Diabetes Disease Prediction Using Machine Learning Algorithms

ABSTRACT

This paper deals with the prediction of Diabetes Disease by performing an analysis of five supervised machine learning algorithms, i.e. K-Nearest Neighbors, Naive Baye, Decision Tree Classifier, Random Forest and Support Vector Machine. Further, by incorporating all the present risk factors of the dataset, we have observed a stable accuracy after classifying and performing cross-validation. We managed to achieve a stable and highest accuracy of 76% with KNN classifier and remaining all other classifiers also give a stable accuracy of above 70%. We analyzed why specific Machine Learning classifiers do not yield stable and good accuracy by visualizing the training and testing accuracy and examining model overfitting and model underfitting. The main goal of this paper is to find the most optimal results in terms of accuracy and computational time for Diabetes disease prediction.

**EXISTING SYSTEM**

In [2], they have used the WEKA tool for data analytics for diabetes disease prediction on Big Data of healthcare. They used the publicly available dataset from UCI and applied different machine learning classifiers on it. The classifiers which they incorporated are Naive Bayes, Support Vector Machine, Random Forest and Simple CART.

Their approach starts with accessing the dataset, preprocess it in Weka tool and then did the 70:30 train and test split for applying different machine algorithms. They did not go with the cross-validation step as it is imperative to get the optimal and accurate results as well.

The authors in [3], also used the publicly available dataset named as Pima Indians Diabetes Database for performing their experiment. Their framework of performing

the prediction starts with the dataset selection and then with data pre-processing. Once the data was preprocessed, they applied three classification algorithms, i.e.

naive Bayes, SVM and Decision tree. As they incorporated different evaluation metrics, they did compare the different performance measure and comparatively analyzed the accuracy. The highest accuracy achieved with their experiment was 76.30%. Like [2] they have also not practised Cross-validation.

In [4], the authors proposed the neural network-based diabetes disease prediction on Indians Pima Diabetes Dataset. They have used several hidden layers to find patterns in the data, and with the help of those patterns, they predicted the outcome. They name their proposed algorithms as ADAP, which is a custom neural network

with multiple partitions and with the set of association weights and units. They managed to achieve a crossover point for sensitivity, and specificity at 0.76 and are trying to precise their result in future.

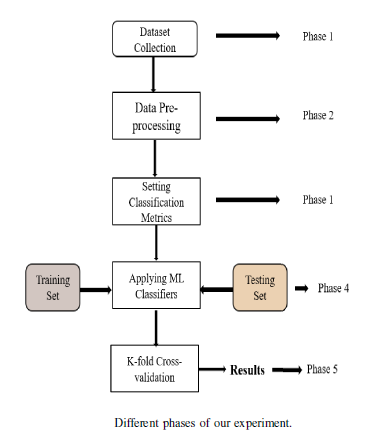
Disadvantages

1). There are no techniques and models for analyzing large scale datasets in the existing system.

2). There is no facility for diabetes dataset in collaboration with a hospital or a medical institute and will try to achieve better results.

**PROPOSED SYSTEM**

To perform our experiment, we have used a publicly available dataset named as Pima Indians Diabetes Database [4]. This dataset includes a various diagnostic measure of diabetes disease. The dataset was originally from the National Institute of Diabetes and Digestive and Kidney Diseases. All the recorded instances are of the patients whose age are above 21 years old. Our proposed model exists of 5 phases which are shown in the proposed system by following Figure.



**Advantages**

* The system more effective due to fitting datasets for different ML Models by Applying Machine Learning Algorithms.
* The Early determination of a disease can be made possible through machine learning by studying the characteristics of an individual in the proposed system.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV

➢ RAM - 4 GB (min)

➢ Hard Disk - 20 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**SOFTWARE REQUIREMENTS:**

* **Operating system :** Windows 7 Ultimate.
* **Coding Language :** Python.
* **Front-End :** Python.
* **Back-End :** Django-ORM
* **Designing :** Html, css, javascript.
* **Data Base :** MySQL (WAMP Server).