RESEARCH ARTICLE OPEN ACCESS

# Prediction Support System for Multiple Disease Prediction Using Naive Bayes Classifier

Selvaraj A[1], Mithra MK[2], Keerthana S[3], Deepika M[4]

[1,2,34](Information Technology, K.L.N.College of Engineering, Pottapalayam, Sivaganga, Tamilnadu, India.)

## Abstract:

The prediction support system extracts the personal data such as user health conditions from day to day life. The lifestyle data are gathered and stored at data repository by using web technology and mobile applications. The user enter their daily health conditions in textual format. The Natural Language Processing (NLP) is used to understand the given input and further forecast the user's illness. The keyword is extracted using text mining algorithm. In this project an effective Multiclass Naive Bayes algorithm is used for predicting the multiple disease by implementing the operations on medical datasets.

*Keywords* – **Prediction, Classification, Naive Bayes, Datasets.**

## 1. INTRODUCTION

Data Mining is an important domain of computer science field. Data mining is used to mine data patterns which are previously undiscovered, novel, valid and potential. Predictions and descriptions are important and focused goals of this field of interest. Prediction uses the previous related knowledge of the subject and uses that knowledge to foresight or predict the unseen subject of the same field. The patterns extracted are supposed to be potential, vivid and Nobel. Predictive approach includes techniques used for prediction purposes like classification and regression methods. Descriptive approach include techniques like clustering and association rules. The predictions and classifications helps to explore relations and patterns in patient medical data in order to improve their health. The problems of data mining represent great challenges for networking and distributed computing. The operation on the datasets were carried out using classification algorithms Decision Tree and Naive Bayes and results proves that Naive Bayes technique outperformed the other techniques used.

## 2. RELATED WORKS

Various data mining algorithms and techniques have been used for study and analysis of various diseases like Hepatitis, Diabetes and Cancer etc. Recent survey shows that heart disease is one of the biggest cause of death in the countries like UK, Canada, France and Singapore. Various classification models like Decision Tree, KNN and Naive Bayes have been used to diagnose the presence diseases in patients. Pattern recognition and data mining methods have been used previously by the practitioners and researchers for prediction purpose in the field of diagnosis and healthcare. The experiments were carried out using data mining algorithms like Decision Tree, K-NN and Naive Bayes out of which Naive Bayes gave better results. Hiroshi in his project delivers an automatic healthcare system which provides personal status, lifestyles and health status, from period of time.By using mobile phone and web technologies, the lifestyle data are stored on a dynamic healthcare system. The data-mining service are provided by the web application server and use mobile phones to inform users of their health and lifestyle data.This system enable users to input their current data through a mobile phone and to transfer these data to a web application server via the Internet. It

automatically stores current data of volunteer users generated some useful rules correspond their lifestyles with body-fat index [2].Administration of numerous kinds of unending sicknesses, for example, diabetes and asthma depends intensely on patients' self-checking of their malady conditions. These frameworks frequently work with just a single or a couple of kinds of medicinal gadgets and consequently are constrained in the sorts of ailment they can screen. In this paper, it portrays a bland information mapping for a tele-observing framework that is pertinent to various kinds of medicinal gadgets and distinctive illnesses, and demonstrate a usage of the pattern in a social database appropriate for an assortment of tele-checking exercises.[3]. The successful application of data mining in highly visible fields like e-business, commerce and trade has led to its application in other industries. There is a wealth of data possible within the medical systems. Heart disease is a term that assigns to a large number of health care conditions related to heart. The data classification is based on MAFIA algorithms, which result in accuracy, the data is estimated using entropy based cross validations and partition techniques and the results are compared. C4.5 algorithm is used as the training algorithm to show rank of heart attack with the decision tree. The heart disease database is clustered using the K-means clustering algorithm, which will remove the data applicable to heart attack from the database [1].The proposed paper centres around the essential ideas of affiliation control mining and the market bin examination of various things. The proficiency of the FP-Growth calculation can be estimated as far as mining of the incessant example. One discrete preferred standpoint is that it maintains a strategic distance from the age of applicant sets, which is computationally comprehensive. This will help in foreseeing future patterns and practices, enabling organizations to settle on information driven choices. The outcomes and conclusions drawn can be utilized as a part of enhancing the market. [8].

## 3. PROJECT DESCRIPTION

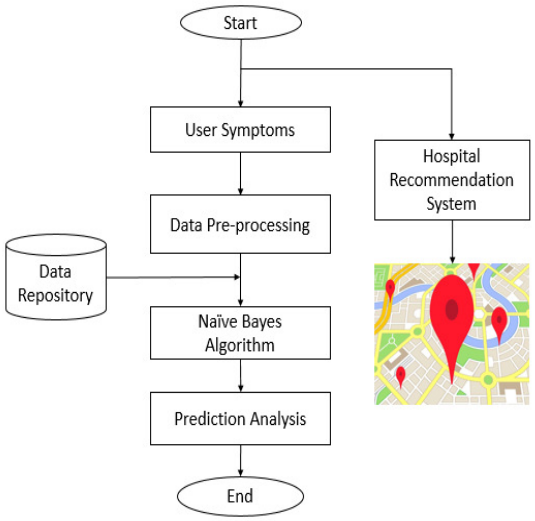
### 3.1 EXISTING SYSTEM

The medical environment is still information rich but knowledge weak. The existing system used data mining techniques like Association rule mining, Classification, Clustering to analyse on specific disease. The algorithms such as J48 and KNN are used to analyse the large data set to predict the heart disease for the user who are more prone to it. It was implemented as the comparative study of different data mining algorithms but not as a user support system.

### 3.2 PROPOSED SYSTEM

Inthis project we gather the user's basic parameters in a query form in which they would explain their health condition known by them. The information that are needed to predict the illness is examined from the given text using the text-mining algorithm. The ideal strategy is to analyse and predict the multiple diseases based on the user's symptoms for end-user. Naive Bayes algorithm is selected which gives out highest degree of accuracy. It is helpful for end-user as a prediction support system. By using these technique, it improves the overall speed and increase the accuracy of algorithm.

### DATAFLOW DIAGRAM




## 4. MODULES DESCRIPTION

In this project we have implemented 4 modules. They are,

- Data Pre-processing
- Naive Bayes Algorithm

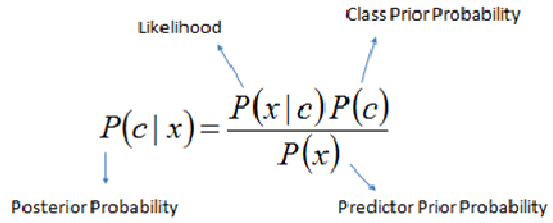
- Prediction Analysis
- Hospital Recommendation System

### 4.1 Data Pre-processing

In this module, the data mining technique that involves converting query data into a structured format. Data pre-processing is a proven method of resolving missing values or resolving the inconsistencies in the data. The user gives their health condition to the application in a textual format. The query from the user is analysed and the symptoms are filtered and given as input to next module**.**

### 4.2 Naive Bayes Algorithm

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian Classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods. Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x) and P(x|c). NB Classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.




$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- **.P (c|x)** is the posterior probability of class (target) given predictor (attribute).
- **P(c)** is the prior probability of class.
- **P (x|c)** is the likelihood which is the probability of predictor given class.
- **P(x)** is the prior probability of predictor.

### 3.3 Prediction Analysis

The symptoms are fetched from the user as input then the relation between the symptoms are found by similar disease which is related to all of the symptom given in the dataset. Thus the disease that is related to all the symptoms fetched by the user is mapped and the disease that's matched is displayed to the user. The results obtained after applying the algorithm will be analyzed on the basis of accuracy.

### 3.4 Hospital Recommendation System

The user's when in need of immediate emergency or in high of risk of illness must consult the doctor. Thus with the help of map provided in this application, the nearby hospitals or pharmacy is located. The user can select one of many hospitals of their convenience at the given time.

## 5. EXPERIMENTAL RESULTS

As mentioned in the above sections, the algorithms used are Naïve Bayes and Decision tree. These two algorithms were tested on datasets built in .arff and .csv formats. Datasets of few disease were downloaded from UCI repository. The datasets of disease selected were Eye disease, heart disease and some symptoms datasets. Every datasets consisted of about 100 to 700 data instances and 14 attributes. The algorithms selected for implementation were judged on the basis of accuracy and time taken for prediction of class labels.

Table 1: Dataset information.

TABLE 1

| S.NO | Dataset | Instance | Number of Attributes |
|---|---|---|---|

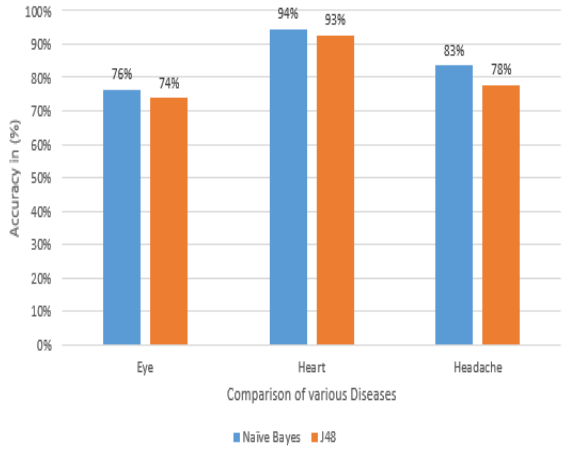
| | | | |
|---|---|---|---|
| 1. | Heart Disease | 600 | 14-17 |
| 2. | Eye Disease | 770 | 14-17 |
| 3. | Symptoms | 700 | 14-17 |

Table 2: It shows data mining techniques used on diseases with accuracy.

TABLE 2

| Algorithm | Accuracy on Eye Dataset (%) | Accuracy on Heart Dataset (%) | Accuracy on Headache Dataset (%) |
|---|---|---|---|
| Naïve Bayes | 76.30 | 94.29 | 83.49 |
| J48 | 73.82 | 92.64 | 77.55 |

**Graph**




## 6. CONCLUSION

Thus, the application predicts the disease with the help of condition of user provided described by themselves. It also recommends nearby Hospitals or pharmacy to the user. Naive Bayes' Classifier provides better prediction results of the user's illness in terms of accuracy when compared with J48 approach.

## 7. FUTURE ENCHANCEMENT

The future scope of the project is to generate a medical report based on the predicted disease and the statistical chart. Which will act as a support system for the doctors to clinically diagnosis the patient before the disease begins to show more symptoms of it. Another enhancement in this work is that all the different doctor's reports can be documented in a repository. From where the document is mined and based on those symptoms from the report the disease can be accurately predicted.

## 8. REFERENCES


1) *M.A.NisharaBanu, B Gomathy's "Disease Predicting System Using Data Mining Techniques" International Journal of Technical Research and Applications e-ISSN: 2320-8163, www.ijtra.com Volume 1, Issue 5 (Nov-Dec 2013), PP. 41-45.*

2) *Hiroshi Takeuchi, Naoki Kodama, Takeshi Hashiguchi and Doubun Hayashi's "Automated Healthcare Data Mining Based on a Personal Dynamic Healthcare System" Proceedings of the 28th IEEE EMBS Annual International Conference New York City, USA, Aug 30-Sept 3, 2006.*

3) *J. Cai, S. Johnson, and G. Hripcsak, "Generic data modeling for home tele-monitoring of chronically ill patients," Proc. AMIA symp. 2000, pp. 116–20.*

4) *NidhiMaheshwari,Nikhilendra K. Pandey and Pankaj Agarwal's "Market Basket Analysis using Association Rule Learning", International Journal of Computer Applications (0975 – 8887).*

5) *Web MD: Better Information, Better Health[online]. Available at http: // Symptoms. Webmd. Com/Symptomchecker, June 10, 2012.*

6) *Svetlana Kiritchenko, Xiaodan Zhu and Saif M. Mohammad's "Sentiment Analysis of Short Informal Texts", Journal Artificial Intelligence Research 50 (2014) 723-762.*

7) *Ambar Dutta's "A Novel Extension for Automatic Keyword Extratcion", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 5, May 2016.*

8) *Beant Kaur, Williamieet Singh "Review On Heart Disease Prediction System Using Data Mining Techniques" International Journal on Recent and Innovation Trends in*

*Computing and Communication ISSN:2321-8169 Volume:2 Issue:10(IJRITCC)October 2014.*

***9)*** *T.Revathi, S.Jeevitha, "Comparative Study On Heart Disease Prediction System Using Data Mining Techniques" International Journal of Science and Research (IJSR) ISSN:2319-7064 Index Copernicus Value (2013).*

***10)*** *M.A.Jabbar, "Computational Intelligence Technique for Early Diagnosis of Heart Disease", ICETECH 2015, IEEE, PP16 (2015).*

***11)*** *Ajinkya Kunjir, Harshal Sawant, Nuzhat F. Shaikh "Data Mining And Visualization For Prediction Of specific Diseases In Healthcare", 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC), 25 March 2017 Pages: 329 – 334.*

***12)*** *Jianchao Han, Juan C. Rodriguez, Juan C. Rodriguez " Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner"Future Generation Communication and Networking, 2008. FGCN '08. Second International Conference on13-15 Dec. 2008.*