Collaborative Analysis of Cancer Patient Data using Rapid Miner

Priyanka Jain
MTech
Computer Science & Engineering
GGITS
Jabalpur, India

Santosh Kr. Vishwakarma
Associate Professor
Computer Science & Engineering
GGITS
Jabalpur, India

ABSTRACT

Breast cancer is one amongst the leading cancers for ladies in developed countries including Asian nation. It is the second most typical explanation for cancer death in women. The high incidence of breast cancer in women has redoubled considerably within the last years. In this paper discussion on varied data processing approaches that are used for carcinoma identification and prognosis. Carcinoma Diagnosis is identifying of benign from malignant breast lumps and carcinoma Prognosis predicts once Breast Cancer is to recur in patients that have had their cancers excised. This study paper summarizes various review and technical articles on carcinoma identification and prognosis additionally tend to concentrate on current analysis being dole out victimisation the info mining techniques to boost the breast cancer identification and prognosis.

General Terms

Data-mining, K-NN algorithm, Naïve Bayes algorithm, Support Vector Machine Algorithm

Keywords

Breast cancer; Data Mining; Classification, Neural Network, Naive.Bayes, Support Vector Machine algorithm

1. INTRODUCTION

Breast cancer has become the leading explanation for death in women in developed countries. The most effective thanks to reduce carcinoma deaths are sight it earlier. Early diagnosing of cancer needs a correct and reliable diagnosis procedure that permits physicians to differentiate benign breast tumours from malignant ones while not going for surgical diagnostics. The objective of those predictions is to assign patients to either a"benign" cluster that's noncancerous or a"malignant" group that's cancerous. The prognosis problem is the long-run outlook for the malady for patients whose cancer has been surgically removed. In this problem a patient is assessed as a 'recur' if the malady is determined at some future time to growth excision and a patient for whom cancer has not recurred and should ne'er recur. The objective of those predictions is to handle cases that cancer has not recurred (censored data) furthermore as case that cancer has recurred at a selected time. Thus, breast cancer diagnostic and prognostic issues square measure principally within the scope of the wide mentioned classification problems.

These problems have attracted several researchers in machine intelligence, data mining, and statistics fields. Cancer research is typically clinical and/or biological in nature, data driven applied mathematics analysis has become a common complement. Predicting the outcome of a disease is one in every of the foremost fascinating and difficult tasks wherever to develop data processing applications. As the use of computers powered with machine-driven tools, large volumes of medical knowledge are being collected and created out there to the medical analysis teams. As a result, Knowledge Discovery in Databases (KDD), which includes data processing techniques, has become a popular analysis tool for medical researchers to spot and exploit patterns and relationships among sizable amount of variables, and made them ready to predict the result of a malady mistreatment the historical cases keep at intervals datasets.

The objective of this study is to summarise various review and technical articles on diagnosing and prognosis of carcinoma. It gives a summary of this analysis being meted out on varied carcinoma datasets mistreatment the info mining techniques to boost the carcinoma diagnosing and prognosis.

2. BREAST CANCER: AN OVERVIEW

Breast cancer is that the most typical cancer disease among ladies, excluding non-melanoma skin cancers. The information concerning the neoplasm from sure examinations and diagnostic tests square measure gathered victimisation staging to work out however widespread the cancer is. The stage of a cancer is one of the foremost important factors in choosing treatment choices, and it uses the Tumour, Nodes and Metastasis (TNM) system. When a patient's T, N, and M categories have been determined then this data is combined in a very method referred to as stage grouping to work out a woman's sickness stage. This is expressed as from Stage 0 (the least advanced stage) to Stage IV (the most advanced stage)[1]. Breast cancer may be a malignance, grew from cells of the breast. Hence, cancer of breast tissue is called carcinoma. Worldwide, it is the foremost common kind of cancer in females that's affecting close to 100 percent of all ladies at some stage of their life within the Western world. Although vital efforts square measure created to reach early detection and effective treatment however scientists don't recognize the precise causes of most carcinoma, they do know a number of the chance factors (i.e. ageing, genetic risk factors, family history, menstrual periods, not having children, obesity) that increase the probability of developing breast cancer in females. The survival analysis that is the study of your time to an occurrence of interest, such as disease prevalence or death additionally offers the physicians and therefore the patient higher data with that to set up treatment and will eliminate the necessity for a prognostic surgical treatment.

Knowledge Discovery and information Mining

This section provides an introduction to data discovery and information mining. The paper lists the numerous analysis tasks that may be goals of a discovery method and lists ways and analysis square measures that are promising in resolution these analysis tasks.

The Knowledge Discovery method

The terms Knowledge Discovery in information bases (KDD) and Data Mining are typically used interchangeably. KDD is the process of turning the low-level information into high-level data. Hence, KDD refers to the nontrivial extraction of implicit, previously unknown and probably helpful data from information in databases. While data mining and KDD are typically treated as equivalent words however in real information mining is a vital step within the KDD process. The following fig. 1 shows information mining as a step in associate degree unvarying data discovery method.

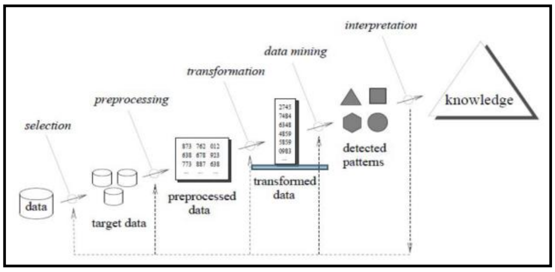


Figure 1: Steps in KDD

The Knowledge Discovery in Databases method contains of a few steps leading from data collections to some type of new information [2]. The iterative method consists of the following steps:

1) Data cleaning: conjointly notable as information cleansing it introduces that noise information and irrelevant information area unit far from the gathering.
2) Data integration: at this stage, multiple data sources, often heterogeneous, may be combined during a common supply.
3) Data selection: at this step, the info relevant to the analysis is determined on and retrieved from the data assortment.
4) Data transformation: conjointly notable as information consolidation, it is a introduce which the chosen information is remodelled into forms applicable for the mining procedure.
5) Data mining: it is the crucial step during which clever techniques area unit applied to extract patterns doubtless helpful.
6) Pattern evaluation: this step, strictly interesting patterns representing information area unit known based mostly on given measures.
7) Knowledge representation: is the final introduced that the discovered information is visually described to the user. In this step visualization techniques area unit wont to facilitate users perceive and interpret the info mining results.

3. DATA MINING PROCESS

In the KDD process, the knowledge mining strategies area unit for extracting patterns from data. The patterns that can be discovered rely on the info mining tasks applied. Generally, there are 2 varieties of data processing tasks: descriptive data processing tasks that describe the overall properties of the present knowledge, and predictive knowledge mining tasks that conceive to do predictions supported accessible knowledge. Data mining are often done on knowledge that area unit in quantitative, textual, or multimedia forms.

Data mining applications will use totally different quite parameters to look at the info. They include association (patterns wherever one event is connected to another event), sequence or path analysis (patterns where one event leads to another event), classification (identification of new patterns with predefined targets) and clustering (grouping of identical or similar objects).Data mining involves a number of the subsequent key steps [3]-

1) Problem definition: The 1st step is to spot goals. Based on the outlined goal, the correct series of tools are often applied to the info to create the corresponding behavioural model.
2) Data exploration: If the quality of knowledge isn't appropriate for Associate in correct model then recommendations on future data assortment and storage methods are often created at this. For analysis, all data wants to be consolidated so it is often treated systematically.
3) Data preparation: The purpose of this step is to scrub and rework the info so missing and invalid values area unit treated and every one notable valid values area unit created consistent for additional strong analysis.

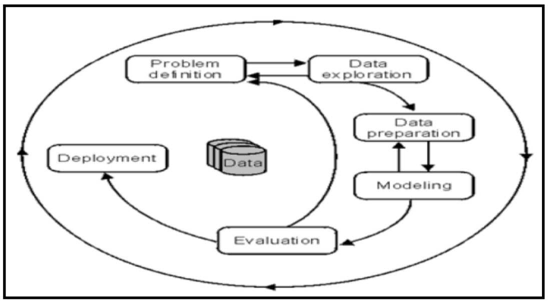


Figure 2: Data Mining Process Representation

The goal of the classifier algorithms is to construct a model from a set of coaching data whose target class labels ar notable and so this model is employed to classify unseen instances. The classifications of Breast Cancer data are often helpful to predict the end result of some diseases or discover the genetic behaviour of tumours.

Breast cancer is one in all the foremost common cancers among women. Breast cancer is one in all the most important causes of death in women compared to any or all alternative cancers. Cancer is a style of diseases that causes the cells of the body to alter its characteristics and cause abnormal growth of cells. Most types of cancer cells eventually become a mass known as neoplasm. The occurrence of breast cancer is increasing globally. It is a serious ill health and represents a big worry for several women [1]. Early detection of breast cancer is crucial in reducing life losses. However earlier treatment needs the ability to observe carcinoma in early stages. Early designation needs associate correct and reliable diagnosis procedure that permits physicians to differentiate benign breast tumours from malignant ones. The automatic diagnosis of carcinoma is a vital, real-world medical problem. Thus, finding a correct and effective designation methodology is terribly vital. In recent years machine learning methods have been wide utilized in prediction, especially in medical designation. Medical diagnosis is one in all major drawbacks in medical application.

The classifications of Breast Cancer data are often helpful to predict the end result of some diseases or discover the genetic behaviour of tumours. A major class of issues in life science involves the designation of illness, based upon numerous tests performed upon the patient. For this reason the use of classifier systems in diagnosis is gradually increasing.

4. CLASSIFICATION TECHNIQUES

Building accurate and economical classifiers for massive databases is one in all the essential tasks of information mining and machine learning analysis. Building effective classification systems is one of the central tasks of information mining. Many totally different sorts of classification techniques are planned in literature that has call Trees, Naive-Bayesian methods, Neural Networks, Logistic Regression, SVM and KNN etc.

1) Decision Tree

Decision tree models are normally used in data processing to look at information and induce the tree and its rules that may be wont to create predictions[10].The prediction could be to predict categorical values (classification trees) once instances are to be placed in classes or categories.

Decision tree is a classifier within the sort of a tree structure wherever every node is either a leaf node, indicating the value of the target attribute or category of the examples, or a decision node, specifying some test to be carried out on one attribute-value, with one branch and sub-tree for each doable outcome of the take a look at. A decision tree are often wont to classify associate example by beginning at the basis of the tree and moving through it till a leaf node is reached, which provides the classification of the instance.

2) Neural Networks

Neural networks are capable of modelling very complicated, typically non-linear functions [10]. It is made from a structure or a network of various interconnected units (artificial neurons). Each of these units consists of input/output characteristics that implement an area computation or operate. The function might be a computation of weighted sums of inputs that produces associate output if it exceeds a given threshold. The output (whatever the result), could serve as associate input to alternative neurons within the network. This process iterates till a final output is made.

3) Naive Bayes (Nb)

The Naive Bayes is a fast methodology for creation of applied mathematics prophetical models [16]. NB is based on the theorem. This classifier technique analyses the relationship between each attribute and also the category for every instance to derive a contingent probability for the relationships between the attribute values and also the class. During coaching, the probability of every category is computed by investigation what percentage times it happens within the coaching dataset. This is called the “prior probability” P(C=c). In addition to the prior chance, the algorithm conjointly computes the chance for the instance x given c with the assumption that the attributes ar freelance. This probability becomes the product of the possibilities of every single attribute. The probabilities will then be calculable from the frequencies of the instances within the coaching set.

4) Logistic Regression (Lr)

LR is considered because the customary applied mathematics approach to modelling binary information [16]. It is a much better alternative for a statistical regression that assigns a linear model to every of the category and predicts unseen instances basing on majority vote of the models. During prediction, instead of predicting the purpose estimate of the event itself, it builds a model to predict the odds of its occurrence. In two category drawbacks for example, when the odds ar larger than five hundredth, then the case is assigned to the category selected as “1” for affirmative and “0” for “YES” and “NO” instead.

5) Support Vector Machine (Svm)

SVMs are a set of connected supervised learning strategies that analyze information and acknowledge patterns, used for classification and regression analysis. SVM is a formula that tries to realize a linear apparatus (hyper-plane) between the info points of 2 categories in three-dimensional area. SVM represents a learning technique which follows principles of applied mathematics learning theory [14]. Generally, the main idea of SVM comes from binary classification, namely to realize a hyper plane as a segmentation of the 2 categories to reduce the classification error. The SVM finds the hyper plane using support vectors (training tuples) and margins (support vectors). The Sequential minimal optimisation (SMO) formula is a straightforward and quick methodology for coaching a SVM.

6) K-Nearest Neighbor (Knn)

K-Nearest Neighbour (KNN) classification [8] classifies instances based on their similarity. An object is classified by a majority of its neighbours. K is always a positive number. The neighbours are elite from a set of objects that the right classification is understood. The training samples ar represented by n dimensional numeric attributes. Each sample represents a purpose in associate n-dimensional area. In this way, all of the training samples ar hold on in associate n-dimensional pattern area. When given associate unknown sample, a k-nearest neighbour classifier searches the pattern space for the k coaching samples that are highest to the unknown sample. "Closeness" is defined in terms of Euclidian distance. The unknown sample is assigned the most common category among its k nearest neighbours. When k=1, the unknown sample is assigned the category of the coaching sample that's highest to that in pattern area. In WEKA this classifier is known as IBK.

5. RELATED WORK

In their paper author present AN analysis of the prediction of survivability rate of breast cancer patient’s mistreatment data processing techniques. The data used is that the SEER Public-Use knowledge. The pre-processed data set consists of 151,886 records, which have all the accessible sixteen fields from the SEER information. Paper has investigated 3 data processing techniques: the Naïve Bayes, the back-propagated neural network, and the C4.5 call tree algorithms. Several experiments were conducted mistreatment these algorithms. The achieved prediction performances are comparable to existing techniques. However, it is found out that C4.5 algorithmic program has a far better performance than the opposite 2 techniques. [1]

Breast cancer is one amongst the leading cancers for girls in developed countries including Asian nation. It is the second commonest explanation for cancer death in women. The high incidence of breast cancer in women has enhanced considerably within the last years. In this paper discussion on varied data processing approaches that are utilised for carcinoma designation and prognosis. carcinoma Diagnosis is identifying of benign from malignant breast lumps and carcinoma Prognosis predicts once Breast Cancer is to recur in patients that have had their cancers excised. This study paper summarizes various review and technical articles on carcinoma designation and prognosis conjointly tend to specialize in current analysis being meted out mistreatment the information mining techniques to reinforce the breast cancer designation and prognosis. [2]

Breast cancer is one amongst the foremost causes of death in women in comparison to all or any alternative cancers. Breast cancer has become the foremost hazardous sorts of cancer among ladies within the world. Early detection of breast cancer is crucial in reducing life losses. This paper presents a comparison among the different data processing classifiers on the information of carcinoma Wisconsin carcinoma (WBC), by using classification accuracy. This paper aims to establish an correct classification model for carcinoma prediction, in order to form full use of the invaluable info in clinical knowledge, especially that is typically neglected by most of the prevailing strategies after they aim for top prediction accuracies. Work has done experiments on blood cell knowledge. The dataset is divided into training set with 499 and take a look at set with two hundred patients. In this experiment, compare six classification techniques in Maori hen software system and comparison results show that Support Vector Machine (SVM) has higher prediction accuracy than those strategies. Different strategies for breast cancer detection square measure explored and their accuracies square measure compared. With these results, infer that the SVM square measure a lot of appropriate in handling the classification downside of breast cancer prediction, and suggest the employment of those approaches in similar classification issues. [3]

Cancer is one of the foremost problems nowadays; diagnosing cancer in earlier stage is still difficult for doctors. Breast cancer is one amongst the foremost death causing diseases of the ladies nowadays everywhere the planet. Every year over million ladies square measure diagnosed with carcinoma worldwide over 1/2 them can die due to the late identification of the illness. So several researchers have undergone for police work the cancer supported data processing technology every approach has its own limitations. This makes us to take up this downside and to implement the information mining primarily based cancer prediction System (DMBCPS). Work has projected this cancer prediction system supported data processing technology. This system estimates the danger of the carcinoma within the earlier stage. This system is validated by scrutiny its foretold results with patient’s previous medical info and it absolutely were analyzed by mistreatment Maori hen system. The main aim of this model is to produce the sooner warning to the users, and it is also price economical to the user. [4]

Breast Cancer Diagnosis and Prognosis square measure 2 medical applications cause a good challenge to the researchers. The use of machine learning and data processing techniques has revolutionized the total process of carcinoma designation and Prognosis. Carcinoma Diagnosis distinguishes benign from malignant breast lumps and carcinoma Prognosis predicts once Breast Cancer is probably going to recur in patients that have had their cancers excised. Thus, these two issues square measure principally in the scope of the classification issues. This study paper summarizes various review and technical articles on breast cancer designation and prognosis. In this paper we a summary of the present analysis being meted out mistreatment the information mining techniques to reinforce the carcinoma designation and prognosis. [5]

6. METHODOLOGY

In this work, the device that is utilized is speedy labourer [6]. Rapid labourer is a product stage created by the organization of identical name that offers a coordinated scenario to machine learning, information mining, content mining, prescient examination and business investigation. This gives over 1500 drag and drop operation with the assistance of that sizable amount of information mining will be performed simply and quickly. Work can be utilized for the text mining, classification, validation, perusing and thus on.

For classification purpose, three classifiers – Naive Bayes classifier, K-NN & Support Vector Machine are used. A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem (from Bayesian statistics) with strong (naive) independence assumptions. A naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Naive Bayes classifiers can handle an arbitrary number of independent variables, whether continuous or categorical. Given a set of variables, X = {x1, x2, x3..., xd}, want to construct the posterior probability for the event Cj among a set of possible outcomes C = {c1, c2, c3..., cd}. In a more familiar language, X is the predictors and C is the set of categorical levels present in the dependent variable. Using Bayes' rule

$p\left(C_j \,\middle|\, x_1, x_2, \ldots, x_d\right) \propto p\left(x_1, x_2, \ldots, x_d \,\middle|\, C_j\right) p(C_j)$ - Eq (1)

Where p(Cj | x1, x2, x3..., xd) is the posterior probability of class membership, i.e., the probability that X belongs to Cj. Since Naive Bayes assumes that the conditional probabilities of the independent variables are statistically independent and can decompose the likelihood of a product of terms:

$p(X \mid C_j) \propto \prod_{k=1}^{d} p(x_k \mid C_j)$ - Eq (2)

And rewrite the posterior as:

$p\left(C_j \,\middle|\, X\right) \propto p(C_j) \prod_{k=1}^{d} p(x_k \mid C_j)$ - Eq (3)

Many classifiers accessible in text mining, here Naive Bayes is used at the side of an additional classifier referred to as K-NN algorithmic program to match the result.

K-Nearest Neighbours makes predictions based on the end result of the K neighbours highest to it purpose. Therefore, to make predictions with KNN, one would like to outline a metric for measure the gap between the question purpose and cases from the examples sample. One of the foremost popular decisions to live this distance is understood as Euclidian.

$D(x, p) = \sqrt{(x-p)^2}$ - Eq (1)

Where *x* and *p* are the query point and a case of the examples sample, respectively.

Since KNN predictions square measure primarily based on the intuitive assumption that objects go on distance are probably similar, it makes good sense to discriminate between the K nearest neighbours once creating predictions. Let the closest purposes among the K nearest neighbours have a lot of say in touching the outcome of the question point. This can be achieved by introducing a collection of weights W, one for each nearest neighbour, defined by the relative closeness of every neighbour with relevance the question purpose.

$W(x, p_i) = \frac{\exp(-D(x, p_i))}{\sum_{i=1}^{k} \exp(-D(x, p_i))}$ - Eq (2)

Where *D(x, p_i)* is the distance between the query point *x* and the *i*th case p_i of the example sample. The weights defined in this manner above will satisfy:

$\sum_{i=1}^{k} W(x_0, x_i) = 1$ - Eq (3)

Thus, for classification problems, the maximum of y is taken for each class variables.

$\max(y = \sum_{i=1}^{k} W(x_0, x_i) y_i)$ - Eq (4)

Figure 3 shows the flow of main method. Process documents from files operator is used for reading text knowledge accessible in any document file. Validation operator is used for providing training and applying totally different data mining algorithms in any process. For three totally different brands have got follow same procedure for 3 times.

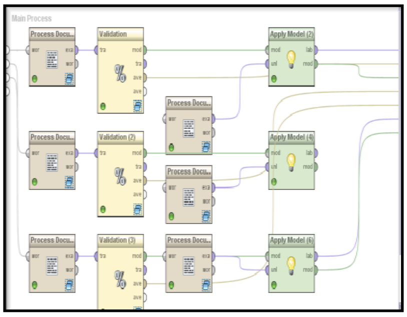


Figure 3: Main Process

7. RESULTS & DISCUSSION SUPPORT VECTOR MACHINE

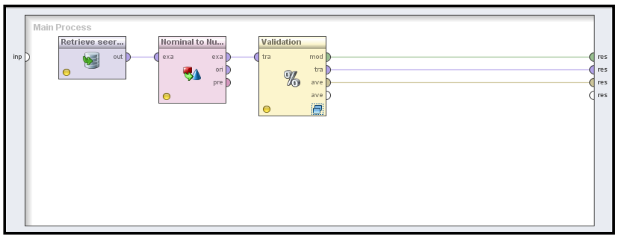


Figure 4: Support Vector Classifier

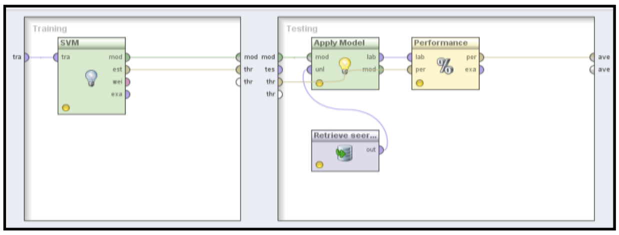
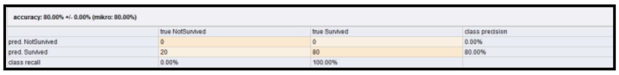


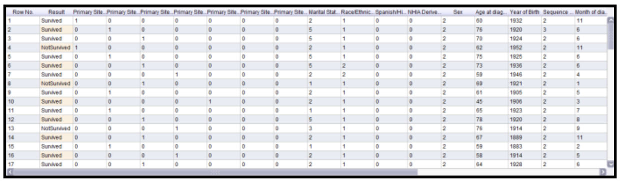
Figure 5: Support Vector Classifier



accuracy: 80.00% +/- 0.00% (mikro: 80.00%)

	true NotSurvived	true Survived	class precision
pred. NotSurvived	0	0	0.00%
pred. Survived	20	80	80.00%
class recall	0.00%	100.00%	

Figure 6: Accuracy calculated using SVM Classifier



Row No.	Result	Primary Site...	Primary Site...	Primary Site...	Primary Site...	Primary Site...	Primary Site...	Primary Site...	Marital Stat...	Race/Ethnic...	Spanish/Hi...	NHIA Derive...	Sex	Age at diag...	Year of Birth	Sequence ...	Month of dia...
1	Survived	1	0	0	0	0	0	0	2	1	0	0	2	60	1932	2	11
2	Survived	0	1	0	0	0	0	0	5	1	0	0	2	76	1920	3	6
3	Survived	0	0	1	0	0	0	0	5	1	0	0	2	70	1924	2	6
4	NotSurvived	1	0	0	0	0	0	0	2	1	0	0	2	62	1952	2	11
5	Survived	0	1	0	0	0	0	0	5	1	0	0	2	75	1925	2	6
6	Survived	0	0	1	0	0	0	0	5	2	0	0	2	73	1936	2	6
7	Survived	0	0	0	1	0	0	0	2	2	0	0	2	59	1946	2	4
8	NotSurvived	0	0	1	0	0	0	0	1	1	0	0	2	69	1921	2	1
9	Survived	0	1	0	0	0	0	0	2	1	0	0	2	61	1905	2	5
10	Survived	0	0	0	0	1	0	0	2	1	0	0	2	45	1906	2	3
11	Survived	0	1	0	0	0	0	0	1	1	0	0	2	65	1923	2	7
12	Survived	0	0	1	0	0	0	0	5	1	0	0	2	78	1920	2	8
13	NotSurvived	0	0	0	1	0	0	0	3	1	0	0	2	76	1914	2	9
14	Survived	0	0	1	0	0	0	0	2	1	0	0	2	67	1889	2	11
15	Survived	0	1	0	0	0	0	0	1	1	0	0	2	59	1883	2	2
16	Survived	0	0	0	1	0	0	0	2	1	0	0	2	58	1914	2	5
17	Survived	0	0	1	0	0	0	0	2	1	0	0	2	64	1928	2	6

Figure 7: Label identified by SVM Classifier

Naïve Bayes

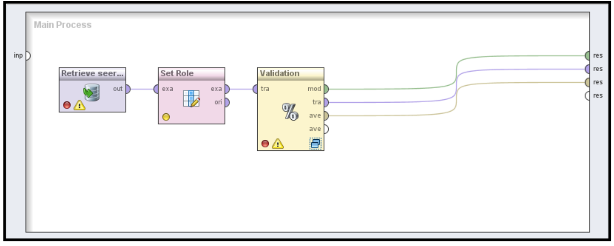


Figure 8: K-NN Classifier

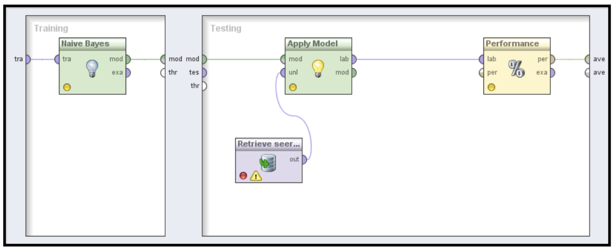
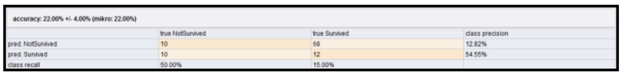


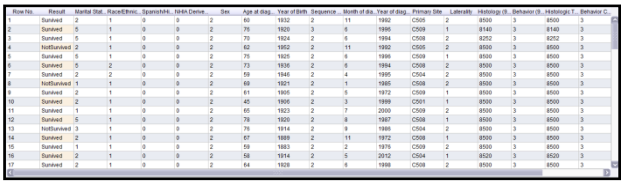
Figure 9: K-NN Classifier



accuracy: 22.00% +/- 4.00% (mikro: 22.00%)

	true NotSurvived	true Survived	class precision
pred. NotSurvived	10	68	12.82%
pred. Survived	10	12	54.55%
class recall	50.00%	15.00%	

Figure 10: K-NN Classifier



Row No.	Result	Marital Stat...	Race/Ethnic...	Spanish/Hi...	NHIA Derive...	Sex	Age at diag...	Year of Birth	Sequence ...	Month of dia...	Year of diag...	Primary Site	Laterality	Histology (9...	Behavior (9...	Histologic T...	Behavior C...
1	Survived	2	1	0	0	2	60	1932	2	11	1992	C505	2	8500	3	8500	3
2	Survived	5	1	0	0	2	76	1920	3	6	1996	C509	1	8140	3	8140	3
3	Survived	5	1	0	0	2	70	1924	2	6	1994	C508	2	8252	3	8252	3
4	NotSurvived	2	1	0	0	2	62	1952	2	11	1992	C505	2	8500	3	8500	3
5	Survived	5	1	0	0	2	75	1925	2	6	1996	C509	1	8500	3	8500	3
6	Survived	5	2	0	0	2	73	1936	2	6	1994	C508	2	8500	3	8500	3
7	Survived	2	2	0	0	2	59	1946	2	4	1995	C504	2	8500	3	8500	3
8	NotSurvived	1	1	0	0	2	69	1921	2	1	1985	C508	2	8500	3	8500	3
9	Survived	2	1	0	0	2	61	1905	2	5	1972	C509	1	8500	3	8500	3
10	Survived	2	1	0	0	2	45	1906	2	3	1999	C501	1	8500	3	8500	3
11	Survived	1	1	0	0	2	65	1923	2	7	2000	C509	2	8500	3	8500	3
12	Survived	5	1	0	0	2	78	1920	2	8	1987	C508	1	8500	3	8500	3
13	NotSurvived	3	1	0	0	2	76	1914	2	9	1986	C504	2	8500	3	8500	3
14	Survived	2	1	0	0	2	67	1889	2	11	1972	C508	1	8500	3	8500	3
15	Survived	1	1	0	0	2	59	1883	2	2	1976	C509	2	8500	3	8500	3
16	Survived	2	1	0	0	2	58	1914	2	5	2012	C504	1	8520	3	8520	3
17	Survived	2	1	0	0	2	64	1928	2	6	1998	C508	2	8500	3	8500	3

Figure 11: K-NN Classifier

K-NN

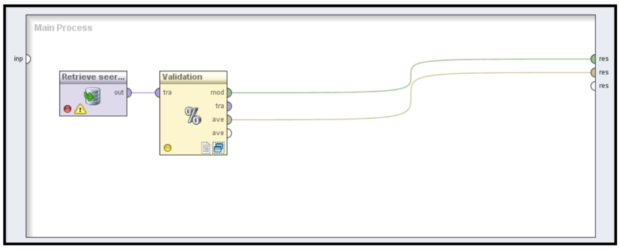


Figure 12: K-NN Setup in Rapid Minor with Training Dataset & Validator

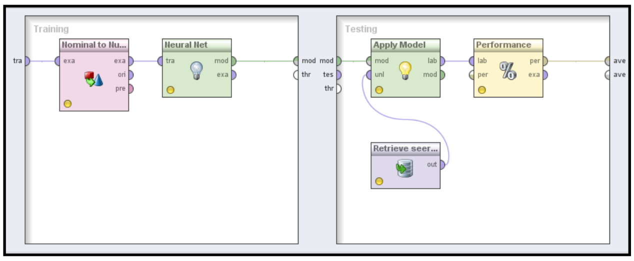
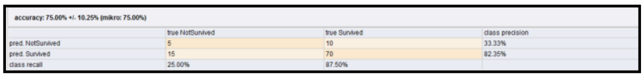


Figure 13: K-NN Setup in Rapid Miner with Application of Model



accuracy: 75.00% +/- 10.25% (mikro: 75.00%)

	true NotSurvived	true Survived	class precision
pred. NotSurvived	5	10	33.33%
pred. Survived	15	70	82.35%
class recall	25.00%	87.50%	

Figure 14: K-NN Results Obtained from

Figure 4 and 14 shows the classification of reviews using the Support Vector classifier in the training & testing dataset into the labels, i.e. survived and not survived of Cancer Patient data according to learning provided to the Support Vector Machine, K-NN and Naive Bayes classifiers respectively during the time of training. These also include the accuracies of the various classifiers for evaluating the results.

Table1- Comparing Three Models using their accuracies obtained from Rapid Miner

Model	Accuracy
SVM	80 %
KNN	75 %
Naïve Bayes	22 %

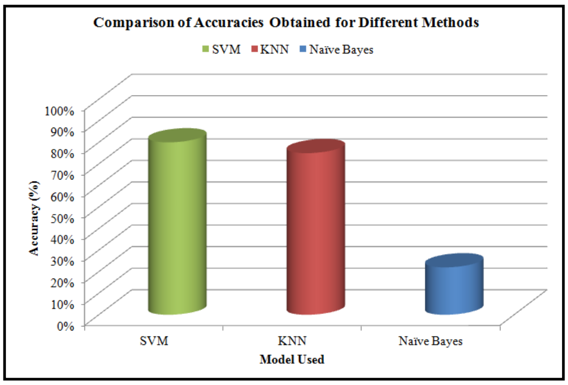


Figure 15: Graph showing accuracies obtained from different methods of mining

Table 1 and figure 15 show the accuracy comparison between the three classifiers i.e. Support Vector Machine, K-NN & Naïve Bayes respectively. The comparison shows that the accuracy of the support vector machine is highest in respect of the other two classifiers. The accuracy of KNN is also comparative with the Naïve Bayes classifiers for the input datasets. Naïve Bayes provides the least accuracy.

8. CONCLUSION

In this paper, the collaborative analysis on cancer patient datasets on different attributes of the values has been performed. For analysis of data i.e. parameter values has been done through the process of mining. Mining is performed by the tool called rapid miner and through the result reached at conclusion that Support vector machine provides best results among the results obtained from other two classifiers i.e. K-NN & Naïve Bayes classifiers. With this essential objective get achieved i.e. to analysis of large dataset of cancel patients which method is best. Trust this research is progressively will be being used as more individuals give their sentiment in analysis of the cancer patients.

In future, research can be utilized with more refine strategy to give more accuracy and manage the some other issue like decide for opinion, additionally build the span of the testing dataset and can look at the more different cancer datasets. Other Labels and application of classification methods can further be applied.

9. REFERENCES

[1] Abdelghani Bellaachia, Erhan Guven, “Predicting Breast Cancer Survivability Using Data Mining Techniques”, The George Washington University, Washington DC 20052

[2] Shweta Kharya, “USING DATA MINING TECHNIQUES FOR DIAGNOSIS AND PROGNOSIS OF CANCER DISEASE”, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012

[3] G. Ravi Kumar, Dr. G. A. Ramachandra, K.Nagamani, “An Efficient Prediction of Breast Cancer Data using Data Mining Techniques”, International Journal of Innovations in Engineering and Technology (IJIET)

[4] A.PRIYANGA, Dr.S.PRAKASAM, “The Role of Data Mining-Based Cancer Prediction system (DMBCPS) in Cancer Awareness”, International Journal of Computer Science and Engineering Communications- IJCSEC. Vol.1 Issue.1, December 2013

[5] Shelly Gupta, Dharminder Kumar, Anand Sharma “Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis”, Indian Journal Of Computer Science And Engineering (Ijcse)

[6] Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric Ringger :Natural Language Processing, Microsoft Research, Redmond, WA 98052, USA “Pulse: Mining Customer Opinions from Free Text”

[7] Minqing Hu and Bing Liu Department of Computer Science from University of Illinois at Chicago 851 South Morgan Street Chicago, IL 60607-7053 “Mining and Summarizing Customer Reviews.”

[8] Handbook of Natural Language Processing, Second Edition, edited by Nitin Indurkhya, Fred J. Damerau

[9] Data Mining Methods for the Content Analyst. By Kalev Leetaru.

[10] Data Mining Techniques, Tenth Edition by Arun K Pujari.

[11] Data Mining Concepts and techniques, Third Edition by Kamber.

[12] RapidMiner.com – Provides details of the RapidMiner Mining Tool and user manual

[13] Data Mining Concepts and techniques, Third Edition by Kamber – Provides good description of data mining concepts and techniques used in this work.