Further Medicine Promotion Prediction: A Machine Learning Approach

Mousumi, Faria Momtaz 12-22289-3

Islam, Nafisa 14-25566-1

Luthfullahel, Mazed 98-0738-02

A thesis submitted in partial fulfilment of the requirement for the degree of Bachelor of Science in Computer Science and Engineering

American International University - Bangladesh

Faculty of Science and Information Technology

Department of Computer Science

Supervisor:
Dr. Khandaker Tabin Hasan
American International University – Bangladesh

May 2017

Further Medicine Promotion Prediction: A Machine Learning Approach

Mousumi, Faria Momtaz 12-22289-3

Islam, Nafisa 14-25566-1

Luthfullahel, Mazed 98-0738-02

A thesis submitted in partial fulfilment of the requirement for the degree of Bachelor of Science in Computer Science and Engineering

American International University - Bangladesh

Faculty of Science and Information Technology

Department of Computer Science

Supervisor:
Dr. Khandaker Tabin Hasan
American International University – Bangladesh

May 2017

Declaration

This is to clarify that this thesis is our original work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledge in the text and a list of references is given.

Mousumi, Faria Momtaz 12-22289-3

Islam, Nafisa 14-25566-1

Luthfullahel, Mazed 98-0738-02

Approval

The thesis titled “Further Medicine Promotion Prediction : A Machine Learning Approach” has been submitted to the following respected members of the board of examiners of the department of computer science by Mousumi , Faria Momtaz(12-22289-3), Islam,Nafisa (14-25566-1) and Luthfullahel, Mazed (98-0738-02) has been accepted as satisfactory.

…….

Dr. Tabin Hasan
Supervisor
Associate Professor and Head
(Graduate Program and MIS)
Department of Computer Science
Faculty of Science & Information Technology
American International University-Bangladesh

…….

Afsah Sharmin
External Examiner
Assistant Professor
Department of Computer Science
Faculty of Science & Information Technology
American International University-Bangladesh

…….

Professor Dr. Tafazzal Hossain
Dean
Faculty of Science & Information Technology
American International University-Bangladesh

…….

Dr. Carmen Z. Lamagna
Vice Chancellor
American International University-Bangladesh

Acknowledgements

First of all, we are grateful to our supervisor, honorable Dr. Tabin Hasan for providing her proper guideline with his knowledge and patience through the whole process. The instructions shared by our supervisor has been acted as a pioneer to create new ideas through brainstorm to find a solution to our problem. Our supervisor has kept track to every update of our work and supported us to modify our work. With his support, we accomplished our goal, thus successfully finished our thesis.

We would like to show our gratitude to respected parents, honorable teachers and fellow classmates who supported and shared knowledge and wisdom which contributed to our thesis and helped us to complete our thesis.

Finally, we would like to thank Almighty help us by blessing us with this opportunity and help us and lessen our difficulty and because of that it was possible to successfully complete our thesis work.

Abstract

Nowadays data analysis in the various sector is opening new possibilities -two of them being medical science and computer science. These two fields have been collaborating and giving tremendous results to help mankind, which is after all the main objective. Through Data mining technique, data analysis has created extraordinary improvement in medical science where it has simplified working with a large amount of data. In Bangladesh, among many pharmaceuticals, Ziska has provided large amounts of data and for data analysis. In this dataset doctors list are categorized area name, prescription code, district, by these data contribute to finding patterns which will help to create a prediction model. Medicines promotion is important and finding which doctor can be approached for promoting can be difficult and is a prolonged process. Our purpose is to predict a model to promote medicines that are useful and safe. Analyzing these data, doctors from various districts and with their diverse thinking of application along with their diagnosis pattern will be collected and their experience will be considered to find a pattern to create our model. We will collect the data of the doctors and there we will consider age, degree, their diagnosis style and furthermore criteria to predict suitable doctors to target. We are going to work with the data of Bangladesh but after our work can be helpful all over the world and contribute to medicine promotion.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

One of the most significant components of the modern age and advancing civilization is the practice of medicine. In this modern era, medical science is developing and with the help of collaborating with technology, this field is gradually reaching a new level. Via technology, medicals have been making breakthroughs in both cure and prevention. Thorough diagnosis and with required intellect, doctors now prescribe more efficiently and with the help of technology, doctors are saving their time and patient's life. The motto of this research is to create a prediction model where newly produced medicines will be promoted to doctors who are experienced enough for promoting the medicines. Medicines are sensitive and usage of medicines need proper guidance and supervision. And of course, that is the most vital part of manufacturing medicine. If a new medicine is produced, pharmaceutical companies should not promote it to doctors who have little experience, as doing that entails the great risk of not misusing the medicine and it can cause disease if not used at the right amount of dose. An experienced doctor can test the drugs himself and follow a procedure for proper confirmation may require making a proper judgment of the medicine. Then comes the added advantage for the pharmaceutical company as they will be benefited as more patients go to the experienced doctors for treatment. Experienced doctors are well-known and them prescribing the medicines just makes the pharmaceutical company reliable and efficient. Overall, this proves to be an ultimate profit to both the pharmaceutical business, doctor's reputation and of course, subsequently the people who take the medication. This prediction model will help to secure public health and moreover, these will contribute to the bioinformatics and create new scope to work with this field.

1.2 Scope:

It is a collaboration of medical science and computer science. The main purpose of this is to promote medicine and create a prediction model that will help doctors. And this will be made by the machine-learning process. And by this process, we will analyze the data and make a prediction model out of it. It will greatly assist the pharmaceutical companies consequently bringing profit and sustaining a good reputation to the people and doctors respectively.

It will be easier for people to go to their preferred doctors -doctors who are experienced and more than well acquainted to their job therefore avoiding any kind of hindrances including overdose, wrong prescription and treatment.

1.3 Research Objectives:

The main objective is to create a prediction model of the doctors for medicine promotion. From the dataset, some initial criteria should be selected first. The sub objective is to specify district and diagnosis, finding few initial criteria, fix the column of doctor selection, determine if doctors are recommended.

1.4 Research Proposal:

The predictive model is a procedure where a large set of data is analyzed with the help of data mining technique by finding a pattern in the dataset. In this dataset, there is doctors list where medicines and doctor's name listed according to their diagnosis. The motto is to find the criteria by which to build the prediction model. In the past years, there have been many types of research where medical data was used for data analysis and for prediction models to make lives better[1][2][3]. A lot of companies make medicines but they are unable to find a suitable doctor to promote their medicines. And finding a suitable doctor is of key importance. On the other hand, many renowned and efficient doctors are unable to find good medicines because they live out of town.

As discussed earlier Medical data is being analyzed with the help of data mining is getting popular day by day**.** Doctors provides treatment to patient and prescribes medicine. Through the process even if all the data are not connected and all the diagnosis is not occurring at the same time these data can give important information, [4] and if analyzed carefully these data create pattern. By analyzing these data, we can build prediction model. Our goal is to create a prediction model where pharmaceutical companies can connect with doctors where two are considered together. This prediction model will make the interactions easier and less time-consuming. Two of the parties will be benefited through this model. Though public health and benefit might not be perceived neglected or deprioritized. But as our model we will focus on doctor's experience and their ability so the medicines they promote can be an assurance and patients will not have any doubt.

For this prediction model, collecting data is hard because these data are assembled in a sequenced way where it is hard to find out the doctors. In the dataset, to find any doctor from these data is time-consuming and the data are not categorized based on very few criteria. Focusing on few initial criteria and creating the prediction model is quite difficult. There are two Ids where prescription Id and doctor's Id, from which it is confusing to which id the whole dataset is going to be organized and dealt with. To find some criteria and also to finding out the pattern of their diagnosis at the same time needs intellect.

There have been many works on the prediction model and medical data. There have been works on disease prediction, medicine prediction, diagnosis prediction but medicine promotion is a different topic.[3][5][6] As it can be considered controversial by some as some doctors are going to get preference for promotion purpose and company can only focus on business through promotion. Moreover, there's a lot of advantages and working potential and capacity in disease, medicine prediction that this sector wasn't considered with respect to other. Our approach will be to first find some initial criteria. In our work, we have decided to focus on three criteria, doctor's age, loyalty, and popularity. Depending on these criteria from the existing dataset through some complex query the production model will be created.

Chapter 2

Related Work

2.1 Critical Literature Review

Technology is making our lives easier, in recent years' technology provided comfort and saving our valuable time. On the other hand, as medical science has improved through time along with another field of science. Many new inventions are made through in medical science and from this invention and diagnosis huge set of data is generated. One of the techniques which computer science has provided is data mining. Through data mining set of data can be developed and new pattern and rules can be found from them. Medical data are very useful in this kind of analysis. In previous years, many works have been done such as, a genetic mutation in a migraine suffer from data mining metabonomic technique, neural network analysis (classifying unknown class for finger motion), cell damage segmentation with unsupervised learning, hierarchical decomposition in diagnosis, heart rate is modeled with data mining and so on. [1]

A research has been done where utilization of data mining in skin disease is analyzed. In this paper [2] various classification is discussed and their accuracy is compared and found out which one is the best for skin disease. In the paper, M. Saraee, A. Mohammadi have discussed association rule and classification.They have analysed and they have suggested gini based decision tree is accurate for skin disease, and a paper [3] where VF15 classification, nearest neighbor classifier, naïve Bayesian classifier discussing doctor's prescription, patient age and with the help of that they have created a visualization tool which is a diagnosis prediction by storing patient data. In medical studies, intern or medical student can examine their knowledge with this tool [2]

A study based on Electronic Prescription Monitoring tool's database of 11 years (1996-2005) focusing on opioids (one kind of morphine) prescription to prevent drug abuse. [5] In the paper dose, prescription and making customer ID and seller ID identifier monitoring opioids drug prescribe was fixed as criteria. They have figured out that doctor's shopping list was suspicious and the reason is found that several prescriptions are overlapping [4]. They have extracted them individually. As they haven't focused on doctor's prescription but not on the doctor. There is no mention of doctor's data were doctor's shopping was specified questionable activity [5] Also they have found another classification accuracy close to 98%. In the research of Nanni ,[6] he has removed age as there was missing value and used 10 fold cross validation estimate error rate in diagnosis. With the help of Random Subspace and feature selection, he built vector support machine which creates classifiers.

Another research found a correlation between population-based pharmacy data and physician's rating for chronic disease, they found a correlation with logistic regression analysis between chronic disease rating and ambulatory visits which were not correlated with gender [7] They have added death and probability of hospitalization later and calculated mortality with priory scoring rule.

A research on breast cancer survivability was carried out using SEER database (1973-2002) using naïve Bayes, the back-propagated neural network and C4.5 decision tree algorithms with weka tool developed in java where C4.5 decision tree algorithm gave best results comparing accuracy from the confusion matrix [8]. E. Guven , the author and his partner classified patient into not survived and survive though haven't worked with the missing value.

The locally frequent disease was predicted with cluster analysis method in hepatitis [9]. They have used the apriori algorithm to find frequent item sets to form association rule. Data was collected from Abha Private Hospital in Saudi Arabia consists of 2000 patients and application was built in java platform. Hernia and Tonsillitis were found frequent from this application. Also, apriori algorithm was modified to preprocess the data. In this research medicine was not an attribute.

A research based on surveys and journals about heart disease used Knn method ,Bayesian method ,Neural networks to predict heart disease[10]. Researchers have used medical profiles such as patient eating habit, sugar intake ,blood pressure to find patterns to relate with heart disease. Analyzing Cleveland database naïve Bayes gave the most accurate prediction. Making numerical and categorical attribute transaction attribute train and test approach was applied to validate the association rule. [10]

In the research of Christos Bellos and other author of the research chronic disease has been categorized with the help of an intelligent system, here pathological pattern in a patient body as well as the mental condition of a patient is well observed and data collected from some with sensor where various information about human body are examined and analyzed. [11] System was built with the help of a shirt and different kinds of sensor was used to collect the ECG, audio sensor, blood pressure and other information of human body. With the help of 10-fold cross validation data was classified and the whole dataset was classified into 11 instances. Mental condition was classified with the help of rule based classifier system and divided into normal, mild, moderate, severe by monitoring pathological and mental stress level information of a human body [11].

Pharmacy data were organized and collected and from there hospital's outcomes were predicted and in the research patients' admission, readmission, discharge and length of stay was observed to predict patient outcomes[12]. Wald X^2 test method was used with the help of logistic regression models to find out the readmission rate. These data were categorized into disease groups and it was not validated thus it makes the data sensitive and was not linked with public database also [12].

A research based on doctor's performance and diagnosis style were examined to find the difference which was significant. Where several information with the help of contextual inquiry was found and They have focused on 5 categories such as innovation, popularity, interpersonal skills, professionalism, respect [13]. 8 kinds of machine learning algorithm were applied to find key points to predict a doctor's performance. After comparing all the scores and results the research has found out respected, professional and interpersonal were the initial criteria for them to data analysis. With the help of R-Text tools and the limitations were some predicted high when performance was low, because the dataset was small [13].

All these researchers processed their data and most of them were for diagnosis and disease prediction. By observing their work medicine and doctor were not combined and doctor's information was not covered. From all this research selected our topic where the focus is on the doctors.

2.2 Research Questions:

This research looks for the answer of following questions. The first one is the main research question, others are the sub question.

- How can we measure the performance of a doctor for the eligibility of medicine promotion?
- What are the initial criteria for a doctor to allow him for medicine promoting by the pharmaceutical companies?
- What is the correlation among the criteria and how to facilitate the prediction model by them?

Chapter 3

Data Source and Domain Analysis

3.1 Data Source

Ziska pharmaceuticals has launched new products and they have made database in the past few years. Data is collected from Ziska pharmaceuticals provided by our supervisor. Our University, American International University -Bangladesh has accumulate data for research purpose from Ziska-4p. In the dataset, all the doctors' information and their diagnosis is collected. Our motto is to create a prediction model, initially we have selected doctors' age, loyalty and popularity.

Ziska pharmaceutical collects data with the help of 4p, a third-party company which helps Ziska to collect information about doctor's such as doctor's prescription, doctor's territory, how many patients he has treated and so on.

3.2 Domain Analysis

Necessary information from various tables has been organized in the dataset. These data have been collected from Ziska's pharmaceutical operations. In the dataset, all the important information like doctor's prescription, doctor's area, patient visit, medicine dosage, human resources and all other medical areas.

In this dataset doctor's prescriptions, noted by themselves are organized for data preprocessing and there are many attributes where all the prescription with doctor's territory, how many patient seek treatment and the location of prescribing are also included here which has helped to find our criteria and to find the main focus from where the prediction model is going to be built with.

Here drugs and their dosage where also included where it might be easier to find how to find a doctor which has reputation and enough experience to provide treatment to secure patient health. All this information helped to find a doctor's popularity calculate and create a prediction model of doctors. Finding how doctors have provided treatment and how they have given. Whether the service was given or not was also found out and calculate from this dataset.

3.3 Data Constraints

3.3.1 Missing Value replacement:

In this dataset, all the information was not given, some rows with missing values were found. To be accurate any row consisting missing value was eliminated.

3.3.2 Lack of proper Documentation:

There were lot of medicine names and generic names were missing or recorded inappropriately in the dataset. For the spelling mistake and typographic error many information could not be retrieved properly. There was few human errors in the documentation processes.

Chapter 4

Algorithm and Tools

4.1 Decision Tree Classifier

Classification is a process of dividing up objects so that each object is assigned to one of a number of mutually exclusive and exhaustive categories (known as class). A classifier gains knowledge from historical dataset and assign any unknown instances to classes.

Decision Trees (DTs) are a non-parametric supervised learning method used for classification. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features [14].

4.2 Development Platform and Tools

The complete dataset was stored in Microsoft SQL Server.

For the accuracy rating WEKA has been used. Weka is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules and visualization [15].

Chapter 5

Building Prediction Model

5.1 Initial Criteria for Doctor Selection

The two characteristics we demanded in a doctor to prefer him for medicine promoting

are popularity and loyalty.

Popularity is the percentage of patience visiting that doctor in a territory. Territory is the smallest unit for medicine distribution by the pharmaceutical companies and the 4P workers collect data territory wise.

Loyalty is the percentage of Ziska prescribed medicine of a doctor among all his prescribed medicine.

- **Doctor Popularity**

$$\text{Popularity} = \frac{TV}{SV} \times 100\ \%$$

TV = Total number of visit to a specific doctor

SV =Total number of visit for all doctors in that territory

- **Doctor Loyalty**

$$\text{Loyalty} = \frac{ZP}{TP} \times 100\ \%$$

ZP = Total number of prescribed Ziska Pharma medicine

TP = Total number of prescribed medicine

5.2 Analytical Approach

The analytical approach shows how we determined the 'Class' of a doctor if it's 'Fair' to promote him for further medicine or not.

Step 1.

Initial format:

For isServiceGiven = Yes, Class: Fair

For isServiceGiven = No, Class: Unfair

Step 2.

If Loyalty >= 5, Class: Fair

Step 3.

For isServiceGiven = Yes

If Loyalty < 5 & TP <200, Class: Fair

Step 4.

For isServiceGiven = No

If Popularity >= 30, Class: Fair

During the class assignment, emphasized on keeping the 'Initial Format' unchanged.

5.3 Sample data

After the analytical approach the sample data is given below

	A	B	C	D	E	F	G	H	I
1	doctorID	TV	SV	Popularity	ZP	TP	Loyalty	isServiceGiven	Class
2	BAR10657	4	9	44	1	137	1	Yes	Fair
3	BAR12169	2	17	12	0	25	0	No	Unfair
4	BAR12200	3	22	14	2	1718	0	Yes	Unfair
5	BAR12404	2	4	50	0	2	0	No	Fair
6	BAR12516	2	12	17	9	503	2	Yes	Unfair
7	BAR14269	2	4	50	2	278	1	Yes	Unfair
8	BAR14774	3	17	18	3	16	19	Yes	Fair
9	BAR14864	1	9	11	11	71	15	Yes	Fair
10	BAR15410	3	11	27	37	286	13	Yes	Fair
11	BAR16459	4	13	31	0	4	0	No	Fair
12	BOG00012	5	10	50	11	476	2	Yes	Unfair
13	BOG00022	5	8	63	16	553	3	Yes	Unfair
14	BOG00027	1	1	100	0	15	0	No	Fair
15	BOG00039	3	11	27	6	280	2	Yes	Unfair
16	BOG00041	2	4	50	3	97	3	Yes	Fair
17	BOG00045	3	8	38	1	255	0	Yes	Unfair
18	BOG00074	1	16	6	7	146	5	Yes	Fair
19	BOG00115	1	1	100	24	110	22	Yes	Fair

Figure 5A : Sample data for classification

5.4 Examining Data into Classifier

After setting the decision rules by the analytical approach the dataset was prepared for the classifier examining.

5.4.1 Feed the Dataset in WEKA

After preparing the dataset it was converted into CSV format. The CSV format was converted into ARFF file to run in WEKA.

5.4.2 Preparing Training Dataset

J48 (Decision Tree) classifier has been used in WEKA for testing the dataset with a 70-30 percentage split. 70% of the dataset was training data and 30% was test data.

5.4.3 Predicting Unseen Instances from Existing Dataset

After developing the model, the prediction for unknown ‘Class’ has been made by WEKA.

Chapter 6

Analyzing Result

6.1 Inspecting Accuracy

After testing the dataset for prediction accuracy the classified instances, TPR, FPR has been determined.

Correctly Classified Instances 97.2912 %

Incorrectly Classified Instances 2.7088 %

TPR (True Positive Rate): 0.969

FPR (False Positive Rate): 0.019

TPR is the proportion of positive instances that are correctly classified as positive.

FPR is the proportion of negative instances that are incorrectly classified as positive.

6.2 Decision Tree View

The decision tree by the classifier result is given below

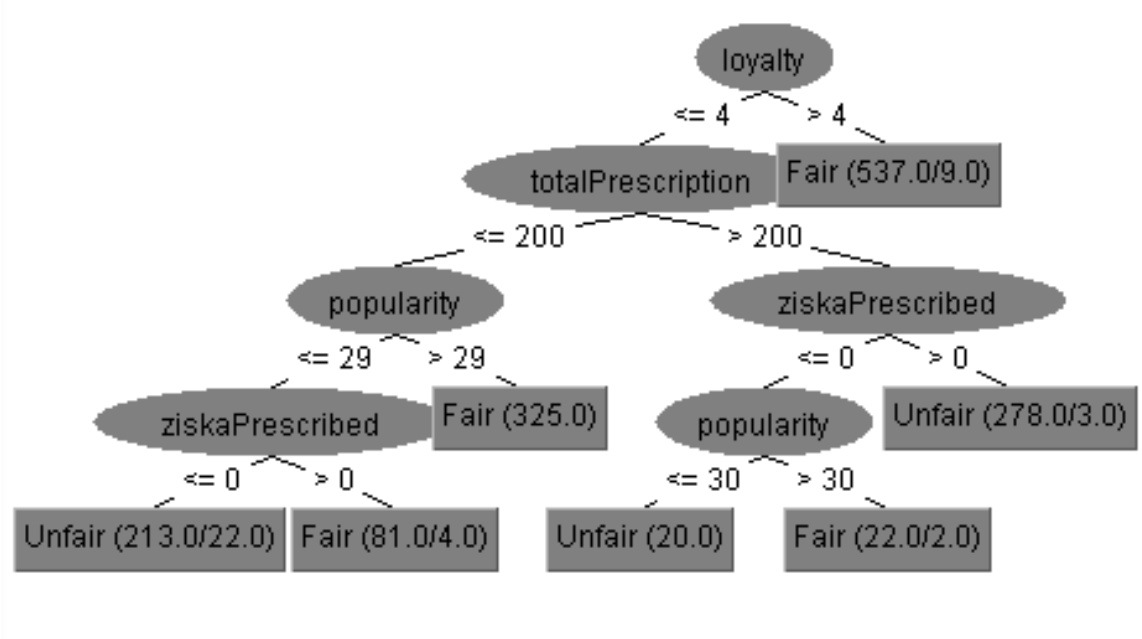

Figure 6A : Decision tree generated by WEKA

A clearer view of the tree is

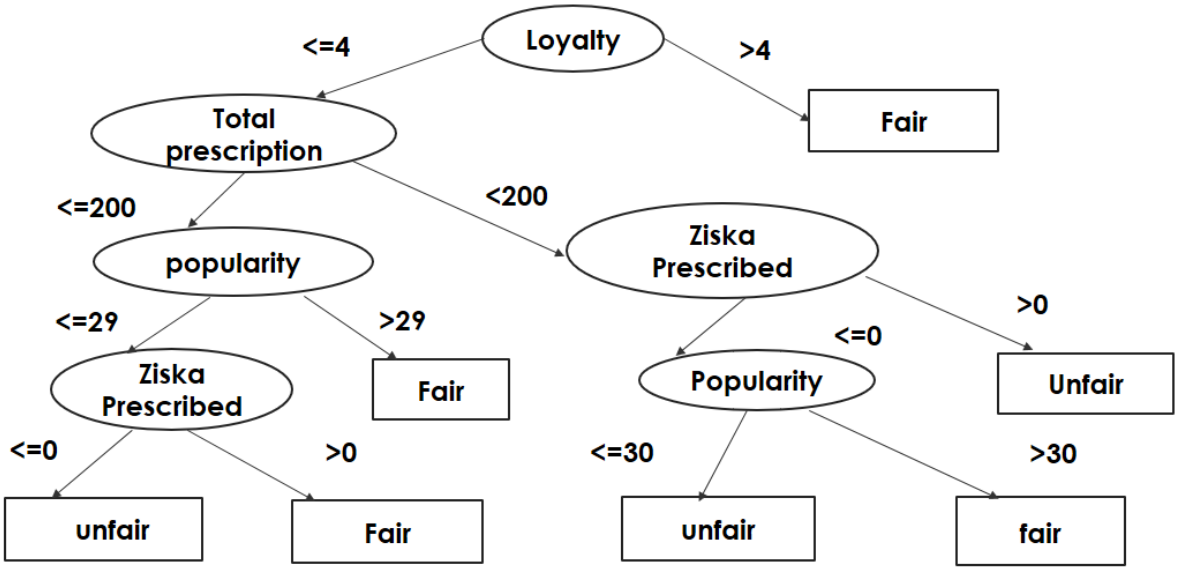

Figure 6B : Decision tree generated by WEKA (Clear View)

6.3 ROC Curve

In statistics, a Receiver Operating Characteristics or ROC curve is a graphical plot that illustrates the performance of a binary classifier system. The curve is created by plotting the TPR against the FPR.

A classifier or prediction model is considered as good if the area under ROC curve is 60% or more. The area under ROC curve of our prediction model is 0.977 which is a good outcome.

Here, X axis = False Positive Rate

and Y axis = True Positive Rate

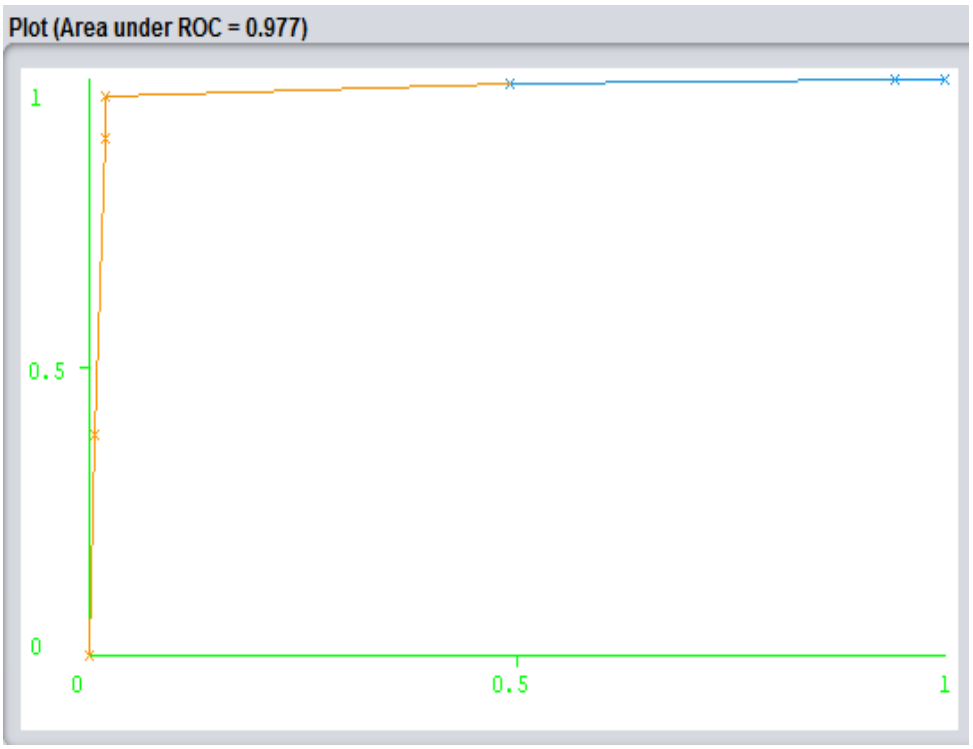

Figure 6C : ROC curve

6.4 Confusion Matrix

Confusion Matrix is a tabular way of illustrating the performance of a classifier. The confusion matrix by our prediction model is

		Predicted Class	
		Fair	**Unfair**
Actual Class	**Fair**	**278**	**9**
	Unfair	**3**	**153**

Table 6A : Confusion Matrix

The confusion matrix result shows only 12 instances were classified incorrectly which is a good prediction result.

Chapter 7

Limitation and Future Work

7.1 Correlation between Popularity and Loyalty

Correlation is the dependence or association in any statistical relationship between two random variables or two sets of data.

7.2 Specifying 'Brand' promoting rate for doctors

In our work we have considered the company of the medicines so far for loyalty detection. Pharmaceutical companies categorize their medicines in different brands based on the characteristics of the elements of the medicine.

The loyalty level can be detected more precisely if we can relate the brands with the prescribed medicine.

7.3 Higher accuracy level

Based on the previous researches we are looking for more criteria's (along popularity and loyalty) for doctor selection, which might enhance our accuracy level.

Chapter 8

Conclusion

The action of our thesis work gives a record of the doctor list with the criteria's of doctors which could be considered for medicine promoting. As the whole process is done manually there are chances to manipulate the records.

Our classification result gives very accurate and spontaneous result for doctor selection. The pharmaceutical company, the third party company (eg. 4P), the MPO (field workers for data collection), there may come diversity among the three entities for the doctor selection process. But when they will already be given an evidence and record of the doctors classification, the deliberate deception or the chance of misappropriating will be decreased for every entities and in every step. Thus million's amount of money can be saved by the pharmaceutical companies.

Reference:

[1] W. Yu, "Data Mining Techniques in Medical Informatics", The Open Medical Informatics Journal, vol. 4, no. 1, pp. 21-22, 2010.

[2] E. Barati,M. Saraee, A. Mohammadi "A Survey on Utilization of Data Mining Approaches for Dermatological (Skin) Diseases Prediction", Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Health Informatics (JSHI): March Edition, 2011.

[3] H.A. Guvenir and N. Emeksiz, "An expert system for the differential diagnosis of erythemato-squamous diseases," Expert Systems with Applications, vol. 18, pp. 43–49, 2000

[4] D. Paperny, Handbook of Adolescent Medicine and Health Promotion, 1st ed. World Scientific Publishing Co. Pte. Ltd, 2011, p. 24.

[5] N. Katz, L. Panas, M. Kim, A. Audet, A. Bilansky, J. Eadie, P. Kreiner, F. Paillard, C. Thomas and G. Carrow, "Usefulness of prescription monitoring programs for surveillance-analysis of Schedule II opioid prescription data in Massachusetts, 1996-2006", Pharmacoepidemiology and Drug Safety, vol. 19, no. 2, pp. 115-123, 2010.

[6] L. Nanni, "An ensemble of classifiers for the diagnosis of erythemato-squamous diseases," Neurocomputing, vol. 69, pp. 842-845, 2006.

[7] M. Von Korff, E. Wagner and K. Saunders, "A chronic disease score from automated pharmacy data", Journal of Clinical Epidemiology, vol. 45, no. 2, pp. 197-203, 1992

[8] E. Guven and A. Bellaachia, "Predicting Breast Cancer Survivability Using Data Mining Techniques", 2007.

[9] M. Khaleel, S. Pradhan and G. Dash, "Finding Locally Frequent Diseases Using Modified Apriori Algorithm", International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, no. 10, 2013.

[10] J. Soni, U. Ansari, D. Sharma and S. Soni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications, vol. 17, no. 8, pp. 43-48, 2011.

[11] C. Bellos, A. Papadopoulos, D. Fotiadis and R. Rosso, "An Intelligent System for Classification of Patients Suffering from Chronic Diseases", FORTH BRI Foundation for Research and Technology - Hellas, Biomedical Research, Ioannina Greece, 2010.

[12] J. Parker, J. McCombs and E. Graddy, "Can Pharmacy Data Improve Prediction of Hospital Outcomes?", Medical Care, vol. 41, no. 3, pp. 407-419, 2003.

[13] C. Gibbons, S. Richards, J. Valderas and J. Campbell, "Supervised Machine Learning Algorithms Can Classify Open-Text Feedback of Doctor Performance With Human-Level Accuracy", Journal of Medical Internet Research, vol. 19, no. 3, p. e65, 2017.

[14] S B Kotsiantis, D Kanellopoulos, and P E Pintelas. \Data preprocessing for supervised learning". In: International Journal of Computer Science 1.2 (2006), pp. 111{117. issn: 1306-4428. doi: 10.1080/02331931003692557.

[15] A. Huppert and G. Katriel. \Mathematical modelling and prediction in in-fectious disease epidemiology". In: Clinical Microbiology and Infection 19.11 (2013), pp. 999{1005. issn: 1198743X. doi: 10.1111/1469-0691.12308.

[16] Courtney D. Corley et al. \Disease prediction models and operational readiness". In: PLoS ONE 9.3 (2014), pp. 1{9. issn: 19326203. doi: 10 . 1371 /journal.pone.0091989.

[17] S. Soni, O.P. Vyas, " Using Associative Classifiers for Predictive Analysis in Health Care Data Mining", International Journal of Computer Application (IJCA, 0975 –8887) Volume 4– No.5, July 2010, pages 33-34.

[18] http://scikit-learn.org/stable/modules/tree.html

[19] http://www.cs.waikato.ac.nz/ml/weka/