Diabetes Prediction Model

Harshit Pandey | Sachin Garg

Department of Information Technology, MAIT, New Delhi, India

ABSTRACT

Diabetes is an illness caused because of high glucose level in a human body. Diabetes should not be ignored if it is untreated then Diabetes may cause some major issues in a person like: heart related problems, kidney problem, blood pressure, eye damage and it can also affects other organs of human body. Diabetes can be controlled if it is predicted earlier. To achieve this goal this project work we will do early prediction of Diabetes in a human body or a patient for a higher accuracy through applying, Various Machine Learning Techniques. Machine learning techniques Provide better result for prediction by constructing models from datasets. In this work we will use Machine Learning Classification and ensemble techniques on a dataset to predict diabetes. Which are Logistic Regression (LR), Support Vector Machine (SVM), Naïve Bayes and Random Forest (RF). We will compare the accuracy of all the above models and predict the result on the basis of accuracy of every model.

INTRODUCTION

Diabetes is an illness caused because of high glucose level in a human body. Dia- betes should not be ignored if it is untreated then Diabetes may cause some major issuesinapersonlike:heartrelatedproblems,kidneyproblem,bloodpressure,eye damage and it can also affects other organs of human body. Diabetes can be controlled if it is predicted earlier. To achieve this goal this project work we will do early prediction of Diabetes in a human body or a patient for a higher accuracy through applying, Various Machine Learning Techniques. Machine learning tech- niques Provide better result for prediction by constructing models from datasets.In thisworkwewilluseMachineLearningClassificationandensembletechniqueson a dataset to predict diabetes. Which are Logistic Reg-ression (LR), Support Vector Machine (SVM), Naïve Bayes and Random Forest (RF). We have been compared the accuracy of all the above models and predict the result on the basis of accuracy ofeverymodel.Afterthecomparingacc-uracyofallthemodelwefoundthatlogistic regression classifier gives more accuracy which is82%.

OBJECTIVES
Algorithms

1.1 Logistic Regression: Logistic regression is also a supervised learning classifi- cation algorithm. It is used to estimate the probability of a binary response based on one or more predictors. They can be continuous or discrete. Logistic regression used when we want to classify or distinguish some data items into categories.

1.2 Support Vector Machine: Support Vector Machine also known as svm is a supervised machine learning algorithm.Svm is most popular classification technique. Svm creates a hyperplane that separate two classes. It can create ahyperplane or set of hyperplane in high dimensional space. This hyper plane can be used for classification or regression also. Svm differentiates

instances in specific classes and can also classify the entities which are not supported by data. Separation is done by through hyperplane performs the separation to the closest training point of any class.

1.3 Naïve Bayes: It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

1.4 Random Forest: It is type of ensemble learning method and also used for classification and regression tasks. This method can easily handle large datasets. Ran- dom Forest is developed by Leo Bremen. It is popular ensemble Learning Method. Random Forest Improve Performance of Decision Tree by reducing variance. It op- erates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the individual trees.

LITERATURE SURVEY

1.Deeraj Shetty et al. [1] proposed diabetes disease prediction using data mining as- semble Intelligent Diabetes Disease Prediction System that gives analysis of diabe- tes malady utilizing diabetes patient's database. In this system, they propose the use of algorithms like Bayesian and KNN (K- Nearest Neighbor) to apply on diabetes patient's database and analyze them by taking various attributes of diabetes for pre- diction of diabetes disease.

2.K.VijiyaKumar et al. [2] proposed random Forest algorithm for the Prediction of diabetes develop a system which can perform early prediction of diabetes for a pa- tient with a higher accuracy by using Random Forest algorithm in machine learning technique. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, efficiently and most importantly, instantly.

3.Tejas N. Joshi et al. [3] presented Diabetes Prediction Using Machine Learning Techniques aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, ANN. This project proposes anf- fective technique for earlier detection of the diabetes disease.

4.Nonso Nnamoko et al. [4] presented predicting diabetes onset: an ensemble super- vised learning approach they used five widely used classifiers are employed for the ensembles and a meta-classifier is used to aggregate their outputs. The results are presented and compared with similar studies that used the same dataset within the literature. It is shown that by using the proposed method, diabetes onset prediction can be done with higher accuracy.

5.[Muhammad Azeem Sarwar et al]. proposed study on prediction of diabetes using machine learning algorithms in healthcare they applied six different machine learn- ing algorithms Performance and accuracy of the applied algorithms is discussed and compared. Comparison of the different machine learning techniques used in this study reveals which algorithm is 5 best suited for prediction of diabetes.

METHODOLOGY

1.Planning and backgroundstudy

Before starting any research or project, basic information of the related topic is re- quired to ensure what the project is all about. In this stage, we have studied about algorithms from which we can predict diabetes in early stage.

2.LiteratureReview

In this stage, it involved the study of the previous research done related to the pro- ject.

3.Data gathering &Analysis

In this stage we will gather all the required data for our machine learning model to be work on. Data mainly consist of multiple attributes which will help us to deter- mining different conditions on which our model can work.
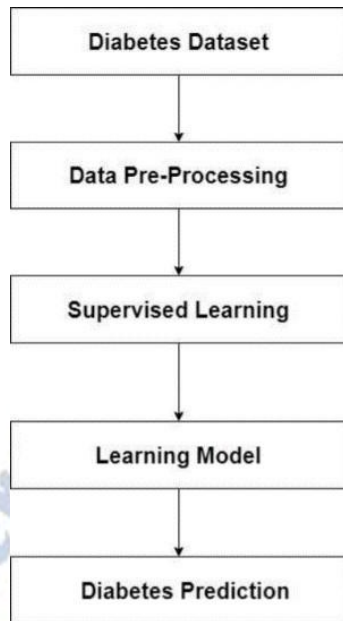
4.Implementing Machine Learning Algorithms

This section provides knowledge about the implementation environment and throws light on the actual steps for the implementation of dataset to get better accuracy to predict diabetes by usingdiffer- ent classifierscombination.

ARCHITECTURE

The system will detect whether the person has diabetes or not using the dataset. If diabetes is detected the classification value will be 1 and if not the value will be 0. We will be using 4 machine learning model in order to detect the disease. The mod- els used are logistic

regression, SVM, Naïve Bayes, Random forest he future too diagnosis diabetes. It will also predict whether someone has diabetes or not on the basis of using a trained dataset.

```
┌─────────────────────┐
│   Diabetes Dataset  │
└─────────────────────┘
           │
           ▼
┌─────────────────────┐
│  Data Pre-Processing│
└─────────────────────┘
           │
           ▼
┌─────────────────────┐
│  Supervised Learning│
└─────────────────────┘
           │
           ▼
┌─────────────────────┐
│   Learning Model    │
└─────────────────────┘
           │
           ▼
┌─────────────────────┐
│  Diabetes Prediction│
└─────────────────────┘
```

RESULTS

The results have been collected of all the used algorithms. The accuracy is different for each of the algorithm used.

Algorithms	Accuracy
Logistic Regression	82.46%
Random forest	81.16%
Kernel SVM	79.22%
Naïve bayes	79.22%

FUTURE SCOPE AND CONCLUSION

The project is aimed at building a "Disease Predication Model" in which it will focus on different diseases such as heart disease, kidney disease, liver disease, as of now it only focuses on diabetes detection. Diabetes is very deadly disease and there are many diseases like this in the field of healthcare and we are striven to contribute in this field by using our knowledge and making it easy for doctors to assist the patients deliberately. So, as we already said that, in this project, we use 4 different Machine Learning Algorithm and trained them on our dataset and compare them. And what we concluded that Kernel SVM & Naïve bayes gave the same accuracy of 79.22% but on the other hand Random Forest worked really well and gave the

accuracy of 81.16% and then Logistic Re- gression which worked even better than all the three of them and gave the accuracy of 82.46%. These models can be built into any online computer system, so that peo- ple can easily predict their diabetes using some commonly lab results with satisfac- tory sensitivity.

REFERENCES

[1] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Im- portance's for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

[2] K.VijiyaKumar, B.Lavanya, I.Nirmala, S.Sofia Caroline, "Random Forest Al- gorithm for the Prediction of Diabetes ".Proceeding of International Conference on Systems Computation Automation and Networking, 2019.

[3] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus". International Confer- ence on Electrical, Computer and Communication Engineering (ECCE), 7-9 Febru- ary, 2019.

[4] Tejas N. Joshi, Prof. Pramila M. Chawan, "Diabetes Prediction Using Machine Learning Techniques".Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-13

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

[6] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes Disease Pre- diction Using Data Mining ".International Conference on Innovations in Infor- mation, Embedded and Communication Systems (ICIIECS), 2017.

[7] Nahla B., Andrew et al,"Intelligible support vector machines for diagnosis of diabetes mellitus. Information Technology in Biomedicine", IEEE Transactions. 14, (July. 2010), 1114-20.

[8] A.K., Dewangan, and P., Agrawal, "Classification of Diabetes Mellitus Using Machine Learning Techniques," International Journal of Engineering and Applied Sciences, vol. 2, 2015.