Abstract

Data mining is a process of extracting information from a dataset and transform it into understandable structure for further use, also it discovers patterns in large data sets. Data mining has number of important techniques such as preprocessing, classification. Classification is one such technique which is based on supervised learning. Diabetic is a life threatening disease prevalent in several developed as well as developing countries like India. The data classification is diabetic patients data set is developed by collecting data from hospital repository consists of 1865 instances with different attributes. The instances in the dataset are two categories of blood tests, urine tests. In this paper we discuss various algorithm approaches of data mining that have been utilized for diabetic disease prediction. Data mining is a well known technique used by health organizations for classification of diseases such as diabetes and cancer in bioinformatics research. In the proposed approach we have used WEKA with 10 cross validation to evaluate data and compare results. Weka has an extensive collection of different machine learning and data mining algorithms.
In this paper we have firstly classified the diabetic data set and then compared the different data mining techniques in weka through Explorer, knowledge flow and Experimenter interfaces. Furthermore in order to validate our approach we have used a diabetic dataset with 108 instances but weka used 99 rows and 18 attributes to determine the prediction of disease and their accuracy using classifications of different algorithms to find out the best performance. The main objective of this paper is to classify data and assist the users in extracting useful information from data and easily identify a suitable algorithm for accurate predictive model from it. From the findings of this paper it can be concluded that Naïve Bayes the best performance algorithms for classified accuracy because they achieved maximum accuracy= 76.3021% correctly classified instances, maximum ROC = 0.819 , had least mean absolute error and it took minimum time for building this model through Explorer and Knowledge flow results.

References

16. J. Han and M. Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann,
An Empirical Comparison by Data Mining Classification Techniques for Diabetes Data Set

2000.


Index Terms

Computer Science Databases

Keywords

Weka, Data mining, Classification, Diabetic Disease Prediction.