RESEARCH ARTICLE OPEN ACCESS

A PERSPECTIVE FOR HEALTHCARE USING DATA ANALYTICS AND MACHINE LEARNING ALGORITHM

A.Kanchana [1], M.Niththila [2], K.Sruthi [3], B.Kumudha [4]
Associate professor [1], U.G.Scholar[2],[3],[4]
Department of Computer Science and Engineering
Panimalar Engineering College
Chennai-India

ABSTRACT

Data analytics is a process of analyzing and identifying different patterns, relations, and trends within a large volume of data.In this project,we apply data analytics techniques and machine learning algorithm to find the factor that leads to blood pressure,by using the real-time data collected about people of village in Karnataka containing records of 1621 patients with 226 fields,where exploratory data analysis is conducted for visualization and trend prediction.Predictive analysis is an important factor in this project,by which prediction can be made more accurate by using datasets.Data analytics helps to analyze the datasets and provide the key insights for the patient care in village.The algorithm used is logistic regression and various machine learning models,which involves univariate analysis(every single fields are analyzed),bivariate analysis,correlation heatmap to check what numerical values are correlated and to what extent and feature selection using Kbest method and to verify the correctness of prediction we use chi square test and association rule.The main goal of this project is to analyze the factors that impacts the health of the society.We produce an enhanced performance level with an accuracy level of 94.2%

***Keywords*:** Data Analytics, Machine Learning Algorithms -Logistic regression, Association rule, Chi-Square test.

I. INTRODUCTION

Everyday,data is generated by a range of different applications,devices and geographical research activities for the purposes of weather forecasting,weather prediction,disaster evaluation,crime detection and health industry. An era of open information in healthcare is now underway. In current days, data analytics is associated with core enterprises including Google, Facebook and IBM, which extract valuable information from the huge volume of data collected.Data analytics is used rapidly in every field including healthcare, with respect to patient care. The main objective is to find the factors impacting the health of the community. The domains used for finding the factors impacting health are the data analytics and machine learning algorithms. In this project,the dataset used is collected from Indian Council Of Medical Research(ICMR) about people of village in karnataka incorporating 1621 patients with 226 fields. In this project, we analyze the factors that lead to blood pressure with many data fields more accurately.There are many factors that impacts the health of the society.So we came up with an objective of finding the factors that affects the health of community.Initially,we have collected the dataset of people of village in karnataka with 1621 patients containing 226 fields and started analyzing the factors that leads to blood pressure.

Inorder to analyze the factors,the data analytics techniques and machine learning algorithms like Logistic regression,Chi-square test and association rules are used. we have used many machine learning models, which involve univariate analysis(every single field are analyzed), bivariate analysis,correlation heatmap to check what numerical values are correlated and to what extent and feature selection using Kbest method.

II. LITERATURE SURVEY:

Data analytics is of paramount importance in healthcare aspects such as patient diagnostics, fast epidemic recognition, and improvement of patient management.

Sunil Kumar and Maninder Singh[1] explore the conceptual architecture of big data analytics for healthcare which involves the data gathering history of different branches, the genome database, electronic health records, text/imagery, and clinical decisions support system.Here they use big data to handle the massive amount of data gathered from healthcare centers of lot of people and their health condition through their lifespan.In Big data analytics they use Predictive analytics , machine learning to identify the data pattern and the program accordingly and Electronic health records(EHR)to track the medical history.And we use in-depth description and conceptual architecture for solving healthcare problems using hadoop and map-

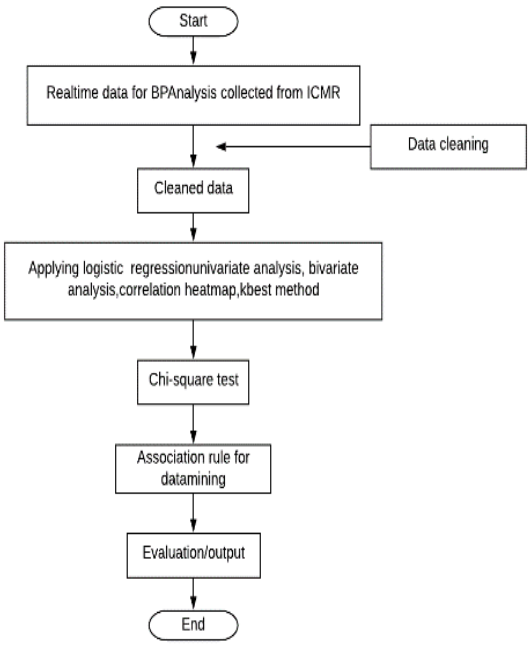
reduce .The combination of big data and healthcare analytics leads to the treatment that are effective to specific patients by providing the ability to prescribe appropriate medications for each individual,rather than those that work for most people.Disadvantage of this paper is most of these problems are unable to solve by the use of Hadoop.

In[2] effective heart disease prediction was done with the help of machine learning techniques which focused on improving the efficiency of heart disease prediction using prediction model,feature selection and classification algorithm with the efficiency of 88.7%.In [3]paper,a novel approach toward the production and consumption of linked data analytics in urban environments is presented.The approach is based on the exploitation of linked data principles, enhancing the ability of managing and processing of data, in ways not available before. In addition to the description of the overall technical approach, the application of the proposed solution into a real-life scenario for examining the health impact of outdoor air pollution in urban areas within an international, national, and regional perspective is detailed. A set of interesting results are produced along with their interpretation toward the provision of suggestions for policy making purposes. Looking forward, a set of opportunities are identified for optimally exploiting the power of linked data analytics for solutions that may have a major impact on quality of life within a city. With regards to the availability of data within a smart city ecosystem, it is foreseen that huge amounts of data is going to be made available in the future through the adoption of crowd-sensing mechanisms as well as the development of low-cost sensors, embedded in many cases in smartphones.

In[4] the literature on various data mining applications, especially applications that are applied to solve the crimes. Survey also throws light on research gaps and challenges of crime data mining. In additional to that, this paper provides insight about the data mining for finding the patterns and trends in crime to be used appropriately and to be a help for beginners in the research of crime data mining.This paper incorporates data mining fundamentals,the data mining technique fo analyzing crime patterns and finally describes the issues and challenges that has been came across.Analysing the crime type such as traffic violation,sex crime,theft,arson,gang/drug offenses,violent crime using data mining techniques and algorithms like decision tree algorithm and apriorty algorithm.In[5] aims to reveal the way of big data approach in extracting data value from data complexity involving variety and velocity into the volume. This study was conducted using contents analysis by reviewing some literature in peer-reviewed journals, chapters, books and proceedings in developing prototypes using data analytics associated from the topic, users and time analytics. The findings reveal that big data emerging technology with analytic process provides particular advantages to transform the pattern of information fitted into the innovative environment of online OLR to enhance in developing the learning resources.

III. PROPOSED WORK:

In the proposed system,the blood pressure analysis is done using the machine learning algorithm and data analytics techniques.The algorithms used is logistic regression and various machine learning models,which involve univariate analysis(every single field are analyzed), bivariate analysis,correlation heatmap to check what numerical values are correlated and to what extent and feature selection using Kbest method and to verify the correctness of prediction we use chi square test and association rule.The datasets used in this project is real time data collected from Indian Council of Medical Research containing 1621 records with 226 fields.Initially,data clean up process is done followed by using logistic regression containing various machine learning models ,chi-square test and association rule.Finally.the performance level with an accuracy level of 94.2% and is calculated based on the model generated results.



IV. EXPERIMENTAL RESULTS:

DATA CLEANUP PROCESS:

In this paper, initially, we have data for 1621 patients with 226 columns. In order to avoid data that are

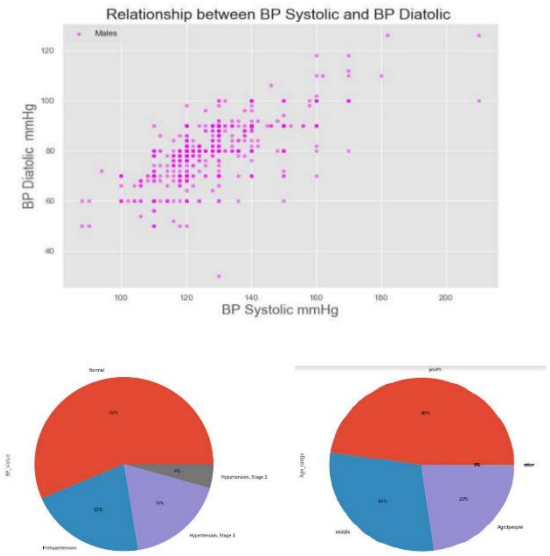
repeated-, we use data cleanup processes using machine learning models. Data cleaning tasks are done using pandas, Numpy and seaborn libraries. After cleaning up the datasets are reduced to 156 columns.

BLOOD PRESSURE ANALYSIS USING LOGISTICREGRESSION: (UNIVARIATE,BIVARIATE,CORRELATE HEATMAP & KBEST CLASSIFIER METHOD)

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.In this project the concept of narrative visualization is done which contains barchart and pie chart

Step 1: To analyze BP, only the required fields are taken for analysis. In order to analyze the required fields we have used many machine learning models, which involve univariate analysis(every single field is analyzed), correlation heatmap to check what numerical values are correlated and to what extent and feature selection using Kbest method. From this analysis, 20 columns are taken for further analysis.

Step 2 -Analyzing BP with 20 columns : Finding the relationship between BP systolic and BP diastolic by regression plot using seaborn and various other relationships like occupation and BPValue for male, Smoking, and BPValue for males and so on.



CHI-SQUARE TEST FOR FINDING RELATIONSHIPS BETWEEN VARIOUS FIELDS:

Chi-square test is commonly used for testing relationships between categorical variables.The null hypothesis of the Chi-Square test is that no relationship exists between the variables and they are independent. Inorder to find the relationship found by logistic regression is correct we use chi-square.The various fields are related to analyze the reason behind BP are relationship between BP systolic and BP diastolic,occupation and BPValue for male,Smoking and BPValue for male and for many other fields

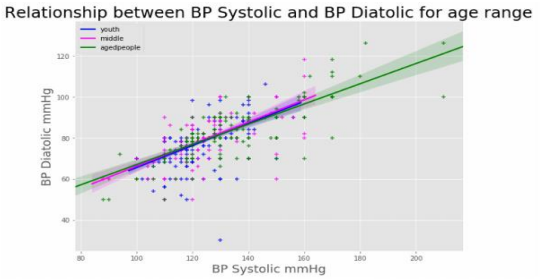


Fig 5

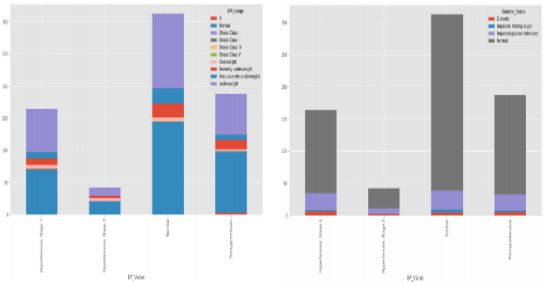


Fig:6 Fig:7

ASSOCIATION RULE FOR BLOOD PRESSURE ANALYSIS:

Association rule mining is a technique to identify underlying relations between different items and it is one of the ways to find patterns in data. It finds features (dimensions) which occur together and features (dimensions) which are “correlated”.In this project, apriori algorithm is used. The Apriori algorithm makes use of libraries such as pandas,numpy,seaborn and matplot .Various fields like smoking and BP,Drinking and BP and age and BP(both male and female) are used for finding out the relationships that causes blood pressure.

Accuracy = (100 + 47) / 156 = 0.9423 (94.2% accuracy)

V. CONCLUSION AND FUTURE SCOPE:

By applying ,association rule ,chi-square test,logistic regression which involves various classifier methods,we can infer that smoking causes blood pressure for male,if male has smoking habit 60% of chance for hypertension,drinking causes blood pressure for male,if male has drinking habit,61% of chance for hypertension and age causes blood pressure for male and female,if people gets older, then their BP level increases with the efficiency of 94.2%.

REFERENCE:

[1] A. S. Abdullah and R. R. Rajalaxmi, ‘‘A data mining model for predicting the coronary heart disease using random forest classifier,’’ in Proc. Int. Conf. Recent Trends Comput. Methods, Commun. Controls, Apr. 2012, pp. 22–25.

[2] A. H. Alkeshuosh, M. Z. Moghadam, I. Al Mansoori, and M. Abdar, ‘‘Using PSO algorithm for producing best rules in diagnosis of heart disease,’’ in Proc. Int. Conf. Comput. Appl. (ICCA), Sep. 2017, pp. 306– 311.

[3] N. Al-milli, ‘‘Backpropogation neural network for prediction of heart disease,’’ J. Theor. Appl.Inf. Technol., vol. 56, no. 1, pp. 131–135, 2013.

[4] C. A. Devi, S. P. Rajamhoana, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika, ‘‘Analysis of neural networks based heart disease prediction system,’’ in Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI), Gdansk, Poland, Jul. 2018, pp. 233–239.

[5] L. Baccour, ‘‘Amended fused TOPSIS-VIKOR for classification (ATOVIC) applied to some UCI data sets,’’ Expert Syst. Appl., vol. 99, pp. 115–125, Jun. 2018. doi: 10.1016/j.eswa.2018.01.025.

[6] Y.Wang,L.Kung,W.Y.C.Wang,andC.G.Cegielski,‘‘Anintegratedbig data analytics-enabled transformation model: Application to health care,’’ Inf. Manage., vol. 55, no. 1, pp. 64–79, Jan. 2018.

[7] A. Gandomi and M. Haider, ‘‘Beyond the hype: Big data concepts, methods, and analytics,’’ Int. J. Inf. Manage., vol. 35, no. 2, pp. 137–144, Apr. 2015.

[8] J. Zakir and T. Seymour, ‘‘Big data analytics,’’ Issues Inf. Syst., vol. 16, no. 2, pp. 81–90, 2015.

[9] Z. Wang, D. Zhang, M. Sun, J. Jiang, and J. Ren, ‘‘A deep-learning based feature hybrid framework for spatiotemporal saliency detection inside videos,’’ Neurocomputing, vol. 287, pp. 68–83, Apr. 2018.