Identify the Patients at high risk of re-admission in Hospital in the next year

Submitted By

Ankur Makwana

13MCEC08

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2015

Identify the Patients at high risk of re-admission in Hospital in the next year

Major Project

Submitted in partial fulfillment of the requirements

for the degree of

Master of Technology in Computer Science and Engineering

Submitted By

Ankur Makwana

(13MCEC08)

Guided By

Prof. Jigna patel

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF TECHNOLOGY

NIRMA UNIVERSITY

AHMEDABAD-382481

May 2015

Certificate

This is to certify that the major project entitled **"Identify the Patients at high risk of re-admission in Hospital in the next year"** submitted by **Ankur Makwana (Roll No: 13MCEC08)**, towards the partial fulfillment of the requirements for the award of degree of Master of Technology in Computer Science and Engineering of Institute of Technology,Nirma University, Ahmedabad, is the record of work carried out by him under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for examination. The results embodied in this project, to the best of my knowledge, haven't been submitted to any other university or institution for award of any degree or diploma.

Prof. Jigna Patel
Guide & Assistant Professor,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Prof. Vijay Ukani
Associate Professor,
Coordinator M.Tech - CSE
Institute of Technology,
Nirma University, Ahmedabad

Dr. Sanjay Garg
Professor and Head,
CSE Department,
Institute of Technology,
Nirma University, Ahmedabad.

Dr. K Kotecha
Director,
Institute of Technology,
Nirma University, Ahmedabad

Statement of Originality

I, **Ankur makwana**, Roll. No. **13MCEC08**, give undertaking that the Major Project entitled "**Identify the Patients at high risk of re-admission in Hospital in the next year**" submitted by me, towards the partial fulfillment of the requirements for the degree of Master of Technology in **Computer Science & Engineering** of Institute of Technology, Nirma University, Ahmedabad, contains no material that has been awarded for any degree or diploma in any university or school in any territory to the best of my knowledge. It is the original work carried out by me and I give assurance that no attempt of plagiarism has been made. It contains no material that is previously published or written, except where reference has been made. I understand that in the event of any similarity found subsequently with any published work or any dissertation work elsewhere; it will result in severe disciplinary action.

_____

Signature of Student

Date:

Place:

Endorsed by

Prof. Jigna Patel

(Signature of Guide)

Acknowledgements

It gives me immense pleasure in expressing thanks and profound gratitude to **Prof. Jigna patel**, Assistant Professor, Computer Science Department, Institute of Technology, Nirma University, Ahmedabad for his valuable guidance and continual encouragement throughout this work. The appreciation and continual support he has imparted has been a great motivation to me in reaching a higher goal. His guidance has triggered and nourished my intellectual maturity that I will benefit from, for a long time to come.

It gives me an immense pleasure to thank **Dr. Sanjay Garg**, Hon'ble Head of Computer Science and Engineering Department, Institute of Technology, Nirma University, Ahmedabad for his kind support and providing basic infrastructure and healthy research environment.

A special thank you is expressed wholeheartedly to **Dr. K Kotecha**, Hon'ble Director, Institute of Technology, Nirma University, Ahmedabad for the unmentionable motivation he has extended throughout course of this work.

I would also thank the Institution, all faculty members of Computer Engineering Department, Nirma University, Ahmedabad for their special attention and suggestions towards the project work.

See that you acknowledge each one who have helped you in the project directly or indirectly.

- Ankur Makwana

13MCEC08

Abstract

Objective is to identify the patients at high risk for the future emergency or unplanned hospital admission. Unplanned hospital affirmation and re-confirmation are considered as a markers of expensive and unacceptable medicinal services and their evasion is principle issue of strategy creators for some nations. In the three years period ,patients data like released from a hospital and re-confessed to hospital expense contain more than a billion every year. Thus, our point is to decreasing unplanned ad- mission rates, the proof for their productivity and lessen the expense. With the specific aim of reduce the future admission or re-admission of patients we build a model to use for distinguish the patients at high hazard for unplanned admission or re-admission in next 12 months. Our target is to utilize an approved calculation to case-nd Medicaid patients at a high danger of hospitalization in one year from now and distinguish obstruction and responsive attributes to lessen hospitalization cost.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

For many years emergency hospital admission have been increasing in several countries, So first question is how we can decreasing or manage patients at high risk for emergency admission in hospital. So we can decrease or contribute for the financial pressure on hospitals and also a national health care budgets. So now days more initiatives are regarding of the improvement of the hospital management of high risk patients admission [1].

For that initiatives include the case management for the patients for a long term condition that might place patients for hospital admission to support the community and its trained by the trained case manager [2].

Case management system will targeted for the hospital admission and condition like patients Personal satisfaction and decline their danger for hospital affirmation [2]. Case management system are used for improved the methods needed for patients who might be profited from careful treatment for patients in primary and secondary condition [2].

In this study, we identify the patients at high danger of crisis hospital admission from the hospital data. We find out the group of patients with the high impact and who had one emergency hospital admission and atleast one emergency hospital admission in next year. So, we can targeted that group of patients identify the size of this group of patients identify the size of this group admission and evaluate the patients effectiveness and admission data before they emergency hospital admission in next year [3].

Focal point of predicting the patients which are at high hazard for one year from now admission in hospital is that this period may allow time to the case managers to contact the patients and captivate with the high-chance patients and it additionally permits time for treatment and changing treatment for the patients [3].

1.2 Definition

Descriptive analysis of inpatient hospital episode statistics and predict model developed using multiple regression techniques.We can develop a decision for patients who will be admitted next year.Using this develop a efficient method for taking good management for patient for improve their health and early prediction of patients who will be admitted in the next year [4].

1.3 Objective of Study

Our Objective is to utilize an approve calculation for case-finding Medicaid patients at a high danger of re-admission in hospital in the following year and distinguish the obstruction and responsive trademark to decrease hospitalization risk [5].

- To use routine information to recognize patients at high risk of future crisis hospital admission[6].
- Improving the management of high cost patients [6].
- Improving the cost of hospital due to emergency admission [6].

Computer implementation and automation is used in all fields of life whether it is entering through the door or measuring nutrition value from a cup of tea. Almost all natural phenomena are digitized now. Advancements in computer technologies are increasing day by day and covering all fields of the life. Implementation of computer technology in the area of health-care is an old story but there are still some areas of health-care where computer applications can make a difference. Improving the quality of life of patients by computer application is one of those.

Our main aim is to identify the patients at who will be admitted within the next year using historical data sets. Based on the sign, symptoms, days in the hospital,primary condition , place of service, drug count month age sex and days since first service etc.

Chapter 2

Literature Survey

Work done in patients identification using historical data sets using different algorithm and techniques are discussed below:

Alex Bottle, Paul Aylin, and Azeem Majeed, this prototype using data mining techniques, namely, Decision Trees, Naive Bayes and Neural Network for identifying the patients with high risk. Using medical profiles such as age, sex, Mental health,different disease it can predict the likelihood of patients getting a re-admission in the next year. It is implemented on the Java platform [2].

Maria C. Raven, John C. Billings developed , the proposed technique involves training a Multilevel Perception with a patients learning algorithm to recognize a pattern for the diagnosing and prediction of patients admitted in the next year. .About 94 cases of different sign and symptoms parameter have been tested in this model.This study exhibits ANN based prediction of neonatal disease and improves the diagnosis accuracy of 75 % with higher stability [3].

Hiroshi Takeuchi, based on this study identify the patients sign and symptoms. this study are the signs, symptoms and the results of physical evaluation of a patient. The proposed system achieved a high accuracy [5].

John Billings, Ian Blunt, Adam Steventon, Theo Georghiou, Geraint Lewis, Martin Bardsley Development of a predictive model to identify inpatients at risk of re-admission a set of records with attributes are used for training and testing. It recommended regulated system for determination of patients sign and symptoms disease and trained it using back propagation algorithm. On the premise of unknown information is entered by doctor the

Table 2.1: Literature survey Table:

Author	Year	Techniques	Suggestion
Sankalp Khanna,Justin Boyle,Norm Good [7]	2014	Filtering the training data for index admission and multiple linear regression method	Give various weights to the presence of condition for other disease
Stephen C. , Michel W.[4]	2013	Regression techniques based on claims	apply regression techniques for supporting people with long term condition.
Joachim Szecsenyi [5]	2013	A tree based algorithm	Not depends on only one algoritham for accurate prediction.
Manisha Rathi,Thierry Chaussalet [1]	2013	Fuzzy regression method using JAVA	It require defination of rules and policies and also require knowledge of medicine for accurate result.
Paul Aylin, Norm Good [8]	2013	Rule based algorithm for hospital level risk for re-admission	Is the all-cause readmission metric sensible? Should we move to stratifying by cause,
Kiyana Zolfaghar, Naren Meadem, Brian Muckian [9]	2013	Risk prediction based on hive structure and mysql	Find the missing value and find the patients for better emergency admission.
Eren Demir, Thierry Chaussalet [6]	2010	Building a longitudinal data set based on sql scripts cleaning rules	using data sets in more efficient way
Hou Zhengkun,Li Suyun,Yu Xueqing,[10]	2010	Derivation of the prediction rule who are middle aged and elder	Many rules are reusable so using that rules within the same system for good accuracy.
Rashedur M. Rahman,Fazle Rabbi Md. Hasan, Brian Muckian [11]	2010	different classification techniques for mining and use of decision tree	Used classification techniques for categorize diagnoses
E. Emam ,L. A.[12]	2010	patients identification using rule based	Data set using neural network and random forest algorithm

Table 2.2: Literature survey Table

Author	Year	Techniques	Suggestion
M. C.,J. C. [3]	2009	based on classification techniques	data sets depends on the attributes
P. A. El [2]	2006	classification techniques analysis for hospital admission of patients	When we try to remove feature it also increase the accuracy and it also contain the loss of information so make sure add or remove data based on data sets.
H. Xie, P H. Aard [13]	2006	based on risk factor and based on features	Based on features apply algorithm

framework will find that unknown information from preparing information and produce run down of conceivable disease from which patient can suffer [1].

Miniati Roberto ; Bonaiuti Roberto, this will be helpful for the inside patients details and identify the patients who will be in the hospital for several days and using this information which patients re-admission in the next 12 days. Also, applying hybrid data mining techniques has shown promising results . so applying hybrid data mining techniques in selecting the suitable treatment for patients sign and symptoms, patients needs further investigation [6].

Teichmann E , Demir E , Chaussalet T , the purpose of this paper is to describe the development of a model for predicting unplanned 30-day readmissions. Our research objective is to develop an all-age, all-cause 30-day readmission risk model for unplanned acute care hospitalization with logistic regression on health plan claims data [10].

Li Jiansheng ; Hou Zhengkun ; Li Suyun ; in this paper Our individual readmission risk scores could be available at the time of admission and thus may have implications for individualized treatment plans and managing the discharge process at an early stage. Another application of our risk scores is to identify patients at high risk of readmission for outpatient care transition management [13].

2.1 Background

Millions of people are getting admitted in the hospital every year and identify the patients who will be admitted within the next year is a biggest challenge. The World Health Organization (WHO) analyzed that twelve million deaths occurs worldwide due to unknown admission rate at hospital.Hospital are not well prepared for unknown or Unplanned admission rate.
Medicinal conclusion plays a important role by undertaking that needs to be executed effectively and precisely. So when we identify the patients who will be admitted then its reduce the cost of hospital. We use data mining techniques for identify patients in efficient manner.Data Mining techniques should be used in all data sets to improve a better prediction by last 10 years.

To identify the patients first we should learn the techniques and sign and symptoms of the patients.This will be useful for hospital or medicaid service to identify the pattern of the patients and identify the patients re-admission. We have to use patients condition like age, sex, month, place of service, length of stay, day since first service, primary condition, days in hospital, drug count etc [14][15].

We use different data mining techniques to help hospital even a medicaid service for exactness of patients identification. There are different methods like Gradient Boosting Machines, Random Forests, Neural Networks, Logistic Regression, Support Vector Regression ,K-Nearest Neighbors these techniques are used for finding the patients at high risk of re-admission in the hospital in the next year. But in the health care prediction it should give a perfect accuracy or higher accuracy then it will be helpful to hospitals. So we try to improve a accuracy for better health care. There are many methods of data mining so using these data mining techniques we can improve our accuracy and accuracy is based on the data sets so that detail study of data sets is also needed for better accuracy. We can also use a mixed method for finding a better accuracy in the patients identification admitted in the next year. There are many models for finding better accuracy and there are many complex models for finding better results but based on the data sets we have to find the which is the better techniques for our data sets and we find more methods for better accuracy based on data sets.

This research uses R-programming. The data sets for hospital re-admission is taken from kaggle site [16][17]. So using these different table first we have to convert into a single table and in csv format which is useful for our data sets using R-Studio. R-Studio is used for applying methods to our data sets and finding the best algorithm based on methods and data sets. Analysis of data sets in different manner for different features and those features are not so useful we eliminate those features and make data sets more accurate. Finally we can compare our accuracy with different algorithm within R-tool find better accuracy and better decision out of our data sets, these methods will help to predict the patients who will be admitted within the next year using historical data sets [18].

2.2 Approach and Methodology

Patients identification prediction who will be admitted within the next year using the below rules.

- The predicted output is not based on the assumption and also not based on the assumption or prior knowledge.
- For these methods should be run on large data sets ,small data sets is not enough for identification of patients.

Implementation is based on R-Studio we are using and R-Studio is an open source tool for data mining methods. R-tool contain the classification techniques, regression techniques, clustering techniques, apply association rules [19][20]. R-tool is used for classification techniques and we used classification techniques in our data sets. The following steps are performed in R-tool.

- Start the R-tool.
- Open data set and upload in CSV format.
- Write classification techniques and install packages for the same.
- Select test data on our data sets
- Click on run and prediction result will be generated.

2.2.1 DATA

For comparing various classification techniques, Download dataset from kaggle site , which is available at http://www.heritagehealthprize.com/c/hhp/data. The dataset has 13 attributes. However, 13 attributes are used for this study and testing as shown in Table 3.

Table 2.3: Selected Patients Attributes

Name	Description
Age	Age in years Generalized into ten year age intervals.
Sex	Biological sex of member: M = Male; F=Female.
PlaceSvc	Generalized place of service.
LengthOfStay	Length of stay (discharge date admission date + 1)
DSFS	Days since first claim, computed from the first claim for that member for each year
PrimaryConditionGroup	Broad diagnostic categories, based on the relative similarity of diseases and mortality rates
ProcedureGroup	Broad categories of procedures, grouped according to the hierarchical structure defined by the Current Procedural Terminology
DrugCount	Count of unique prescription drugs filled by DSFS.
LabCount	Count of unique laboratory and pathology tests by DSFS
Year	Year in which the drug prescription was filled: Y1; Y2; Y3.
CharlsonIndex	The slope of the peak exercise segment .
PCP	Primary care physician pseudonym.
Grouping of Specialty	Specialties were grouped in consultation with hospital physicians, and informed by the specialty definitions provided.

Chapter 3

Methods for finding Patients using Data sets

3.1 Case-finding algorithm

Case-finding algorithm methods are used at hospital levels so that we avoid unwanted hospitalization.Briefly for the three years of data of the hospital study the hospital financial problem or unwanted hospitalization. We applied regression techniques for the hospital service like admission in the hospital, inpatients, outpatients, emergency department and a clinic visit. Identify the all parameter and diagnose the for every patients for three years of data and identify the patients who will be admitted within the next year. Coefficient from the algorithm applied for past three years and generates the score of each patients also find the risk of the re-admission in the next year [3]. Algorithm generates the risk score of each patients and patients having the highest risk score will be admitted next year [3].

3.2 Patients Sampling

Patients sampling is based on the age of the patients and who is getting the highest risk score are eligible for the further inclusion.we conduct a daily computer query for the patients hospital admission and find the patients was on the high risk for the future re-hospitalization and met the criteria.how ever many states are used to focus on the reducing the cost of hospitalization patients needed help at home and patients who are in the jail are excluded.as a reduce of the unwanted hospitalization collect quantitative

data as well as quality data collected from each patients and work upon them.eligible patients were enroll for the further details and study [3].

3.3 Quantitative Interview Tools

It has been illustrated that other than the disease factor like mental problem, housing illness, use of medical at home, use of health service etc are calculated as a general health status. For that information we use validate tools for collect data. For further information like patients participants in their health, social support, mental health , patients hunger and medical data, these information collected by the user for further investigation. User want detail description of the quantitative interview tool [7].

3.4 Claims Data

Toward the completion of the study period, we extricated symptomatic of patients and outpatient administrations data for our future reason for a long time information set and study healing facility's monetary information used to create our case-discovering calculation [7].

3.5 Data analysis

We utilized engaging dissect devices for the information sets and that information sets contrasted and the quantitative instrument for the mean score of the information and contrasted and the general information on every measure. Claims information were dissected utilizing illustrative insights including means and frequencies [7].

3.6 Data Characteristics

The Data set contain all the necessary information like patient id,day since first service, length of stay, age at first claim, age, sex, disease during the stay in a hospital.So, when we cover all the data there may be a chance of interpreting the result. In the data set there should be a missing value,missing data elements due to not applicable data or not relevant data, high risk of the data and data which are not relevant. Some data has their individual value information and that information should not be ignored so modeling of missing data is required[8]. The data set should be in a proper manner or in a proper structure and provide meta-data that provide syntax of all data record.Identifying the patients for re-

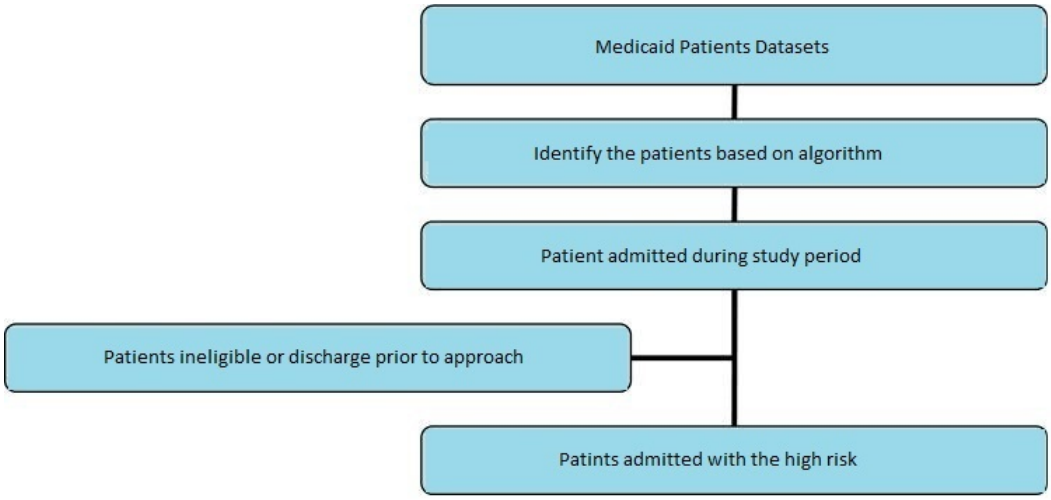

Figure 3.1: Complete scenario of Identify the patients of admitted in Hospital [1].

admission depends on the corresponding same visit to underlying condition and different visit may contain different episodes. For example, sometimes second visit of patients might identified with the last visit. The level of connection of primary determinations was dictated by taking a gander at the category for patients [9].

3.6.1 Mental Health

Estimation of a mental well being is a vital piece of the patients well being care. There are more than 50 percent of patients that are experience the ill effects of the mental well being and misery than the general population.Taken together, substance use and mental well being analyses recognized from the Medicaid claims information were more pervasive in our study populace than other therapeutic diagnoses. From the overview more patients are conceded because of the mental well being issues. The extent of patients with no less than one inpatient affirmation every year because of a substance use related condition or mental well being condition was higher than the extent utilizing any outpatient dependence treatment or mental well being administrations [3].

Chapter 4

The Data Preparation Frame-work

4.1 Overview

Figure 4.1 below is showing the Data Preparation Framework. Beginning with the raw data set which are described in the figure and simplify it :

Data cleansing

e.g.It remove records without an essential analysis found in the information sets or clean the information where release date is before the admission date [1].

Variable reduction

e.g. remove variables that are populated short of what 1 of the record set and The reason for appointment is to disentangle the elucidation of a complex information set. Notwithstanding, this intention is vanquished if there are a substantial number of natural variables. What is implied by a "vast number" is to a great extent a matter of taste, and the goals of the examination.

Cardinality reduction

As name suggest numerous variables have a high number of classifications that won't lose any huge esteem through gathering e.g. post codes of patients and Classification of Diseases and Related Health Problems based on the data sets and codes [1].

longitudinal data set

An information set is longitudinal in the event that it tracks the same sort of data on the same subject at different focuses in time. Utilization of longitudinal information is that we can appraise the impact of different components on change and we can measure the effect of different arrangements with sensible exactness. Linking and transformation

rules are kept in a configurable flat file [11]. The process of collecting sample observation from a larger population over a given time period.

To build a longitudinal data set , we looked at the three year of data set and find out the patients whose aged is 65 and admitted as an emergency in the hospital. Deaths were excluded from the data sets. To assemble a longitudinal information set, all the relating records of patients were gathered and union over the monetary years utilizing the framework created id. All the records were linked using data manipulation techniques as described in the fig and also performed many of the data transformation techniques which are used for modeling [12].

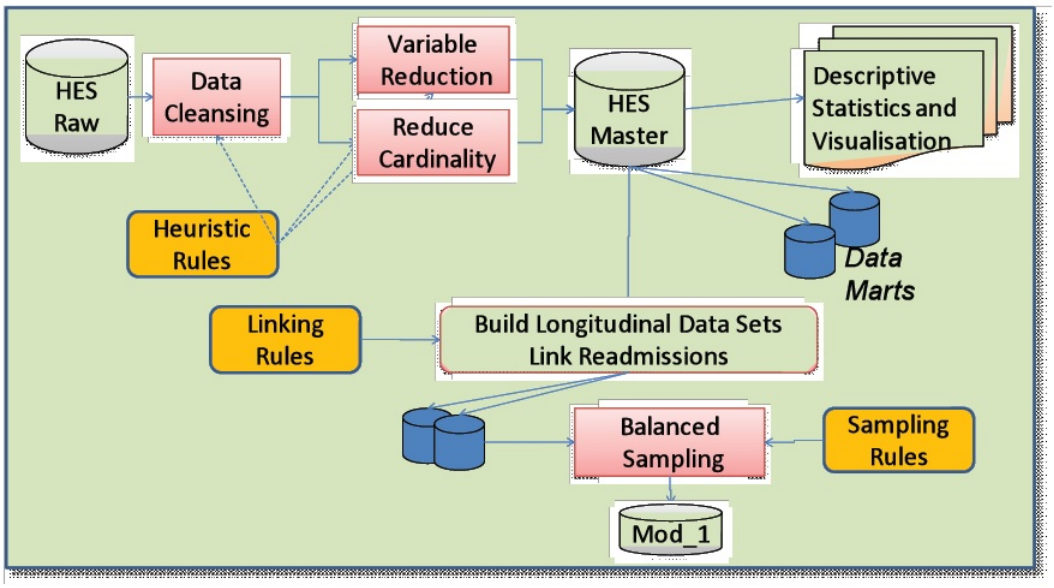

Figure 4.1: Data preparation Frame-work. [1]

In the above figure 4.1 Raw data is consumed as a data cleansing procedure,data cleaning procedure remove records without an essential analysis found in the information sets or clean the information then heuristic rules, reduce cardinality and variable reduction is applied in the data cleaning set so that it obtain the other table with reduction values and use in the future purpose and these data are kept in a file. The process of collecting sample observation from a larger population over a given time period. For linking the data set we build a longitudinal data set for that we looked at the three year of data set and find out the patients whose aged is 65 and admitted as an emergency in the hospital.

Table 4.1: Additional Variable

Variable	Description
Count	number of re-admission in the hospital
Predisdate	Previous discharge date
Nxtdisdate	Next discharge date
break	Time to re-admission

Deaths were excluded from the data sets To assemble a longitudinal information set,all the relating records of patients were gathered and union over the monetary years utilizing the framework created id. All the records were linked using data manipulation techniques as described in the fig and also performed many of the data transformation techniques which are used for used for model the frame of patients.

In the above table it describe as a count the number of re-admission in the hospital based on the previous discharge date and next discharge date. So ,based on the count, Predisdate and Nxtdisdate we can identify the patients who are admitted in the next year.

Chapter 5

Classification Algorithm Used:

This paper has underlined particularly on Gradient Boosting Machines classifiers for patients identification who will be admitted within the next year. Gradient Boosting Machines is used here for Data mining techniques because it is very useful. Gradient Boosting Machines techniques definitely not hard to execute and direct. Gradient Boosting Machines is a method used in R-tools for better accuracy. Gradient Boosting Machines is data mining techniques which gives a better accuracy. Gradient Boosting Machines are exceedingly powerful tools in numerous regions, for example, information and data mining based on the data sets. First find out the pattern from our data sets and extract that information for better use. Make sure that it should be deal with the missing values based on the feature are used [19].

Another techniques in classification we are used is Random Forest.Random Forest is another classification techniques which based on the attribute and based on the attribute find out the best accuracy. This classification techniques is based on the supervised classification techniques because of the dependent attribute in our data sets [21].
Random Forest is used for better accuracy for a patients identification so that it impact on the exactness. Random Forest complexity is based on the variables like : nodes, leaf nodes, tree size and other property that can effect the accuracy. Tree size ought to be moderately little that can be controlled by utilizing a strategy called pruning [22].To have a sensible relationship between these calculations, its base on the time and size of the tree for our data sets. The general approach took after for Random Forest request for satisfying the objective is.

Algorithms Used are :

In this research data have highlighted a number of predictors, each with varying levels of accuracy. The predictors that we will focus on in this paper include:

- Gradient Boosting Machines
- Random Forests
- Neural Networks
- Logistic Regression
- K-Nearest Neighbors
- Support Vector Regression

We use several algorithms and its accuracy then from which several algorithms are combined to create a superior predictor.

5.1 Results of Classification Algorithm used :

This data sets provides a several tables to get information about patient information. The data sets includes a three years of hospital records. However, the amount of days spent within the hospital by every patient is just provided for two of the three years [16]. The job of the predictor is to see however long the patients spent in the hospital within the third year.

The tables provided by these data sets include data regarding Days since first service, prescriptions, lab results, primary conditions, and other relevant information. In order to facilitate the prediction models, it is necessary to reduce this data into one consistent data set. We use different method to derive our task to identify the patients who will be admitted within the next year using these data sets [23]. More about this method will be explained in later sections.

Predictions are evaluated using root mean squared logarithmic error, referred as a RMSLE.

$$\varepsilon = \sqrt{1/n \sum_{i=1}^{n} log(p_i + 1) - log(a_i + 1)^2} \tag{5.1}$$

Where:

- patient's MemberID= i.
- total number of patients = n.
- prediction made for patient i = p(i).
- actual number of days spent in the hospital by patient i = a(i).

We include 70 percent of the data as a train data and only 30 percent of the data as a test data is used. The root mean squared logarithmic error (RMSLE) calculated on the entire test set is and that data sets is compare with the year three data sets [24].

A maximum accuracy threshold has been set at $\varepsilon = 0.4$. In order to achieve a good results , their predictor output must have an root mean squared logarithmic error (RMSLE) below this threshold. So far we get a accuracy threshold value between 0.45 to 0.5.

After getting results of all method then While simply predicting the mean of all predictions may be quite effective, we can use the feedback from data scientists to improve our blending algorithm significantly [25]. The RMSLE indicates however correct a predictor is, that makes it possible for us to assign weights to every predictor. This ought to issue us an even more exact ensembles indicator.
This strategy includes discovering an ideal weight vector for the arrangement of classifiers through improvement.

5.1.1 Data Pre-processing:

Although data sets includes the information about with several tables of information about medical records, the prediction models we utilize require the information to be consolidated into a one uniform information table, where every individual is represented to as a highlight vector.

We apply different method and released their results for this problem [20]. We use different method to pre-process the data tables provided by Data sets and reduce them to two consistent matrices. The first matrix consists of training information, that the amount of days spent within the hospital is known to us and there area unit 147,473 members during this set. The second matrix consists of 70,942 members, that the amount

of days in hospital don't seem to be known to us and we have to find out that results. These are the individuals on which our expectation is scored. This technique diminishes each individual's credits to 139 remarkable features. When this system for represent to every individual as a vector was made, it then got to be conceivable to produce expectation models for the test set [17].

5.1.2 Gradient Boosting Machines :

Gradient boosting machines deliver an expectation model as an outfit of weaker forecast models, which for our situation are decision trees. We construct the model for a Gradient Boosting Machines.

In a stage-wise style wherein every stage an cost and accuracy is derived, much like other boosting systems do. By issuing one our training set which comprises of the features of X patients and their comparing Y days spent in the hospital, it can find an approximation $\widehat{F}(x)$ to $F^*(x)$ that minimizes the error.

$F = arg_F min E_{x,y} C(y, F(x))$

In Gradient Boosting Machines method we include Its parameters like include the number of trees to group together the measure of these trees, the minimum number of observations and shrinkage.We generate the different results for this method and Through cross-validation, we have discovered that expanding the quantity of trees, increment shrinking the upgrades,keeping the depth value within six to eight, and expanding the number of observation all assistance toward discovering the ideal results or prediction while at the same time preventing over-fitting.

Presently, our best usage of this model in R utilizes 8000 trees, a shrinkage of 0.002, a depth of 7, and 100 base observation for each leaf, with a specific end goal to acquire a RMSLE of 0.462998, making this our generally exact predictor for this method.

5.1.3 K-Nearest Neighbours

The K-Nearest Neighbor indicator discovers individuals in the preparation set that are most like individuals from the test set inside of the feature space, and use a blend of the known number of days spent in the hospital to make prediction for the year three values in our data sets which are unknown [20]. In order to calculate a correlation value between

two users, we use the below equation for that:

$$C_{ij} = \frac{1}{N}\sum_{i=1}^{n}(X_{in} - X_i) - (X_{jn} - X_j)^T \tag{5.2}$$

where C_{ij} = correlation coefficient, N = the number of features, $X_{in} - X_i$ = the centralized training set, $X_{jn} - X_j$ is the centralized test set We utilize this to make sense of which individuals from the training set are most nearly identified with the objective individual from the test set. This method of comparing users is, however, very difficult to understand and very costly.

With a specific end goal to find the nearest matches between individuals from the test set and, training set, every user in the test set must be thought about the every part of the test set. Moreover, numerous features are exceedingly connected to others also, accordingly, can bring about the correlation algorithm to place an excessive amount of weight on specific features. We tackle this issue utilizing eigenvalue decomposition. Utilizing this strategy, we decrease the quantity of features and spherize the information. We join and break down the training set and test set utilizing the following mathematical statement:

$$X = U\lambda U^T \tag{5.3}$$

5.1.4 Logistic Regression:

Logistic regression fits a logistic bend onto the preparation set and afterward utilizes this bend as a model to make expectations on a test set. Our forecast is given by the following formula :

$$\sigma(X; w) = \frac{1}{1 + e^{w^T X - w_0}} \tag{5.4}$$

We then find the subsidiary of the expense capacity and utilization a gradient descent to find an ideal estimation of w. Since the data set contains such a large number of individuals, it is unreasonable to figure the inclination for all individuals at every step. Subsequently, we randomly select a part at every step and compute the gradient descent that member only [20]. After calculating the cost gradient for member i, we upgrade w by subtracting the gradient multiplied by a step size constant , which is gotten through

cross-acceptance.

$$w^{t+1} = w^t - \eta g(w^t) \tag{5.5}$$

In this comparison, w^t is the current cycle of w and w^{t+1} is the redesigned w vector. We repeat this procedure more than an extensive number of ventures until the cost gradient meets on zero, significance no further steps are essential. In the wake of streamlining for $\eta = 0.001$ and the quantity of steps T = 100,000, our R-tool execution of this technique accomplished a RMSLE of 0.46672.

5.1.5 Support Vector Regression:

Support Vector Regression is a linear regression method, similar in many ways to logistic regression. What separates it is that, keeping in mind the end goal to avoid over-fitting. furthermore, decrease the calculation time, all information focuses not as much as a certain separation from the best-fit line are overlooked when ascertaining the cost gradient. Through cross-validation, we verified that the optimal satisfactory separation from this line is $\varepsilon = 0.02$. Using this value of ε and Liblinear's L2-regularized support vector regression, we obtained an RMSLE of 0.467152

5.1.6 Random Forests

A random forest prediction model uses an outfit of randomly generated decision trees with a specific end goal to create its expectations. Generally, it makes numerous decision trees with the expectation that the individual errors of every decision tree will counteract one another when joined in a an entire forest.

On the other hand, all together for this forecast model to work, a component of irregularity must be kept up all through the whole era of the woods. On the other hand, all together for this prediction model to work, a component of randomness must be kept up all through the whole era of the forest. In our usage, this component of randomness is presented by feature determination at the distinctive nodes. In addition to this, the decision trees were bagged [26].

This permits the diverse decision trees to be based upon different subsets of the training set, accordingly making a difference to counteract over-fitting. Our R-tool implementation of this prediction model used 500 trees, each with a maximum depth of 15, and achieved an RMSLE of 0.464918. This is the second best individual model behind

gradient boosting Machines.

5.1.7 Neural Networks

Artificial Neural Networking may be a process model that is galvanized by the biological workings of neurons. The model consists of layers of nodes, the artificial corresponding to a nerve cell. These nodes cascade info to every other through Associate in Nursing abundance of weighted lines. The model utilized in our analysis is composed of 3 totally connected layers. The primary layer inputs vectored member options. The second layer of nodes combines the input file with different weights. The third combines the nodes once more to provide a prediction [25].

We use the technique of back propagation to change the feature weights. The process found the error between the model with verity information. This result was then backtracked through the pathways to seek out that weights were most responsible for the error. These weights were then modified within the direction of negative gradient of the price perform. In our implementation of this predictor, we used seven neurons within the hidden layer, and completed 3000 cycles. Our most eminent neural network predictor, that we tend to enforced with R-tool, obtained an RMSLE of 0.465705.

5.1.8 Summary of all Predictors method that are used:

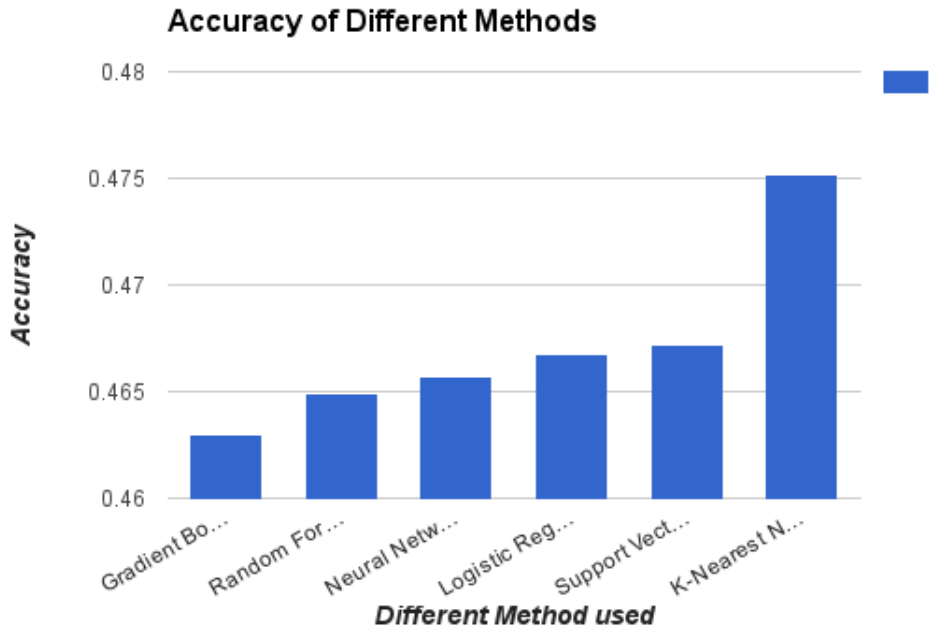

Figure 5.1: Analysis of Methods

- Gradient Boosting Machines = 0.462998
- Random Forests = 0.464918
- Neural Networks = 0.465705
- Logistic Regression = 0.466726
- Support Vector Regression = 0.467152
- K-Nearest Neighbors = 0.475197

When we have an adequate number of predictions and a RMSLE for each of them, we are then ready to ensemble the different predictors.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Based on the prediction results , we can say that Gradient Boosting Machines technique is the better techniques for classification using our data sets for patients identification prediction who will be admitted within the next year. Gradient Boosting Machines give us a better accuracy. We can say that based on the accuracy its clearly shows that Gradient Boosting Machines methods give us a better accuracy compare to other methods. The method Gradient Boosting Machines based on kaggle data has the highest accuracy of 0.4629% while at the same time we get the accuracy of Random Forest algorithm has the lowest of 0.464948%

In conclusion, as distinguished through the literature survey, we acknowledge just a negligible achievement is accomplished in the production of prediction model for patients and hence there is a necessity for combinational and more complex models to build the exactness of prediction. construct the precision of anticipating exactness.

By upgrading our expectations through blending, we have possessed the capacity to reliably minimize our error and get the minimum accuracy. This reaffirms the idea that ensembling numerous indicators together creates preferred results over any person one can. In spite of the fact that this attestation is an accomplishment all by itself, there is considerably more work that should be possible to enhance our accuracy.We could investigate advancing our mixing mathematical statement considerably further through the utilization of a regularization constant. Since this criticism is just based upon a subset of the test data, we run the high risk of over-fitting to it toward the end of the

finding accuracy of all methods, we will be scored on a different subset of the test data and find out the accuracy.By the by, ideal regularization can be determined through cross-validation.

6.2 Future Work

There are numerous conceivable upgrades that could be investigated to enhance the precision of this expectation framework. Because of time constraint, the accompanying examination requirements to be performed later on.

- Like to make utilization of testing different data mining strategies, numerous classifiers procedure.
- This paper proposes a framework using the methods like Gradient Boosting Machines and Random Forest to arrive at an accurate prediction of patients identification in the next year. Further work includes improvement of framework utilizing the said approach to be use for checking the inconsistency with other Data mining models.
- In future we used different classification techniques for better accuracy.

Further upgrades could be made in the zone of feature configuration. On the other hand, it might be a smart thought to include or exclude a few features. For instance, we may include linear or quadratic of distinctive features that we watch may have interesting connections with each other On the inverse side of the range, we could likewise remove a few features that may simply behave like a noise for some of our prediction models.

In both cases, we would need to run tests data and examinations to guarantee that the progressions that we make on our features are actually empowering us to create better results. In conclusion, we could upgrade our expectation models much further and include new ones that may work well on our data set.

References

[1] T. C. Eren Demir, "Data preparation for clinical data mining to identify patients at risk of readmission," IEEE Health and Social Care Modelling Group, 2010.

[2] P. A. Alex Bottle and A. Majeed, "Identifying patients at high risk of emergency hospital admissions: a logistic regression analysis," Jouranal of Royal society Medicine, November 29, 2006.

[3] J. C. B. Maria C. Raven, "Medicaid patients at high risk for frequent hospital admission: Real-time identification and remediable risks," IEEE International Conference on Healthcare Informatics, 2009.

[4] M. W. Stephen Campbell and J. Szecsenyi, "Comparison of physician referral and insurance claims data-based risk prediction as approaches to identify patients for care management in primary care: an observational study," October, 2013.

[5] M. W. Stephen Campbell and J. Szecsenyi, "Penn study shows automated prediction alert helps identify patients at risk for 30-day readmission," Penn Medicine, November, 2013.

[6] M. Rathi, "Risk prediction model using fuzzy regression method for predicting unplanned hospital admissions," IEEE, 2013.

[7] N. G. Sankalp Khanna, Justin Boyle, "Precise prediction for managing chronic disease readmissions," IEEE,Health and Social Care Modelling Group, 2014.

[8] D. B. Alex Bottle, Paul Aylin, "Predictors of readmission in heart failure patients vary by cause of readmission," IEEE International Conference on Healthcare Informatics, 2013.

[9] B. M. Kiyana Zolfaghar, Naren Meadem, "Big data solutions for predicting risk-of-readmission for congestive heart failure patients," IEEE International Conference on Big Data, 2013.

[10] Y. X. W. M. Hou Zhengkun, Li Suyun, "Ccerw: a new prediction rule to identify treatment failure patients with community-acquired pneumonia for middle aged and elderly," IEEE, 2010.

[11] H. W. Chao Ou-Yang, Sheila Agustianty, "Developing a data mining approach to investigate association between physician prescription and patient outcome ? a study on re-hospitalization in stevens?johnson syndrome," Computer Methods and Programs in Biomedicine,, October,2013.

[12] K. E. Emam and L. Arbuckle, *Anonymizing Health Data*. O'REILLY.

[13] H. X. P. H. M. Eren Demir, Thierry Chaussalet, "A method for determining an emergency readmission time window for better patient management," IEEE,Health and Social Care Modelling Group, 2006.

[14] I. Saxon, "intrade prediction market accuracy and efficiency: An analysis of the 2004 and 2008 democratic presidential nomination contests," *University of Nottingham. Dissertation*, 2010.

[15] A. Töscher, M. Jahrer, and R. M. Bell, "The bigchaos solution to the netflix grand prize," *Netflix prize documentation*, 2009.

[16] "Heritage provider network health prize description," 2012.

[17] "Heritage provider network health prize round 2 milestone leaderboard," 2012.

[18] F. Galton, "Vox populi," *Nature*, vol. 75, pp. 450–451, 1907.

[19] A. Khemphila and V. Boonjing, "Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients," pp. 193–198, 2010.

[20] I. Kurt, M. Ture, and A. T. Kurum, "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease," *Expert Systems with Applications*, vol. 34, no. 1, pp. 366–374, 2008.

[21] Y. Xing, J. Wang, Z. Zhao, and Y. Gao, "Combination data mining methods with new medical data to predicting outcome of coronary heart disease," pp. 868–872, 2007.

[22] B. Deekshatulu, P. Chandra, *et al.*, "Classification of heart disease using k-nearest neighbor and genetic algorithm," *Procedia Technology*, vol. 10, pp. 85–94, 2013.

[23] J. M. D. J. E. Dayhoff., "Artificial neural networks: Opening the black box. cancer," vol. 91, pp. 1615–1635, April 2001.

[24] N. Cheung, "Machine learning techniques for medical analysis," *School of Information Technology and Electrical Engineering, BsC thesis, University of Queenland*, vol. 19, 2001.

[25] S. K. Biswas, N. Sinha, B. Purakayastha, and L. Marbaniang, "Hybrid expert system using case based reasoning and neural network for classification," *Biologically Inspired Cognitive Architectures*, vol. 9, pp. 57–70, 2014.

[26] F. R. M. H. Rashedur M. Rahman, "Using and comparing different decision tree classification techniques for mining icddr,b hospital surveillance data," Science Direct, Expert System with Application, 2011.