ANALYSIS OF CLASSIFIERS DISEASES PREDICTION USING WEKA TOOL

Rikndra

M.TECH(CSE),Research Scholar

Department of Computer Science & Engineering

OM Institute of Technology and Management

Hisar ,India

Er.Deepika

Assistant Professor

Department of Computer Science & Engineering

OM Institute of Technology and Management

Hisar ,India

1. ABSTRACT:

Binary classification is the task of classifying the members of a given set of objects into two groups on the basis of whether they have some property or not. A typical binary classification task in health care management could be diagnosis of medical testing to determine if a patient will die or live. We have used HEPITITIES database from UCI Machine Repository. The database is containing 153 instances and 20 attributes on which various binary classifiers have been applied, we have used mainly J48,NB TREE AND AD TREE classifiers. We have compared these algorithms on various parameters of performance evaluation; our focus will be on mainly four parameters namely: precision, sensitivity, accuracy and error rate. For classification task we have used WEKA and TANAGRA data mining tools. The results of experiment show that AD Tree gives a promising classification result on the basis of sensitivity, precision ,error rate and accuracy.

Keywords: Data mining, Weka tools ,Classification

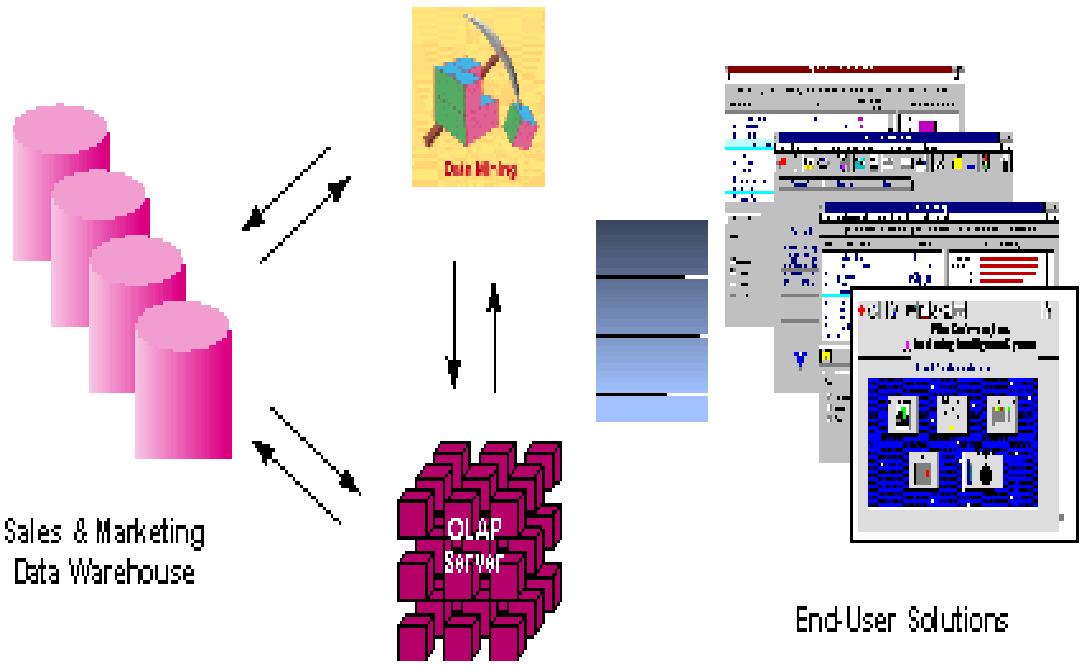
INTRODUCTION:

Over the last few years, there has been an extensive growth in the amount of data collection from different resources. It may be from transport, library, financial, hospitals telecom, medical, and shopping records. This data can be taken on a single platform and analyzed digitally. All this is possible only due to the rapid growth in data science, communication media, networking, and computing technologies; despite of this, another technology due to this advancement came onto the market cover the progress of data mining tools that aim to infer valuable fashion from this data. But this simple access to private data causes a risk to the human being confidentiality. In this thesis, we discussed the piecewise quantization approach, which is used for confidentiality reserving clustering.

Data Mining :

This technique deals with the pulling out of secret predictive information from big database. Data mining make use of complicated algorithms for the practice of sorting by huge amounts of data sets and expose related information. It can also be defined as “The Analysis step of the Knowledge Discovery in Databases process, or KDD”, a relatively new and combination of different domains in computer science. Data mining is used to extract the Patterns from large data sets. For this we are combining different techniques from data analysis, statistics, and artificial intelligence with database organization.

With new technical move forward in many domains it becomes more significant tool. It can transform from current business scenario to extraordinary quantities of digital data into business intelligence. Data mining is presently used in a broad array of profile practice, such as advertising, observation and technical discovery. The emergent agreement that it can hold real cost that has led to a sudden increase in order for fresh data mining technology .



Integrated Data Mining Architecture

Scope of Data Mining :

Data mining, it get its name by the likeness involving findings for main business information in a vast database – i.e. , receiving related item in gigabytes of store scanner data and mining a pile for a layer of valuable ore. Processes for data mining require changing all over huge amount of material, or smartly penetrating it to locate accurately where the value exists. This technology may provide new business prospects by providing a variety of features in databases of adequate size and quality. *Automated prediction of trends and behaviors.* The way of find out extrapolative information in huge databases is computerized by data mining. Questions that need wide investigation usually be able to answer directly and quickly with data mining method.

Data mining usually have four classes of responsibilities:

- **Association Learning Rule** – this learning rule seek out for the interaction among variables. Let's take an example of a market might collect data for purchasing habits of a customer. Using this learning rule, the market can decide that products are regularly purchased together and can apply this information for marketing purposes.
- **Clustering** – it is the responsibility of discovering groups and structures in the data that are in some way or a different "similar", with no using identified organization of the data.
- **Classification** – in this technique we generalize the known organization to pertain to new data. i.e., an email program capacity endeavors to categorize an email as genuine or spam. Ordinary algorithms contain different techniques as SVM, NN, Adjoining neighbor, decision tree learning, and naive Bayesian classification.
- **Regression** – attempt to find out a job that models the data with the minimum error.
- **Data Mining Technologies:**

- The main data mining techniques are as follows:
- *ANN (Artificial Neural Networks):* it is non-linear analytical models which learn by the training and it mimics the behavior of biological neural networks.
- *Decision Trees:* in this technique the sets of decisions are in tree-shaped structures. For the classification of dataset rules are generated with the help of these decisions.
- *Genetic Algorithms: it is an o*ptimization technique which has the concepts of evaluation design with the help of genetic combination, mutation, and natural selection.
- *Nearest Neighbor Method*: this technique categorize each record in a dataset on the behalf of the combination of the classes of the k record(s) most similar to it in a historical dataset.
- *Rule Induction*: in this technique the mining of useful if-then rules from data by statistical significance.

LITERATURE REVIEW:

CLASSIFICATION-:

Data mining Classification technique is applied for the predicting group membership for data instances [1]. For example, classification techniques may be used to determine whether it will be rainy or sunny weather outside. Some well known classification techniques are decision trees and neural networks.

Clustering and classification analyses are the two very common data mining techniques of finding hidden patterns in data. Though these two techniques are also considered as the two sides of the same coin; but in fact these are two entirely dissimilar

approaches in data analysis. Just similarity is that clustering and classification segments customer records into different data segments called classes. But contrasting clustering, classification is known in advance by the user or analyst that how the classes are defined.

TYPES OF CLASSIFICATION TECHNIQUES [1]:

There are following types of classification –

1. Classification based on Decision tree induction
   i. Decision tree induction
   ii. Tree pruning
   iii. Decision trees based on extraction from classification rules
   iv. Induction based on Scalability and decision tree

2. Classification based on Bayesian
   i. Bayes theorm
   ii. Naïve Bayesian classification
   iii. Bayesian belief classification

3. Classification by backpropogation
4. Association based classification
5. Other classification schemes
   i. "KNN"
   ii. "Case based reasoning"
   iii. "Genetic algorithms"
   iv. "Rough set theory"
   v. "Fuzzy set approaches"

Software used WEKA:-

WEKA (The *Waikato Environment for Knowledge Analysis*) is a machine learning tool, that being initiated by University of Waikato. It was planned to permit the consumer to access a diversity of machine learning process for the purposes of testing and identification by accurate world data sets. The boards at present dash on Sun workstations under X-windows, with machine learning tools and equipment written in a collection of programming languages. The board is not a single program, however relatively a sum of device joined mutually by a common user attachment.

WEKA at present includes seven various machine learning plans and schemes-----

1. In a normal session, a user have potency to choose a data set, execute on different machine learning methods on it, exclude and include various sets of aspect, and build evaluation between the resulting approach. Output from different scheme can be summarizing in an appropriate manner. To permit the user to focus on testing and analysis of the results, which are covered from the application particulars of the machine learning algorithms and the input format which they require?

ADVANTAGES OF WEKA --

- It offer various kind of algorithms for data learning and machine mining
- WEKA is an open source and without any cost accessible to the user.
- It is a autonomy type platform.
- WEKA tool is simply applicable by people who are not data mining expert

It offered various flexible factors for scripting experiments; this tool has some special feature in it that it kept up-to-date, with latest algorithms being added as they available in the research literature work

CHALLENGES— Geo-spatial data warehouse be likely to be very large. besides, existing GIS data-sets are often split into traits and feature parts, that are unsurprisingly archived in hybrid data management system. Algorithmic supplies vary significantly for relational (attribute) data management and for topological (feature) data management. Associated to this is the array and variety of geographic data format which also show exclusive challenges. The digital geographic data disorder is creating new types of data formats beyond the conventional "vector" and "raster" format. Geographic data warehouse more and more consist of ill-structured data such as images and geo-referenced multi-media.

CONCLUSION

According to above classification techniques result, we can find the best technique for our hepatitis dataset by comparing output of confusion matrix and summary statistic. So the following results are achieved,

Name of the algorithm	Summary	Confusion Matrix
SMO	Correctly Classified Instances 130 90.2778 % Incorrectly Classified Instances 14 9.7222 % Ignored class unknown instance 2	a b 112 5 | a = DIE 9 18 | b = LIVE
NB TREE	Correctly Classified Instances 140 90.2222% Incorrectly Classified Instances 4 2.7772 % Ignored class unknown instances 2	a b 115 2 | a = DIE 2 25 | b = LIVE
NAÏVE BAYES	Correctly Classified Instances 125 86.8056 % Incorrectly Classified Instances 19 13.1944 % Ignored class unknown instances 2	a, b 106 11| a = DIE 8 19 | b = LIVE

From above conclusion we can say that NB Tree gives the more efficient result than others classifiers .But we may get also more promising result by applying other classifiers ,

Result:-

(A) Table: 1 ACCURACY OF BINARY CLASSIFIERS

Parameters	SMO	NB TREE	NAÏVE BAYES
TP	112	115	106
FP	5	2	11
FN	9	2	8
TN	18	25	19
Accuracy	0.9027	0.9722	0.8680

TP/FP/FN/TN/Accuracy of classifiers

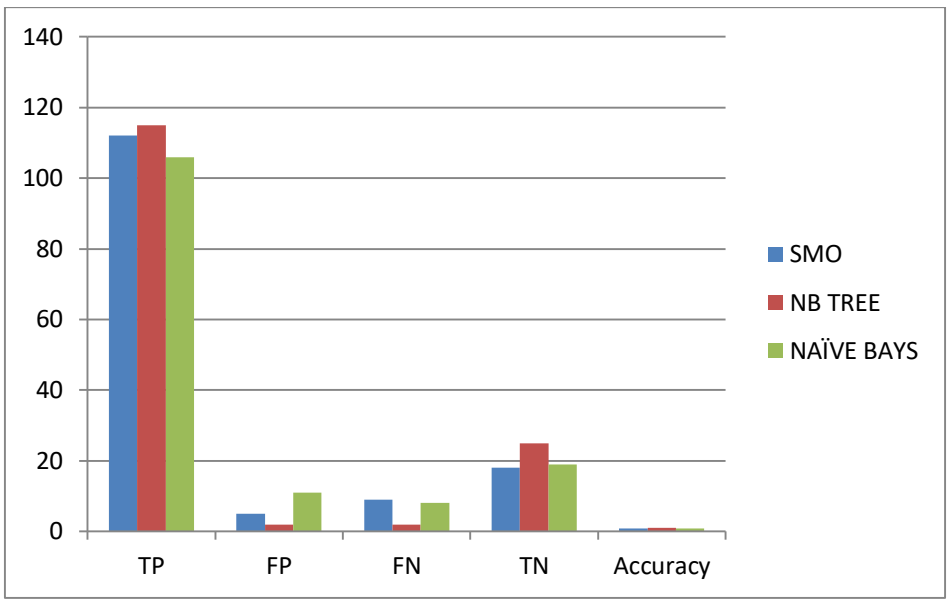


FIGURE 22**:** TP/FP/FN/TN/Accuracy of Classifiers

(B) **Table 2: LIVE Class - Precision/TPR/TNR/FPR**

Parameters	**SMO**	**NB TREE**	**NAÏVE BAYES**
TPR(Recall/Sensitivity)	0.957	0.983	0.906
TNR(Specificity)	0.7826	0.9259	0.6333
FPR	0.333	0.074	0.296
PRECISION	0.926	0.983	0.93

❖ **Please note that all the values shown above in tables have been obtained by applying mentioned classification techniques in weka.**

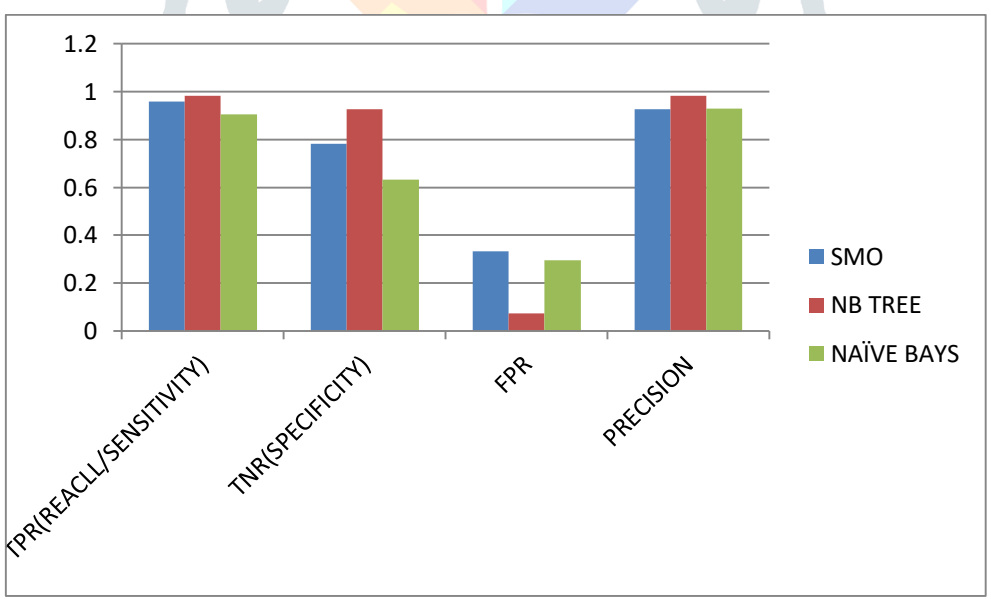


FIGURE 23: Bar chart showing parameters among classifiers

Table 3: DIE Class Precision/TNR/TPR/FPR

Parameters	**SMO**	**NB TREE**	**NAÏVE BAYES**
TPR(Recall/Sensitivity)	0.667	0.926	0.704
TNR(Specificity)	0.7826	0.9259	0.45
FPR	0.043	0.017	0.094
PRECISION	0.783	0.926	0.633

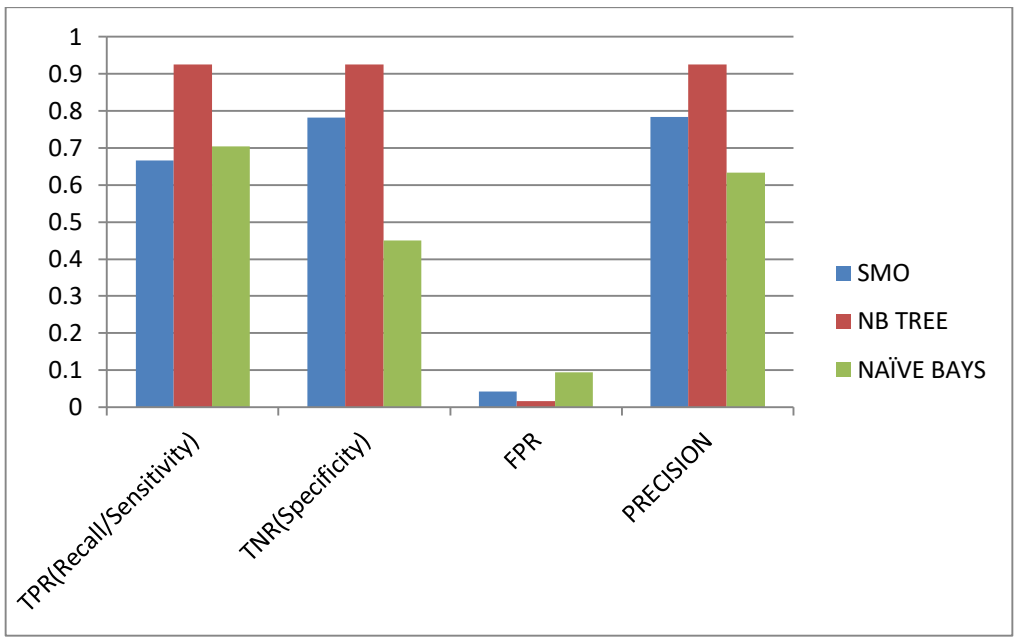


FIGURE 24: Bar chart for die class

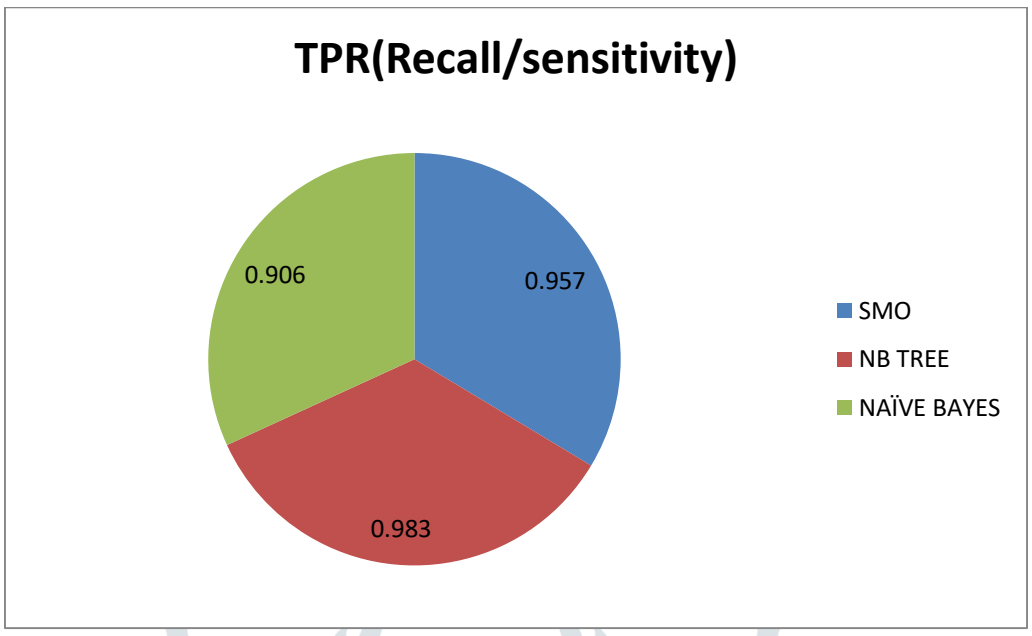
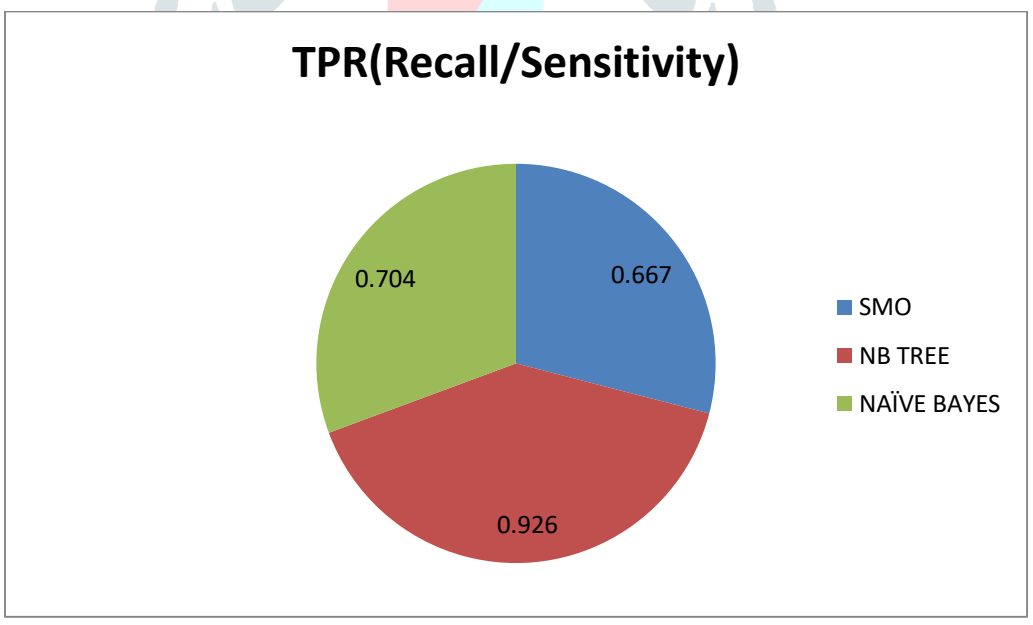


FIGURE 25: Pi chart for Live Class Sensitivity (TPR)



.

FIGURE 26: Pi chart showing TPR(Recall/Sensitivity) for DIE Class

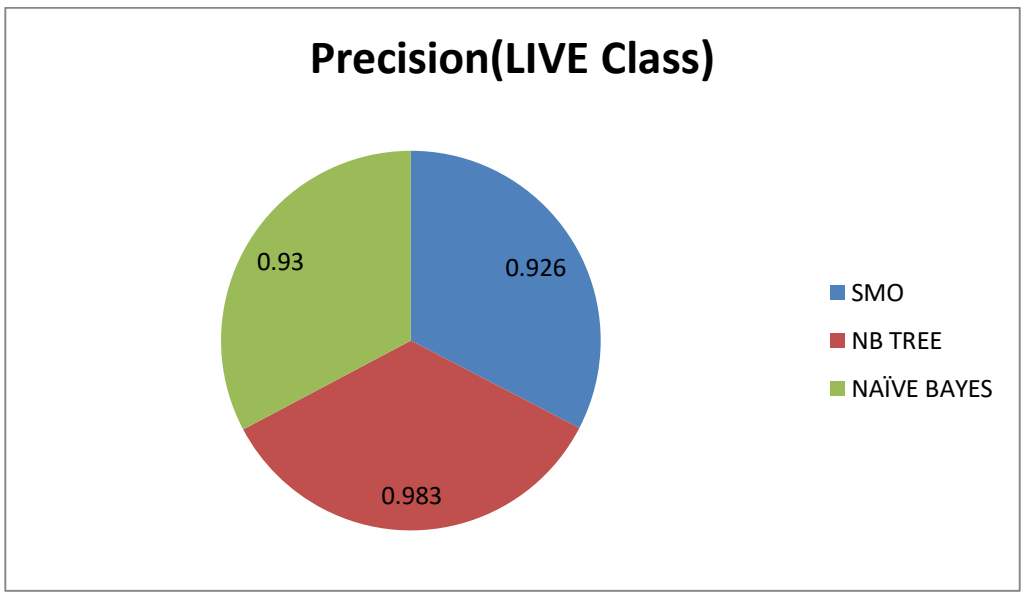


FIGURE 27: Pi chart showing precision for live class (Hepatitis)

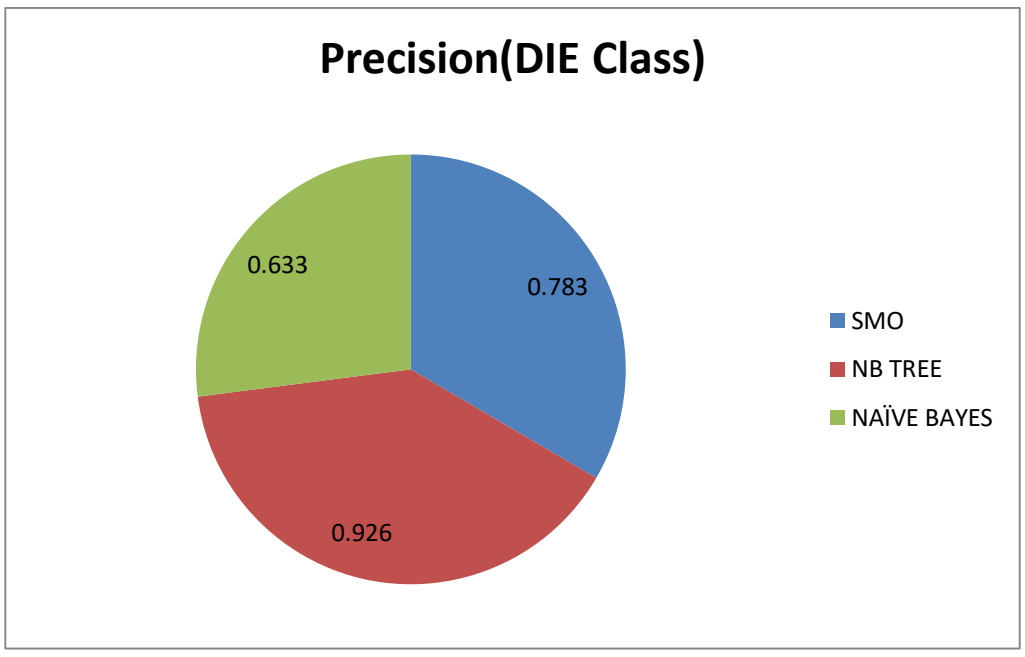


FIGURE 28: Pi chart precision for die class

DISCUSSION (OBSERVATION & ANALYSIS):

- It has been observed (from Table :1)that TP and TN values of NB Tree are high than other classifiers, which is a desirable point,
- The accuracy of NB Tree (0.9722) is high in compare to SMO and NAÏVE BAYES Classification algorithms, which is a important point during classification of Hepatitis patients, we may easily guess that NB Tree is quite sufficient to judge Hepatitis symptoms.
- It is giving very less false positive, it has shown only 2 out of 153 instances which is a desirable point.
- TPR (Recall/Sensitivity) is high(0.983) than others i.e. there are very few chances of Disease undetection, means maximum cases have been detected(refer Table:2,3).
- Since TNR(Specificity) of NB Tree is high(0.9259),means this algorithms has given maximum true result, means very few patients will be labled as sick(refer table:2,3).
- In ideal condition FPR should be zero,NB Tree has only 0.074 false positive rate which indicates that this algorithm does very minor mistake to judge Heaptitis, means most of the time it has given true result.
- Precision answers that" how likely it is that patients have the disease, given that there test's results were positive"?,from table2 and 3 we can say that precision of NB Tree is high which shows that results are positive out of all positive results however result is affected as number of Hepatitis patients increases time to time in world.
- According to experiments and results in this work ,NB Tree gives a satisfying and promising results for the task of classification of class label (LIVE/DIE) in Hepatitis dataset in Healthcare industry. It gives better result in both cases(live class as well as die class).

REFERENCES:

1. Data Mining concept and Techniques jiawei Han and Micheline Kamber :2000,Simon Fraser University
2. Dr. Varun Kumar, 2Luxmi Verma Department of Computer Science and Engineering, ITM University, Gurgaon, India." Binary Classifiers for Health Care Databases: A Comparative.
3. Study of Data Mining Classification Algorithms in the Diagnosis of Breast Cancer" IJCST Vol. 1, Iss ue 2, December 2010, I S S N : 2 2 2 9 - 4 3 3 3 (P r i n t) | I S S N : 0 9 7 6 - 8 4 9 1 (On l i n e).
4. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines.
5. John C. Platt Microsoft Research jplatt@microsoft.com Technical Report MSR-TR-98-14 April 21, 1998 © 1998 John Platt.
6. Wikipedia NAÏVE BAYES CLASSIFIER :www.wikipedia.org/en/classification.htm
7. An Introduction to the WEKA Data Mining System Zdravko Markov Central Connecticut State University markovz@ccsu.edu Ingrid Russell University of Hartford irussell@hartford.edu.
8. uci machine repository for dataset: http://www.ics.uci.edu/~mlearn/databases/hepatitis/hepatitis.names Web Documents: About Hepatitis domain database.
9. Benchmark results of Naive Bayes implementations (http:/ / tunedit. org/ results?d=UCI/ & a=bayes).
10. http://wekadocs.com/node/6 Web Documents: WEKA Software.
11. BAHÇEŞEHİR UNIVERSITY: APPLYING CLASSIFICATION METHODS ON HEPATITIS – DOMAIN DATASET(pdf) BY: Ergin DEMİREL (0569841)

AUTHOR'S BIOGRAPHIES

Rikendra is a MTECH student in department of Computer science & engineering from OM Institute of Technology and Management ,Haryana(Hisar).

Er. Deepika is presently work as assistant professor in the department of Computer science & Engineering at OM Institute of Technology , Hisar since 2012. She has 8 years experience in the field of Computer science & Engineering. She did M.tech (CSE) from GJU, Hisar in 2010.She has publishesd 2 paper presented in national conference, 4 paper publish in international journal.