HUMAN HEART DISEASE PREDICTIION SYSTEM USING DATAMINING TECHNIQUES

Submitted to

SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL SCIENCES

In partial fulfilment for the award of the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE ANDENGINEERING

by

A. MANOJ KUMAR

(191611264)

Supervisor

MR. R SENTHIL KUMAR

SAVEETHA SCHOOL OF ENGINEERING

SAVEETHA INSTITUTE OF MEDICAL AND TECHNICAL

SCIENCES, CHENNAI – 602 105

APRIL 2019

BONAFIDE CERTIFICATE

Certified that this project report **“HUMAN HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES”** is the bonafide work of **“A. MANOJ KUMAR (Reg. No. 191611264)”**who carried out the project work under my supervision.

SIGNATURE	**SIGNATURE**
Dr. CHOKALINGAM	**MR.R. SENTHIL KUMAR**
HEAD OF THE DEPARTMENT	**PROJECT SUPERVISOR**
Professor, Dept. of CSE	Assistant Professor, Dept. of CSE Saveetha
School of Engineering	Saveetha School of Engineering
SIMATS, Chennai - 602 105	SIMATS, Chennai – 602105

INTERNAL EXAMINER	**EXTERNAL EXAMINER**

DECLARATION BY THECANDIDATE

I declare that the report entitled **“HUMAN HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES”** submitted by me for the degree of Bachelor of Engineering is the record of the project work carried out by me under the guidance of **“MR.R. SENTHIL KUMAR”** and this work has not formed the basis for the award of any degree, diploma, associateship, fellowship, titled in this or any University or other similar institution of higher learning.

SIGNATURE

A. MANOJ KUMAR

(Reg. No. 191611264)

ABSTRACT

In day to day life the Heart Diseases have been became very common because of hereditary. Heart diseases are very dangerous in our human body. For each and every individual has been different values for Blood pressure, cholesterol and pulse rate. According to medical survey reports the range of Blood pressure is 120/90, the cholesterol and pulse rate is 72. This paper describes the various techniques to predict the heart diseases. Some of the risk factors for heart diseases are Smoking, Diabetes, Stress, Drinking, Poor diet, Hyper tension, Obesity. Based on risk factors the heart diseases are predicted for Every year. Data Mining techniques are be used to discover hidden pattern form these data. Nowadays Data mining techniques has become advanced. Techniques used to predict the risk level of each person. Classification techniques are Decision trees, Naive Bayes, k-means algorithm. The Scope of this paper is to predict more accuracy in the presence of heart disease with reduced number of attributes. These techniques are efficiently compared through specificity, accuracy, sensitivity and error rate.

Keywords – Heart Diseases, Predict, Techniques, Data Mining, Risk level

ACKNOWLEDGEMENT

This project work would not have been possible without the contribution of many people. It gives me immense pleasure to express my profound gratitude to our honorable Chancellor, **Dr.N.M.Veeraiyan**, Saveetha Institute of Medical and Technical Sciences, for his blessings and for being a source of inspiration. I sincerely thank our Vice Chancellor,**Dr. Jawahar Nesan,** for his visionary thoughts and support.I am indebted to extend my gratitude to our Director Madam, **Mrs. Ramya Deepak,** Saveetha School of Engineering, for facilitating us all the facilities and extended support to gain valuable education and learning experience.

I register my special thanks to **Dr. D. Dhanasekaran,** Principal, Saveetha School of Engineering and **Dr. Chokalingam,** HoD, Department of Computer Science and Engineering, for the support given to me in the successful conduct of this project.I wish to express my sincere gratitude to my supervisor, **MR.R. SENTHIL KUMAR**, for her inspiring guidance, personal involvement and constant encouragement during the entire course of this work.

I am grateful to Project Coordinators Dr. K.Sashi Rekha and Mrs T. Devi, for their great support, Review Panel External, Internal Members and the entire faculty of the Department of Computer Science and Engineering, for their constructive criticisms and valuable suggestions which have been a rich source to improve the quality of this work.

A.MANOJ KUMAR

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

In current scenario, most of the patients complain about the various test conducted by hospitals for diagnosis, which cause them both money and time loss. In some cases, the delay in diagnosing the patient correctly results in the delay of starting the proper treatment. It may lead to disastrous consequences in case of deadly diseases. Sometimes, after conducting so many tests, the patient's results are negative and results in both money and time loss. All these are caused due to the doctor's wrong intuitions and inexperience. But we cannot blame doctors for this, since they suggest each diagnosis test based on their intuitions and experience after examining the patient which may go wrong. This project suggests a solution for overcoming these problems by utilizing the large amount of patient records collected by various hospitals or health care centers with the help of data mining techniques. The system thus developed can be used as a decision support system to seek a second opinion as from an experienced doctor. This will help to avoid unnecessary test conducted for diagnosis thereby saving both time and money. This helps hospitals to provide quality services at affordable cost. Doctors, medical students and nurses can use this system for second opinion. Patients can use this system if they have their test results.

Nowadays, some hospitals use decision support systems for simple queries, such as what is the average age of patients having a particular disease, whether it is more prevalent among men or women, whether it is more common among young or old people etc. They cannot run complex queries such as whether a patient is affected by a particular disease or not, which treatment is more effective among patients with deadly diseases after crossing a particular stage etc. Currently, the large amount of data collected from patients are simply stored in hospitals or health care centers and is not used for any other purpose. If we utilize the knowledge hidden in these databases, we can find solutions for many other problems that exist in health care field regarding the services provided to the patients.

This project presents a solution for diagnosing patients with heart disease. This decision support system uses data mining technique Naïve Bayes algorithm for predicting whether a patient have heart disease or not and uses smoothing technique Laplace smoothing for increasing prediction accuracy.

There are number of factors to increase the heart disease.

- Poor diet
- Obesity
- High blood pressure
- Hyper tension
- High cholesterol
- Family history

TYPES OF HEART DISEASE:

The disease of heart and blood vessels within it. The various types of heart disease given below. They are

- Cardiovascular disease
- Arythametic disease
- Coronary artery disease
- Disease of Heart valve

DATA MINING ALGORITHM:

Classification

Classification is based on machine learning. Classification is used to classify every item in a set of data into one of predefined set of classes or groups. Classification technique makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics.

Prediction

Prediction is a data mining technique that discovers the relationship between independent variables and relationship among dependent and independent variables.

1.2 OBJECTIVES OF PROJECT:

The main objective of this research is to develop a prototype Health Care Prediction System using, Naive Bayes. The System can discover and extract hidden knowledge associated with diseases (heart attack, cancer and diabetes) from a historical heart disease database. It can answer complex queries for diagnosing disease and thus assist healthcare practitioners to make intelligent clinical decisions which traditional decision support systems cannot. By providing effective treatments, it also helps to reduce treatment costs. To enhance visualization and ease of interpretation, it displays the results in tabular and PDF forms.

Most hospitals maintain a hospital information system. This system contains a large amounts of patient data. This information is largely not accessed. Data mining techniques are used to convert a data into useful information. The main objective of this research is o developed a one data mining modeling technique namely as improved naïve bayes. It can extract the hidden knowledge from heart disease historical dataset.

Three different supervised machine learning algorithms i.e. Naive Bayes, K-NN, Decision List algorithm have been used for analyzing the dataset. The tool is used to classify the data and the data is evaluated using 10-fold cross validation and the results are compared. Decision Tree is a popular classifier which is simple and easy to implement. It requires no domain knowledge or parameter setting and can handle high dimensional data. The results obtained from Decision Trees are easier to read and interpret. The drill through feature to access detailed patients‟ profiles is only available in Decision Trees. Naïve Bayes is a statistical classifier which assumes no dependency between attributes. It attempts to maximize the posterior probability in determining the class. The advantage of using naive bayes.

The k-nearest neighbour algorithm (k-NN) is a method for classifying objects based on closest training data in the feature space. k-NN is a type of instance-based learning. The k-nearest neighbour algorithm is amongst the simplest of all machine learning algorithms.

1.3 SCOPE OF THE PROJECT

Here the scope of the project is that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g significantly improve the quality of clinical decisions .The main objective of this research is to develop a prototype Heart Disease Prediction System (HDPS) using three data mining modeling techniques, namely, Decision Trees, Naïve Bayes and Neural Network .So it provides effective treatments, it also helps to reduce treatment costs and also enhances visualization and ease of interpretation. With immense knowledge and accurate data in that field. Large corporations invest heavily in this kind of activity to help focus attention on possible events and risks that are involved. Such work brings together all available past and current data, as a basis on which to develop reasonable expectations about the future.

Nowadays, some hospitals use decision support systems for simple queries, such as what is the average age of patients having a particular disease, whether it is more prevalent among men or women, whether it is more common among young or old people etc. They cannot run complex queries such as whether a patient is affected by a particular disease or not, which treatment is more effective among patients with deadly diseases after crossing a particular stage etc. Currently, the large amount of data collected from patients are simply stored in hospitals or health care centers and is not used for any other purpose. If we utilize the knowledge hidden in these databases, we can find solutions for many other problems that exist in health care field regarding the services provided to the patients. This paper presents a solution for diagnosing patients with heart disease. This decision support system uses data mining technique Naïve Bayes algorithm for predicting whether a patient have heart disease or not and uses smoothing technique Laplace smoothing for increasing prediction accuracy.

CHAPTER 2
LITERATURE SURVEY

TITLE 1: Intelligent heart disease prediction system using data mining techniques
AUTHOR: Sellappan Palaniappan
YEAR: 2010
DESCRIPTION:

The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined"; to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation. This research has developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Naive Bayes and Neural Network. Results show that each technique has its unique strength in realizing the objectives of the defined mining goals. IHDPS can answer complex ";what if"; queries which traditional decision support systems cannot. Using medical profiles such as age, sex, blood pressure and blood sugar it can predict the likelihood of patients getting a heart disease. It enables significant knowledge, e.g. patterns, relationships between medical factors related to heart disease, to be established. IHDPS is Web-based, user-friendly, scalable, reliable and expandable.

TITLE 2: Effective heart disease prediction system using data mining techniques

AUTHOR: Poornima Singh,[1] Sanjay Singh,[2]
YEAR: 2017
DESCRIPTION:

The health care industries collect huge amounts of data that contain some hidden information, which is useful for making effective decisions. For providing appropriate results and making effective decisions on data, some advanced data mining techniques are used. In this study, an effective heart disease prediction system is developed using neural network for predicting the risk level of heart disease. The system uses 15 medical parameters such as age, sex, blood pressure, cholesterol, and obesity for prediction. The EHDPS predicts the likelihood of patients getting heart disease. It enables significant knowledge, eg, relationships between medical factors related to heart disease and patterns, to be established. We have employed the multilayer perceptron neural network with backpropagation as the training algorithm. The obtained results have illustrated that the designed diagnostic system can effectively predict the risk level of heart diseases.

TITLE 3: **Data Mining Approach to Detect Heart Diseases**
AUTHOR: Vikas Chaurasia
YEAR: 2013
DESCRIPTION:

Globally, heart diseases are the number one cause of death. About 80% of deaths occurred in low- and middle income countries. If current trends are allowed to continue, by 2030 an estimated 23.6 million people will die from cardiovascular disease (mainly from heart attacks and strokes). The healthcare industry gathers enormous amounts of heart disease data which, unfortunately, are not “mined” to discover hidden information for effective decision making. The reduction of blood and oxygen supply to the heart leads to heart disease. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. This research paper intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques which will be useful for medical practitioners to take effective decision. The objective of this research work is to predict more accurately the presence of heart disease with reduced number of attributes. Originally, thirteen attributes were involved in predicting the heart disease. Thirteen attributes are reduced to 11 attributes. Three classifiers like Naive Bayes, J48 Decision Tree and Bagging algorithm are used to predict the diagnosis of patients with the same accuracy as obtained before the reduction of number of attributes. In our studies 10- fold cross validation method was used to measure the unbiased estimate of these prediction models.

TITLE 4: **Predicting the Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods**
AUTHOR: V. Manikantan & S. Latha
YEAR: 2014
DESCRIPTION:

Medicinal data mining methods are used to analyze the medical data information resources. Medical data mining content mining and structure methods are used to analyze the medical data contents. The effort to develop knowledge and experience of frequent specialists and clinical selection data of patients collected in databases to facilitate the diagnosis process is considered a valuable option. Diagnosis of heart disease is a significant and tedious task in medicine. The term Heart disease encompasses the various diseases that affect the heart. The exposure of heart disease from various factors or symptom is an issue which is not complimentary from false presumptions often accompanied by unpredictable effects. Association rule mining procedures are used to extract item set relations. Item set regularities are used in the rule mining process. The data classification is based on MAFIA algorithms

which result in accuracy, the data is evaluated using entropy based cross validations and partition techniques and the results are compared. Here using the C4.5 algorithm as the training algorithm to show rank of heart attack with the decision tree. Finally, the heart disease database is clustered using the K-means clustering algorithm, which will remove the data applicable to heart attack from the database. The results showed that the medicinal prescription and designed prediction system is capable of prophesying the heart attack successfully.

TITLE 5: Prediction of heart disease using data mining techniques

AUTHOR: Rithika chanda

YEAR: 2016

DESCRIPTION:

The healthcare industry is a vast field with a plethora of data about patients, added to the huge medical records every passing day. In terms of science, this industry is ’information rich’ yet ’knowledge poor’. However, data mining with its various analytical tools and techniques plays a major role in reducing the use of cumbersome tests used on patients to detect a disease. The aim of this paper is to employ and analyze different data mining techniques for the prediction of heart disease in a patient through extraction of interesting patterns from the dataset using vital parameters. This paper strives to bring out the methodology and implementation of these techniques-Artificial Neural Networks, Decision Tree and Naive Bayes and stress upon the results and conclusion induced on the basis of accuracy and time complexity. By far, the observations reveal that Artificial Neural Networks outperformed Naive Bayes and Decision Tree.

TITLE 6: Analysis of Heart Disease Prediction Using Datamining Techniques

AUTHOR: S.SHARMILA,

YEAR: 2012

DESCRIPTION:

Data mining is the very vast area in research field. Health care is most important organization in our world. The various data mining techniques are used to predict the heart disease. Heart disease is very dangerous disease in our human body. Heart is important part in our body. Data mining prediction tool is play on vital role in healthcare. This paper analysis the various technique to predict the heart disease.

CHAPTER 3
PROBLEM STATEMENT AND METHODOLOGY

3.1 PROBLEM DEFINATION:

The heart is very important part of human body. Which pumps blood into the entire body? If circulation of blood in body is inefficient the organs like brain suffer and if heart stops working altogether, death occurs within minutes. Life is completely dependent on efficient working of the heart. The term Heart disease refers to disease of heart & blood vessel system within it[9,10,11].

Some of the risk factors for heart disease are

1. **Smoking**: Smokers risk a heart attack twice as much as non smokers.
2. **Cholesterol:** A diet low in cholesterol and saturated Tran's fat will help lower cholesterol Levels and reduce the risk of heart disease.
3. **Blood pressure**: High BP leads to heart Attack
4. **Diabetes**: Diabetes if not controlled can lead to significant heart damage including heart attack and death.
5. **Sedentary life style**: Simple leisure time activities like gardening and walking can lower our risk of heart disease.
6. **Eating Habits:** Heart healthy diet, low in salt, saturated fat, Trans fat, cholesterol and refined sugars will lower our chances of getting heart disease.
7. **Stress**: Poorly controlled stress an danger can lead to heart attacks and strokes.

3.2 METHODOLOGY

3.2.1EXISTING SYSTEM:

Since the knowledge gained from the different experts are a high-level description of the problem from the medical point of view, a literature review was carried out and relevant works related to
data mining and heart disease have been reviewed to have more knowledge about the domain. Furthermore, a real time observation of the system was performed to understand the business process of the hospital. After recording, the new database now contains 7,339instances each instance resembling a single file. In the next step I have selected appropriate data mining technique for developing a predictive model.

3.2.2PROPOSED SYSTEM:

The decision support system developed uses the above techniques to predict heart disease. Users can use either the prediction with 13 attribute if they have the test results of fluoroscopy, thallium test, ECG, ST depression etc. or the prediction with 6 attributes if they don't.

The 6 attributes used were age, gender, blood pressure, fasting blood sugar, cholesterol and exercise induced angina. These were selected since their results can be easily inputted by users without much help. Then the performance of the classifier thus formed, is evaluated. Then all 13 medical attributes selected from data source were used and its performance is evaluated.

3.3TECHNIQUES:

Data mining is defined as "a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database" or as "a process of selection, exploration and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of database".

Data Mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

Data mining uses two strategies: supervised and unsupervised learning. In supervised learning, a training set is used to learn model parameters whereas in unsupervised learning no training set is used (e.g., k means clustering is unsupervised). Each data mining technique serves a different purpose depending on the modeling objective. The two most common modeling objectives are classification and prediction. Classification models predict categorical labels (discrete, unordered) while prediction models predict continuous-valued functions. Decision Trees and Neural Networks use classification algorithms while Regression, Association Rules and Clustering use prediction algorithms.

3.3.1 Naïve Bayes Algorithm

Naive Bayes or Bayes Rule is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. The Naïve Bayes Classifier technique is mainly applicable when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naïve Bayes model recognizes the characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state. Naive Bayes algorithm is preferred in the following cases. When the dimensionality of data is high.

When the attributes are independent of each other. Otherwise, attributes are assumed to be independent in order to simplify the computations involved and, in this sense, is considered "naïve".

- When we expect more efficient output, as compared to other methods output.
- Exhibits high accuracy and speed when applied to large databases.

1) Bayes Rule:

A conditional probability is the likelihood of some conclusion say C, given some evidence/observation, E, where a dependence relationship exists between C and E.

This probability is denoted as P(C |E) where

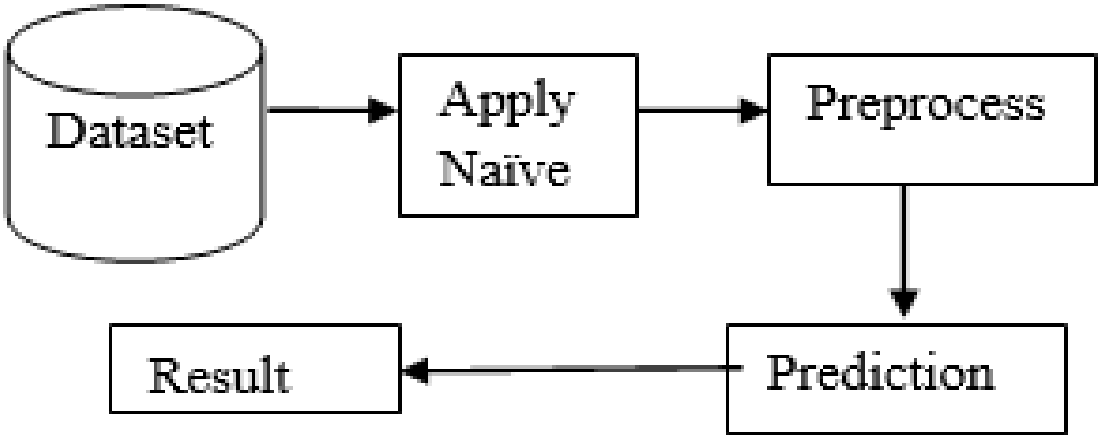
$$P(C|E) = \frac{P(E|C)P(C)}{P(E)}$$

3.3.2Decision Tree

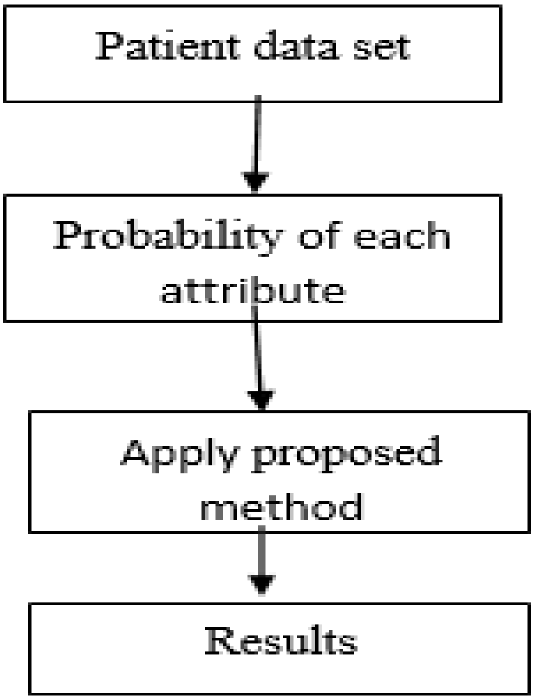
It is also based on Hunt’s algorithm. J48 handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes, J48 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. J48 uses Gain Ratio as an attribute selection measure to build a decision tree. It removes the biasness of information gain when there are many outcome values of an attribute. At first, calculate the gain ratio of each attribute. The root node will be the attribute whose gain ratio is maximum. J48 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification.

3.4 SYSTEM ARCHITECTURE:

-



3.5 STEPS TO BE PROCESSED:



CHAPTER4
SYSTEM IMPLEMENTATION

4. 1MODULES DESCRIPTION

- Collect data set
- Data integration
- Data transformation
- Apply Naïve bayes
- Preprocess
- Prediction
- Accuracy
- Results

4.1.1 DATA SET:

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang
63	1	3	145	233	1	0	150	0
37	1	2	130	250	0	1	187	0
41	0	1	130	204	0	0	172	0
56	1	1	120	236	0	1	178	0
57	0	0	120	354	0	1	163	1
57	1	0	140	192	0	1	148	0
56	0	1	140	294	0	0	153	0
44	1	1	120	263	0	1	173	0
52	1	2	172	199	1	1	162	0
57	1	2	150	168	0	1	174	0
54	1	0	140	239	0	1	160	0
48	0	2	130	275	0	1	139	0
49	1	1	130	266	0	1	171	0
64	1	3	110	211	0	0	144	1

4.1.2APPLY NAÏVE BAYES:

Naïve Bayes Algorithm Naive Bayes or Bayes Rule is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. The Naïve Bayes Classifier technique is mainly applicable when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naïve Bayes model recognizes the characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state. Naive Bayes algorithm is preferred in the following cases.

- When the dimensionality of data is high
- When the attributes are independent of each other. Otherwise, attributes are assumed to be independent in order to simplify the computations involved and, in this sense, is considered “naïve”.
- When we expect more efficient output, as compared to other methods output.
- Exhibits high accuracy and speed when applied to large databases.

1) Bayes Rule: A conditional probability is the likelihood of some conclusion say C, given some evidence/observation, E, where a dependence relationship exists between C and E. This probability is denoted as P(C |E) where (1) 2) Naive Bayesian Classification Algorithm :
2) The Naive Bayesian classifier, or simple Bayesian classifier, works as follows:
    - Let D be a training set of tuples and their associated class labels. As usual, each record is represented by an ndimensional attribute vector, X=(x1, x2…, xn-1, xn), depicting n measurements made on the tuple from n attributes, i.e. A1 to An.
    - Suppose that there are m numbers of classes for prediction, C1, C2… Cm. Given a record, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the Naïve Bayesian classifier predicts that tuple X belongs to the class Ci if and only if P (Ci|X) >P (Cj|X) for 1≤ j≤ m and j≠ i (2) Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes theorem (3)
    - As P(X) is constant for all classes, only P (X|Ci)* P(Ci) need be maximized. If the class prior probabilities are not known, then it is often assumed that the classes are equally likely, that is, P(C1) =P(C2) =…P(Cm-1) =P(Cm) and we would therefore maximize P(X|Ci). Otherwise, we maximize P (X|Ci) P(C
    - i). Note that the class prior probabilities may be estimated by P (Ci) = |Ci, D| / |D| (4) where |Ci, D| is the number of training tuples of class Ci in D.
    - Given data sets with many attributes, it would be extremely computationally expensive to compute P(X|Ci). To reduce computation in evaluating P(X|Ci), the naïve assumption of class conditional independence is made.
    - This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes). Thus, We can easily estimate the probabilities
    - P(|Ci), P(|Ci)… P(|Ci) from the database training tuples. Recall that here refers to the value of attribute for tuple X. For each attribute, we will see that whether the attribute is categorical or continuous-valued. For instance, to compute P(X|Ci).

HARDWARE SET UP:

- PROCESSOR SPEED: 2GHZ
- TOTAL RAM: 3GB
- OPERATING SYSTEM: WINDOWS 10
- HARD DISK:250 GB
- TOOLS: R STUDIO R-3.5

ALGORITHMS:

CLASSIFICATION USING RANDOM FOREST

Random forests (RF) are combination of tree predictors using decision tree such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. They are more robust with respect to noise. It is a supervised classification algorithm used for the prediction and it is considered as the superior due to its large number of trees in the forest giving improved accuracy than decision trees. Typically, the trees are trained independently and the predictions of the trees are combined through averaging. Random forest algorithm can use both for classification and the regression based on the problem domain. The algorithm for random forest is given below:

Step 1: Randomly select k features from entire m features, where k << m.

Step 2: Surrounded by the k features, calculate the node "d" using the best split point.

Step 3: Split the node into daughter nodes using the best split.

Step 4: Repeat 1 to 3 steps until l number of nodes has been reached.

Step 5: Construct forest by repeating steps 1 to 4 for n number times to create n number of trees. Firstly, the k features are taken out of total m features. In the next stage, in each tree randomly select k features in order to find the root node by using the best split approach. The next stage involves calculating the daughter nodes using the same best split approach for the heart disease dataset. Similarly, the tree is formed from the root node and until all the leaf nodes are generated from the attributes. This randomly created tree forms the random forest that is used for making heart disease prediction in patients.

CLASSIFICATION USING DECISION TREE

Decision Tree (DT) is a simple and easy to implement classifier. The bit through feature to access in depth patients' profiles is only obtainable in Decision Trees. Decision tree builds classification or regression models in the structure of a tree making it simple to debug and handle. Decision trees can handle both categorical and numerical data. The algorithm works by finding the information gain of the attributes and taking out the attributes for splitting the branches in threes. The information gain for the tree is identified using the below given

EG: E(S)= -P(P)log2P(P)-P(N)log2P(N)

The algorithm for the decision tree is given below:

Step 1: Identify the information gain for the attributes in the dataset

Step 2: Sort the information gain for the heart disease datasets in descending order.

Step 3: After the identification of the information gain assign the best attribute of the dataset at the root of the tree.

Step 4: Then calculate the information gain using the same formula.

Step 5: Split the nodes based on the highest information gain value.

Step 6: Repeat the process until each attributes are set as leaf nodes in all the branches of the tree.

CLASSIFICATION USING NAÏVE BAYES

Naïve Bayes (NB) is a statistical classifier which assumes no enslavement between attributes. Naive Bayes [15] is based on Bayes rule and it assumes that attributes are independent of each other. The working principle of naïve Bayes classifier is as follows:

- Training Step: By assuming predictors to be conditionally independent given for a class, the method estimates the parameters of a probability distribution known as the prior probability from the training data.
- Prediction Step: For unknown test data, the method computes the posterior probability of the dataset which is belonging to each class. The method finally classifies the test data based upon the largest posterior probability from the set.

CHAPTER 5
OUTPUT SCREEN SHOTS WITH EXPLANATION

5.1 PERFORMANCE MEASURE:

The following measures were used to analyze the performance of the prediction system. E. Classifier Evaluation Measures Neg are the negative tuples that were correctly labeled by the classifier. False positives (F Pos) are the negative tuples that were incorrectly labeled by the classifier, while false negatives are the positive tuples that were incorrectly labeled by the classifier.

The sensitivity and specificity measures can be used for calculating performance and precision is used for the percentage of samples labeled as "diseased" or "1".

1) Sensitivity:

It means recognition rate or true positive rate. It is used for measuring the percentage of sick people from the dataset.

Sensitivity=TRUE POSITIVE/POSITIVE

Where TruePos is the number of true positives (i.e. "Present" samples that were correctly classified) and Pos are the number of positive samples.

2) Specificity:

It means true negative rate. It is used for measuring the percentage of healthy people who are correctly identified from the dataset.

Specificity=TRUE NEGATIVE/NEGATIVE

TrueNeg is the number of true negatives (i.e." Absent" samples that were correctly classified) and Neg is the number of negative samples.

3) Precision:

It is used for the percentage of samples labeled as "diseased" or "1". It is also known as positive predictive value. It is defined as the average probability of relevant retrieval.

Precision=TRUE POSIIVE/TRUEPOSITIVE+FALSE POSITIVE

FPos is the number of false positives ("Absent" samples that were incorrectly labeled as "diseased" or "1")

4) Accuracy:

Accuracy = Number of correctly classified samples/Total number of samples

The true positives, true negatives, false positives and false negatives are also useful in assessing the costs and benefits (or risks and gains) associated with a classification model.

5) Confusion Matrix:

It is used for displaying the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data.

The matrix is represented in the form of n-by-n, where n is the number of classes. The accuracy of classification algorithm can be calculated using this matrix.

5.2 ACCURACY:

TO PREDICT ACCUARCY FOR NAÏVE BAYES:

```
heart <- read.csv("C:/Users/Dell/Downloads/heart.csv")
View(heart)
```

FINDING THE CONDITIONAL PROBABILITIES

```
> library(e1071)
> model <- naiveBayes(chol ~ ., data = heart)
> class(model)
[1] "naiveBayes"
> summary(model)
          Length Class  Mode
apriori   152    table  numeric
tables     13    -none- list
levels      0    -none- NULL
isnumeric  13    -none- logical
call        4    -none- call
> tbl_list <- sapply(heart[-10], table, heart[ , 10])
>tbl_list <- lapply(tbl_list, t)
>cond_probs <- sapply(tbl_list, function(x) {
       apply(x, 1, function(x) {
          x / sum(x) }) })
>cond_probs <- lapply(cond_probs, t)
>print(cond_probs)
```

PREDICTI THE MODEL:

```
>preds <- predict(model, newdata = heart)
>conf_matrix <- table(preds, heart$class)
> conf_matrix
< table of extent 0 x 0 >
> lvs <- c("with heart disease", "without heart disease")
> chol<- factor(rep(lvs, times = c(86, 258)),
+                               levels = rev(lvs))
> pred <- factor(
+       c(
+               rep(lvs, times = c(54, 32)),
+                rep(lvs, times = c(27, 231))),
+         levels = rev(lvs))
> xtab <- table(pred, chol)
> library(caret)
> confusionMatrix(xtab)
```

OUTPUT:

```
Confusion Matrix and Statistics

                      chol
pred                     without heart disease with heart disease
  without heart disease                  231                32
  with heart disease                    27                54

               Accuracy : 0.8285
                95% CI : (0.7844, 0.8668)
               No Information Rate : 0.75
               P-Value [Acc > NIR] : 0.0003097
               Kappa : 0.5336
               Mcnemar's Test P-Value : 0.6025370
               Sensitivity : 0.8953
               Specificity : 0.6279
               Pos Pred Value : 0.8783
               Neg Pred Value : 0.6667
               Prevalence : 0.7500
               Detection Rate : 0.6715
               Detection Prevalence : 0.7645
               Balanced Accuracy : 0.7616
        'Positive' Class : without heart disease
```

FINDING THE SENSITIVITY, SPECIFICITY, PRECISION, RECALL, F MEASURE

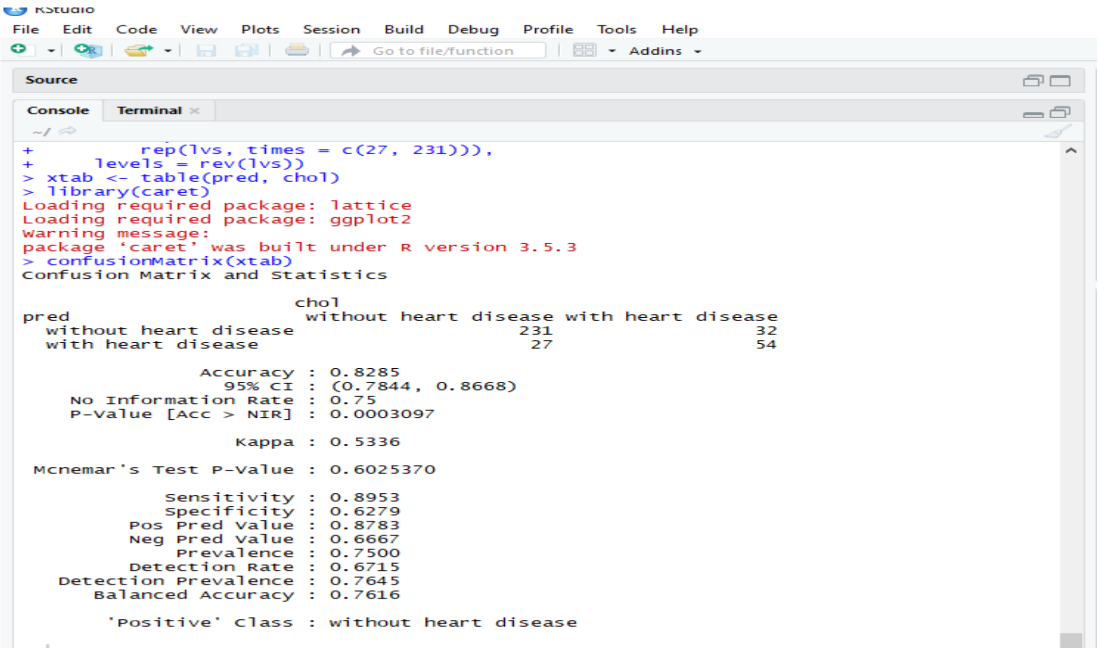
```
> library(caret)
> result <- confusionMatrix(pred,chol)
> precision <- result$byClass['Pos Pred Value']
> recall <- result$byClass['Sensitivity']
> recall <- result$byClass['Specificity']
> f_measure <- 2 * ((precision * recall) / (precision + recall))
```

OUTPUT:

```
> precision
Pos Pred Value
     0.878327
> recall
Specificity
   0.627907
>recall
Sensitivity
```

```
  0.8953488
>f_measure
Pos Pred Value
     0.7323001
```

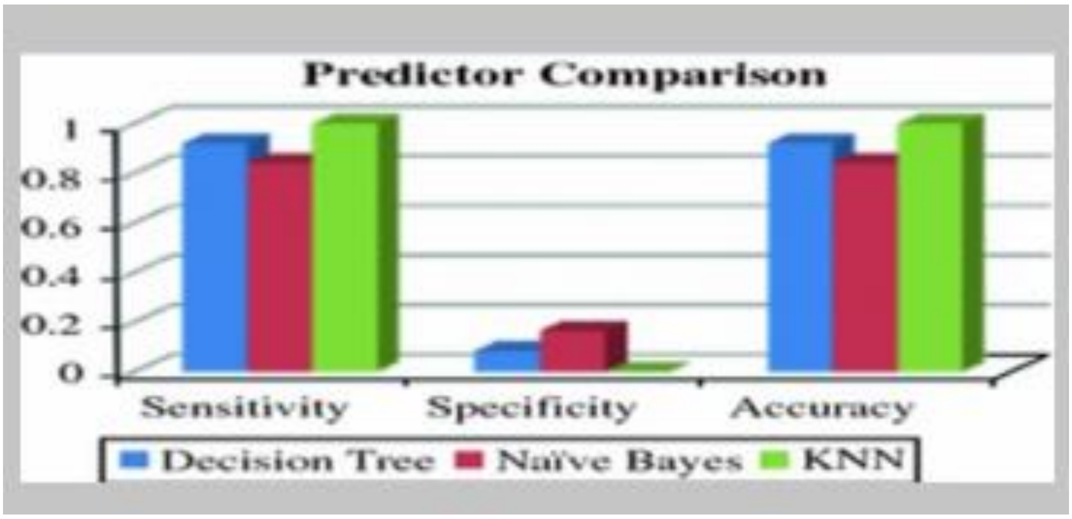
SCREENSHOT:


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Source
Console
Terminal
~/

```
+           rep(lvs, times = c(27, 231))),
+       levels = rev(lvs))
> xtab <- table(pred, chol)
> library(caret)
Loading required package: lattice
Loading required package: ggplot2
Warning message:
package ‘caret’ was built under R version 3.5.3
> confusionMatrix(xtab)
Confusion Matrix and Statistics

                       chol
pred                     without heart disease with heart disease
  without heart disease                    231                 32
  with heart disease                        27                 54

               Accuracy : 0.8285
                 95% CI : (0.7844, 0.8668)
    No Information Rate : 0.75
    P-Value [Acc > NIR] : 0.0003097

                  Kappa : 0.5336

 Mcnemar's Test P-Value : 0.6025370

            Sensitivity : 0.8953
            Specificity : 0.6279
         Pos Pred Value : 0.8783
         Neg Pred Value : 0.6667
             Prevalence : 0.7500
         Detection Rate : 0.6715
   Detection Prevalence : 0.7645
      Balanced Accuracy : 0.7616

       'Positive' Class : without heart disease
```

5.3 Output Analysis

5.3.1 Performance Graph



CHAPTER6
CONCLUSION AND FUTURESCOPE

6.1 CONCLUSION:

Heart disease is one of the leading causes of deaths worldwide and the early prediction of heart disease is very important. In this study prove that the proposed new algorithm is a highest accuracy compare with another algorithm. From the experiment it clear that proposed method is more accurately classify the recodes as compared to other method. Proposed method considers all attribute given to heart attack condition. Proposed method is also simple to under stands and calculation is also easy. We have taken only ten attribute which are mainly responsible for heart attack, in future we have consider more than ten attribute which are also responsible for heart attack. The main motivation of this paper is to study the various data mining techniques available to predict the heart disease and to compare them to find the best method of predictions. We studied about various techniques using in Heart Disease prediction system. We focus on classification method and prediction method of data mining using Naive Bayes and Improved K-means algorithm. The accuracy of the algorithm used in each technique can be enhanced by hybridizing or combining algorithm to single algorithm. (Naive Bayes and Improved K-means algorithm). The accuracy predicts in naïve bayes algorithm is 0.86%.

6.2 FUTURE ENHANCEMENT:

As a future work, the researcher has planned to perform additional experiments with more dataset and algorithms to improve the classification accuracy and to build a model that can predict specific heart disease types. The performances of the models were evaluated using the standard metrics of accuracy, precision, recall and F-measure the training and test data samples. All eight models performed well in predicting heart disease cases. The most effective model to predict patients with heart disease appears to be a J48 classifier implemented on selected attributes with a classification accuracy of 95.56%.

REFERENCES

[1] Mrs. G. Subbalakshmi and Mr. K. Ramesh Decision Support in Heart Disease Prediction System using Naive Bayes Journal of Computer Science and Engineering (IJCSE) ISSN: 0976-5166 Vol. 2 No. 2 Apr-May 2015

[2] V. Manikantan & S. Latha Predicting the Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods International Journal on Advanced Computer Theory and Engineering (IJACTE) ISSN 2319 – 2526, Volume-2, Issue-2, 2013.

[3] Rajkumar, A. and G.S. Reena, Diagnosis of Heart Disease Using Datamining Algorithm. Global Journal of Computer Science and Technology, 2012. Vol. 10 (Issue 10).

[4] K.Sudhaka, Dr. M.Manimekalai, Study of Heart Disease Prediction System using Data Mining, ISSN: 2277 128X, 2015.

[5] Latha Parthiban and R.Subramanian, Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm, International Journal of Biological and Medical Sciences, 2010.

[6] K. Srinivas, B. Kavitha Rani and Dr. A. Govrdhan, "Application of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", International Journal on Computer Science and Engineering, Vol. 02, No. 02, pp. 250 - 255, 2016.

[7] T.Georgeena.S. Thomas, Siddhesh.S. Budhkar, Siddhesh.K. Cheulkar, Akshay.B.Choudhary, Rohan Singh" Heart Disease Diagnosis System Using naïve bayes Algorithm"; International Journal of Advanced Research in Computer Science and Software Engineering Volume 5, Issue 2, February 2016.

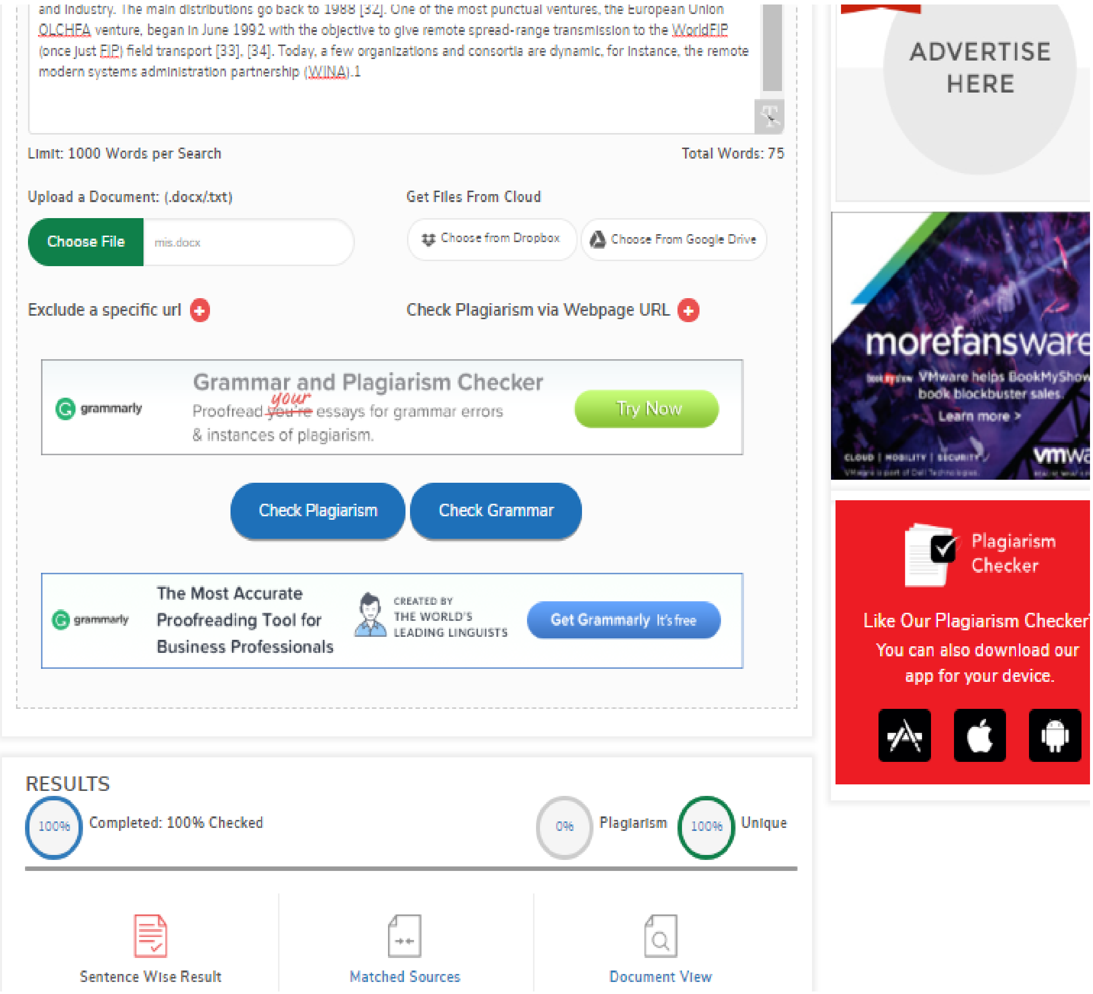
[8] Shruti Ratnakar, K. Rajeswari, Rose Jacob., Prediction Of Heart Disease Using Genetic Algorithm For Selection Of Optimal Reduced Set Of Attributes, International Journal of Advanced Computational Engineering and Networking, ISSN (PRINT): 2320-2106, Volume – 1, Issue – 2, 2013.

[9] BVenkatalakshmi, M.V Shivsankar, "Heart Disease Diagnosis Using Predictive Data mining", International Journal of Innovative Research in Science, Engineering and Technology ISSN 2319-8753 Vol.3, Special Issue 3, pp. 1873-1877 ©2015ICIET.

[10] D. Shanthi, G. Sahoo and N. Saravanan "Input Feature Selection using Hybrid Neuro-Genetic Approach in the Diagnosis of Stroke Disease", International Journal of Computer Science and Network Security, Vol. 8, No.12, pp. 99 - 106, 2015.

APPENDIX I

PLAGRISIM REPORT



and industry. The main distributions go back to 1988 [32]. One of the most punctual ventures, the European Union OLCHFA venture, began in June 1992 with the objective to give remote spread-range transmission to the WorldFIP (once just FIP) field transport [33], [34]. Today, a few organizations and consortia are dynamic, for instance, the remote modern systems administration partnership (WINA).1

Limit: 1000 Words per Search

Total Words: 75

Upload a Document: (.docx/.txt)

Choose File mis.docx

Get Files From Cloud

RESULTS

Completed: 100% Checked

0% Plagiarism

100% Unique

APPENDIX II

SAMPLE CODE

TO PREDICT ACCUARCY FOR NAÏVE BAYES:

```
heart <- read.csv("C:/Users/Dell/Downloads/heart.csv")
View(heart)
FINDING THE CONDITIONAL PROBABILITIES
> library(e1071)
> model <- naiveBayes(chol ~ ., data = heart)
> class(model)
[1] "naiveBayes"
> summary(model)
          Length Class  Mode
apriori   152    table  numeric
tables     13    -none- list
levels      0    -none- NULL
isnumeric  13    -none- logical
call        4    -none- call
> tbl_list <- sapply(heart[-10], table, heart[ , 10])
>tbl_list <- lapply(tbl_list, t)
>cond_probs <- sapply(tbl_list, function(x) {
       apply(x, 1, function(x) {
          x / sum(x) }) })
>cond_probs <- lapply(cond_probs, t)
>print(cond_probs)
```

PREDICTI THE MODEL:

```
>preds <- predict(model, newdata = heart)
>conf_matrix <- table(preds, heart$class)
> conf_matrix
< table of extent 0 x 0 >
> lvs <- c("with heart disease", "without heart disease")
> chol<- factor(rep(lvs, times = c(86, 258)),
+                                levels = rev(lvs))
> pred <- factor(
+        c(
+                rep(lvs, times = c(54, 32)),
+                 rep(lvs, times = c(27, 231))),
+          levels = rev(lvs))
> xtab <- table(pred, chol)
> library(caret)
> confusionMatrix(xtab)
```