IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-14, May 2019 E-ISSN: 2347-2693

BigData Analytics Predicting Risk of Readmissions of Diabetic Patients

G. Amrutha Varshini[1*], Nafisa.S[2], Priyanka[3], S.Sri Vishnupriya[4], Ambika B.J[5]

[1, 2, 3,4]School of Computing and Information Technology, REVA University Bangalore, India
[5] School of Computing and Information Technology, REVA University Bangalore, India India

*Corresponding Author: Ambikabj@reva.edu.in,

DOI: https://doi.org/10.26438/ijcse/v7si14.99102 | Available online at: www.ijcseonline.org

Abstract- Healthcare has huge impact on the society and also holds more importance in which analytics are applied to achieve accurate results about patients and to identify bottlenecks and to increase the business efficiency. Hospital readmissions are way too expensive and reflect the insufficiency in the healthcare system. Since readmission into hospitals has become unaffordable necessary measure needs to be taken to make them preventable [1] Readmissions rate decides the quality of treatment provided by the hospitals. Mostly readmissions are caused due to improper medication, early discharge, unmonitored discharge and poor care of hospital staff. In USA alone treatment of readmitted diabetics patients has exceeded over 250 million dollars per year. Advance identification of patient having high risk of readmission can allow the healthcare providers to perform additional investigations and also provides possibility to prevent readmissions. This method improves the quality of care and also reduces the medical expenses caused due to readmission .Number of patient visits, discharge order, type of admission were identified as the predicators of readmission. It was found that based on number of laboratory tests and discharge order both together predict whether the patient will be readmitted shortly after being discharged from the hospital (i.e. <30 days) or after a longer period of time (i.e. >30 days).These accurate results help the healthcare providers to improve care taken for diabetic patients.

Keywords*— *Machine Learning, Analysis on Medical data, Data collection, Data preprocessing, Data labeling, Predictive modeling, Model training, Prediction

I. Introduction

Agency for Health care Research and Quality (AHRQ) has conducted survey in the year 2011 which was held in United states and found that more than 3.3 million patients has been readmitted within 30 days of their discharge[3]. The readmissions are caused due to improper care provided to the patient at the time of their first medication. Improper care leads to the damage of patient’s life and treating of readmitted patients leads to increased cost of healthcare. During 2011 approximately 40 billion dollars has been spent on treating readmitted patients. A diabetic is the seventh leading disease that leads to cause of death and also affects about 23 million people in United States. In case of diabetics care hospital readmission has become a major concern, during 2011 more than 250 million dollars has been spent for treating readmitted diabetic patients [3].

Patients having high risk of readmission has to be identified before getting discharged from the hospital, to provide improved treatment to reduce the chances of their readmission. Patient getting readmission within 30 days of their discharge has been widely used as a survey for studying readmissions [3]. However a remarkable number of diabetic patients are readmitted after 30 days of their discharge. By considering the previous work that has been done in the domain in opposition to that here we are going to consider both short-term and long-term readmission scenarios. Identifying risk factors, corresponding to readmission will help in considering these factors with better documentation for future medical records of patients and with better care.

A diabetic is defined as clinical syndrome that is distinguished by hyperglycaemia, due to inadequacy of insulin in the human body [2]. This syndrome has become quite normal in today’s life irrespective of age. The disease is persistent and has no specific cure. Depending on the symptoms and levels of blood sugar in human body it differs from person to person. This turns the disease in to a considerate way so that the awareness has to be increased among the society in order to reduce the higher risk of readmission of patients in the hospital since the diabetics is a chronic disease .therefore its necessary to identify the risk of readmission of diabetic patients is important.

The key idea is to provide a complete data solution to readmission problem to make easier implementations at the healthcare organisations to undertake an outstanding improvement in the patient diabetic care. This solution provides all the information that is needed for implementing healthcare organisation along with the cost analysis.

Labelling these kinds of critical problems involves several challenges of data which are needed to be considered throughout the research. The main contribution of this work involves:

- Identifying diabetic patients with high risk of readmission by considering the patient's medical records using machine learning classifiers.
- Examining the characteristics of short-term (within 30 days) and long-term (after 30days) readmissions using different classifiers.
- Discovering the critical risk features using ablation (surgical removal of a body part or tissue) study.
- Using Association rule mining for identifying the critical risk features.
- Using cost analysis to determine the effective cost that has been saved by implementing the work in the real world.

The rest of the paper is arranged as follows: Section 2 provides a brief overview of the previous work. Section 3 describes the dataset used and the proposed methodology covering all the enumerated points mentioned above. Results and discussions are presented in Section 4 with respected to each part, followed by conclusion and future work in section 5.

II. Related Work

The dataset that we are using in this project has been en from the UCI machine learning repository [6] The dataset consists of more than 100000 hospital admissions from patients suffering diabetics which have been taken from 130 US hospitals. Various previous studies have been examined the risk factors that predict the readmissions rates of diabetic patients [7], [8], [9] out of these studies only the useful ones are discussed here. [8] Examined the readmission risk for a dataset consisting more than 52,000 patients in the humedica network.

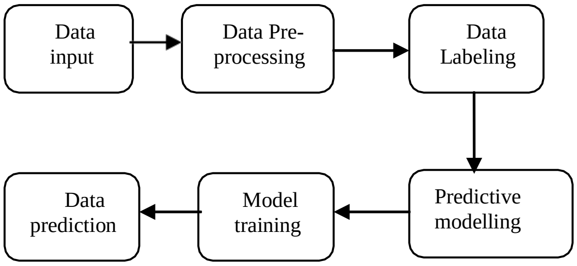


Fig. 1

The dataset that has been considered in this project covers clinical procedure and diagnostic related features along with drug information of patients around 65 years of age in order to predict the readmission within 30 days (short-term readmission).This contains related results on feature reduction methods, but the performance of these prediction models are adequate. In this project we are also going to talk about the meaning of diabetics.

Yasodha[4] has used the classification on various types of datasets that are used to decide whether a person is diabetic or not. The diabetic patient's data set is taken from hospital warehouse which contains various instances. These instances of this dataset which we considered are referring to two groups i.e. blood tests and urine tests. Aiswarya [5] aims to discover solutions to detect the risk of readmission of diabetes by investigating and examining the patterns that are originated in the data via classification analysis by using Decision Tree and Naïve Bayes algorithms.

With the best of our knowledge, our project work is based on analyzing the diabetic patients facing problems with the risk of both short and long term of readmission. We used a larger and a more balanced dataset (i.e. data consisting across all age groups and across 130 US hospitals) as compared to previous works that had been done. In the previous studies both short-term and long-term readmission scenarios were not considered and also not documented the performance of various kinds of machine learning classifiers.

In order to address the gaps in the research, this work covers various methods to identify the risk factors for predicting the rate of readmission. We hope the results that have been presented in this project work serve as a good baseline for any future project work to compare against.

III. Methodology

In this project we are going to demonstrate how to build a model for predicting risk of readmissions in python using the following modules.

1) Data Collection
2) Data Preprocessing
3) Data Labeling
4) Predictive Modeling
5) Model Training
6) Prediction

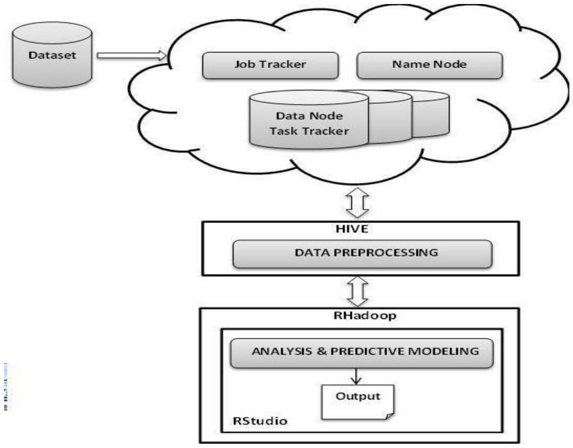


FIG:2 OVERVIEW OF THE METHODOLOGY

1. Data collection

To obtain accurate results for any study we need to consider data collection which plays an important role for collecting appropriate data from reliable source. While collecting the data there is chance of getting highly noisy data which is variant in nature. So it is necessary to check the data after collecting and also necessary to examine the data. It is required to choose the correct kind of attributes which may help us in getting accurate results. The dataset we are using is obtained from the UCI repository of machine learning databases. For a larger dataset that has been held by the national institutes of diabetes and digestive and kidney diseases this dataset has been selected. First this dataset is loaded into Hadoop File System. To modify the data into classifiable data, prepossessing of data is required.

2. Data Preprocessing

For modeling, we need to prepare the data for that we need to convert the noisy data into classified data. For training predictive models data preprocessing is premonitory step to construct the data which is suitable .Because of existence of inconsistencies and heterogeneity in the data, data preprocessing poses several challenges .By running Hive queries data is cleaned and null, missing values, outliers are eliminated .Thus, for predictive modeling the analysis data is prepared.

3. Data labeling

Class labels should be defined to proceed with the classification process .The status of readmission is denoted by dichotomous variable .The values 0 and 1 which are binary response variables are taken where the value 0 means tested negative for readmission and 1 means tested positive for readmission.

4. Predictive modeling

The suitable classification prediction technique is selected to build a model for predicting readmission in this module. Logistic-regression, gradient-descent, naïve-bayes, decision tree random forest tree, gradient boosting classifier, K nearest neighbors are the machine learning models that are applied to perform the experiment .Through common model quality measures such as error rate% ,confusion matrix, accuracy and ROC In the ratio of 7:3 model quality the data is classified in to training and testing .The basic step of predictive modeling is to identify the appropriate classification method. Error rates classifications are calculated for various methods based on our diabetes dataset.

5. Model training

In the ratio of 7:3 the data is split into training and testing sets and the training set is sent to models for training in this module. Based on all the machine learning algorithms for which the analysis will be carried out. This module creates a predictive model.

6. Prediction

This is the final phase where the actual prediction of the system will take place. The analysis on the test set part of data is performed in this module. It is the user end of the system .Here in this module any modifications to the data or any adaptive measures will be taken place. A short description of all the machine learning algorithms that are considered in this work is provided below.

(i) K nearest neighbours

KNN is one of the simplest machine learning models. This model is pretty easy to implement and to understand. For any given sample input the model looks for the k closest data points and determines their probability by counting the set of positive labels and then divides them by k.

(ii) Logistic Regression

It is a traditional machine learning model that best fits for a linear decision boundary between the positive and negative sample values. For features that are linearly separable logistic regression is preferred. The best advantage of logistic regression is that the model is interpretable

(iii) Gradient descent

It is similar to Logistic Regression. Both these methods use gradient descent in order to optimise the coefficients of a linear function. In gradient descent only a small set of samples are used at each iteration, where as in Logistic Regression all the samples are used at each iteration.

(iv) Naive Bayes

Naïve Bayes [10] is also a machine learning model which is occasionally used. The naïve part assumes that all the features of the samples are independent.

(v) Decision tree

It is one of the popular machine learning models and is a tree based method. The simple tree based method is Decision tree. This method is basically used to know which variable and threshold to use at every split of data.

(vi) Random Forest

These are created to reduce the over fitting. In these kinds of models multiple trees are created and their results are collected. The trees in a forest are arranged using random samples.

(vii) Gradient boosting Classifier

In this method we are going to create a bunch of shallow trees which helps us in improving the errors of previously trained data.

IV. RESULTS

Correlation matrix result is shown in the graph below. In this process of calculating correlation; we find that feature time in hospital is highly correlated with readmission feature. We can also find that A1cresult_1 is least correlated with readmission. But for building the decision tree classification model all the below mentioned features are required.

V. CONCLUSION AND FUTURE ENHANCEMENT

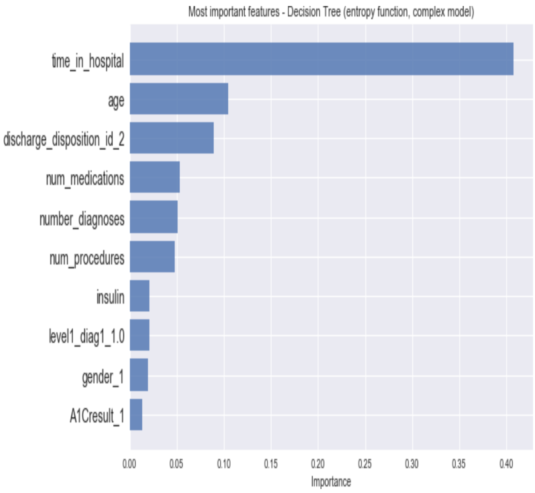
In this project work we developed a scheme which is used to identify the high risk of readmissions and also we have used different machine learning algorithms. In this work we considered both the short-term and long-term readmissions which are going to happen within 30 days of discharge of particular patient. And in this project we are particularly concentrating on disease called diabetics.

Using this project we have created a machine learning model which is used to predict the patients with diabetics who has higher risk getting readmitted within 30 days. Out of all the models that we considered Gradient Boosting classifier has most optimized hyper parameters. The model has the ability to catch 50 percent of readmissions and is better than just randomly picking patients.

To evaluate the risk of readmission for diabetes patients Big data analytics has been used .By using the above

ACKNOWLEDGMENT

We would like to express our deepest gratitude and thanks to Dr.P.Shyama Raj, Chancellor, S.Y.Kulkarni, and Vice Chancellor, Dr.M.Dhanamjaya,Register Evaluation, and Dr.Sunilkumar S Manvi ,REVA University.



Our deep and profound gratitude to our guide Mrs.Ambika B.j,Assistant Professor for her keen interest and boundless encouragement which we received during entire period of our project. Finally, we thank to all those who helped us directly or indirectly with their suggestions.

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] T. D. Briefing. Ahrq: The conditions that cause the most readmissions. The Daily Briefing. Web,2014..

[4] P.Yasodha and M. Kannan, "Analysis of a Population of Diabetic Patients Databases in WekaTool", International Journal of Scientific & Engineering Research.

[5] [2]A. Iyer, J. S and R. Sumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques".

[6] UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html

[7] T. D. Briefing. Ahrq: The conditions that cause the most readmissions. The Daily Briefing. Web, 2014.

[8] K. M. Dungan. The effect of diabetes on hospital readmissions. Journal of diabetes science and technology, 6(5):1045–1052, 2012.

[9] E. Eby, C. Hardwick, M. Yu, S. Gelwicks, K. Deschamps, J. Xie, and S. George. Predictors of 30 day hospital readmission in patients with type 2 diabetes: a retrospective, case-control, database study. Current Medical Research & Opinion, 31(1):107–114, 2014.

[10] H. Zhang. The optimality of naive bayes. AA, 1(2):3, 2004.