Predicting the length of stay in a hospital for a particular disease using machine learning algorithms

by

Saif Imam
Student ID- 17373001

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
Master of Engineering in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
October 2021

Declaration

It is hereby declared that

1. The thesis submitted is my own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. I have acknowledged all main sources of help.

Student's Full Name & Signature:

Saif Imam
Student ID-17373001

Approval

The project titled "Predicting the length of stay in a hospital for a particular disease to help improve seat management in a hospital" submitted by

Saif Imam (ID-17373001)

of Summer, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Master of Engineering in Computer Science and Engineering on 07/10/21

Examining Committee:

Supervisor:
(Member)

Amitabha Chakrabarty, PhD
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Every single day a large number of patients go to the hospital. But the fact is, facilities available in hospitals are not sufficient in comparison with the number of patients. The idea of this paper is to offer a prediction system that will be able to say how many days a patient may stay in a hospital. So that the hospital authority may be able to make a better plan to support a large number of patients. In this project, the focus is to give a statistical overview of the recovery time of different diseases in Bangladesh and provide a predictive knowledge based on machine learning algorithms about the possible treatment duration for those diseases. We are hopeful that this prediction system will be a great thing for any hospital authority. They will be able to know the estimated staying duration of a patient in hospital and based on that they will prepare plans to provide support to a larger number of patients. In traditional computing, implementing a system like this is quite impossible because the data here does not follow any algorithmic pattern and that's the reason behind introducing machine learning for this particular task. We tried to accumulate all possible treatments and records of patients and run machine learning algorithms like Linear Regression, Boosted Decision Tree, and Bayesian Regression. We compared the accuracy, mean absolute error and root mean squared error for the results we generated from ML Studio using Linear Regression (LR), Bayesian Regression (BR) and Boosted Decision Tree (BDT) and the results are as follows:
Accuracy: LR=0.79, BDT=0.72, BR=0.71
Mean absolute error: LR=0.21, BDT=0.24, BR=0.27
Root mean squared error: LR=0.32, BDT=0.36, BR=0.37

Keywords: LOS: Length of Stay (in a hospital),
LR: Linear Regression,
BR: Bayesian Regression,
BDT: Boosted Decision Tree

Dedication

I would like to dedicate this work to my late grandmother who has been a constant source of motivation for me and my parents who are always there for me as a guiding light, My uncle for his continuous support, and my best friend for motivating me to study further.

Acknowledgement

Firstly many thanks to my respected supervisor Dr. Amitabha Chakrabarty Sir for his inspiration and guidelines. Even I didn't know if I could do such a good job but my supervisor believed in me, pushed hard for every detail, and with his clear vision and perfect instruction, I was able to complete this project. Thanks to Dr. Shirin who helped to collect the data from Bangabandhu Sheikh Mujib Medical Hospital (BSMMU) and the director of BSMMU hospital for allowing to collect data from different wards.

Table of Contents

List of Figures

List of Tables

Nomenclature

BDT Boosted Decision Tree

BR Bayesian Regression

LOS Length of stay

LR Linear Regression

Chapter 1

Introduction

1.1 Introduction

Every single day a large number of patients go to the hospital. But the fact is, facilities available in hospitals are not sufficient in comparison with the number of patients. The idea of this paper is to offer a prediction system based on machine learning techniques that will be able to say how many days a patient may need to stay in a hospital. So that the hospital authority may be able to make a better plan to support a large number of patients.

1.2 Problem Statement

In this project, the focus is to give a statistical overview of the recovery time of different diseases in Bangladesh and provide a predictive knowledge based on machine learning about the possible treatment duration for those diseases. There were lots of challenges collecting the data as we still lack advancements in recording data in computers or servers and keeping them in a hazardous way in hand written format in hard copies. In this project, the program will accumulate all possible treatments and records of patients and run different statistical analysis and compare the results. For the restriction of data, the experiment is on a small scale but in the future, it can be expanded by collecting data from different hospitals from different cities.

1.3 Aim of Study

We are hopeful that this prediction system will be beneficial to a great extend for any hospital authority. They will be able to know the estimated staying duration of a patient in hospital and based on that they will prepare plans to provide support to a larger number of patients. In traditional computing, implementing a system like this is quite impossible because the data here does not follow any algorithmic pattern and that is the reason behind introducing machine learning for this particular task.

1.4 Research Methodology

In this project, we tried to give a statistical overview of the patients in Bangladesh and provide a predictive knowledge about the possible length of stay they need in a hospital. We tried to accumulate all possible treatments and records of patients and run machine learning algorithms like Linear Regression, Boosted Decision Tree, and Bayesian Regression and compare the results. In all the algorithms applied we have got an accuracy of around 70-80 percent.

1.5 Thesis Outline

The main purpose of the project is to predict how long (in days) a patient needs to stay in a hospital for a specific disease which will help the hospital to allocate seats accordingly and also the patients get to know how long they need to stay in the hospital so that the can manage work leave or other household matters. The Project contains a large database consisting of hospital entries (date of admission and date of discharge) for a specific disease. This project will greatly help the hospital to allocate seats and also they can inform the patients how long they will be staying out even before getting admitted to the hospital. The project report focuses on these steps:

Firstly, the introduction part (Chapter 1) states the motivation behind the research which inspired the author to address this particular problem statement. The goals of this research and a summary of the work are briefly discussed here.

Secondly, literature review section (Chapter 2) we have discussed different papers from the field of computer science that have addressed the similar issues. In addition to that, the purpose of the background study was to find out the shortcomings of previous researches.

In the data collection phase (Chapter 3) the data collection and conversion process to use in this project have been discussed. This portion also includes feature extraction and a detailed study on each feature and how they contribute to the outcome.

In the Methodology and Model Selection (Chapter 4) We discuss the data types and which approach is better for different sections of patients data and select appropriate algorithm.

Furthermore, Algorithm and Result (Chapter 5) included the proposed models and comparative study of prediction rate among respective models. Also, outcomes are summarized with a demonstration.

The disease database phase has been moved to Appendix A. as it is not part of the primary purpose, we learned more about the diseases and treatments and recovery time. The theoretical recovery time has been noted to compare with predicted outcomes to justify the validity of the program.

Chapter 2

Related Work

In [6] a model was presented for the effects of multi-drug syntheses for anti-cancer medicines and antibiotics. The approach is rested on moderately few measurements on medication pairs. The pairs are approached using regression to extrapolate from data, warranting reprise rounds of forecasting to home in on results. Experimentally measured the dosage reaction matrix for eight dosages of three ordinarily applied chemotherapy medicines-doxorubicin, taxol and cisplatin on survival of A549 natural lung cancer cells. Ways used to measure data are Regression Model, Estimation of Model Parameter, Reckoning of Effective dosages in the Model and cell Lines and culture Conditions. The concluding model precisely describes data offered now on dosage – response of compounds of three cancer medicines and the parameters describe how the effective dosage of a cure is affected by the company of the other medication.

In [2], it uses enrichment score for grading symptoms and additionally the classifier is conditioned with that data to distinguish various kinds of cancers. The classifier employed in this article is J48 tree with the class label generating pruned or unpruned C4.5 decision tree for categorizing the cancer classes. With the target to minimize the tumor size under a set of constraints, the correlation feature selection algorithm and a local search algorithm called Iterative dynamic programming (IDP) were combined to form a new Memetic algorithm (MA-IDP) to answer the problem. Pathways used for classifications are Classification of the Cancer classes employing Decision Tree in this case J48 tree and Memtic Algorithm for choosing the tumor cells using local search and IDP constructing MA-IDP which has been applied successfully to break the multi-drug scheduling optimization challenge.

In another paper [4], a comparison between three distinct machine learning methods in a class setting where learning and prediction follow an instruction schedule to mimic the medication discovery procedure has been showed. The methods are standard SVM classification, SVM-based multi kernel classification and SVM classification based on learning using privileged facts to derive developmental in-vitro data and compound structure descriptors. Strategies used to approach the challenges here were Multi kernel learning (MKL), one approach to integrate different types of data in a kernel learning setting. MKL has shown to be equal to combine kernels from 4 distinct kinds of chemical structure descriptors, outperforming the classification performance of applying a single descriptor at a moment. Also, examination in learning using privileged information (LUPI) which in recent times has been offered as a original machine learning paradigm. Thus, because the PI is just applied during training and not during forecasting, LUPI is of particular interest to this environment. Using a dataset with 5639 compliances each accompanied with the date of when the trials were performed and ultimately employing the median of the initial training cases of hmc as a threshold they balanced dataset of 2491 versus in 3148 observances across-the-board.

Determining how long a patient will stay during a medical center is vital in healthcare to supply a better look after the patient and thus to extend the reputation of the hospital. This paper [5] focuses on the length of stay forecasting predicated on distinct approaches so that the medical center can assume when their bed will be obtainable for new patients and the hospital can deliver better service. The duration of staying is low in a community predicated society than in a hospital predicated system and the danger of re-hospitalization is self-reliant of the characteristics of the healthiness system.

Hospitals learned to optimize their healthcare planning and association to attenuate charges. The index that is constantly used to scale the hospital effectiveness is that the average length of stay. Numerous studies [8] display a strong and striking correlation between the costs of patients and the Length of Stay (LOS). The article proposes to employ data processing approaches to forecast the LOS. An evidential variant of knowledge mining called also evidential data processing are custom to reduce the impact of uncertainty and skipping data.

A paper [1] uses DBSCAN clustering to create the training set for classification. The forecasting models are compared employing exactness, perfection, recall and located that applying DBSCAN as a precursor to classification gives better conclusions. It gave us a conception on how a set to data (well-defined) can assist us fabricating an applicable prediction system.

Data mining methods are universally referred within the sector of healthcare. At an alternate occasion, it's recognized that the maturity of healthcare datasets are stuffed with skipping values. In this paper [3] they apply decision trees, Naive Bayesian classifiers, and feature selection approaches to a geriatric hospital dataset to forecast patient's extent of stay, specifically for lengthy-stay cases. Cases with psychological issues require to stay in hospital for a fairly long time in comparison with the cases with physical issues. Their readmission rate is likewise high.

This paper [6] describes an audit during which 273 surgical cases at St. Thomas' clinic, London, were studied to decide the limit to which their discharge might be predicted. Initial length of stay estimates were registered for 57 percent of the cases; the excess of concrete over the predicted length of stay possessed a mean of 2.71 and a typical deviation of 10.2 days. The length of stay of the other 43 percent had greater variability. The chances of readmission is low if a patient receives a lengthy and correct treatment at the first occasion.

In another paper, [7], Five methodologies for forecasting Length of stay in hospitals were developed and compared. Then a subjective Bayesian forecaster and a regression forecaster correspondingly measured the approximate weight of the characteristic and demographic factors in reading length of stay. The reading of the procedures was valuated with several benchmarks of conclusiveness and one among cost. The number of pattern and association rule is substantial due to the clustering of the attributes and the order.

Chapter 3

Data Collection and Feature Extraction

3.1 Data Collection

The data is collected from the different wards (sections) of Bangabandhu Sheikh Mujib Medical Hospital (BSMMU). The data is in hard form which is a handwritten register book. The data is photographed as I didn't have permission to take the register book outside the hospital to scan or make photocopies.

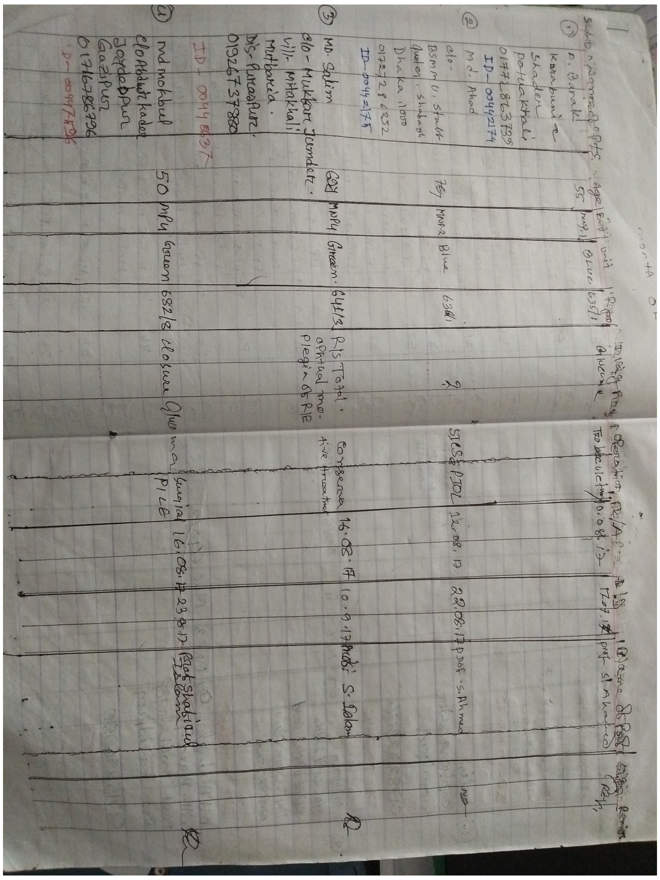

Figure 3.1: Hospital Entry Book

3.2 Data Conversion

The Data is manually entered into Excel worksheets from the picture of the register entries. Later the Excel data is converted into CSV file to use it as an input on Microsoft Azure Machine Learning Studio and perform various analysis and predict the final outcome which is the length of stay of a patient.

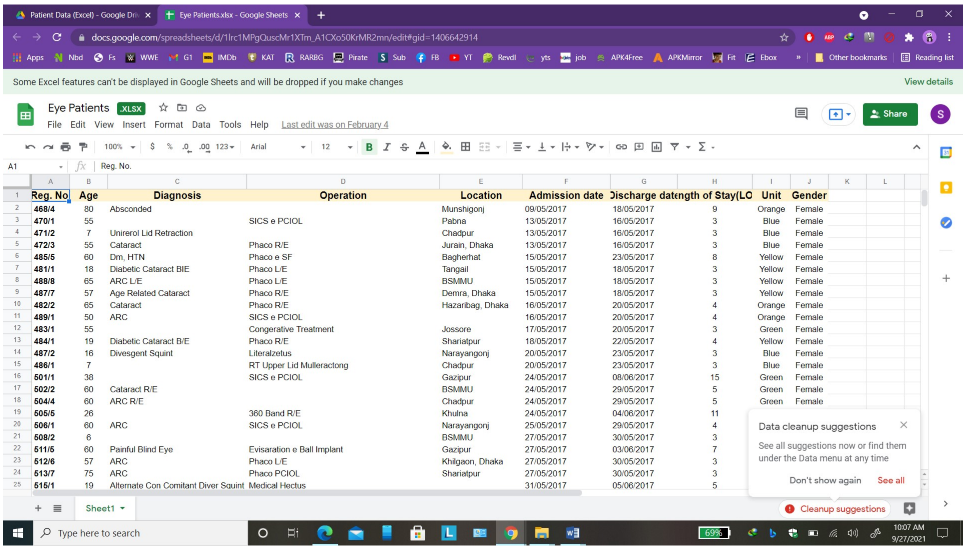

Figure 3.2: Entry in Excel

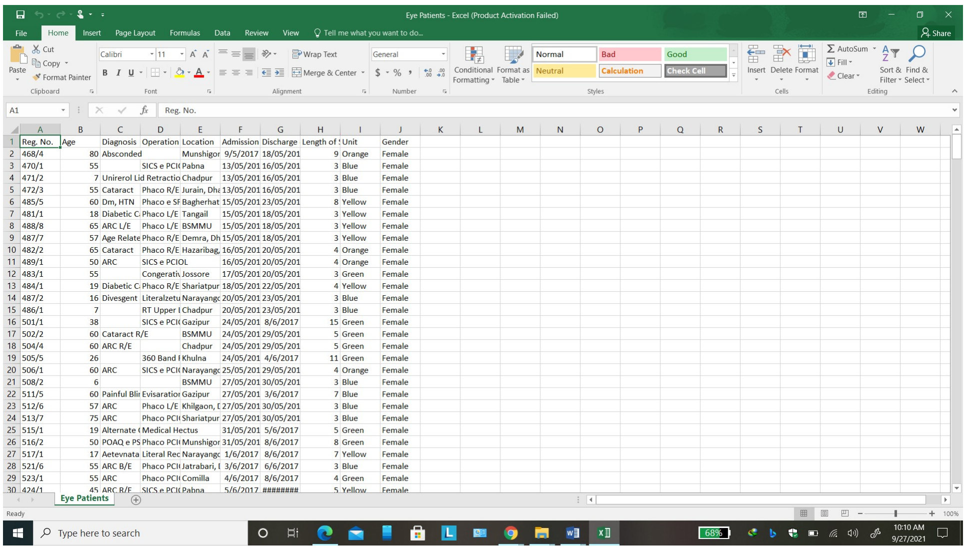

Figure 3.3: Excel to CSV

3.3 Analyzing Data

The patient data is categorized into different groups based on age, gender, types of diseases etc. These are the basic properties of any patient data. Important features like age group, gender ratio and other conditions (blood pressure, heart conditions, diabetes) are analyzed further to have a better understanding of each features and their relation with respect to LOS. Exceptional cases are omitted from the general classes to predict better for a large number of people more accurately.

3.4 Feature Extraction

3.4.1 MRMR feature extraction

Since our data had a significant quantity of attributes, it was imperative to reduce numerous dimensions of data space. Our data initially withheld dimension (N * F) where N = x1,x2,.. xn. The number of samples in our observance F denotes to f1,f2,.. fn. the number of features esteemed for partition of target variable. Though there were multiple ways accessible for identifying selection, we employed Minimum Redundancy Maximum Relevance algorithm for tagging significant features based on familiar facts score to reduce declassification violations. In our data set, all of the features does not have equal impact on target variable. Consequently the ambition of employing this algorithm was to select out the features fi which had agreeable familiar facts with target variable ci cognate that the operating features were elected on common ground facts score M (fi , ci). The equation which was applied to acquire common facts between characteristic fi and class variable ci is formulated below:

$$M(f,c) = P(fi,ci) * log(P(fi,ci)P(fi) * P(ci)) \tag{3.1}$$

Now P denoted computed likelihood estimate of corresponding features. The algorithm performed in a descending list T of top k features where target variable ci was astronomically dependent on the features f1,f2,.. fk belonging to T. Initially, we reaped a huge list of characterizing features having significant impact on target variable. Though the set of features which possessed utmost relevancy could have accelerated preciseness of classification, they tend to have redundancy between them. The set of features was acquired by computing the average of the common facts gathered from the presented equation for N samples. Still, the shared likelihood of the features affecting the target variable would not reduce significantly if we could minimize redundancy by electing mutually exclusive set of features. Therefore, it has been tested to figure out a trade-off between the highest relevancy and the minimal redundancy to obtain an optimal set of features which can be stated as the max value of the following function.

$$mRMR(F) = max(D) - min(R) \tag{3.2}$$

The function has held F features as parameters and returned a set of ordered features T where the primary specific had most dynamic effect on label variable. It could be said that for each specific fi belonging to T, they have high correlation with the class variable. Either, they recreate lower correlation between themselves at an equal time conforming to the affecting arrangement of features.

3.5 List of Features

Features	Affect on LOS
Age	high
Gender	high
History	high
Religion	low
Food Diet	low
Location (Urban or Rural)	mid
Profession	high
Martial Status	low
Section (Ward)	low
Social Status (Low,middle,high class)	mid

Table 3.1: List of Features

3.5.1 Age

Age is the most important feature while determining the length of stay. It affects the outcome greatly. As a person's immune system is related to ones age, thus this feature contributes to the length of stay directly.

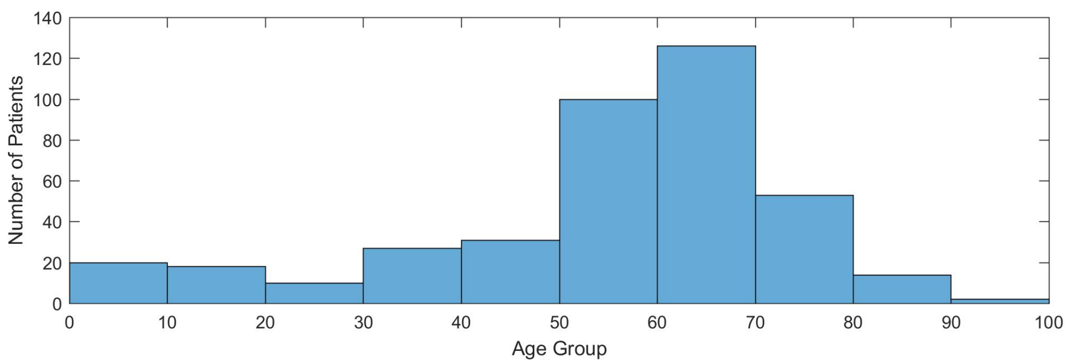

Figure 3.4: Age (Histogram)

Here in Figure-3.4 on the horizontal axis age is shown and on the vertical axis numbers of patients are shown. It can be said that people from 50 years to 80 years old, are suffering from the problem most.

3.5.2 Gender

Gender is the second most important factor in predicting the length of stay. As we can see females are less affected or may be they are not getting proper treatment or check-up. A separate research is needed to comment on this.

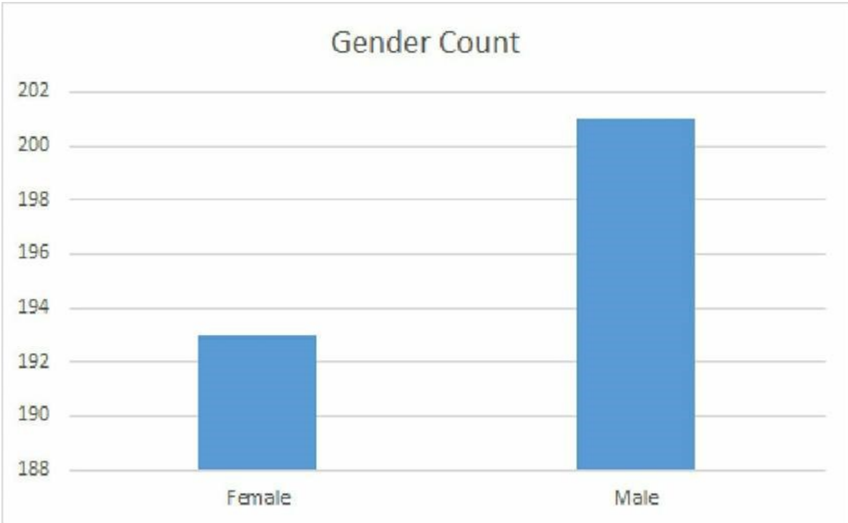

Figure 3.5: Male Female Ratio

From Figure-3.5 we can observe that the number of female patients are half the number of the male patients.

3.5.3 Ward Units

Patients are assigned into different wards based on their gender and the severity of their disease. Doctors can be assigned based on the severity (ie. patients with heart condition or other critical case) of the patients.

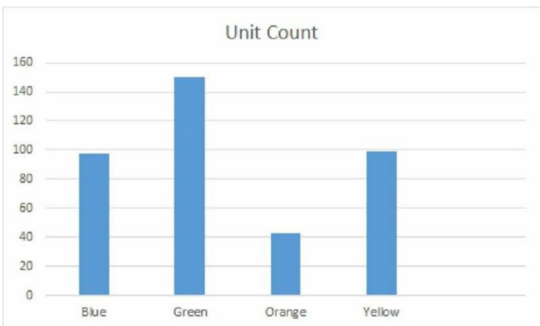

Figure 3.6: Patient Count (Ward Basis)

3.6 Length of Stay (LOS)

In most cases, it is seen that people are staying in the hospital between 0 to 5 days for an eye problem. The following graph shows the same in Figure 3.7

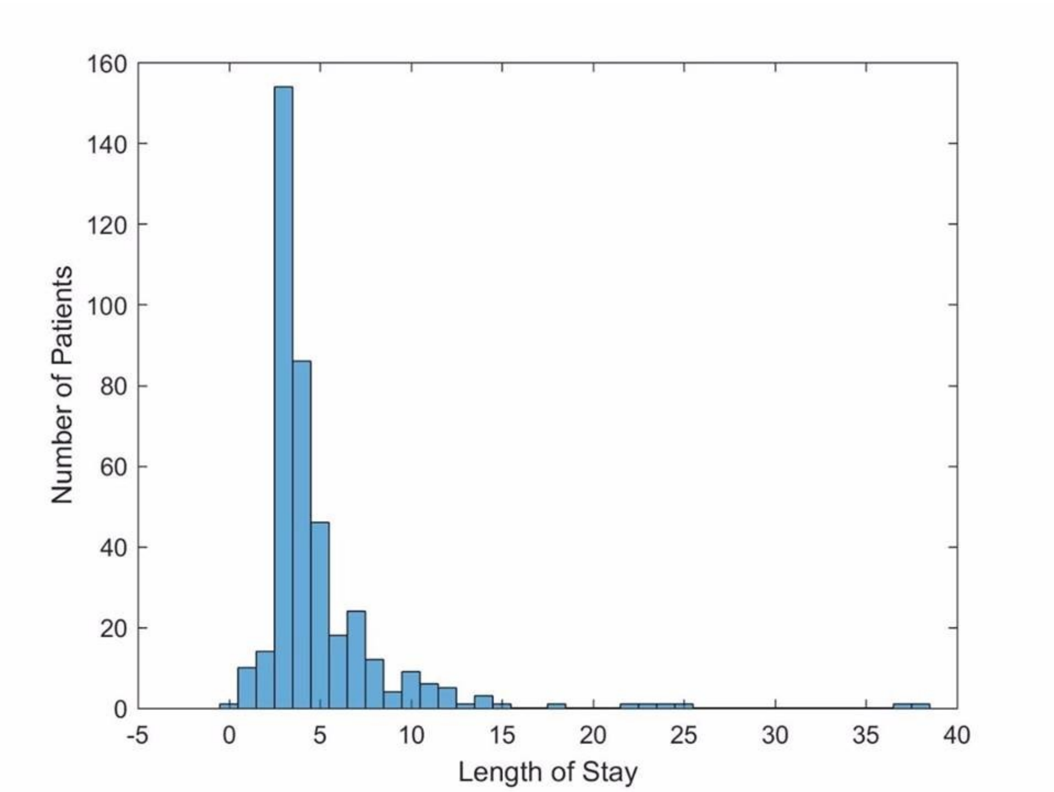

Figure 3.7: Length of Stay (LOS)

Chapter 4

Methodology and Model Selection

As there are data from data sections of a hospital and these diseases doesn't follow any specific patterns, thus selecting the right approach is very important. Although, similar diseases follow some common patterns but different disease especially heart related conditions follows a different path. General diseases also follow their own pattern but there are so many disease that general section of the hospital treats and to determine the length of stay for each disease are very troublesome and of course we need a huge amount of data for this distinct analysis.

In the project there are different kinds of patient data and we will use different approaches for these variant data and determine the best approach for each section. Machine learning has made revolutionary progress in recent times and we can implement various machine learning algorithms to different sets of data but the goal for each methods is to find the length of stay for a disease.

4.1 Machine Learning

The term Machine Learning was first assumed by Arthur Samuel in 1959. Machine Learning is the methodical education of algorithms and statistical models. A computer uses Machine Learning in order to execute a task successfully without applying express commands, preferably hold advantage of patterns and interpretation. It is accounted as a part of Artificial Intelligence. Conforming to author Tom Mitchell Machine Learning is outlined as, " A computer program is affirmed to discover from experience E with reference to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E". Machine learning algorithms are getting used during a broad selection of exercises. It is employed in every region of Artificial Intelligent resembling as Image Processing, Computer Vision. Generally, Machine Learning is preferable where it is infeasible to construct an algorithm of unequivocal instructions for accomplishing the task. Similarly, Machine Learning is stated to computational statistics in some ways, which largely employed for creating predictions applying computers. In our research we have integrated a machine learning method to forecast the length of stay of a case.

4.2 Supervised Learning

Machine Learning is categorized as different types. For example, supervised learning, unsupervised learning, reinforcement learning etc. For our study we have used supervised learning as our problem is related to both input and output. A mathematical model is made consisting both input and wanted output in supervised learning. The data is recognized as training data which has multiple training samples. One training sample contains one or more input and a desired output. Each training data may be a matrix and every training example is an array or vector within the mathematical model of the supervised learning. An iterative optimization of a function helps to seek out the way to learn the task so as to predict the output. Finally, an input from outside which wasn't a part of the training data is used to be predicted by the system. This is how supervised learning find out the prediction of a problem. However, unsupervised learning algorithms works with data which only consider inputs and discover different types of structures in data. Again, task of reinforcement learning is to how software agents are getting to take measures during a situation so as to take full advantage of some concept of cumulative return.

4.3 Naïve Bayes Algorithm

Naïve Bayes is one of the fastest algorithm according to the computational cost and a widely used one as well in prediction problems. In one inquisition, authors have examined the modeling of battery descent under distinct application provisions and ambient temperatures and eventually read the online state of health (SoH) estimation and abiding applicable life of lithium ion batteries. They have employed Naïve Bayes algorithm for arithmetic and also compare the result with the result of Support Vector Machine outcome. But in our study we've decided to use Naïve Bayes for heart patients data as it is a fast approach to work with. We've particularly used Naïve Bayes approach on heart patient's data and discuss the result of it. We all understand the Bayes Theorem as follows:

$$P(A|B) = (P(B|A) * P(A))/P(B) \tag{4.1}$$

currently, A is class, B is Data. In our example, we have one observation to predict and two possible classes which are 1. LOS less than a week, 2. LOS more than a week. As a result, we will determine two posteriors: one for less than a week and one for more than a week.

P(LOS less than a week — person's data) = [P(person's data — LOS less than a week)*P(LOS less than a week)]/ P(person's data)

P(LOS more than a week — person's data) = [P(person's data — LOS more than a week)*P(LOS more than a week)]/ P(person's data)

For the heart patience the data were not conclusive and all the attributes were not defined and most of the entries were missing some values. So instead of finding the exact value of LOS we find whether LOS is more or less than a week. Later for eye patients with enough relevant data we will predict the exact length of stay with great accuracy.

4.3.1 Na¨ıve Bayes Algorithm on Heart Patients

We can use the Naïve Bayes Classifier theorem to take a decision of LOS based on dataset with 50 tuples.

Class:
C1: StayInHospital more than a Week= 'yes'
C2: StayInHospital more than a Week= 'no'

Data to be classified:
X = (age = 55, Sex=Male, Religion=Islam, Marital Status=Married, Occupation=Service Holder, Social Status=Middle, Place to Live=City) Social Status = poor, middle, rich. Place to live = city, village.

Now from the rules and from the dataset we can define
P (Ci) = P (StayInHospital more than a Week= 'Yes') = 15/50 = 0.5
P (StayInHospital more than a Week= 'No') = 35/50 = 1.17

Now we have to compute P (X—Ci) for each class then the next step depending on
P(Age Group=AG-3 (40 +) — StayInHospital more than a Week= 'Yes') 10/15= 0.67
P(Age Group=AG-3 (40 +) — StayInHospital more than a Week= 'No) 3/35= 0.09
P(Sex=Male — StayInHospital more than a Week= 'Yes') 12/15= 0.80
P(Sex=Male — StayInHospital more than a Week= 'No) 23/35= 0.66
P(Religion=Islam — StayInHospital more than a Week= 'Yes') 11/15= 0.73
P(Religion=Islam — StayInHospital more than a Week= 'No) 31/35= 0.89
P(Marital Status=Married — StayInHospital more than a Week= 'Yes') 12/15= 0.80
P(Marital Status=Married — StayInHospital more than a Week= 'No) 13/35= 0.37
P(Occupation=Service Holder — StayInHospital more than a Week= 'Yes') 3/15= 0.20
P(Occupation=Service Holder — StayInHospital more than a Week= 'No) 3/35= 0.09
P(Social Status=Middle — StayInHospital more than a Week= 'Yes') 7/15= 0.40
P(Social Status=Middle — StayInHospital more than a Week= 'No) 18/35= 0.51
P(Place to Live=City — StayInHospital more than a Week= 'Yes') 7/15= 0.40
P(Place to Live=City — StayInHospital more than a Week= 'No) 19/35= 0.54
P(X—Ci) * P(Ci):
P(X — StayInHospital more than a Week= 'Yes') * P(StayInHospital more than a Week= 'Yes') =0.0500
P(X — StayInHospital more than a Week= 'No') * P(StayInHospital more than a

Week= ‘No’)=0.0006

Therefor X belongs to class (StayInHospital more than a Week= ‘Yes’)

By applying the Naïve Bayes Classification on sample data, we can predict the LOS of new patient accurately.

4.4 Version Space on General Patients

Now we have to figure out the specific hypothesis and general hypothesis. These two hypothesis will lead us to create the final version space. So based on the attributes we have from General Patients data, the specific hypothesizes we have derived is as follows:

S0: <ø,ø,ø,ø,ø,ø >
S1: <Male, 50, GOO, Outsider, Yes, 8 >
S2: <Male, 50, ?, Outsider, Yes, 8 >
S3: <Male, 50, ?, Outsider, Yes, 8 >
S4: <Male, ?, ?, Outsider, ?, 8 >

Needless to mention that if we try to figure out the specific hypothesis of all our collected instances, it will take more iterations and the result will differ from this one. Now lets see the general hypothesizes:

G0: <?, ?, ?, ?, ?, ? >
G1: <?, ?, ?, ?, ?, ? >
G2: <?, ?, ?, ?, ?, ? >
G3: <Male, ?, ?, ?, ?, ? ><?, 50, ?, ?, ?, ? ><?, ?, ?, ?, ?, 8 >
G4: <Male, ?, ?, ?, ?, ? ><?, ?, ?, ?, ?, 8>

For the specific hypothesis, we have used symbol ø for the first iteration where no instance from the attribute has taken. In the second iteration as the result is “Yes” (for first instance) all the element of first instance got place in specific hypothesis. In third iteration “Disease” got “?” simply because it has no impact on output. Changing of this value can also produce a positive value. As the third instance of our table produced a negative result, no changes occurred in forth iteration. But at the fifth iteration, we again encountered some significant changes which changed the specific hypothesis from its previous state.

For the general hypothesis, we have used “?” for first three iterations because till the second instance of our table, we got positive results. So general hypothesis just told us that for any value we will get a positive result. But significant change came on our way while we started working on G3. In this stage, general hypothesis pointed out three major factor which must be maintained if we want to get a positive outcome. Though this scenario changed at the last iteration and we finally got a generalized hypothesis with two major constraints. Now if we just take the most specific hypothesis and most general hypothesis, we will find our version space at the very middle of them and that will ultimately lead us toward our desired state. The version space has been illustrated in the below figure"

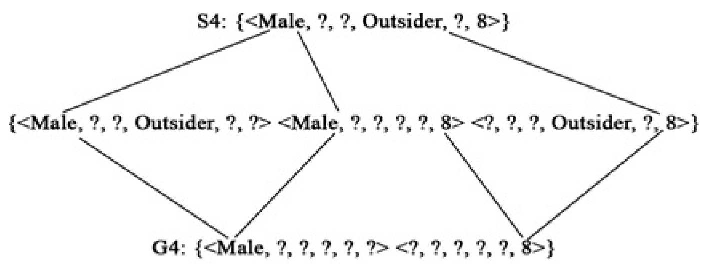

Figure 4.1: Version Space

4.5 Algorithm on Eye Patients

As the eye patients data were most consistent and complete and each eye disease has enough data to predict the length of stay for each specific disease.We have used three algorithms on the data which are Bayesian Regression, Boosted Decision Tree and Linear Regression. The Implementation of the algorithms and result comparison for Accuracy and Mean Squared Error were briefly discussed in the next chapter.

Chapter 5

Implementation and Results

5.1 Implementation of Machine Learning on Final Dataset

In Table-5.1 some statistical analysis is presented based on our collected data. We can observe that in Bangladesh a patient needs to stay a minimum if 1 day in the hospital.

Statistical Analysis	Results
Mean	2.9565
Median	3
Min	1
Max	5
Standard Deviation	0.6953

Table 5.1: General Patient's Length of Stay

In Figure-5.1 the statistical analysis of our collected data is presented. In this graph we can see that a patient will have to stay a minimum of 1 day and maximum of 5 days.

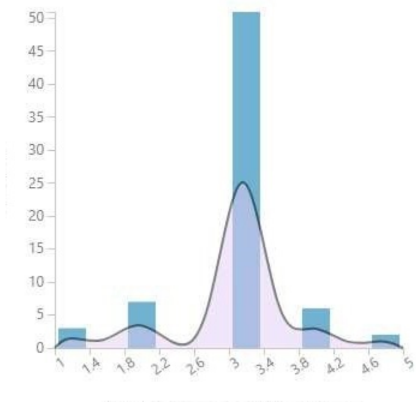

Figure 5.1: Length of Stay vs Frequency

5.2 Bayesian Regression Results

Project Setup in Azure Machine Learning Studio:

• We converted our data to a CSV file to feed it to the experiment.

• Later we have used the "Select Dataset in Columns" module to select the necessary columns. Unnecessary columns are ignored such as patient id etc.

• Then we used a "Train Model" module which used a dataset and training method as parameters. Our filtered dataset and "Bayesian Linear Regression" modules are passed on to the Train Model.

• The "Score Model" uses the trained data and initial raw data to score the training. We used the results of the "Train Model" and initial filtered data as the parameters.

• The results of the "Scored Model" are passed on to two modules. One is the "Convert to CSV" which converts the data to CSV for later use and to an "Evaluate Model" module which helped us visualize the results as well as the errors.

Figure-5.2 is a Screenshot from Machine Learning Studio from the experiment.

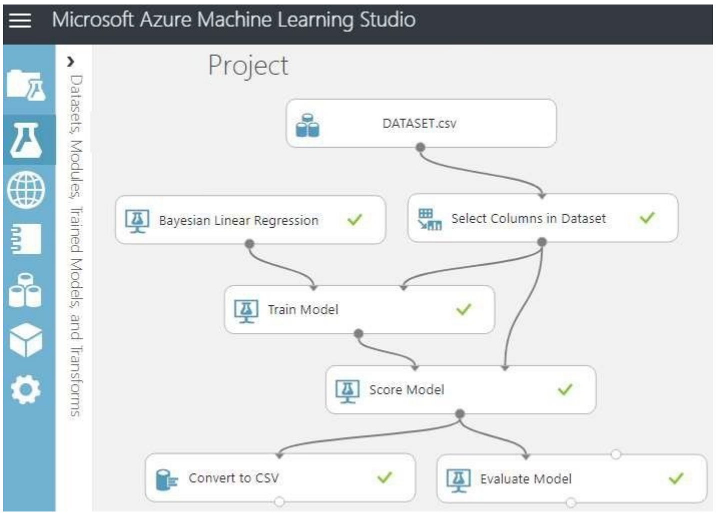

Figure 5.2: Bayesian Regression Model from the Azure Machine Learning Studio

After successfully running the Bayesian Regression on our dataset we received the following statistical data.

Statistical Analysis	Results
Mean	2.9483
Median	2.9513
Min	1.6325
Max	4.9037
Standard Deviation	0.5664

Table 5.2: Statistical Analysis of Bayesian Regression

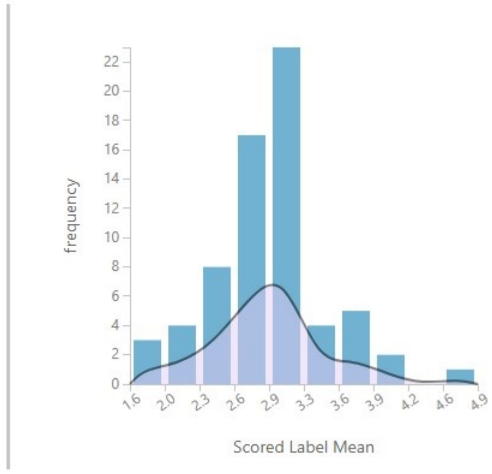

Figure 5.3: Length of stay vs Frequency after running Bayesian Regression

5.3 Evaluation of Result(Bayesian regression)

the following errors were found after running Bayesian regression. After a successful run of Linear Regression the following errors were found:

Statistical Analysis	Results
Mean Absolute Error	0.270297
Root Mean Squared Error	0.375658
Coefficient of Determination	0.712315

Table 5.3: Errors from Bayesian Regression

5.4 Boosted Decision Tree

Project Setup in Azure Machine Learning Studio:

• We converted our data to a CSV file to feed it to the experiment.

• Later we have used the “Select Dataset in Columns” module to select the necessary columns. Unnecessary columns are ignored such as patient id etc.

• Then we used a “Train Model” module which used a dataset and training method as parameters. Our filtered dataset and “Boosted Decision Tree” modules are passed on to the Train Model.

• The “Score Model” uses the trained data and initial raw data to score the training. We used the results of the “Train Model” and initial filtered data as the parameters.

• The results of the “Scored Model” are passed on to two modules. One is the “Convert to CSV” which converts the data to CSV for later use and to an “Evaluate Model” module which helped us visualize the results as well as the errors.

Figure-5.4 is a Screenshot from Machine Learning Studio from the experiment.

Figure 5.4: Decision Tree Model from the Azure Machine Learning Studio

After successfully running the Boosted Decision Tree on our dataset we received the following statistical data.

Statistical Analysis	Results
Mean	2.9552
Median	3.0047
Min	1.5437
Max	4.4986
Standard Deviation	0.4677

Table 5.4: Statistical Analysis of Boosted Tree

5.5 Evaluation of result(Boosted Decision Tree)

Figure-5.5 is the presentation Patient’s length of stay after running the Boosted Decision Tree.

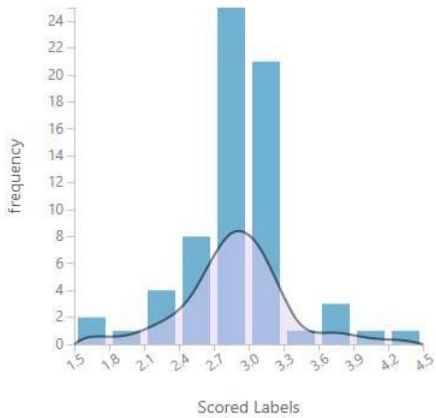

Figure 5.5: Length of stay vs Frequency after running Boosted Tree

We received the following errors after running boosted decision tree. After a successful run of Linear Regression in our Dataset we received the following errors.

Statistical Analysis	Results
Mean Absolute Error	0.246491
Root Mean Squared Error	0.369061
Coefficient of Determination	0.72233

Table 5.5: Errors from Boosted Decision Tree

5.6 Linear Regression

Project Setup in Azure Machine Learning Studio:

• We converted our data to a CSV file to feed it to the experiment.

• Later we have used the "Select Dataset in Columns" module to select the necessary columns. Unnecessary columns are ignored such as patient id etc.

• Then we used a "Train Model" module which used a dataset and training method as parameters. Our filtered dataset and "Linear Regression" modules are passed on to the Train Model.

• The "Score Model" uses the trained data and initial raw data to score the training. We used the results of the "Train Model" and initial filtered data as the parameters.

• The results of the "Scored Model" are passed on to two modules. One is the "Convert to CSV" which converts the data to CSV for later use and to an "Evaluate Model" module which helped us visualize the results as well as the errors.

Figure-5.6 is a Screenshot from Machine Learning Studio from the experiment.

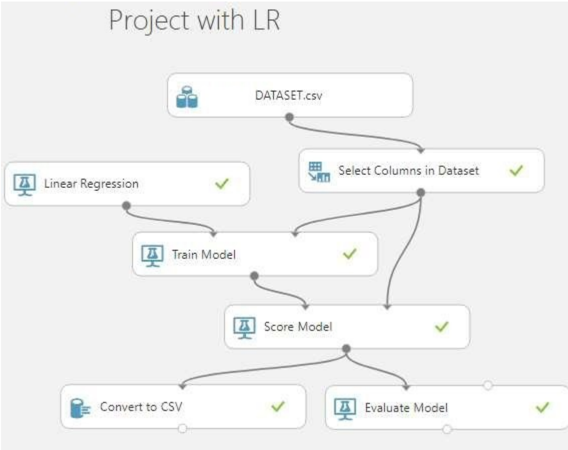

Figure 5.6: Linear Regression Model from the Azure Machine Learning Studio

5.7 Evaluation of result(Linear Regression)

After successfully running the Linear Regression on our dataset we received the following statistical data.

Statistical Analysis	Results
Mean	2.9552
Median	3
Min	1.0021
Max	4.9788
Standard Deviation	0.6227

Table 5.6: Statistical Analysis of Linear Regression

These data are visually presented in Figure-5.7

We received the following errors after running linear regression. After a successful run of Linear Regression in our Dataset we received the following errors.

Statistical Analysis	Results
Mean Absolute Error	0.217287
Root Mean Squared Error	0.323033
Coefficient of Determination	0.787271

Table 5.7: Errors from Linear Regression

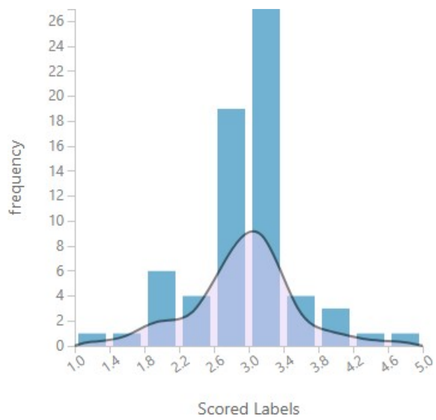

Figure 5.7: Length of stay vs Frequency after running Linear Regression

5.8 Result Comparison

We compared the accuracy, mean absolute error and root mean squared error for the results we generated from ML Studio using Linear Regression, Bayesian Regression and Boosted Decision Tree in the following figures 5.8, 5.9, 5.10

Figure 5.8: Accuracy Comparison

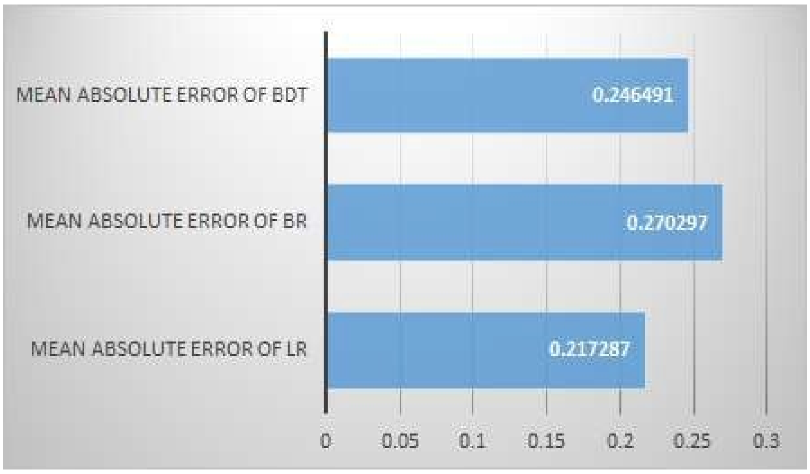

Figure 5.9: Comparison of Mean Absolute Error

Figure 5.10: Comparison of Root Mean Squared Error

Chapter 6

Conclusion and Future Work

In the era of modern computing, machine learning is spreading its branches in almost every possible sectors. It is making things easier for us. In the tasks like prediction or recognition, machine learning is a first choice for front line technologists and scientists. Through this paper we are proposing a prediction system which has conceptualized on the foundation of machine learning that will be able to predict how many days a patients may stay in hospital.

Predicting the length of stay (LOS) of patients during a hospital is vital in providing them with better services and better satisfaction, also as helping the hospital management plan and manage hospital resources as meticulously as possible.

The length of stay (LOS) varies for different people depending on the healing power of their bodies. Many people live in a village or unhealthy environment and take non-hygienic food. Some people are rickshaw pullers or work in factories. They need more time to recover because of their way of livings. Also, high aged people need more time to recover.

The project contains great potential and can be improved by using different machine learning algorithms. More data from different hospitals and a digital record book system can help this project. The program can be integrated into hospital servers to provide a real-time update. More features can be selected to classify the data like urban or rural location, hygiene and food nutrition in formations can provide more accurate predictions.

Bibliography

[1] J. C. Robinson, "Hospital competition and surgical length of stay," *JAMA: The Journal of the American Medical Association*, vol. 259, p. 696, Feb. 1988. DOI: 10.1001/jama.1988.03720050032019.

[2] L. Appleby, N. Desai, D. Luchins, R. Gibbons, and D. Hedeker, "Length of stay and recidivism in schizophrenia: A study of public psychiatric hospital patients," *Am J Psychiatry*, vol. 150, p. 1, 1993.

[3] P. R. Hachesu, M. Ahmadi, S. Alizadeh, and F. Sadoughi, "Use of data mining techniques to determine and predict length of stay of cardiac patients," *Healthcare Informatics Research*, vol. 19, p. 121, 2013. DOI: 10.4258/hir.2013.19.2.121.

[4] S. Mani, Y. Chen, X. Li, L. Arlinghaus, A. B. Chakravarthy, V. Abramson, S. R. Bhave, M. A. Levy, H. Xu, and T. E. Yankeelov, "Machine learning for predicting the response of breast cancer to neoadjuvant chemotherapy," *Journal of the American Medical Informatics Association*, vol. 20, pp. 688–695, Jul. 2013. DOI: 10.1136/amiajnl-2012-001332.

[5] N. Radhika, "A novel approach for predicting the length of hospital stay with dbscan and supervised classification algorithms," *Journal of Intelligent Computing*, vol. 5, 2014.

[6] M. Elbattah and O. Molloy, "Using machine learning to predict length of stay and discharge destination for hip-fracture patients," *Proceedings of SAI Intelligent Systems Conference (IntelliSys) 2016*, pp. 207–217, Aug. 2017. DOI: 10.1007/978-3-319-56994-9_15.

[7] B. Nandi, A. Kumkum, S. Nazrin, and E. Baral, *Prediction of number of days staying at hospital for fractured people.*

[8] A. Perez, W. Chan, and R. Dennis, *Predicting the length of stay of patients admitted for intensive care using a first step analysis predicting the length of stay of patients admitted for intensive care using a first step analysis.*

Appendix A.

A disease database is created based on medical science references. The information will also be visible upon running the program for a specific disease.

6.1 Anomalous Retinal Correspondence (ARC)

Arc eye or welder's flash is a state where the outer layer of the attention is broken due to ultraviolet radiation. The name came from one of the foremost common causes of corneal burns which is exposure to the discharges that created during welding metals. This can produce pain and blurred vision. The symptoms generally stay only a few days if the burn is mild. In case of severe injury, eye should be checked at ER to promote accurate diagnosis and proper treatment.

Welder and peoples who are within the vicinity of a welding project but do not use proper eye protection are at high risk of developing Arc. During strong welding, the UV radiation can burn the cornea within a fraction of seconds. Many other sources of UV and longtime exposure to sunlight my also cause arc. Generally, the symptoms arc does not appear immediately after the injury. In most cases the signs and symptoms are observed within three to six hours of injury. Among the symptoms, itchiness, bloodshot and excessive tear production is common. Other complications including blurry or spotty vision, temporary vision loss may develop if remain untreated. Long term exposure to UV may cause permanent vision loss.

In case of minor arc OTC ointment is prescribed. For further recovery oral anti-inflammatory medicines can be used. Vision generally returns to normal state within 3 days after injury. An ophthalmologist plays a crucial role in arc not only by providing treatment to current injury but also by making them aware related to eye injury.

6.2 Lid Retraction

The displacement of upper and lower eyelid is termed as lid retraction. Lower-eyelid retraction is an outcome of shallow orbits, hereditary conditions, or certain characteristics of the eyelid. Vision discomfort and corneal disturbance are most common complication of lid retraction.
Repair Lid retraction:

Repairing of the eye lid depends on the cause and severity of lid retraction. The doctor will stretch the loosened tissue that pulls the eyelid downwards and will place a spacer graft boost the eyelid upwards. When the space between two eyelid is reduced then the eye starts performing better.

6.3 Hypertensive retinopathy

The retina is a membrane which is located at the back of eye. The light when enters through the lens of the eye are converted into nerve signal which are sent to brain for interpretation. The retina's vessel walls may thicken due to high vital sign causing narrow blood vessels which ultimately reduce the blood flow to the retina. In some case retina become swollen. The high vital sign for a long time may be a cause of damage of retina's blood vessel, reducing retina's function and vision problem for patients. This situation is termed as hypertensive retinopathy.

6.4 Congenital Eye Diseases

Generally the congenital eye disease is originated from the pregnancy. It may also originate s from gene mutation, exposure to drugs or alcohol during pregnancy. Any part of the cornea may be affected by this disease. The symptoms of this disease are cloudiness of front surface of retina, increased eye pressure, irregular shaped pupil, opacities within the lens etc. Due to the severity of this disease significant vision loss may occur.

6.5 Excisional Biopsy and Incisional Biopsy

The procedure used to remove the whole tumor of a patient is termed as excisional biopsy. In case of incisional biopsy some portion of the tumor is removed. As a treatment of melanoma, excisional biopsy is preferred as in this procedure whole tumor along with some tissues around the tumor removed. Based on the size of the tumor, the number of surrounding tissues is considered. Skin grafting is used in case of possible melanoma. Rotation flap of skin is also used to cover the wound. After the surgery the eye may become red or swollen. It may take few weeks for full recovery.

6.6 Nuclear sclerosis

Nuclear sclerosis is very common disease particularly observed in aging people which is also the indication of old age. This disease is also observed in dogs, cats and horses. The main symptom of this disease is the cloudiness, hardening and yellowing of the central region of the lens of the eye, particularly the area of nucleus. The severity of this disease is termed as the nuclear cataract. The treatment of this disease is to remove the cloudy lens by surgery and replace lens by artificial lens.

6.7 Narrow-angle glaucoma (angle-closure glaucoma)

The drainage angle which helps the fluid to empty normally from the eye becomes obstructed and then narrow-angle glaucoma which is also one type of glaucoma happens. Due to this disease the drainage angle become narrower and the eye pressure increases and the patient losses eye vision. This situation is termed as acute glaucoma. The drainage angel may close gradually, occasionally or intermittently with some symptoms. Immediate medical action is required in case of acute glaucoma.

6.8 Meibomian gland dysfunction

The very common eye problem is Meibomian Gland Dysfunction (MGD) due to which inflammation and clogging of small glands in eyelids occur. These glands produce the oily layer of tears. Less amount of oil is secreted into the tear film as the meibomian glands are inflamed and clogged. As a result of this, the tears become evaporated vary fast resulted dry and irritated eye. The Tear Breakup Time (TBUT) test is a common test for this disease. A small amount of dye is administered to the tear film of the eye and examined with a cobalt blue light which indicates the loss of stability of tear film.

6.9 Dacryocystitis

The infection of lacrimal sac which is originated due to the blockage of tear duct is known as Dacrocytes. Tears drain into the small chamber- the lacrimal sac. The blockage of the duct originated from the lacrimal sac into the nose may be the cause of Dacrocytes.

6.10 Grading Posterior Subcapsular Cataract

By using modern technology cataract may be removed from the eyes though the development of cataract is still a challenge for the ophthalmologist. The treatment of cataract depends on the grading of cataract. The Posterior Subcapsular Cataract typically feather shaped. The pupillary margin is blurred due to the concentration of PSC and in this case only retro illumination opacity is concentrated and graded. In case of multiple PSCs, the leading visible opacities with distinct boarders may be considered.

6.11 Retinoblastoma

Retinoblastoma is one type of cancer originates within the retina of the eye. The young are the mostly affected patients where as the adults are less. The retina which is formed from the nervous tissue that senses light comes from the front of eyes. The brain receives the signals from the retina through the nervous opticus and creates image. The retinoblastoma which is very rare eye cancer affects the attention in children. This disease can affect any of the eyes or both.

6.12 Treating tumors

Small tumors within eye can be treated in two ways: laser treatment (photocoagulation or thermotherapy) and freezing the tumor (cryotherapy). General anesthesia is used to administer these treatments, as a result during this procedure the patient do not feel any pain or discomfort. Based on the necessity chemotherapy is used after of before the procedure.
Treatment of larger tumors:
For treating larger tumors, the following procedures are taken: Brachytherapy: In case of reasonable size, plaques- small radioactive plates are stitched over the tumor and left it for few days until the removal. In case of larger tumor where this procedure is not applicable radiotherapy is administered. As a treatment, chemotherapy is also administered if there is a chance of cancer spreading. In case of very large tumor eye surgery is done to remove the eye and chemotherapy is directly given to the eye. In this case there is no vision from the eye and the eye is replaced by artificial eye.

6.13 Pathophysiology of Preseptal and Orbital Cellulitis

Orbital Cellulitis creates from a large nearby focus of fulminant infection which is separated by a thin bone. This infection may be extensive and severe. Subperiosteal abscesses can accumulate subperiosteal fluid which are sterile initially.

6.14 Preseptal and orbital cellulitis

Vision loss (3-11 percent) is observed with the complication of orbital cellulitis. Increased Intra orbital pressure are the cause of retinopathy and neuropathy. Also, cerebral abscess, meningitis, nusus thrombosis, intracranial sequelae from a central spread of infection are caused by restricted ocular movement(ophthalmoplegia).

6.15 Rhegmatogenous retinal detachment (RRD)

Retinal rapture, vitreous liquefaction and the traction are the three factors on which the pathogenesis of rhegmatogenous detachment of the retina depend. Among these retinal rapture are the most important factor due to which retina break away the pigment epithelium. Patient's visual function is severely damaged by retinopathy which is developed by definite segment between the neurosensory retina and the retinal pigment epithelium. Selection of appropriate treatment by early detection of the disease as well as prevention is important.

6.16 Esotropia

Eso means inward where as trope means turn which is the disease Esotropia due to which one or both eyes turned inward. According to the College of Optometrist in Vision Development 1 to 2 percent of people suffers this disease. Both the infants and adult may suffer this disease. This may affect in several ways. Infants with infantile esotropia are unable to use both eyes together. The infants become expose to the risk of amblyopia or Lazy eye when one of the eyes turn inward more often than the other one. The treatment for this disease is surgery, eye glass or Botox injection is used sometimes. The correction of esotropia is very successful if it takes place during the very early time of infancy better before age 2 years, then it may happen that few infants may suffer other eye problem during their old age.
Recovery Period: In case with the mild symptoms the infants at the age of 5 months old the Esotropia may resolve with no interventions and the misalignment of eyes become intermittent.