Symptom based clinical document clustering and Cancer prediction using data mining

Meenakshi Pandey, Prachi dalvi, Radhika Desai

Post- Graduate Student, Post- Graduate Student, Project Guide

Information Technology,

Thakur College of Science & Commerce, Mumbai-400101, India

Abstract : in this research paper detection of the cancer disease based on the symptoms .Medical records which contains vast amount of data includes unidentified patterns .Storing these vast amount of medical records is very necessary and important. Managing these records helps your work done very easily and faster

So here data mining is one of the technique used for managing these records. Data Mining is a technique which is used for mining the data. The term mining means bringing out patterns which are hidden and previously unidentified for better grasp of the particular problem. Several data mining techniques such as the Classification algorithms like , Genetic algorithm, Neural network, Artificial intelligence, Naïve Bayes and Clustering algorithms like SVM, .Hence in this paper we are going to make prediction and it will include part of clustering documents based on the symptoms that patient will provide. So if Symptom based clinical document clustering is done properly then a good report can be presented to the client. Proper symptoms with proper details of the patient must be given to the doctor so that the client or the patient will be aware of the disease on the time and hence the patient can be given the proper treatment.

Early detection of cancer plays a very important role in reducing deaths caused by cancer.

After that the prediction is done we need to make the document clustering based on the most common cluster we make cluster based on the type of the heterogeneous datasets and The technique which is used is python for the actual datasets implementation and accuracy is calculated through the algorithm for better accuracy

IndexTerms - **clustering, k-mean algorithm, naïve Bayes algorithm, cluster**

I. INTRODUCTION

II. Cancer is one of the dangerous disease in the world. Early detection and its prevention plays a very important role in reducing deaths caused by cancer. . Identification of genetic and environmental factors is very important in developing novel methods to detect and prevent cancer. Another part of our project includes cancer prediction which will also be using data mining technique. Here it will predict the cancer according to the symptoms, details the patient will provide to the doctor. If diagnosed with cancer then the patient will come to know the survivability rate and the risk status This will help a lot in future as many people does not comes to know the exact disease and hence they are diagnosed at such a stage where it comes to be the last stage of the cancer. This research helps in detection of a person's predisposition for cancer before going for clinical and lab tests which is cost and time consuming

III. Data mining is the process of is the process of extracting the useful information from the large amount of data .It is one of the essential method that is used to discovered the interesting pattern from the data

II. LITERATURE REVIEW

IV. In this research, we will mainly focus on the work carried out by different research work done by people on the detection of cancer based on the symptom and to detect the cancer by applying different prediction algorithm that so here we are applying prediction algorithm to find out the accuracy of the algorithm so for this purpose we have refer different research work what people have done previously for the cancer prediction and what are their finding

V.

VI. Pavithra R* S.Y. [1] Suggests that none of the data mining and statistical learning algorithms applied to breast cancer dataset outperformed the others in such way that it could be declared the optimal algorithm and none of the algorithm performed poorly as to be eliminated from future prediction model in breast cancer survivability tasks

VII. Rritu Chauhan [2] focuses on clustering algorithm such as HAC and K-Means in which, HAC is applied on K-means to determine the number of clusters. The quality of cluster is improved, if HAC is applied on K-means.

VIII.

IX. V.Krishnaiah [3] developed a prototype lung cancer disease prediction system using data mining classification techniques. The most effective model to predict patients with Lung cancer disease appears to be Naïve Bayes followed by IF-THEN rule, Decision Trees and Neural Network. For Diagnosis of Lung Cancer Disease Naïve Bayes observes better results and fared better than Decision Trees.

X. Sahar A. Mokhtar [4] have analyzed three different classification models for the prediction of the severity of breast masses namely the decision tree, artificial neural network and support vector machine

III.METHODOLOGY

Handling the large and complex datasets produced by cancer experiments is one of the prime challenges in the cancer research field. Data handling can be considered as a pipeline of successive steps like data pre-processing and data analysis . Some of the main considerations for the choice of the 95 appropriate data handling procedure are the analytical platform used to generate the data, the biological question to be answered and the inherent properties of the data. x Data Pre-processing (condensing and extracting features from the raw data)

Data condensing:-

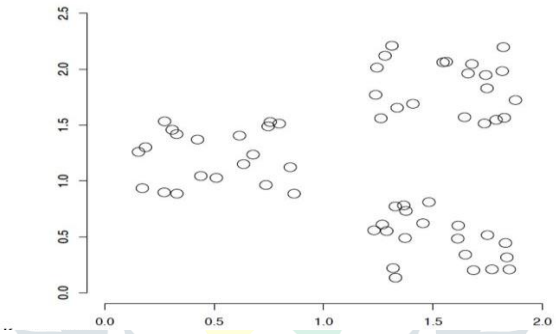
In cancer analysis, a raw dataset may contain tens or hundreds of spectra, each of them containing many hundreds or thousands of intensity measurements. Low-level pre-processing is often necessary in order to make sense of this large volume of data. Data pre-processing constitutes the initial step in data handling and its main goal is to extract all the relevant information from the raw data and summarize them in a single table

Data extraction: -

Here we have applied the preprocessing method to cancer dataset to prepare the raw data. Preprocessing is an important step that is used to transform the raw data into a format that makes it possible to apply data mining techniques and also to improve the quality of data. It can be noted from the related work that attribute selection plays an important role in identifying parameters that are important and significant for proper cancer detection and prevention

CLUSTERING FOR CANCER DISEASE DETECTION AND PREVENTION

Clustering is the process of dividing the object into similar groups . Each group, called cluster, consists of objects that are similar between
themselves and dissimilar to objects of other groups.it is also called as unsupervised learning the goal of a document clustering is to minimize intra-cluster distances between documents, while maximizing inter-cluster distances . A distance measure thus lies at the heart of document clustering. While the term segmentation and partitioning are sometime use as synonyms for clustering . No super-vision means that there is no human expert involve in this who has assigned documents to classes



IV.RESULT

To measure the probability of the cancer disease based on the symptom that patient provides in detail for prediction purpose algorithm like k-mean have used and it gives the proper prediction based on the patient datasets and gives detail of the benign and malignant cancer stage

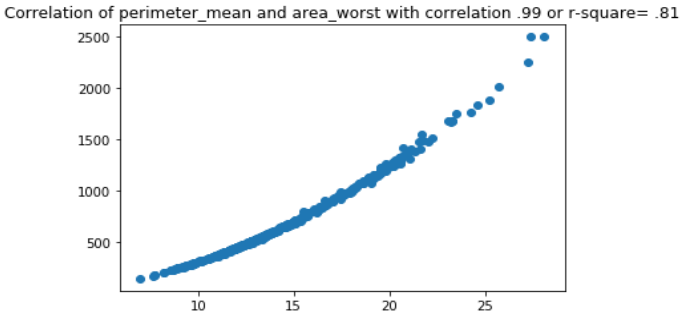


Fig. 1. Prediction graph Image.

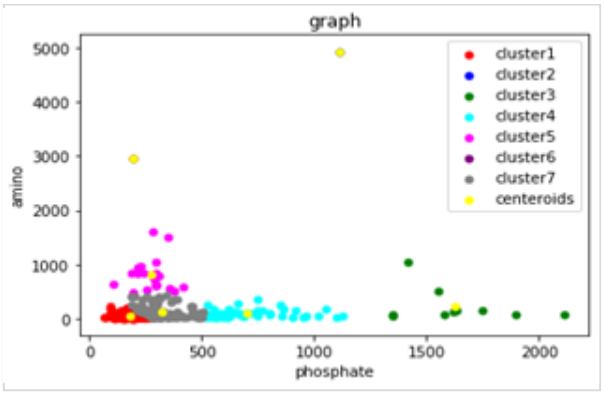


Fig. 2. Clustering graph Image.

Name	Type	Size	Value
X	int64	(583, 2)	[[187 18]
dataset	DataFrame	(583, 11)	Column names: A…
i	int	1	1
wcss	list	1	[82930799.71526…
y_kmeans	int32	(583,)	[0 3 6 ... 0 0 …

In python we have Variable explorer that shows the detail information about the dataset that is used for the clustering purpose it shows its data type and size of the dataset

Here we have defined the group of eight cluster defined in different colors and it will shows the different grouping of clusters

Algorithm	Accuracy
Naïve Bayes	93.70
SVM	62.93

Prediction Algorithm Analysis

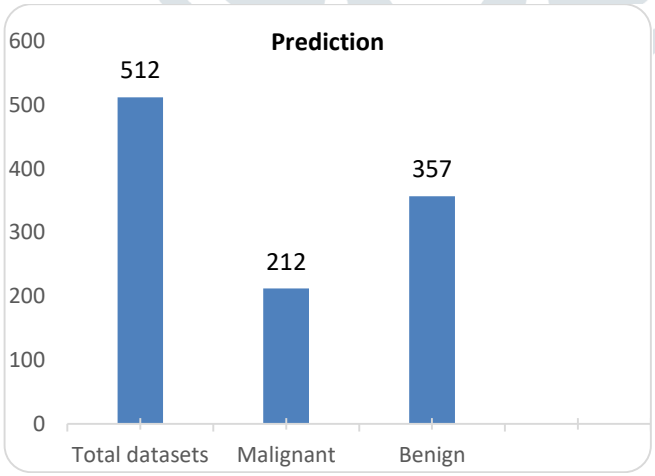


Fig. 3. Graphical Representation of Prediction model for cancer prediction

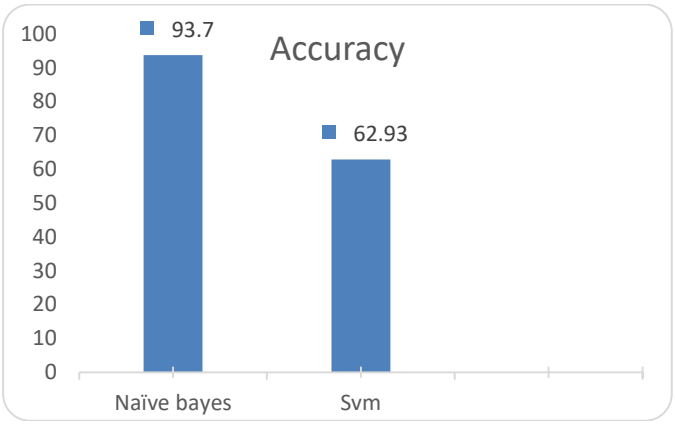


Fig.4. Graphical Representation of Accuracy calculated

V.CONCLUSION

This research paper proposes a method for early detection of cancer disease based on the prior symptom. After applying the K-mean algorithm for clustering features we had done prediction by applying Naïve Bayes algorithm on the cancer datasets,the disease was detected and the accuracy of the algorithm is also calculated.Basically, naïve Bayes algorithm gives the higher accuracy results, different algorithm we have used to calculate the accuracy and to find out which one is less time consuming and gives better accuracy these effective classification data helps to find the treatment to the patient In future a better method to predict the cancer disease can be found out with improvements in existing methods

Future Enhancement

Thus the survey helps to identify the data mining techniques to predict the cancer disease at earlier stage.
This application can be used by all patients or their family members who need help in emergency. Doctor can check patent information and past history at any time. Here it will predict the cancer according to the symptoms, details the patient will provide to the doctor. If diagnosed with cancer then the patient will come to know the survivability rate and the risk status This will help a lot in future as many people does not comes to know the exact disease and hence they are diagnosed at such a stage where it comes to be the last stage of the cancer it will save the time for the patient and they will get proper treatment on time

.

REFERENCES

[1] K.Arutchelvan1, Dr.R.Periyasamy CANCER PREDICTION SYSTEM USING DATAMINING TECHNIQUES International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056

[2] Neelam Singh and Santosh Kumar Singh Bhadauria Early Detection of Cancer Using Data Mining. International Journal of Applied Mathematical Sciences. pp. 47-52

[3] Roseline Jecintha Poonguzhali Study on Data Mining Techniques for Cancer Prediction System International Journal of Data Mining Techniques and Applications SSN: 2278-2419

[4] P. Saranya and B. SatheeskumarA Survey on Feature Selection of Cancer Disease Using Data Mining Techniques International Journal of Computer Science and Mobile Computing. IJCSMC, Vol. 5, Issue

[5] .P. Saranya and B. SatheeskumarA Survey on Feature Selection of Cancer Disease Using Data Mining Techniques International Journal of Computer Science and Mobile Computing. IJCSMC, Vol. 5, Issue

[6] G.Vijaya and Dr.A.Suhasini Early Detection of Lung Cancer using Data Mining Techniques international journal of engineering research & technology "ICSEM'13"

[7] S.Vijiyarani , S.Sudha Disease Prediction in Data Mining Technique – A Survey, International Journal of Computer Applications & Information Technology Vol. II, Issue I, January 2013 (ISSN: 2278-7720)