A

Project Report

On

"Data Mining for Breast Cancer Detection using RapidMiner"

Prepared By,

Ankita Dutta

Pratishtha Ashok

Tanuka Nayak

Sharmili Nag

Department of Information Systems

NEW JERSEY INSTITUTE OF TECHNOLOGY, NEWARK

Year 2016-2017

ABSTRACT

Breast cancer poses a serious threat to the lives of the people and is the second leading cause of the death of women today. Due to the increase in life expectancy, urbanization and adoption of unknown western lifestyle the incidence of breast cancer is constantly increasing in this developing world. Although there are many treatments to cure breast cancer, these strategies cannot eliminate majority of the cancer that develop in lower and lower-middle class families where breast cancer is diagnosed in very later stage. In order to improve the diagnosis of breast cancer and survival outcome, early detection remains the cornerstone of breast cancer control.

In this paper, we have tried to analyze the breast cancer data available on kaggle.com with the aim of developing accurate prediction model using Data mining techniques. For the purpose of this research, we have used RapidMiner as the software platform and evaluated the dataset using Decision Tree, Naïve Bayes and k-NN classification techniques. Experimental results show that the proposed algorithms reduce data transmissions significantly and incur only small constant rounds of data communications. The experimental results demonstrate the superiority of the supervised algorithm – Naïve Bayes, which achieves a near-optimal performance under various conditions.

CONTENTS

I. INTRODUCTION

Breast cancer is the most common cancer in women both in developed and under developed countries in the world. As per Global Health Estimate WHO 2013, it is estimated that over 508,000 women died in 2011 due to breast cancer. Today in U.S., approximately one in every eight women over their lifetime has a risk of developing breast cancer. A recent study of SEER (Surveillance Epidemiology and End Results) which is a unique, reliable and essential resource of analyzing and investigating different aspects of cancer reveals that the survival rate of any breast cancer patient is 88% after 5 years of diagnosis and 80% after 10 years of diagnosis. Along with that the study of American Society also concluded that early detection of breast cancer can help in reducing the possibility of mitigating full growth of tumor. Normally the cells of whole body grow and divide to form new cells which result in the death of the old one. But with the addition of unwanted new cells with the existing old cell result in a formation of mass of tissue called tumor. Tumors can be malignant (harmful, easily spread) or benign (don't spread to any other part and less harmful). And diagnosis is the process of predicting the presence of breast cancer as either in benign or malignant form.

Classification is the data mining technique that involves the use of supervised machine learning techniques which predict the categorical class labels. And its accuracy depends on the percentage of the test samples that are correctly classified.

In this study, we focus on the dataset of breast cancer and apply three data mining algorithms to predict the most accurate model for diagnosis of breast cancer.

II. METHODOLOGY

Data Mining is the process of extracting information from a huge set of data. There are two basic types of data mining task. One which involves the task needed to understand the characteristics of dataset i.e. descriptive data mining and the other is used to perform predictions based on available dataset i.e. predictive data mining. Here, in this study, we are focusing on predictive data mining.

The dataset which we have used to perform this research is downloaded from kaggle.com. It consists of a total of 32 attributes consisting of I.D. number, Diagnosis (M= malignant and B= benign) and other 30 real valued input features with a class distribution of 357 benign and 212 malignant cells.

The three data mining techniques we have used are Naïve Bayes, Decision Tree and k-NN (k Nearest Neighbor). Our aim is to find out the most suitable algorithm to predict breast cancer.

We have used Rapid miner tool to implement these algorithms. RapidMiner is a software platform that provides an integrated environment for data mining, predictive analytics and is used for business, commercial applications as well as for research, education and training.

III. MINING ALGORITHMS

The first technique Naïve Bayes depends on Bayesian approach following a simple, clear and fast classifier. This method has been used to represent, utilize and learn probabilistic knowledge and significant results have been achieved in machine learning.

The second technique used is Decision Tree which builds classification or regression models in the form of a tree structure.

And the third technique k NN is a simple algorithm which stores all available cases and classifies new cases based on similarity measure.

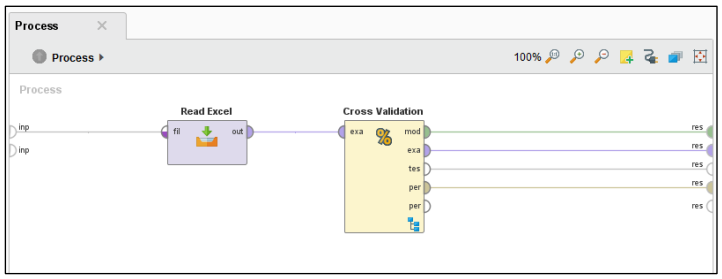
3.1 Supervised Learning

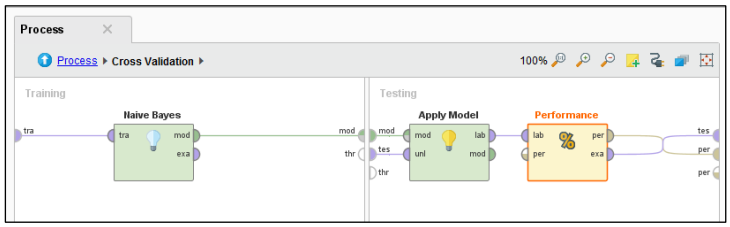
Supervised learning is a machine learning task of inferring a function from supervised training data. The training data consists of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which is called a classifier or a regression function. The inferred function should predict the correct output value for any valid input object.

3.1.1 Naïve Bayes

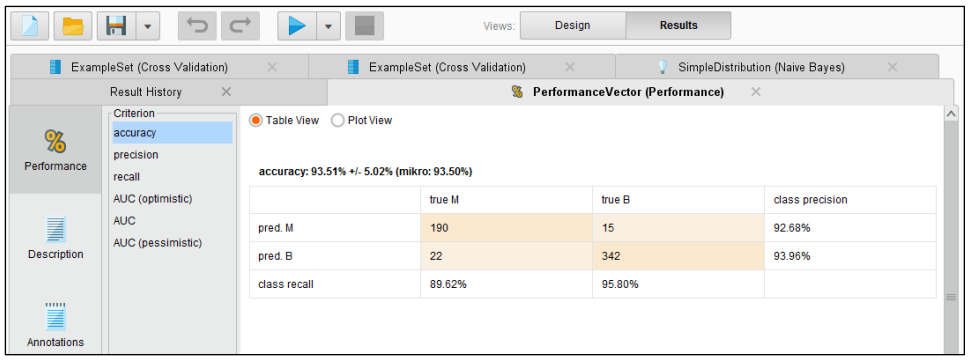
Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter.

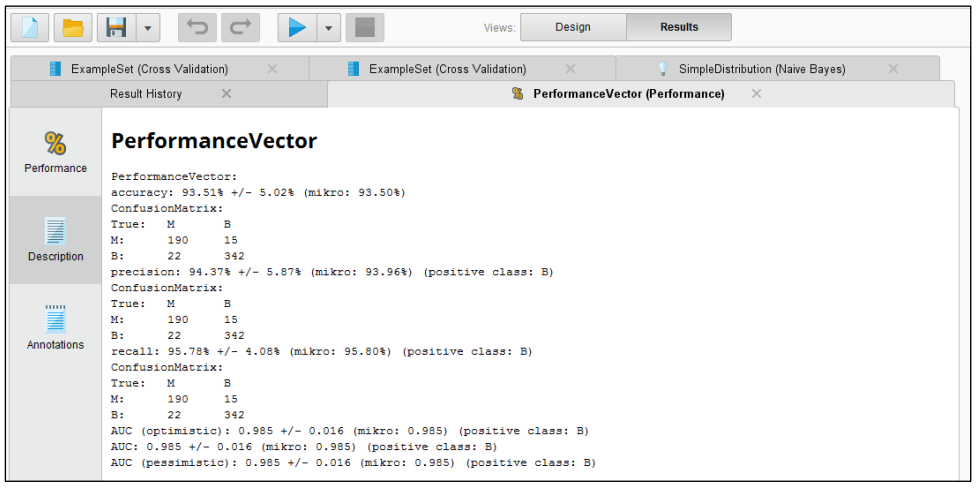
Rapid Miner Design: -

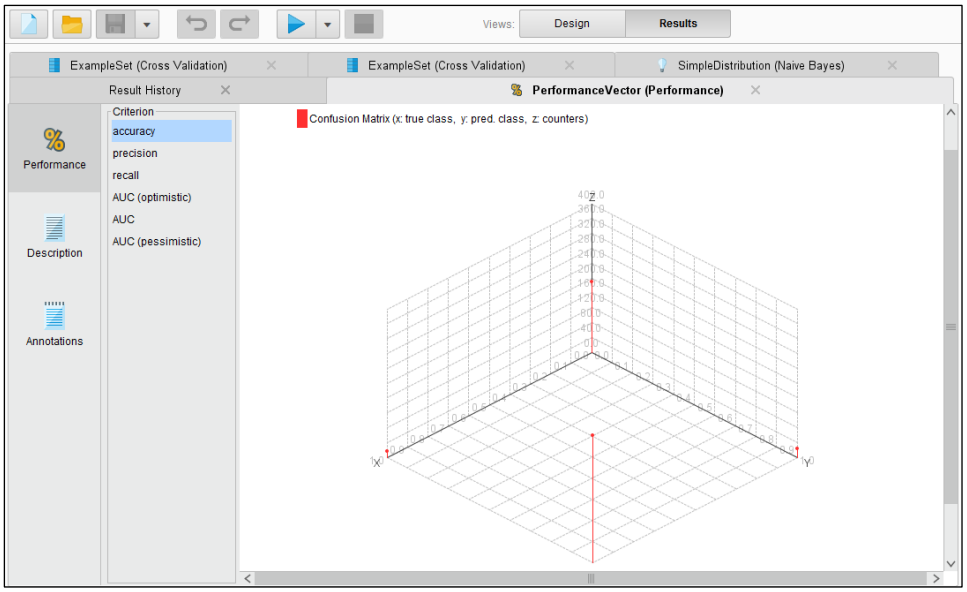




Rapid Miner Results: -





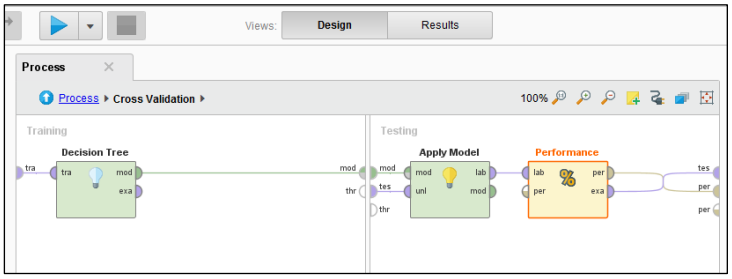


3.1.2 Decision Tree

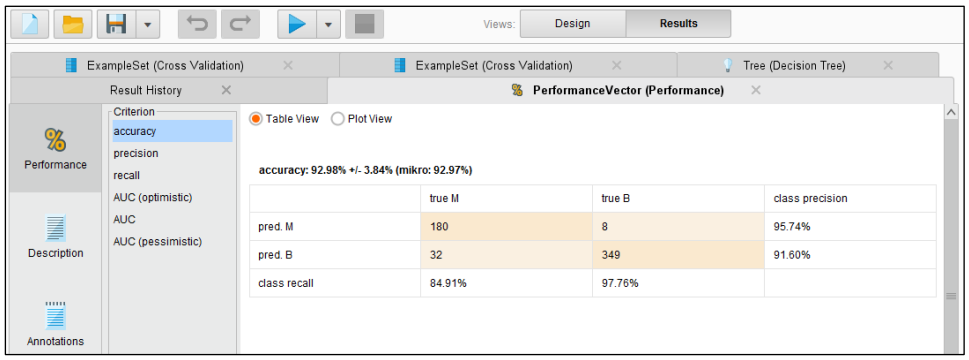
The decision tree algorithm assumes that each attribute is categorical, that is containing discrete data only, in contrast to continuous data such as age, height etc. The principle of the decision tree algorithm is as below: -

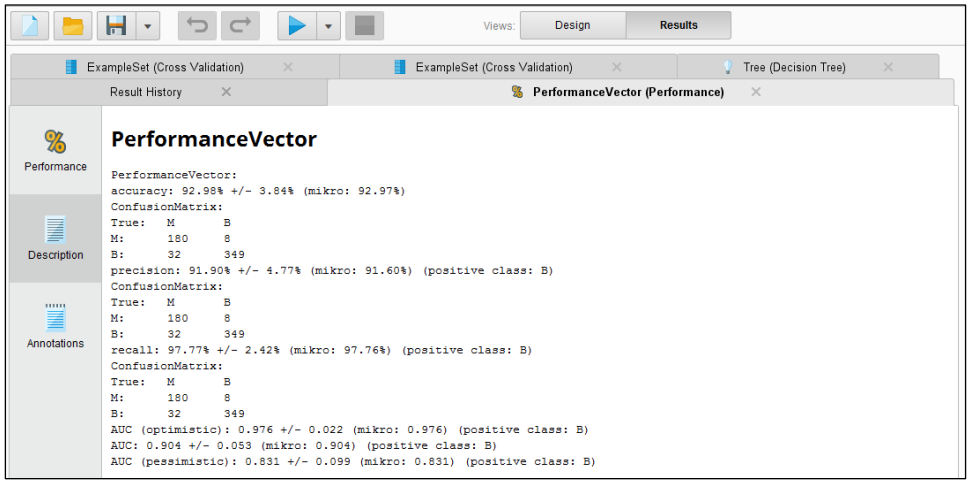
The tree is constructed top-down in a recursive fashion. At the root, each attribute is tested to determine how well it alone classifies the transactions. The “best” attribute is then chosen and the remaining transactions are partitioned by it. It is then recursively called on each partition (which is a smaller database containing only the appropriate transactions and without the splitting attribute).

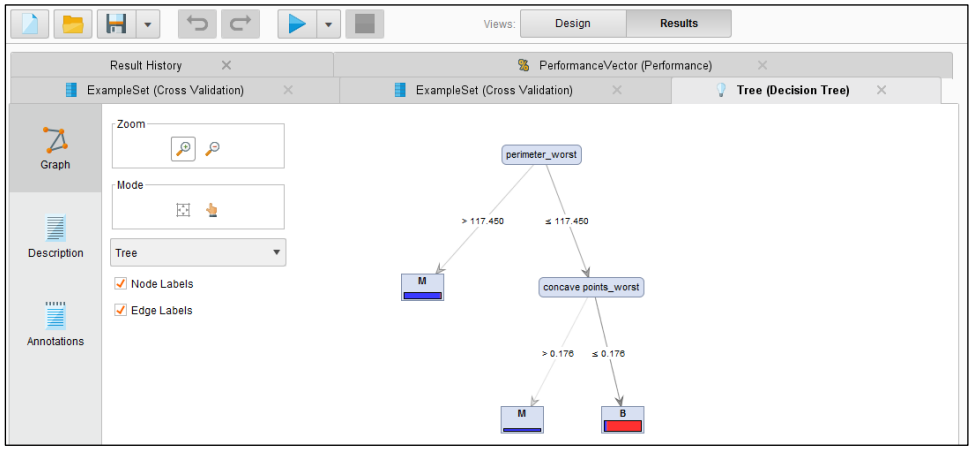
Rapid Miner Design: -



Rapid Miner Results: -





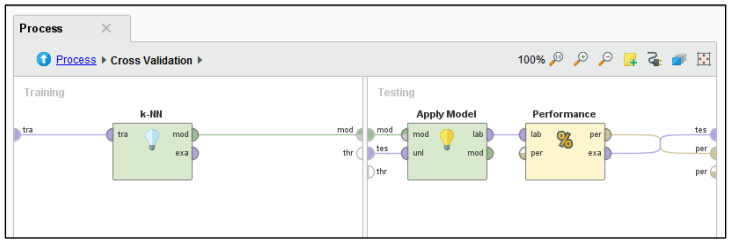


3.2.3 K-NN Classification

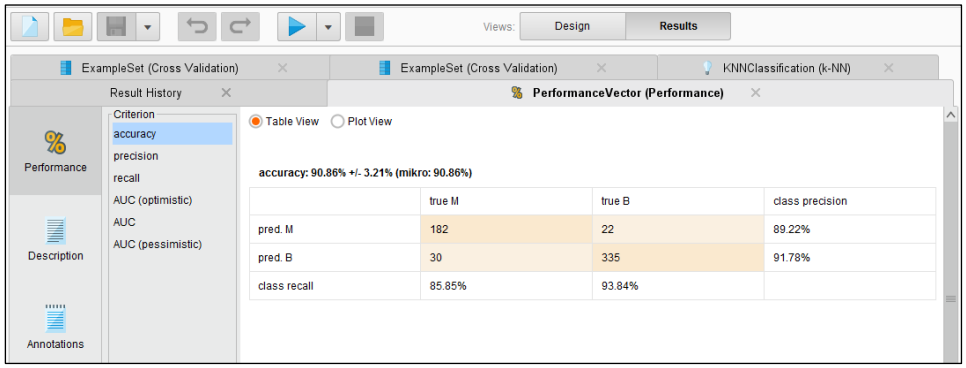
k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms. Both for classification and regression, it can be useful to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a

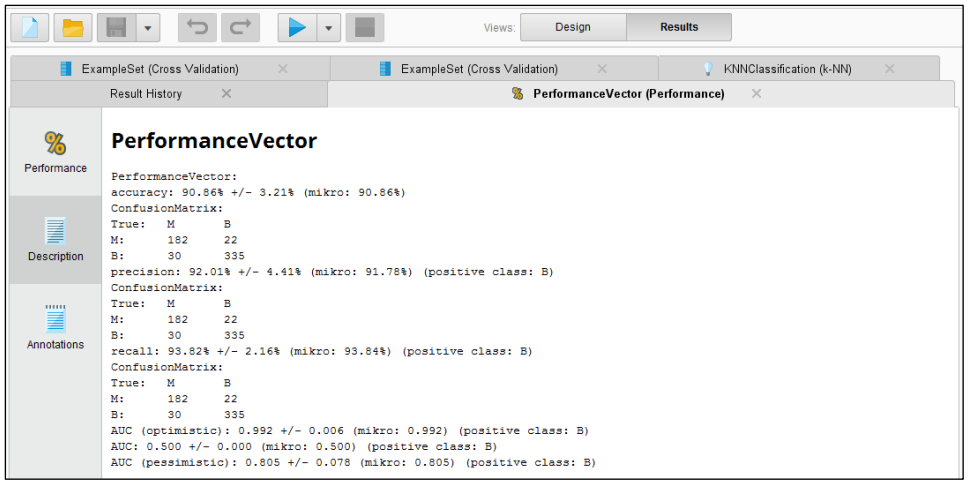
common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor.

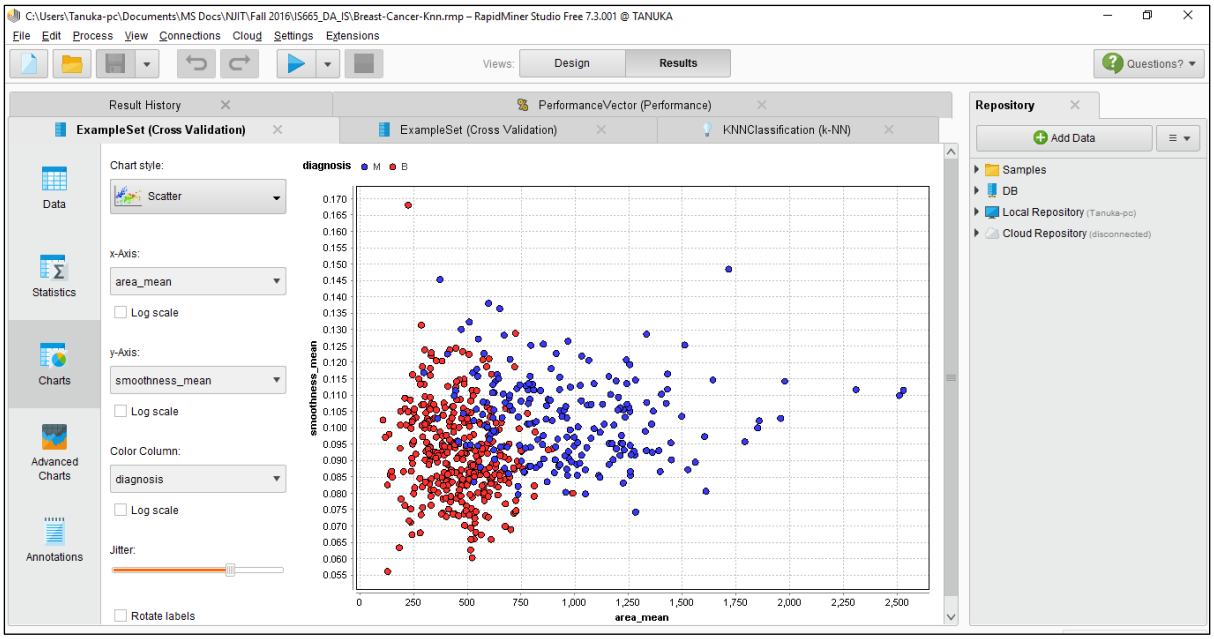
Rapid Miner Design: -



Rapid Miner Results: -







IV. RESULT COMPARISON

In this study, we have predicted the accuracy of three data mining techniques. The goal is to have high accuracy beside precision and recall metrics. The experimental results of the techniques are as given below: -

Classification Technique	Accuracy	Precision (positive class: B)	Recall (positive class: B)
Naïve Bayes	93.50%	93.96%	95.80%
Decision Tree	92.97%	91.60%	97.76%
k NN	90.86%	91.78%	93.84%

V. CONCLUSION & FUTURE WORK

As per our research, we analyzed that Naïve Bayes provides the most accurate result compared to decision tree and K-NN classification techniques.

This research has outlined, discussed and analyzed the best technique to predict the type of tumor in human body based on the dataset available from kaggle. Our analysis does not include any record consisting of null or invalid data. For our future assignments, we would want to include such data for a better exposure. In this study, we have focused only on the detection of the type of breast cancer whereas in our future study we would also like to concentrate on other aspects of breast cancer research to find out the survival rate of the cancer, implementing different data mining techniques.

VI. LIMITATION

The analysis to predict the breast cancer is achieved totally based on the above three data mining algorithms and the dataset of around 789 data of breast cancer, downloaded from kaggle.com. The whole study of breast cancer however involves a wide arena which could not be covered due to time constraints.

VII. REFERENCES

[1] https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

[2] http://www.medicalnewstoday.com/articles/37136.php

[3] https://www.ncbi.nlm.nih.gov/pubmed/25417087

[4] www.siam.org/meetings/sdm06/.../Scientific%20Datasets/bellaachia.pdf?q=data-mini

[5] https://pdfs.semanticscholar.org/85f6/a72dfc83ee65597c844c779bd67ccac36f84.pdf