Applying Machine Learning Algorithms to find correlation between socioeconomic data and diseases

[1] Ms. Pranita A Patil, [2] Ms. Seemab Nadaf, [3] Dr.M.S. Karyakarte , [4] Prof.Y.V.Dongre

[1]PG Student, [2]PG Student, [3]Professor, [4]Assistant Professor
COMPUTER ENGINEERING
Vishwakarma Institute Of Technology, Pune

Abstract : Today's healthcare industries are moving from volume-based business into value-based business, which requires doctors and nurses to be more efficient. This will improve healthcare industry, life style and provides longer life, which prevent diseases, illness. Data analytics tools are used to improve healthcare in areas like reports, decision making, and prediction and prevention system. Healthcare data is more complex because abundance data is available. Healthcare systems collect large amount of textual and numeric data about patients, visits, prescriptions, physician notes etc. Therefore, in healthcare data analytics tool is important to manage a huge amount of complex data, which is used to improve healthcare industry and provide accuracy and efficiency. Machine Learning in healthcare is used to analyze different type of data and predict possible outcomes, risk prediction, etc. WEKA uses the statistical procedures such as simple linear regression, logistic regression models and machine language techniques like decision trees, clustering, multi-layer perceptron When socioeconomic parameter is combined with patient's electronic health records, providers will have a more detail view of their patients, allowing them to make better decisions.

IndexTerms - Machine learning algorithms, WEKA, Correlation, Linear Regression.

I. INTRODUCTION

With the rapid development of hospital information technologies, more hospitals build electronic heath record (EHR) systems, which provides a complete source for medical data mining and analysis, nowadays technology is allowing healthcare specialists develop alternative and reducing administrative and supply costs. Machine learning in healthcare is one such area that is seeing gradual acceptance within the health care industry. Nowadays, machine learning is implemented in different applications by using various tools and methodologies. Machine learning tools are user-friendly and permit the data driven decisions. For data design and visualizations these tools are applied.

The increasingly range of applications of machine learning in healthcare permits to glimpse at a future where knowledge, analysis, and innovation work hand-in-hand to support infinite patients without them ever realizing it.It is common to find ML-based applications embedded with real-time patient data available from different healthcare sources and increasing the efficiency of new treatment options which were unavailable before. socioeconomic data and advanced analytics can be used to determine health risks. Socioeconomic data consists of information on social parameters that provides insights into social, economic and environmental factors that affects healthcare of individual. Socioeconomic factors and health are inseparably linked. The definition of two are related to socioeconomic concepts. Healthcare represent most attractive application areas. Predicting future health risks has never been easy. Medical and pharmacy claims data have traditionally been the main source of information for healthcare industries. Classification algorithms like Naive Bayes, Linear Regression; Random Forest, Logistic Regression, Hidden Markov Models are used for developing a predictive model. WEKA is a machine learning tool which is easily implemented for any data streams such as medical, environmental, spatial, text, web, etc. Socioeconomic status refers to the resources possessed by an individual to obtain what he or she wants and needs, and the perception of those resources by society. Socioeconomic status is an important risk factor for the

development of specific diseases and for the occurrence of disability among patients with the disorder. Socioeconomic significances of health refer to the effects of disorder both on the resources of individual and on the resources of society.

Data analytics give rise to many medical applications such as remote health monitoring fitness programs to overcome chronic diseases. Compliance with treatment and medication by medical field experts is another important potential application.

II. LITERATURE REVIEW

Nowadays, Data Mining process is implemented in different applications by using various tools and methods. Data mining tools are easy and permit the data decision systems. For effective and economical data design and visualizations these tools are applied. WEKA is such tool which is easily implemented for any data streams such as medical, environmental, spatial, text, web, etc. This paper throws some light on prediction of air pollutants in environmental data for forthcoming year using data processing tool weka. The air pollutants data was collected from the power industry in Andhra Pradesh and forecasting the pollutants for forthcoming year. Using WEKA, data is analyzed by correlations and linear regression model within a short time period. Data mining is a bright and relatively new technology. [1]

Characteristics of unstructured EMR data of real in-patient EMR of chronic disease is analysed from a hospital in Qingdao, and then proposed a novel association rule mining algorithm, called Un-Apriori, based on Apriori. Data and information recorded in the EMR system includes the patient's complete and complex information, such as patient's personal information, history of illness, , family medical history, physical checkup, examination and laboratory reports, records of daily round, nursing and treatment, etc. Therefore, this type of information included in the EMR system is different. there is not only a structured data, but also many fuzzy and unstructured data. Except a little numerical information, most of information in the EMR system is unstructured data. Un-Apriori adopts a categorized preprocess scheme for EMR data to satisfy requirements of Apriori algorithm. There are two types of experiments on Hadoop to validate efficiency and performance of the algorithm. Un-Apriori adopts a categorized preprocess scheme for EMR data to satisfy requirements of Apriori algorithm.[2].

Ischemic Heart Disease (IHD) is difficult to diagnose because most of the symptoms are similar to other diseases. It is a very common, harmful disease, which is identified mostly during the mortality of an individual. The system builds a clinical decision support system, which diagnose the presence of IHD with an integrated automated classifier using Artificial Intelligence (AI) techniques. The primary work of suggest that IHD can be diagnosed using the CDSS (Clinical Decision Support System). A total of 59 models were used for classification of ischemic disease, and 16 models were found to be more efficient with accuracy when compared with the physician's diagnostic impression, KSTAR separate out the negative and positive cases more appropriately than the other methods. Sensitivity and accuracy square measure higher than the opposite strategies. It is essential to continue studying classifier accuracy, including attribute selection for IHD, in order to develop an electronic protocol along with CDSS in future. This study helps people to determine their heart disease risk, as it involves a simple procedure for decision making in an effective way by extracting hidden knowledge from a historical database. various climatic, environmental and habits may also impact the case. Samples collected from one area may differ from others and can result in varied sensitivity and accuracy.[3]

With big data growth in biomedical and healthcare industry, detailed analysis of medical data benefits early disease detection. However, the analysis accuracy is reduced when the quality of medical data is incomplete. Moreover, different regions shows unique characteristics of certain regional diseases, which may weaken the prediction of disease occurrences. System experiment the modified prediction models over real-life hospital data collected from central China in 2013-2015. To overcome the difficulty of incomplete data, use of latent factor model is used to reconstruct the missing data. A new convolutional neural network (CNN)-based multimodal disease risk prediction algorithm using structured and unstructured data from hospital is proposed. In conclusion, for disease risk model, the accuracy of risk prediction depends on the diversity

feature of the hospital data, i.e., the better is the feature description of the disease, the higher the accuracy will be. Therefore, this system uses structured data and text data of patients based on the proposed CNN-MDPR algorithm. [4]

Electronic Health Records (EHR) is growing at an exponential rate that is being stored in original databases or cloud storages. These records have now grown to be called as Big Data. Most of these data are unstructured. The data can be efficiently processed on cloud for lowering the processing costs. Predictive analytics help the doctors to identify the patient admission to hospital at early stage. To perform predictive analytics various factors with demographic data, different parameters, patient past medical history and various symptoms for a specific disease. A predictive model using scalable Random forest classification algorithm which can correctly identify the classifier rate for risk of diabetes. In a distributed computing environment processing the huge data is done based on MapReduce model. To find the accuracy of the patient data the classification model is helpful. The aim is to find the nearest accuracy of the classifier by using the scalable random forest algorithm. The CART model and Random forest is built for the data set and the accuracy of the classifier is calculated. Results shows that by using the Scalable Random forest algorithm we can get the best possible results for accuracy of the prediction.[5]

III.RESULTS

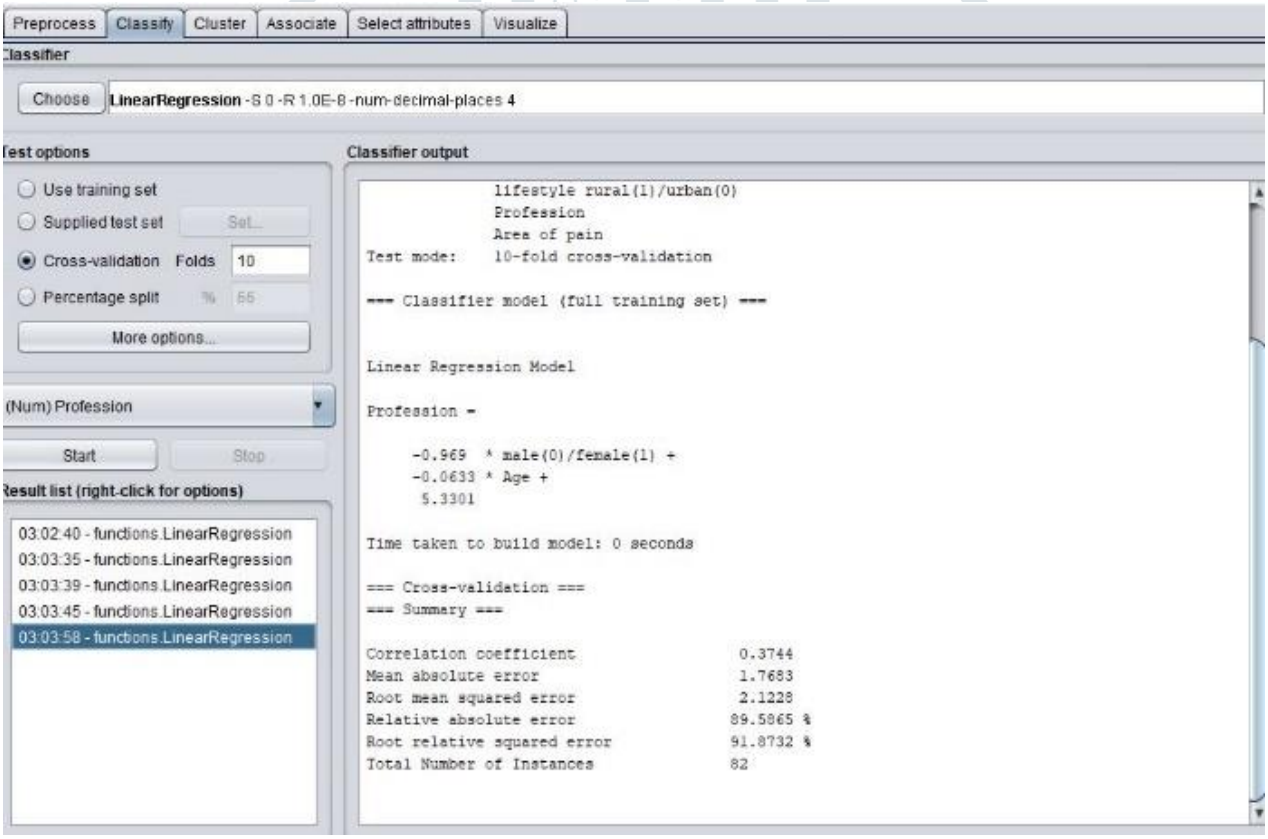

Fig.1.Linear regression Statistics

Correlation coefficient 0.3744 implies 37.44% of the variance in our data is explained by model. For R^2 for more. A low value isn't bad if it truly is the best model available. Mean absolute error is the average distance the model's predictions are from the actual data points. Absolute in the title indicates that predictions below data points ate not treated as negative distances. Root mean squared error is a different way of calculating the mean absolute error. From this value we can compare between models constructed with larger/smaller valued data. The last line is the number of data points in the data set.

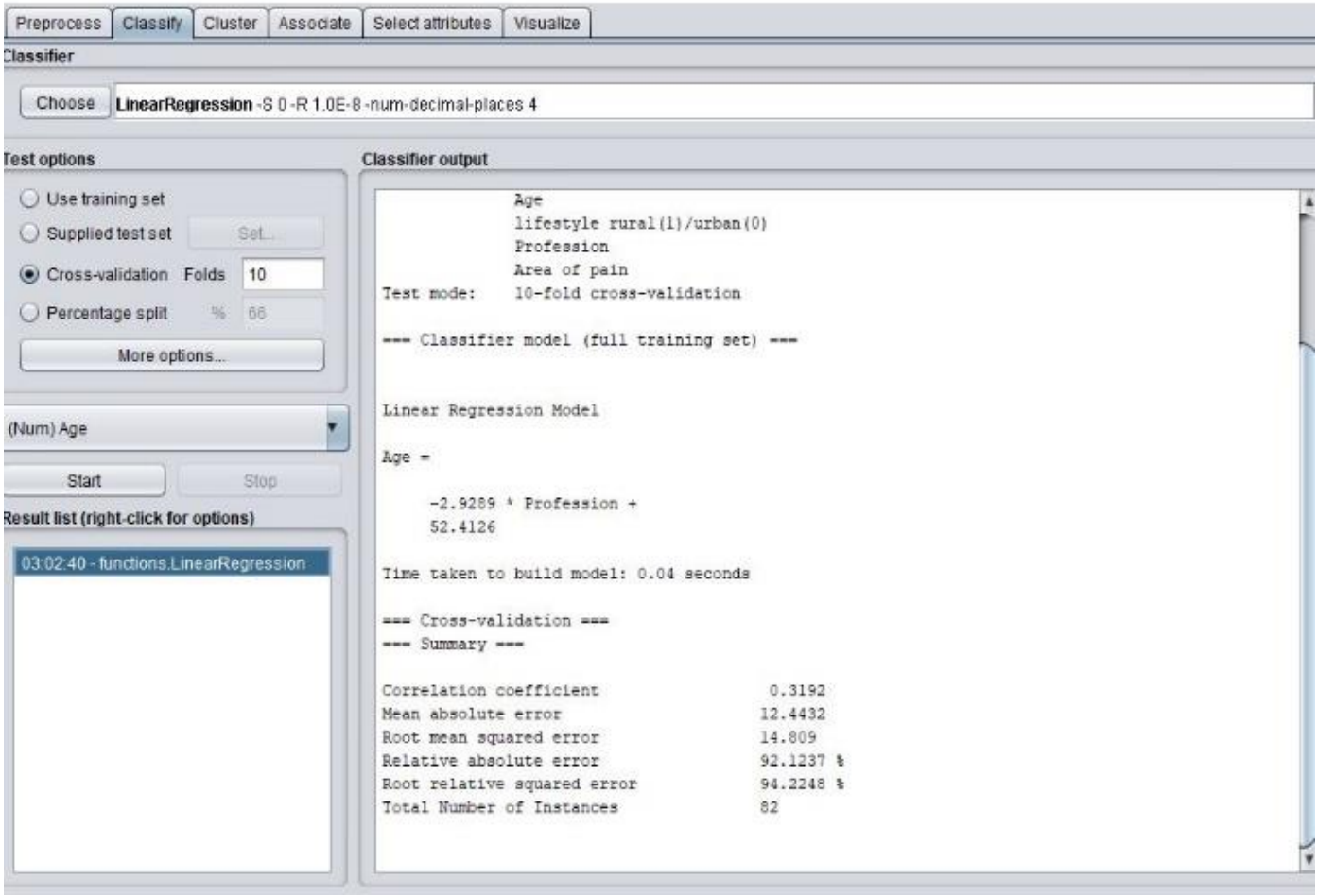

Fig.1.Linear regression Statistics

The result presents the obtained coefficients of the variables representing the regression output. The. Positive value of correlation coefficient shows that area of pain and occupation are depends upon each other. WEKA uses only columns that statistically contribute to the accuracy of the model.

IV. CONCLUSION

In healthcare industry data analytics is important tool in order to manage a large amount of complex data, which can lead to improve healthcare industry and help medical practice to reach at a high level of efficiency and work accuracy. Machine learning in healthcare helps to analyze thousands of various data points and Predicted outcomes, provide risk, exact resource allocation, and has many applications. WEKA uses the statistical procedures such as simple linear regression. Therefore, when socioeconomic parameter is combined with patient's electronic health records, providers will have a more detail view of their patients, allowing them to make better decisions.

REFERENCES

[1] Marian Cristian Mihăescu, "Proceedings of the Federated Conference on Computer Science and Information Systems" pp. 717–721 ISBN 978-83-60810-22-4

[2] Bo Song1, Yunxia Feng1;_, Xu Li2, Zhen Sun1 and Yanli Yang1 , "Un-Apriori: a Novel Association Rule Mining Algorithm for Unstructured EMRs", IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom) 2017.

[3] K.Rajeswari V.Vaithiyanathan Deepa Abin,." Artificial Intelligence (AI) Techniques Applied for the Development of a Clinical Decision Support System (CDSS) for Diagnosing Ischemic Heart Disease", International Journal of Computer Applications (0975 – 8887) Volume 101– No.9, September 2014.

[4] MIN CHEN1, (Senior Member, IEEE), YIXUE HAO1, KAI HWANG2, (Life Fellow, IEEE),LU WANG1, AND LIN WANG3,4 "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities " IEEE 2017.

[5] Sreekanth Rallapalli Dr. Suryakanthi T, "Predicting the Risk of Diabetes in Big Data Electronic Health Records by using Scalable Random Forest Classification Algorithm", 2016 IEEE,

[6]].Ian H.Witten and Elbe Frank, "Datamining Practical Machine Learning Tools and Techniques," Second Edition, Morgan Kaufmann, San Fransisco.(2005).

[7] Z. Huang, W. Dong, P. Bath, et al. On mining latent treatment patterns from electronic medical records[J]. Data mining and knowledge discovery, 29(4): 914-949, 2015.

[8]T. Hernandez-Boussard, S. Tamang, D. Blayney ,J. Brooks and N. Shah, "New Paradigms for Patient-Centered Outcomes Research in Electronic Medical Records", 4(3), 2016.