Comparative Study of Heart Disease Prediction Using Naïve Bayes and Linear Regression

Prof. Deven D. Ketkar[1], Mr. Amol A. Waghmode[2], Mr. DarpanSawant[3]
[1]Assistant Professor, Dept of Information Technology
[2, 3]Dept of Information Technology
[1, 2, 3] Finolex Collage of Management and Technology, Ratnagiri

Abstract- *Technological rapid growth in biomedical applications generate high volume of personal data. This biomedical data raises privacy concerns as it reveals sensitive data such as health status and peoples living style. Information generated by biomedical mobile applications need to keep private. The proposed system keeps private data locally on mobile and only data required for heart disease prediction is uploaded to server.*

Data analytics for Heart disease prediction is implemented using two algorithms Logistic Regression and Naive Bayes. This paper proposes a rule based model to compare the accuracies of applying rules to the individual results of naive bayes and logistic regression on the mHealth web application database in order to present an accurate model of predicting heart disease.

Keywords- Data Analytics, Naive Bayes, Logistic Regression, mHealth web application.

I. INTRODUCTION

Data mining is a process of extracting useful information from large amount of data set. The resulted data is used for prediction purpose so that it can be beneficial for various purposes like improving business process, finding causes of diseases likewise. Different techniques involved in data mining are classification, clustering, association etc. Data mining has immense applicability in diverse area like Biomedical Analysis, Telecom Industry, Intrusion Detection System, Financial Data Analysis etc. In biomedical analysis, different algorithms are used to design model for healthcare prediction. The algorithms used are Naive Bayes, SVM, logistic regression etc.

Around 17.3 billion people die in the world for year of 2008[9]. In spite of the fact that cardiovascular diseases are controllable by taking cautions which were analyzed using different data mining techniques. Last two decades, a lot of research is going on in health care industry to find out the causes of various diseases as precaution is always better than prevention. Cardiovascular diseases includes heart failure, cardiomyopathy, coronary heart disease etc. and the common causes for these diseases are diabetes, smoking, high cholesterol, hypertension. Data set used for the data analytics is downloaded from UCI Machine Learning Repository.

II. REVIEW OF LITERATURE

Nowadays multiple people can use forward biomedical sensors and mobile application. That technology generate large amount of biomedical data which include some personal information of patient about lifestyle. In this paper, personal information can be kept secretly and logistic regression technique can be used to predicate disease prediction and treatment [1]. In today’s environment, large companies or hospital can be store the patients sensitive information on host data center. An efficient algorithm / techniques are used for designing predictive model for disease diagnosis and treatment. In this paper the information can be kept locally and use ofholomorphic encryption is suggested. Holomorphic encryption is a type of encryption which cannot be decrypted and requires no decryption key [3].

The Support Vector Machine (SVM) is a state-of-the-art classification method introduced in 1992. The SVM classifier is widely used in bioinformatics due to its high accuracy, ability to deal with high-dimensional data such as gene expression, and exibility in modeling diverse sources of data. In this paper, PPSVC techniques can be used to tackle the privacy violation problem of the classification model of the SVM [5]. MediNet is a system that can be used to develop to personalize the self-healthcare process for patients with diabetes and cardiovascular disease using a mobile phone network. It can be use current and past information from monitoring devices to recommendations. It can provide the uniqueness of each patient by personalizing its recommendations based on personal level characteristics of the patient, as well as groups of patients share that characteristics [2].

Cloud-assisted mobile health (CAM) monitoring mobile communications and cloud computing technologies to provide feedback decision support, has been considered as a revolutionary approach to improving the quality of healthcare service while lowering the healthcare cost. CAM, which can effectively provide security for privacy of clients and the intellectual property of mHealth service providers [4].

III. SYSTEM ARCHITECTURE

For heart disease prediction, this paper proposes a combination of models which is shown in Fig 1. This secure data analytics approach is divided into five modules involving mHealth application, Preprocessing, Training Model, Applying Logistic Regression, applying Naive Bayes, and Result Comparison and Heart Disease prediction. Patient database is collected from mHealth application and also taken from UCI repository for training purpose.

Proposed System: The above mentioned five models are described as below.

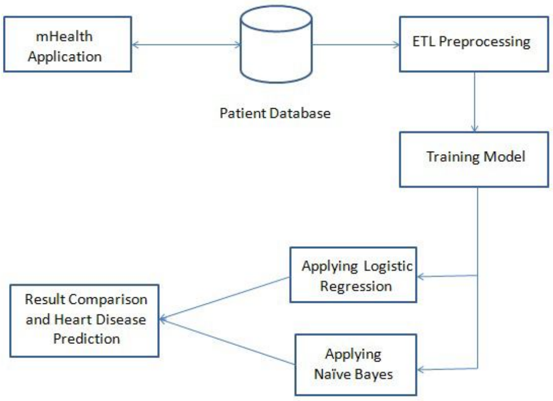


Fig. 1. Block Diagram of Proposed System

A. mHealth web Application

Web health technologies are rapidly growing which includes wearable devices. As technology growing rapidly,mHealth web application development using mobile health technology is also growing in a same way. mHealth web applications record activity of individual and inform if any cautions need to take.But it comes with privacy concern.

B. ETL Preprocessing

ETL stands for Extract, Transform, Load. It is a tool whichis combination of three functions which is mentioned above.It is used to get data from one database and transfer to otherdatabase. Data preprocessing is a data mining technique used to transform sample raw data into an understandable format. Real world collected data may be inconsistent, incomplete or contains an error.This paper proposes ETL and preprocessing combined together to process on existing patient data and loadit into mHealth server database. Following are data processing techniques.

Data Cleaning: Resolve inconsistency and eliminate noise in data.

Data Integration: Incorporate data from different sources into one rational source such as data warehouse.

Data Transformation: In data transformation, data transform from one source to another source. It involves following terms.

1. Normalization
2. Aggregation
3. Generalization

C. Training the Model

Each of the two models has been trained using different methods.For logistic regression, the first step for training is to find the significant attributes by calculating their individual Pvalues. As a rule of thumb, if it is below 0.05, only then is the attribute significant. The Hosmer-Lemeshow test is also required to check for goodness fit of the model. The corresponding P-value must abide by a 5 Naive bays are a classification method works on bayes theorem which assumes independence among attributes. The basic assumption is that presence of a particular feature in a class unrelated to presence of any other .Naive bayes requires less amount of data as compared to other method for estimation of parameters for classification. D. Applying Logistic Regression and Naive Bayes For prediction of categorical dependent variable outcome, from set of independent variables [8]. Logistic regression is mainly used to for prediction and also calculating the probability of success. Logistic Regression involves fitting an equation of the form to the data[8].

y = e_(b0 + b1 _ x)=(1 + e_(b0 + b1 _ x)) (1)
x=input values
y=output values
b0 = bias or intercept term
b1=coefficient for single input value (x)

The Naive bayes classifier is based on Bayes theorem with independent assumptions between predictors Continuous values associated with each class are distributed according to a Normal distribution [6]. Naive Bayes classification algorithm is based on Bayes theorem.

P(Ak=B) = P(Ak \ B)P(A1 \ B) + P(A2 \ B) + :: + P(An \ B)(2)
Ak = set of mutually exclusive events

B = any event from the sample space such that P(B)¿0

Here,In proposed system, Bayes theorem can be written in below given way.

P(D=S) = P(D=S) _ P(D)=P(S) (3)
D = Disease
S= Symptom

IV. IMPLIMENTATION

During this project, the proposed solution are investigating and presenting the new framework for addressing the problem of finding relevant result. The aim of this project was to improve the performance of algorithm presented in base paper. The results demonstrated inthis project are showing the current state of work done over practical implementation of this algorithm.

V. RESULT COMPARISON AND HEART DISEASE PREDICTION

This modules includes the comparison of results from Logical Regression and Naive Bayes algorithm. Also this systemfind out the accuracy percentage from both algorithms.

A. User interface and screen shots

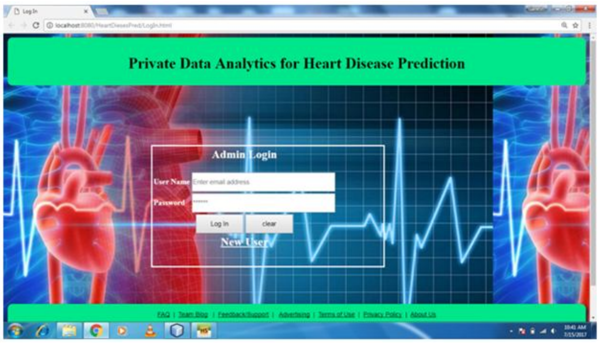

Private Data Analytics for Heart Disease Prediction
Admin Login
User Name
Enter email address
Password
Log In
clear
New User
FAQ | Team Blog | Feedback/Support | Advertising | Terms of Use | Privacy Policy | About Us

Fig 2. Login page

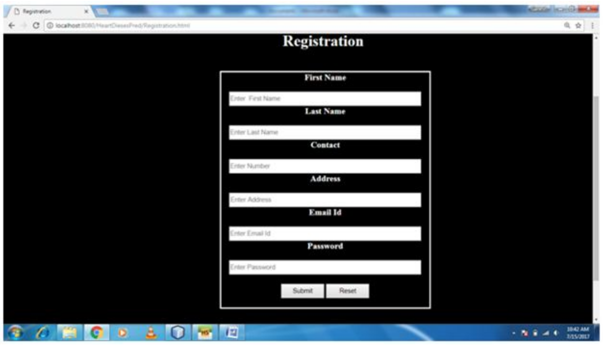


Fig 3. Registration page

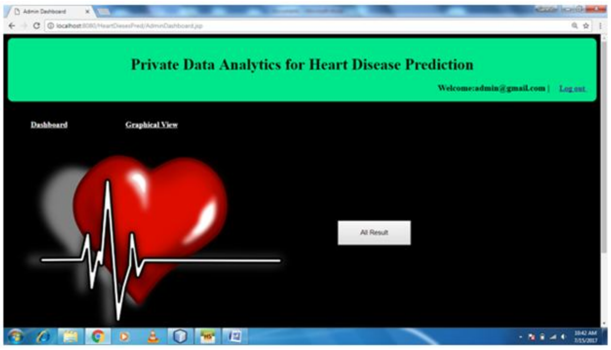
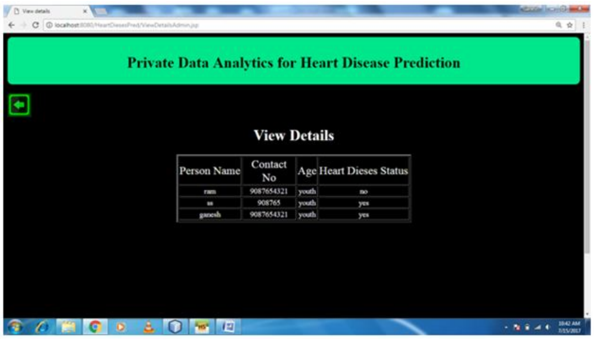

Private Data Analytics for Heart Disease Prediction
Welcome:admin@gmail.com | Log out
Dashboard
Graphical View
All Result

Fig 4. Admin dashboard page



View Details

Person Name	Contact No	Age	Heart Dieses Status
ram	9087654321	youth	no
ss	908765	youth	yes
ganesh	9087654321	youth	yes

Fig 5 View Details page

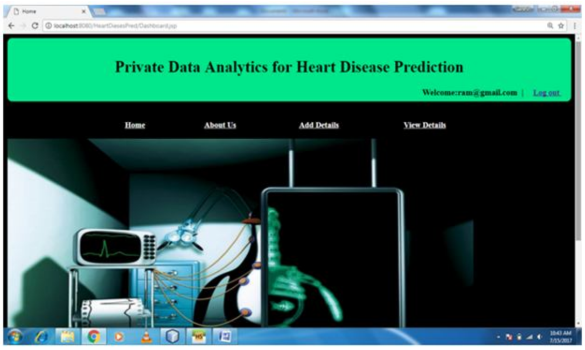

Private Data Analytics for Heart Disease Prediction
Welcome:ram@gmail.com | Log out
Home
About Us
Add Details
View Details

Fig 6 User dashboard page

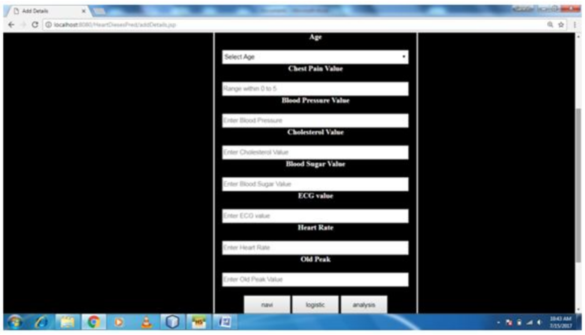
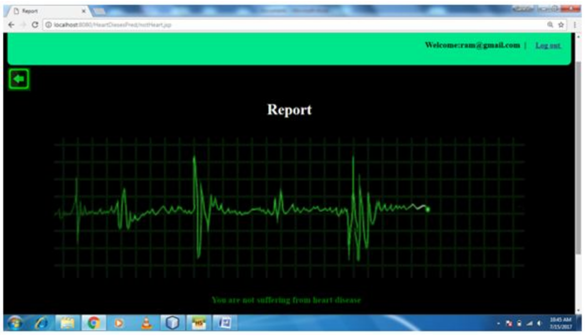


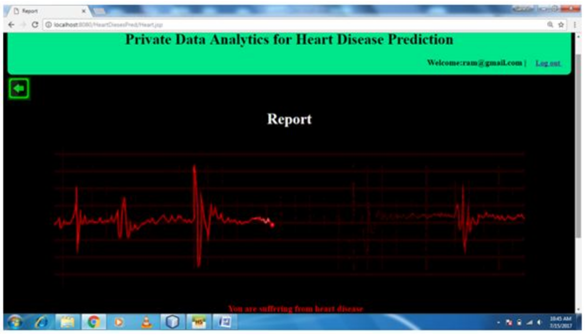
Fig 7 User Details page


Report
localhost:8080/HeartDiesesPred/notHeart.jsp
Welcome:ram@gmail.com | Log out

Report

You are not suffering from heart disease

Fig 8 Result for heart disease


Report
localhost:8080/HeartDiesesPred/Heart.jsp
Welcome:ram@gmail.com | Log out

Private Data Analytics for Heart Disease Prediction

Report

You are suffering from heart disease

Fig 9Result for not heart disease

VI. CONCLUSION

This system proposes confidential scheme for predicting heart disease using two different models, Naive Bayes and Logistic Regression.As per our result, accuracy calculated using Naïve bayes classifier is 90 % and by using linear regression, it is 85%. Hence, Naïve bayes classifier showing better results in this case.

VIII. ACKNOLEDGMENT

I express my sense of gratitude towards my project guide Prof. DevenKetkar for his valuable guidance at every step of study of this project, also his contribution for the solution of every problem at each stage. I am thankful to Dr .Vinayak A. Bharadi head of the department of B.E. Information Technology. I am very much thankful to respected Principal Dr. Kaushal K. Prasad for his support and providing all facilities to complete the project report. Finally I want to thank to all my friends for their support suggestions. Last but not least I want to express thanks to my family for giving me support and confidence at each and every stage of this project

REFERENCES

[1] Yanmin Gong , Yuguang Fang , YuanxiongGuo ,Private Data Analytics on Biomedical Sensing Data Via Distributed Computation 1545-5963 (c) 2015 IEEE.

[2] P. Mohan, D. Marin, S. Sultan, and A. Deen, Medinet: personalizing the self-care process for patients with diabetes and cardiovascular disease using mobile telephony, in Engineering in Medicine and Biology Society,2008. EMBS 2008. 30th Annual International Conference of the IEEE. IEEE, 2008, pp. 755758.

[3] J. W. Bos, K. Lauter, and M. Naehrig, Private predictive analysis on encrypted medical data, Journal of biomedical informatics, vol. 50, pp.234243, 2014.

[4] H. Lin, J. Shao, C. Zhang, and Y. Fang, Cam: cloud-assisted privacy preserving mobile health monitoring, Information Forensics and Security, IEEE Transactions on, vol. 8, no. 6, pp.985997, 2013.

[5] K.-P. Lin and M.-S. Chen, On the design and analysis of the privacy preserving svm classifier, Knowledge and Data Engineering, IEEE Transactions on, vol. 23, no. 11, pp. 17041717, 2011.

[6] R. Agrawal and R. Srikant, Privacy-preserving data mining, in ACM Sigmod Record, vol. 29, no. 2. ACM, 2000, pp. 439450.

[7] Shalabi, L.A., Z. Shaaban and B. Kasasbeh, Data Mining: A Preprocessing Engine, J. Comput. Sci., 2: 735-739, 2006.

[8] Mythili T., Dev Mukherji, Nikita Padalia, and Abhiram Naidu, A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL), International Journal of Computer Applications (0975 8887), Volume 68 No.16, April 2013.

[9] http://www.world-heart-federation.org/cardiovascular-health/global-factsmap/

[10] Robert Detrano 1989 Cleveland Heart Disease Database V.A. Medical Center, Long Beach and Cleveland Clinic Foundation.