis of diabetes. In Type 2, they will solve the problem of regression and analyze the distance between varieties of input data points. The training process is much better when compared to others. Furthermore, the results can be obtained and the performance is good, as well as the accuracy will be analyzed.

Based on the algorithms which have been discussed above, we can predict and analyze the diabetes mellitus. At the same time, we can found out the missing values by using a machine learning algorithm.

LITERATURE SURVEY

Gauri D. Kalyanakar et al. (2017) used a technique for a Diabetic patient as Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop. They used an algorithm to predict the types of diabetes are widespread as well as to get the missing values in it by using Hadoop-Map Reduce environment. In this approach, we can use the Hadoop Distributed File System (HDFS) which is to store as well as process the data. By using this algorithm, we can access the massive data with less in time as well as efforts too. From the results of the experiments, we can easily predict the data's which is lost and to discover the patterns. As well as we can also find the types of diabetes. [1]

VeenaVijayan V and Anjali C (2015) proposed a technique for the diagnosis of Diabetic patients such as Prediction and Diagnosis of Diabetes Mellitus – A Machine Learning Approach. They used this technique to provide high accuracy based on the AdaBoost algorithm. We can collect the local dataset by using the relating mean value as a part of the global dataset. As well as we can train and validate the dataset collection by using four base classifiers as a Decision tree, Support Vector Machine, Native Bayes and Decision stump. After that, we can also calculate Body Mass Index (BMI) by using the height and weight of a person. By analyzing all these techniques as well as by using the AdaBoost algorithm, we can easily get the performance accuracy, sensitivity, specificity and also error rate. [2]

Deeraj Shetty, KishorRit, Sohail Shaikh and Nikita Patil (2017) introduced a system based on Diabetes Disease Prediction Using Data Mining. In this proposed system, they used an algorithm to apply on diabetes patient's database as well as analyze them for the prediction of diabetes diseases. This process is mostly related to the hospital's report. The admin of the system will collect the personal data's about diabetes patient's so that we can easily analyze the report, as well as the whole pieces of information, will provide to the patient's as a report which is likely given in the hospital and for diabetes, this factor is mostly responsible. The available prediction is performed by utilizing the algorithm as the Bayesian and K-NN algorithm. By the usage of an algorithm, we can conclude and analyze the performance of the diabetes patient's database using data mining. [3]

Aparimita Swain, SachiNandanMohanty and Ananta Chandra Das (2016) discussed Comparative Risk Analysis on Prediction of Diabetes Mellitus Using Machine Learning Approach. This system proposed an algorithm as Artificial Neural Network (ANN) and hybrid Adaptive Neuro-Fuzzy Inference System (ANFIS). By the usage of these two algorithms, we can predict the diabetes mellitus as well as validates accuracy prediction. In this research, ANFIS is much better when compared to ANN and also the accuracy performance of the ANFIS is good when compared to ANN. ANFIS system presentation is considered by Root Mean Squared Error (RMSE). By using RMSE we can reduce the error in the given samples. To increase the accuracy of the neural network, the number of an unseen coating as well as concealed neurons can be enlarged. [4]

Mukesh kumara, DrRajan Vohra and Anshul Arora (2014) proposed a system based on Prediction of Diabetes Using Bayesian Network. Here, to predict the diabetic persons or not by using the Bayesian Network. WEKA (Waikato Environment for Knowledge Analysis) tool is used to do collect the information of persons from the hospital and analyze that the person is having diabetes mellitus or not. An algorithm is applied as Classification algorithm which is to collect the dataset of the persons from the hospital. As well as, by applying data mining technique, we can easily predict the diabetes mellitus. Weka is an advantage tool when compared to others because it is based on the collection of machine learning algorithm as well as easy to solve real-world data mining problems and the results are obtained. [5]

Ms K Sowjanya, DrAyushSinghal and MsChaitaliChoudhary (2015) introduced a test which is called MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices. To overcome the deficiency of awareness based on diabetes, here we used a mobile/android application based solution. The above application is to predict the diabetes level for uses by using novel machine learning techniques. Here the Decision Tree (DT) of machine learning algorithm is second-hand to propose the equipment used for the mobile/android application in favor of diabetes forecast by real-world dataset collections. As well as, we can give awareness about the disease and by the result, we can easily predict the database analysis. [6]

Vrushali R. Balpande and Rakhi D. Wajgi (2017) developed a system based on Prediction and Severity Estimation of Diabetes Using Data Mining Technique. Here they introduced Elcatalgorithm which is used to calculate the control and un-control condition of diabetes. By including data set, the also identify the parameters like BMI, HbAIC, FBG, PMBG. Then estimate the ranges for each test and record it. After that by comparing the range with the standard datasets forming Prediction of severity estimations on organs such as heart, kidney, eye, etc., Elcat algorithm entail less time for prototype generation than apriori and it is well apt intended for petite datasets. The result shows the performance of the estimation. [7]

Raid M. Khalil and Adel Al-Jumaily (2017) discussed Machine Learning Based Prediction of Depression among Type 2 Diabetic Patients. They used four techniques as, support vector machine (SVM), K-Mean, F-Cmean and the Probabilistic Neural Network (PNN). By using SVM, we can solve the classification as well as the regression problem and it is less prone when compared to over fitting. The aim of the K-means algorithm is to partition n observations into k clusters. And also it is used towards segregate an exact mysterious dataset keen on a constant number (k) of the clusters. Fuzzy C-Mean is used to analyze the distance among an assortment of input data points. PNN is a trendy optimization procedure which is used to diminish factual error among the factual as well as envisage yield. Its training process is fast when compared to Back-Propagation Neural Networks and the results are obtained by the performance of these algorithms. [8]

Rahul Joshi and MinyechilAlehegn (2017) proposed a theory based on Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. To make this system as an ensemble hybrid model, the following algorithms are used: KNN, Nave Bayes, Random forest and J48 which is used to increase the performance and accuracy. J48 is one of the most popular as well as better accuracy. All these algorithms are used to enhance the accuracy and all these are advanced when compared to others. The random forest provides better accuracy than J48 as well as Nave Bayes in 10 cross-validation splitting method. The fuzzy rule was developed to reduce the wrong treatment. We can analyze the performance by using the result of this proposed theory. [9]

Dr D. Asir Antony Gnana Singh, Dr E. JebamalarLeavline, B. ShanawazBaig introduced a system as Diabetes Prediction Using Medical Data. Here their aim is to improve the accuracy in diabetes prediction. Initially, a dataset is given to the data pre-processing module. The pre-processing module gives the dataset with relevant features only to the machine learning algorithm. For removing irrelevant features, correlation-based feature selection technique is used. Here the pre-processing technique is used to enhance the accuracy of the model. Finally, the machine learning algorithm develops a learning model from the pre-processed dataset and we can say this learning model as a knowledge model. Furthermore, the results are obtained as well as the person’s medical report was predicted by using the learning model. [10]

PROPOSED METHOD

PREPROCESSING

The Algorithms which are used for the process of preprocessing are Logistic Regression, Decision Tree Classifier, Linear Discriminant Analysis, KNN Classifier, GNB and SVM. The concepts of all the algorithms are given below.

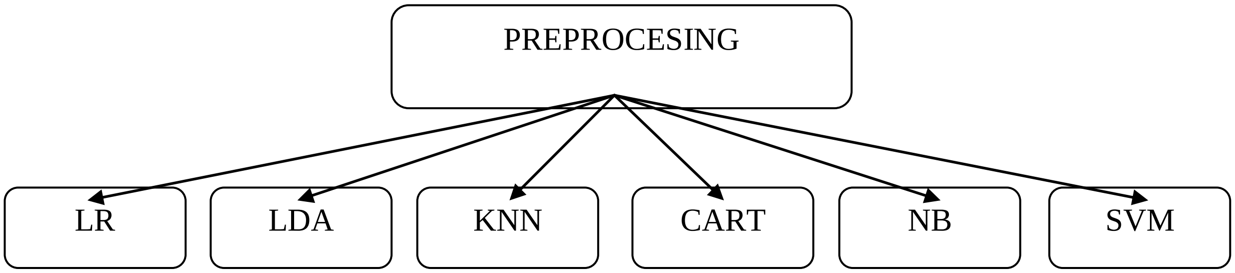


Figure 1 – Preprocessing Algorithms

PREDICTION

For the process of Prediction, the Advanced Support Vector Machine method is used. The concept of this algorithm is given below.

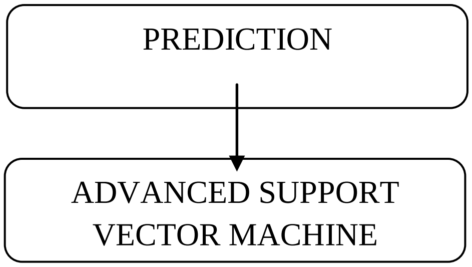


Figure 2 – Prediction Algorithm

RESULTS AND DISCUSSION

The parameters which are taken for the process of prediction of diabetics are Times of pregnant, Plasma glucose concentration, BP (mm Hg), Triceps skin fold thickness (mm), Insulin (mu U/ml), BMI (weight in kg/(height in m)^2), Diabetes pedigree function, Age and results. The Histogram values for the input data are given below.

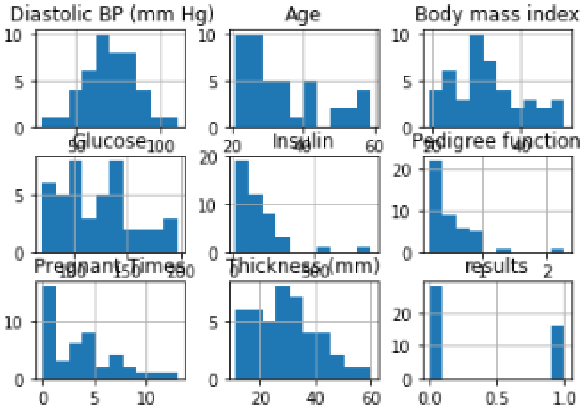


Figure 3 – Histogram of Input Data

For the process of preprocessing, the used techniques are Logistic Regression, Decision Tree Classifier, Linear Discriminant Analysis, KNN Classifier, GNB and SVM. For preprocessing, ten folded matrix method is used for calculation. The comparative results of all the above-mentioned algorithms are given below.

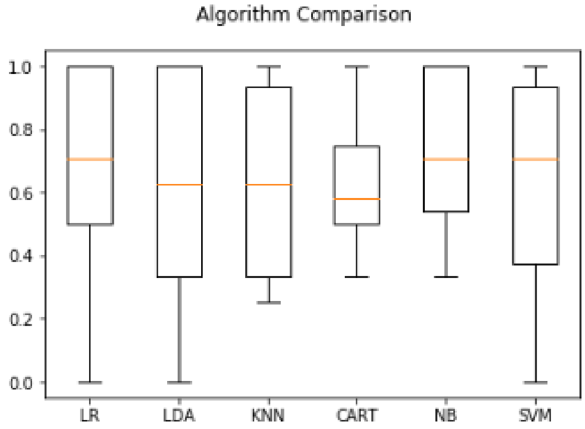
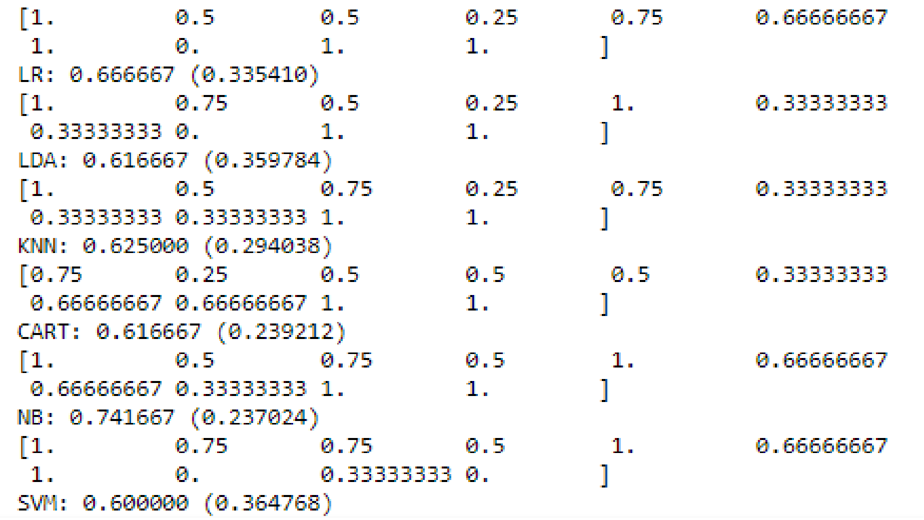


Figure 4– Comparative Results for Preprocessing

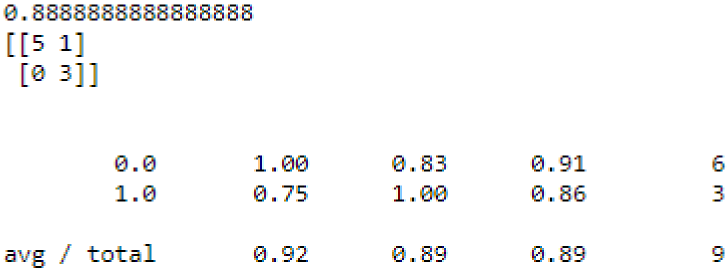
The 10 folded matrixes provide ten different values as preprocessed values. From that value, the final preprocessed values are calculated by two methods. The first method is the selection of preprocessed values from the repeated values of 10 folded matrixes. Otherwise secondly the average value of 10 folded matrixes can also be calculated as the preprocessed values. The ten values of all the algorithms with the finalized preprocessed values are given below.



```
[1.         0.5        0.5        0.25       0.75       0.66666667
 1.         0.         1.         1.        ]
LR: 0.666667 (0.335410)
[1.         0.75       0.5        0.25       1.         0.33333333
 0.33333333 0.         1.         1.        ]
LDA: 0.616667 (0.359784)
[1.         0.5        0.75       0.25       0.75       0.33333333
 0.33333333 0.33333333 1.         1.        ]
KNN: 0.625000 (0.294038)
[0.75       0.25       0.5        0.5        0.5        0.33333333
 0.66666667 0.66666667 1.         1.        ]
CART: 0.616667 (0.239212)
[1.         0.5        0.75       0.5        1.         0.66666667
 0.66666667 0.33333333 1.         1.        ]
NB: 0.741667 (0.237024)
[1.         0.75       0.75       0.5        1.         0.66666667
 1.         0.         0.33333333 0.        ]
SVM: 0.600000 (0.364768)
```

Figure 5 –10 Folded Values and Finalized Preprocessed Values of all the algorithms

The next process is the prediction. Here for prediction process, Support Vector Machine method is used and the prediction values such as accuracy score, confusion matrix, classification reports as well as Average and Total prediction values are given below.



```
0.8888888888888888
[[5 1]
 [0 3]]


        0.0       1.00      0.83      0.91         6
        1.0       0.75      1.00      0.86         3

avg / total       0.92      0.89      0.89         9
```

Figure 6 - Accuracy Score, Confusion Matrix, Classification Reports, Average and Total prediction values

From the figure, we identified that the accuracy score is 0.92%, the confusion matrix is 0.89%, classification reports is 0.89% and the Average & Total Prediction is 90%.

CONCLUSION

The maximum score for the predictive analysis and diagnosis for diabetics are received by the above-mentioned process called preprocessing and Prediction. Here, 10 folded matrix concepts are used for the confusion matrix in the process of preprocessing. Therefore by using test and training datasets, maximum performance of preprocessing is achieved which leads to improving the prediction ratio. And also by using the advanced support vector machine concept the average and total prediction percentage are improved.

REFERENCES

1. Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar “Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop”, International conference on I-SMAC 2017.
2. VeenaVijayan V. And Anjali C. “Prediction and Diagnosis of Diabetes Mellitus – A Machine Learning Approach”, IEEE Recent Advances in Intelligent Computational Systems (RAICS) 2015.
3. Deeraj Shetty, Kishore Rit, Sohail Shaikh and Nikita Patil “Diabetes Disease Prediction Using Data Mining”, International Conference on Innovatiojns in Information, Embedded and Communication Systems (ICIIECS) 2017.
4. Aparimita Swain, SachiNandanMohanty and Ananta Chandra Das “Comparative Risk Analysis on Prediction of Diabetes Mellitus using Machine Learning Approach”, International Conference on Electrical, Electronics and Optimization Techniques (ICEEOT) 2016.

5. MukeshKumari, Dr.Rajan Vohra and Anshularora "Prediction of Diabetes Using Bayesian Network", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5(4), 2014.
6. Ms. K Sowjanya, Dr.AyushSinghal and Ms.ChaitaliChoudhary "MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices", IEEE International Advance Computing Conference (IACC) 2015.
7. Vrushali R. Balpande and Rakhi D. Wajgi "Prediction and Severity Estimation of Diabetes Using Data Mining Technique", International Conference on Innovative Mechnaism for Industry Applications (ICIMIA 2017).
8. Raid M. Khalil and Adel Al-Jumaily "Machine Learning Based Prediction of Depression among Type 2 Diabetic Patients", 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE).
9. Rahul Joshi and MinyechilAlehegn "Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach", International Research Journal of Engineering and Technology (IRJET) 2017.
10. Dr. D. Asir Antony Gnana Singh, Dr. E. JebamalarLeavline, B.ShanawazBiag "Diabetes Prediction Using Medical Data", Journal of Computational Intelligence in Bioinformatics ISSN 0973-385X Volume 10, Number 1 (2017).
11. Dr.Saravanakumar, Eswari, Sampath, Lavanya "Predictive Methodology for Diabetic Data Analysis in Big Data," ELSEVIER, ISBCC 2015.
12. R. Bellazzi and B.Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines", International Journal of Medical Informatics, vol 77, pp 81-97, 2008.
13. Y. Cai, D. Ji,D. Cai, "A KNN Research Paper Classification Method Based on Shared Nearest Neighbor", Proceedings of NTCIR-8 Workshop Meeting, 2010.
14. A.M. Aibunu, M.J.E. Salami and A.A.Shafie "Application of Modelling Techniques to Diabetes Diagnosis", IEEE EMBS Conference on Biomedical Engineering & Sciences (IECBES),vol.8, pp.194-198,2010.
15. Rohanizadeh.s "A proposed data mining methodogy application to industrial procedures".
16. SeemaAbhijeetKaveeshwar and Jon Cornwall, "The current state of diabetes mellitus in India," Australas Med J; PMCID: PMC3920109, pp: 45–48, January 2014.
17. G.J. Simon,P. J. Caraballo,T. M. Therneau,S. S. Cha, M. Regina Castro and Peter W.Li, "Extending Association Rule Summarization Techniques to Assess Risk Of Diabetes Mellitus," IEEE Transactions on Knowledge and Data Engineering, vol.27, no.1, January 2015.
18. Habtewold, T.D., Alemu, S.M. & Haile, Y.G. 2016, 'Sociodemographic, clinical, and psy-chosocial factors associated with depression among type 2 diabetic outpatients in Black Lion General Specialized Hospital, Addis Ababa, Ethiopia: a cross-sectional study', BMC psychiatry, vol. 16, no. 1, p. 103.

19. Song, Y., Liang, J., Lu, J., & Zhao, X. (2017). An efficient instance selection algorithm for k nearest neighbour regression. Neurocomputing, 251, 26-34.
20. Kaveeshwar, S.A., and Cornwall, J., 2014, "The current state of diabetes mellitus in India". AMJ, 7(1), pp. 45-48.