International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Big Data Analytics Predicting Risk of Readmissions of Diabetic Patients

Saumya Salian[1], Dr. G. Harisekaran[2]

[1]SRM University, Department of Information and Technology, SRM Nagar, Chennai 603203, India

[2]Professor, SRM University, Department of Information and Technology, SRM Nagar, Chennai 603203, India

Abstract: ***Healthcare holds paramount importance where analytics can be applied to achieve insights about patients, identify bottlenecks and enhance the business efficiency. Readmission rates cater to the quality of treatment provided by the hospitals. Readmission results from improper medication, early discharge, unmonitored discharge, meager care of hospital staff. To identify high risk of readmission through data analytics leads to accessibility to healthcare providers to develop programs to improve the quality of care and institute targeted interventions. The proper implementation of these analytic methods aid in proper utilization of resources in hospitals thus reduces the readmission rate and the cost incurred due to re-hospitalization. Evolving predictive modeling solutions is highly challenging for recognizing risks of readmission in healthcare informatics. The procedure involves integration of numerous factors such as clinical factors, socio-demographic factors, health conditions, disease parameters, hospitality quality parameters and various other parameters that can be specific to requirement of each individual health provider. Big data consists of large data sets that require high computational processing to procure the data patterns, trends and associations. The effectiveness of big data and its analytics in predicting the risk of readmission in diabetic's patients has been dealt in the research. The aim of this project is to determine the risk predictors that can cause readmission among diabetic patients and detailed analysis has been performed to predict risk of readmission of diabetic patients.***

Keywords: Healthcare, Diabetes, Analysis, Risk Prediction

1. Introduction

Readmission into hospitals has highly become unaffordable nowadays and necessary measure needs to be employed to make them preventable [1]. A patient being admitted into the hospital frequently in avers short period of time is known as readmission. The frequency of readmission is generally high in hospitals that cater to a vast population. Readmission may be accounted to patient diagnosis, severity of illness, adherence to discharge instruction, improper medication, quality of post discharge care etc. Some readmission may also be accounted to follow-up surgery or rehabilitation or transfer of hospital to another. A key care parameter that is measurable is reducing the risk of readmission. In order to follow up this task, it is highly predominant to develop an accurate tool to predict and analyze the pattern of readmission in hospital. To develop a predictive model is very challenging task as it requires dealing with voluminous unstructured data. In today's arena of health informatics, big data health care implementation can help increase the focus on understanding and development of predictive tools to manipulate and analyze data sets to correlate and collate insights to facilitate better understanding of the issues that can lead to readmission.

Diabetes is defined as a clinical syndrome that is characterized by hyperglycemia, due to inadequacy of insulin in the human body [2]. The syndrome has become quite customary in today's life irrespective of age group. The disease is chronic and does not have any specific cure .It varies from person to person depending on the symptoms and levels of blood sugar in human body. This makes the disease very substantial in a way that the awareness needs to be increased among the population diabetes mellitus being a chronic disease, higher the risk of readmission of patients in the hospital. Thus, the necessity to predict the risk of readmission of diabetes affected patients is obligatory.

Analytics were being performed and created using structured data that have been obtained from consolidated data systems. The significant increase in the global digital content and growing volume of information over the last ten years accelerated the need for efficient tools to analyze the large data evolving analytical processing technologies never possible before. Big Data includes new data management systems, improved analytics capabilities, faster hardware [3].

Hadoop [4] is an open source software framework for storage and large scale processing of data sets on a MapReduce programming model for large scale datasets on clusters of commodity hardware. It has a MapReduce programming model for large scale data processing and archives performance, scalability and fault tolerance. Hive [5] is an open source data warehousing solution built on top of Hadoop. It supports queries expressed in a SQL like declarative language –HiveQL. RHadoop acts as a bridge between Hadoop and R. R is a language that is employs to analyze the data statistically and explore data sets. Hadoop is a framework that allows distributed processing of large data sets across groups of computers. RHadoop is built out of 3 components which are R packages: rmr, rhdfs and rhbase. The rmr package offers Hadoop MapReduce functionalities in R. The rhdfs package offers basic connectivity to the Hadoop Distributed File System. It comes with convenient functions to browse, read, write and modify tables stored in HBase. Such tools enable important quality of care metrics to be developed across numerous dimensions. In the proposed system methodology we build a predictive model that can identify the patients with diabetes chronic diseases

who are most likely to get readmitted.

a. In the proposed system the raw data is first loaded into the Hadoop File System (HDFS).
b. Using Hive queries, all the selected predictive variables are retrieved into a coherent dataset which is used for modeling.
c. Actual model building task is done by selecting and applying various classifications, prediction method using RHadoop.

2. Methodology

The architecture of the methodology is shown in Figure.1

2.1 Data Collection

Data collection plays a significant role in obtaining accurate results for any study being carried out. Yielding accurate results is highly dependent on collecting the appropriate data from reliable source. The data that is obtained is highly noisy and variant in nature. It is necessary to closely examine the data and select the right kind of attributes that can help in obtaining accurate results. This data set was obtained from the UCI Repository of Machine Learning Databases [2]. The data set was selected from a larger data set held by the National Institutes of Diabetes and Digestive and Kidney Diseases. The data set is first loaded into Hadoop File System (HDFS).The preprocessing of data is required to modify the data into classifiable data.

2.2 Data Preprocessing

Data needs to be prepared for modeling where noisy data is converted into classifiable data. Thus it is precursory step to construct the data suitable for training predictive models. Due to presence of heterogeneity in the data and existence of inconsistencies, this stage poses several challenges. Hadoop is a distributed framework that implements MapReduce. But the method of selection in Hadoop is slow and Hadoop do not provide any query functionality. It is necessary to stimulate a scalable data warehouse on top of Hadoop. Hive is used as an open source data warehousing solution built on top of Hadoop. It is used for providing data summarization, query and analysis. Hive supports queries which are read and transparently converts queries to map-reduce in a SQL-like declarative language called HiveQL. HiveQL facilitates users to plugin custom map-reduce scripts into queries. It consists of two user interfaces of Command Line Interface (CLI) and Web User Interface (UI).

The dataset is first loaded into HDFS. A table is created with attributes present in the diabetic dataset. Next, the data is cleaned and the null and missing values, outliers are eliminated by running Hive queries. Thus, data is ready for analysis for predictive modeling.

2.3 Class Label Definition

To move ahead with the classification process, class labels need to be defined. It is a dichotomous variable which denotes the status of readmission. The binary response variables takes the values ‗0' and ‗1', where ‗1' means ―tested positive‖ for readmission and ‗0' means ―tested negative‖ for readmission. Class Label can be defined by running HiveQL queries on the data set stored in Hive as diabetes table.

2.4 Predictive Modeling

Predictive modeling has been executed through RHadoop. Analysis is being carried out on RStudio, where all folders and files of HDFS are accessible for further analysis. The diabetes table from Hive warehouse is chosen for building a predictive model. In this phase the appropriate classification prediction technique is selected to build a model for prediction of readmission. The classification models that are applied to perform the experiments are Logistic Regression, Decision Tree, Support Vector Machine (SVM). The data is classified into training and testing data in the ratio of 7:3.Model quality was assessed through common model quality measures such as Miss-Classification (error rate %), Confusion Matrix, Accuracy and ROC curve [6]. To identify the appropriate classification method, is the basic step of predictive modeling. Based on our diabetes dataset miss classification error rates are calculated for various methods i.e. Logistic Regression, SVM, K-NN and Decision Tree.

Table 1: Miss Classification

Algorithm	Miss Classification (Error Rate %)
Logistic Regression	43.61
Decision Tree	28
Support Vector Machine	32.05
K-Nearest Neighbor	32.4

It has been observed from the Table 1, that Decision Tree and Support Vector Machine classification methods prove to give less miss classification error rates. For further examination of accuracy, confusion matrix of classification methods is evaluated. Confusion matrix represents the classification results in terms of matrix. The information regarding actual and predicted classification is contained in confusion matrix. To build a valid model we use Correlation Matrix which signifies the relevance of the features among those in data to label namely ―readmission‖. After the process of correlation, we use Decision Tree as our classification model. In a Decision Tree, the leaves represent classification and the branches represent classification and branches represent features that can lead to classifications [8]. A classification model is obtained based on the input attributes to predict the values of the output class. The Decision Tree predicts whether or not the patient is readmitted. After we build the model using Decision Tree classification method, the next step is find the importance of each variables in our diabetes dataset. We also use random forest classification method to identify the predictors in our dataset.

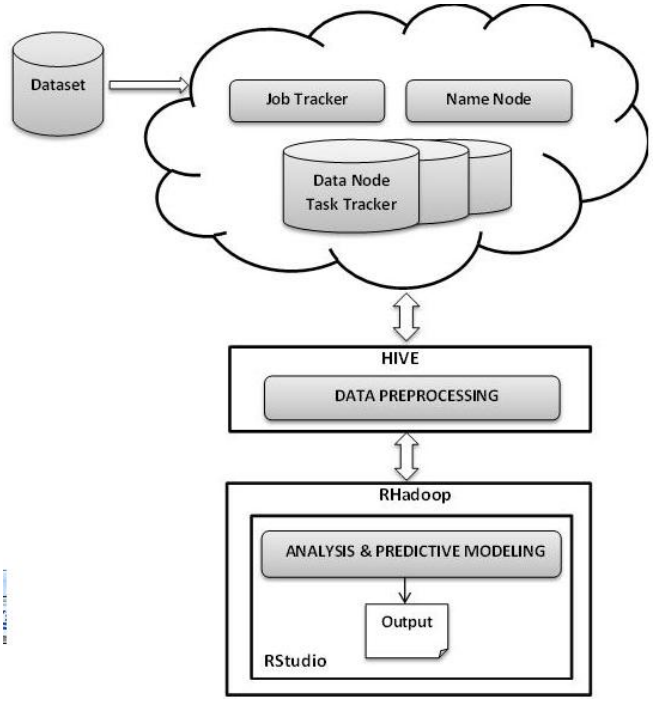

Figure 1: Architecture

Some classification rules for diabetes datasets from the decision tree are as follows,

1. Plasma glucose <127.5 AND Age <28.5 AND Body mass index <26.35 then 'tested_negative' for readmission (39/2).
2. Plasma glucose <127.5 AND Age <28.5 AND Body mass index <26.35 AND plasma glucose <99.5 AND pedigree <0.561 then 'tested_positive' for readmission (9/25).
3. Plasma glucose <127.5 AND Body mass index < 29.25 AND Plasma glucose <145.5 then 'tested_negative' for readmission (35/6).
4. Plasma glucose <127.5 AND Body mass index < 29.25 AND Plasma glucose <145.5 AND AGE <25.5 then 'tested_negative' for readmission (4/0).
5. Plasma glucose <127.5 AND Body mass index < 29.25 AND Plasma glucose<157.5 AND Age <30.5 then _tested_positive‘ for readmission (18/47).

To improve our model performance, next step is to find out the importance of each variable in diabetes dataset. All measures of importance are scaled to have a maximum value of 100. As show in Figure.3 Plasma glucose have highest variable importance i.e. 0.7881 followed by age (0.687), body mass index (0.686) while the insulin attribute is the least important. Now we detect the risk predictors in our diabetes dataset by using recursive feature elimination function. We rank each predictor based on its importance in the model. Based on recursive feature elimination function, we discover top 5 variables out of 8 variables i.e. plasma glucose, body mass index, age, pregnant, pedigree function are top predictors in our model. These are the top 5 readmission predictors in diabetic dataset.

3. Result

The result of correlation matrix is shown in Table.2. In the process of calculating correlation, we find that feature plasma glucose is highly correlated with readmission feature. From the Table.2 we can also find that Blood Pressure (0.065) is least correlated with readmission. But for building the decision tree classification model all the 8 features are included. Based on these features a Decision Tree has been generated. The decision tree is shown in Figure 2.

Table 2: Correlation between 8 features and readmission

Feature	Value of correlation
Pregnant	0.2218982
Plasma glucose	0.4665814
Blood pressure	0.06506836
Skin	0.07475223
Insulin	0.130548
Body Mass Index	0.2926947
Pedigree Function	0.1738441
Age	0.238356

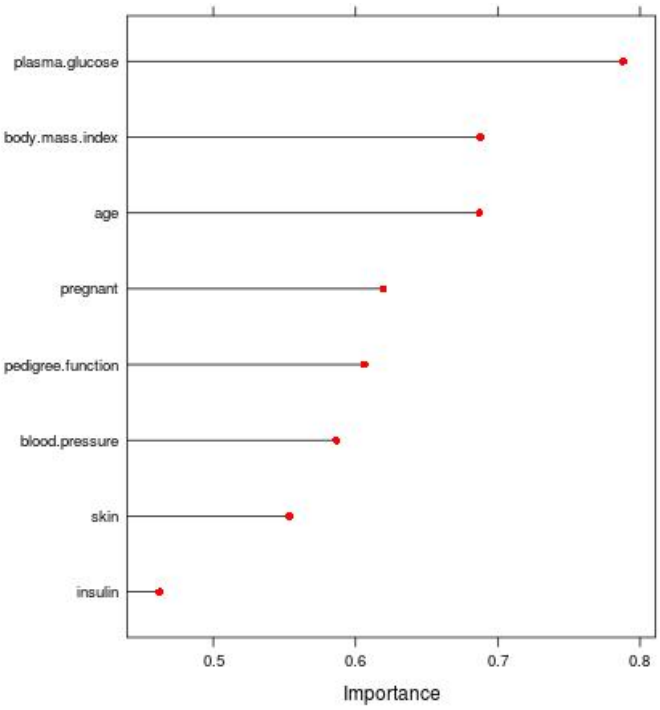

Figure 3: Variable Importance

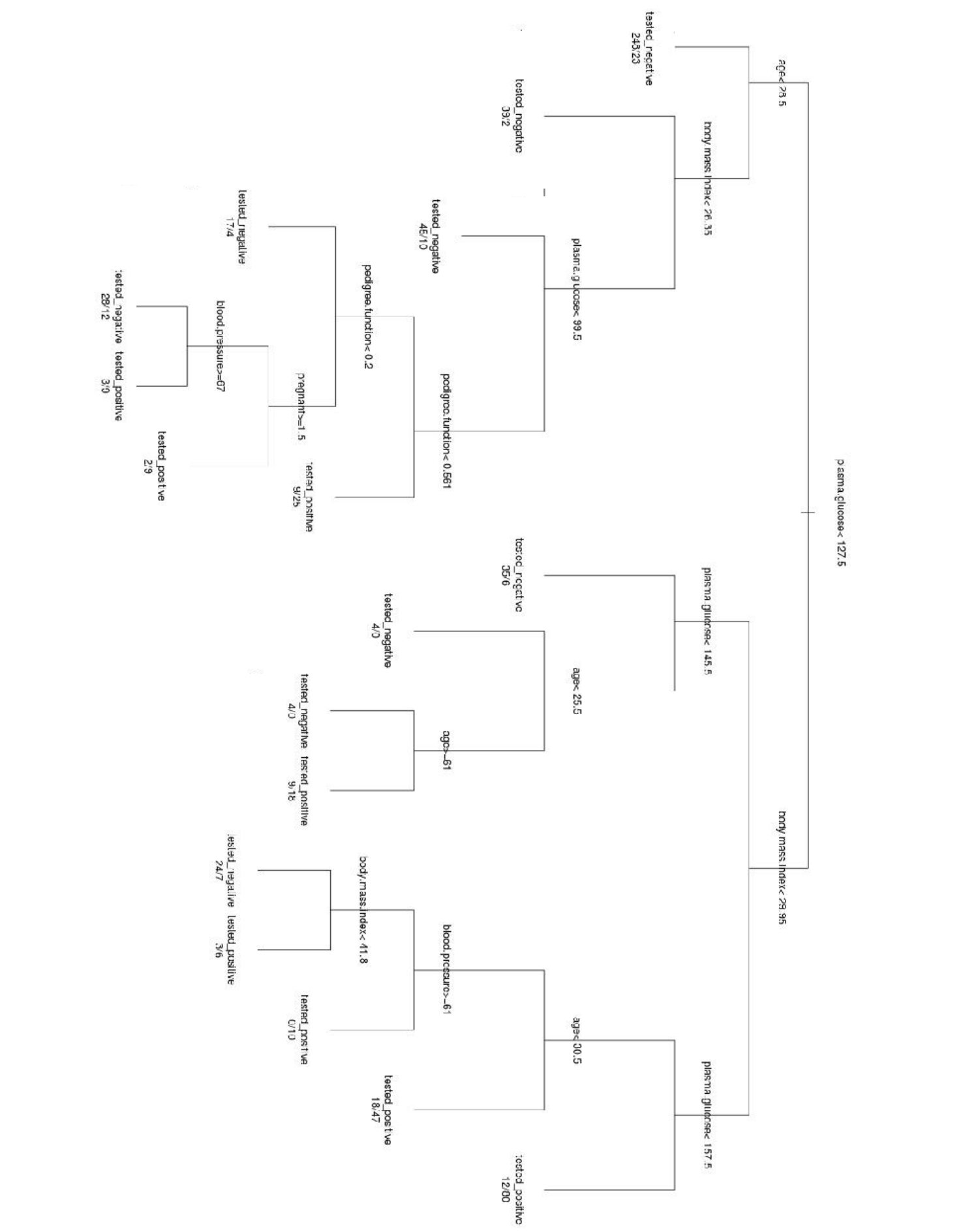

Figure 2: Decision Tree

4. Conclusion

Big Data analytics have been applied to evaluate the risk of readmission for diabetes patients. Predictive modeling has been employed by applying decision tree classification method. It has been observed that chance of readmission in diabetic patient is successfully predicted using the above analysis. Many analysis methods can be explored to improve the accuracy of the existing system.

References

[1] Donzé J. Aujesky D., Williams D., Schnipper J.L, MD. ―Potentially avoidable 30-day hospital readmissions in medical patients: Derivation and validation of a prediction model. JAMA Internal Medicine,‖173(8):632-638, Apr. 2013.
[2] ―Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications,‖ Report of a WHO Consultation Part 1: Diagnosis and Classification of Diabetes Mellitus World Health Organization Department of Non communicable Disease Surveillance, Geneva, 1999.
[3] UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html.
[4] Manyika J., Chui M., Brown B., and Bughin J. and Dobbs R. ―Big data: The next frontier for innovation competition and productivity,‖ McKinsey Global Institute, 2012.
[5] The Apache Software Foundation, http://hadoop.apache.org/common/credits.html.
[6] Thusoo, A., Sarma, J. S., Jain, N., Shao, Z., Chakka, P., Anthony, S., Liu, H., Wyckoff, P., And Murthy, R Hive—a warehousing solution over a Map-Reduce framework. In VLDB, 2009.
[7] H.W. Ian, E.F., ―Data mining: Practical machine learning tools and techniques,‖ 2005: Morgan Kaufmann.
[8] Sung-Hyuk Cha, and Charles Tappert, "A genetic algorithm for constructing compact binary decision trees,‖ Journal of Pattern Recognition Research, vol. 4, no.1, pp. 1-13, 2009.
[9] Ottenbacher K., Smith P., Illig S., Linn R., Fiedler R., and Granger C. ―Comparison of logistic regression and neural networks to predict rehospitalization in patients with stroke,‖ Journal of clinical epidemiology, 54(11):1159-1165, 2001.
[10] Murdoch T., Detsky A., ―The Inevitable Application of Big Data to Health Care,‖ JAMA. 2013; 1351- 1352. doi:10.1001/jama.2013.393.