Data Mining Techniques Implementation To Improve Healthcare Among Diabetic Patients

تطبيق تقنيات التنقيب في البيانات من اجل تحسين العناية الصحية لدى مرضى داء السكري

By

Student Name Ghania Aref AlBanna

Student ID number 2014128042

Dissertation submitted in partial fulfilment of the requirements for the degree of MSc Informatics (Knowledge and Data Management)

Faculty of Engineering & Information Technology

Dissertation Supervisor
Dr. Sherief Abdullah

December - 2016

DISSERTATION RELEASE FORM

Student Name	**Student ID**	**Programme**	**Date**
Ghania Aref AlBanna	2014128042	Masters in Informatics	Dec, 2016

Title

Data Mining Techniques Implementation To Improve Healthcare Among Diabetic Patients

I warrant that the content of this dissertation is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that one copy of my dissertation will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

Electronic Submission Copyright Statement

Please choose one of the following two licenses and check appropriate box.

☒ I grant The British University in Dubai the non-exclusive right to reproduce and/or distribute my dissertation worldwide including the users of the repository, in any format or medium, for non-commercial, research, educational and related academic purposes only.

Public access to my dissertation in the Repository shall become effective:

☒ Immediately ☐ 24 months after my submission

☐ 12 months after my submission ☐ 48 months after my submission

☒ I grant The British University in Dubai the non-exclusive right to reproduce and/or distribute my dissertation to students, faculty, staff and walk-in users of BUiD Library, in any format or medium, for non-commercial, research, educational and related academic purposes only.

Signature
Ghania Aref Albanna

Abstract

Medical data mining is an emergent field and, on overcoming its facing challenges such as privacy of documentation and ethical use of information about patients, voluminous and heterogeneous data, and imprecise and erroneous data, medical data mining can be as powerful as that in any other common field such as ecommerce and marketing. Traditional research could not overcome completely these challenges and only hypotheses based on anthropological approaches are tested.

Unlike traditional research, this dissertation discusses predictive analysis and knowledge discovery of trends and patterns from databases in the medical field. Retrieval of clinical medical data is helpful in conducting different learning techniques. Performance of different classification techniques is compared and ensemble learning of best classifiers is tested.

The analysis showed that ensemble learning via bagging predicts best the percentage of diabetic adolescents who are most prone to hospital re-admission and more susceptible to join the "Diabetic Self-Management Educational Support Program". This predictive classification helps in leveraging the healthy psychological status of the patients (social and medical), reducing readmission costs (economic), and pre-hypothesizing (scientific) relationships between different parameters based on different patterns and trends predicted by machine learning techniques.

ملخـــــص

يعدّ التنقيب في البيانات في المجال الطبي من الحقول الناشئة في يومنا هذا. وفي ظلّ التحديات التي يواجهها هذا الحقل مثل خصوصية الوثائق واستخدام المعلومات الخاصة بالمرضى وفقاً لأخلاقيات المهنة والبيانات الضخمة غير المنظمة والبيانات غير الدقيقة والخاطئة، يبرز بقوةٍ لا تقلّ عن الحقول الأخرى الشائعة مثل التجارة الإلكترونية والتسويق. فأساليب البحث التقليدية لم تتمكن من التغلب بشكل كامل على تلك الصعوبات إذ تمّ فقط اختبار الفرضيات القائمة على المقاربـــات الأنتروبولوجيـــة.

وعلى خلاف أساليب البحث التقليدية، يتناول هذا البحث التحليل التنبؤي والاكتشاف القائم على المعرفة للتوجّهات والأنماط من قواعد بيانات في المجال الطبي. إن استرجاع البيانات الطبية السريرية يساعد على تطبيق العديد من تقنيات التعلم المختلفة حيث تجري مقارنة بين أداء تقنيات التصنيف المختلفة واختبـار التعلـم الجمعـي لأفضـل المصنّفيـن.

وقد بيّنت الدراسة التحليلية أن التعلّم الجمعي عبر خوارزميات التعلّم الآلي (bagging) هي أفضل طريقة للتنبؤ بالنسبة المئوية للمصابين بالسكري من المراهقين وهم الأكثر عرضةً للدخول إلى المستشفى بصورة متكرّرة والأولى بالانضمام إلى "برنامج التثقيف والدعم حول الإدارة الذاتية لمرض السكري". يساعد هذا التصنيف التنبؤي على تعزيز الوضع النفسي الصحّي لدى المرضى (من الناحية الاجتماعية والطبية)، وتخفيض تكاليف التردّد إلى المستشفى (من الناحية الاقتصادية) والافتراض المسبق (من الناحية العلمية) للعلاقات بين المعايير المختلفة استناداً إلى أنماط وتوجهات مختلفة تمّ التنبؤ بها عبر تقنيات التعلّـــم الآلـــي.

Table of Contents

1. Introduction 1
1.1 Overview of Data Mining in The Medical Field 1
1.2 Problem Statement 2
1.3 Goals and Objectives 5
1.4 Research Questions 7
1.5 Uniqueness of the research 7

2. Literature Review 9
2.1 Hospital Readmission and Diabetes: 9
2.2 Data mining approaches in the field: 11

3. Research Design 15
3.1 Data Collection and Description 17
3.2 Data Preparation, Transformation, and Preprocessing 22
3.2.1 Description of Attributes 22
3.2.2 Possible weights assigned to attributes in each subset of features: 23
3.3 Features Selection 28
3.4 Size of Sample Under Study: 32
3.5 Predictive Modeling: Classification 32
3.6 Predictive Model Evaluation 33

4. Classification Models And Performance Evaluation 34
4.1 Decision Tree: 34
4.1.1 Description of the Model: 34
4.1.2 Feature Selection and Model Performance: 35
4.2 Artificial Neural Network Model: 48
4.2.1 Description of the Model 48
4.2.2 Feature Selection and Model Performance: 49
4.3 K-nearest Neighbor: 52
4.3.1 Description of the Model: 52
4.3.2. Feature selection and performance: 53
4.4 Naïve Bayes Classification: 56
4.4.1 Description of the Model: 56
4.4.2 Feature selection and performance: 56
4.5 Comparison of The Models: 58
4.6 Ensemble Learning 60
4.6.1 Random Forest: 61
4.6.2 Ensemble classifier combining the two models, decision tree and Naïve Bayes, via voting 62
4.6.3 Ensemble classifier via bagging 64
4.6.4 Ensemble classifier via boosting 65
4.6.5 Comparison of the Four Ensemble Learners 65

5. Discussion 69

6. Conclusion..... 71
7. References: 72

List of Figures

Figure 1: Proportion of Diagnosed Diabetic Patients based by age 3
Figure 2: Proportion of people with diagnosed diabetes, by gender.....4
Figure 3: United Arab Emirates versus world prevalence with diabetes 4
Figure 4: Comparison of p(y) and p(o). 14
Figure 5: Data Mining Approach 17
Figure 6: The proportion of diabetic patients based on gender 21
Figure7: the frequency of patients' readmissions.....22
Figure 8: Weight of Identity Attributes Based on Info Gain, Chi Square, and Info Gain Ratio 23
Figure 9: Weight of Drug Attributes by Chi Square Used in Rapid Miner 25
Figure 10: Summary indicating the number of the most relevant features 28
Figure 11: Clustering of diabetic patients based on readmission to hospital and gender.....36
Figure 12: Decision Tree Process Design with Optimization of features 38
Figure 13a: Performance Evaluation of Decision Tree of Discretized Input Data. 41
Figure 13b: Performance Evaluation of Decision Tree of Non-discretized Input Data. 41
Figure16a: Part of the Best Preforming Decision Tree When Number_Patient $\geq$ 0.5. 46
Figure16b: Part of the Best Preforming Decision Tree When Number_Patient $\leq$ 0.5. 47
Figure 17: Performance Evaluation of Decision Tree of Discretized Transformed Input Data.....47
Figure 18a: Performance Evaluation of Decision Tree of Non-discretized Balanced Transformed Input Data.....48
Figure 18b: Performance Evaluation of Decision Tree of Discretized Balanced Transformed Input Data.....48
Figure 19: Model Topology of Artificial Neural Network 49
Figure 20: Artificial Neural Network Performance 51

Figure 21: Confusion Matrix of ANN Model Performance 51

Figure 22: Comparison of Highest Performance (accuracy) of Different Models 59

Figure 23: Comparison of Highest Performance (% Mean Recall) of Different Models 60

Figure 24: Ensemble Learner is a Combination of Weaker Learners 61

Figure 25: The Best Automatically Selected Decision Trees by Random Forest Ensemble Learner 62

Figure 26: Rapid miner Vote Modeling Process 63

Figure 27: Cross Validation Performance of the Vote Model 63

Figure 28: Cross Validation Performance of the Ensemble-Learning Model via Bagging 64

Figure 29: Cross Validation Performance of the Ensemble-Learning Model via Bagging Using the Balanced Dataset 64

Figure 30: Comparison of the performance of the Ensemble Learners with the Best Performing Individual Classifiers 66

Figure 31: Statistical Analysis of the Predicted Readmissions 66

Figure 32: Box Plot (Quartile) Chart of the above distribution 67

Figure 33a: One of Detected Decision Trees and Relationships via Bagging 68

Figure 33b: Another Detected Decision Tree and Relationships via Bagging 68

List of Tables

Table 1: The proportion of readmissions based on gender 21

Table 2: Division of Attributes 22

Table 3: the most relevant drug attributes based on Information Gain 24

Table 4: the Most Relevant Drug Attributes Based on Weight by Chi-Square 25

Table 5: Attribute weight using forward and backward selection techniques 26

Table 6: Description of Relevant Attributes 30

Table 7: Values of Relevant Features 32

Table 8: Results of the decision tree without discretizing the input data 39

Table 9: Results of the decision tree after discretizing the input data 40

Table 10: Results of the decision tree without discretizing the transformed input data 44

Table 11: Results of the decision tree after discretizing the transformed input data 45

Table 12: ANN Model Performance 50

Table 13a: Model Performance Where the Readmission is Categorized into Three Labels 55

Table 13b: Model Performance Where the Readmission is Categorized into Two Labels 55

Table 14: Percent Accuracy of Naïve Bayesian Classification Model 57

1. Introduction

This research is unique not only in its approaches, but also in its aims. The research elaborates on carefully different data mining approaches and answers constructively the research questions. This research discusses how data mining can be a good tool to contribute to the medical field from different perspectives. The research is divided into six sections. The first section is the introduction that comprises overview of medical data mining, the problem statement, goals and objectives of research, research questions, and uniqueness of research. The second section describes the related work done in the area of research, the third is the research design, the fourth contains the methodology and results of the used data mining techniques, the fifth is the discussion and the sixth section is the conclusion.

As an introduction, this section highlights the importance of data mining in medicine by overviewing what data mining targets and detects. It also includes the problem statement, the goals and objectives, the research questions, and the uniqueness of the study. This part of the dissertation emphasizes the significance of the problem that needs to be studied by data mining techniques and other statistical methods for future.

1.1 Overview of Data Mining in The Medical Field

Data mining, in many fields such as web-based businesses and commerce, is highly popular and valuable. Knowledge discovery database (KDD) is helpful not only to target retails, fast-moving products, current customers, but also potential customers. Data mining is now developing to be implemented in new fields such as banking, education, and medicine. This dissertation discusses knowledge discovery of trends and patterns from

databases in the medical field. Traditionally, research in medicine is merely statistical where a hypothesis is stated based on researchers' perspectives and experiences in the field. The data is collected in their search and statistically proven or rejected based on the implemented experimental results. Data mining is peculiar in many aspects in medicine: 1) Hypothesis is not deduced from experiences and trends, but rather from data analysis using Machine leaning and data mining techniques such as clustering, association, and classification. 2) Medical data mining usually starts with pre-hypothesis and the results are adjusted to fit the hypothesis. 3) Data mining discovers trends and patterns, which are hard to overview by traditional methods. 4) Anomalies and minorities are more important for study in medicine than general trends; Anomaly trends are hard to observe through experiences, but are easily inspected by medical data mining modeling techniques (Shillabeer, & Roddick, 2007).

Data mining in medicine is even different from standard data mining in other fields. While, in the former, researchers are not only interested in describing the patterns and trends like other fields of commerce, marketing, banking, and telecommunications, but they are also detecting exceptions and explaining interesting patterns and trends. Medicine needs these explanations as a slight difference may cause a change in patient's life destiny (Hofmann, & Klinkenberg, 2013). In this study, data mining techniques are used to find patterns and trends among the diabetic patients in United Arab Emirates (U.A.E).

1.2 Problem Statement

U.A.E is a young developing country in the Middle East. UAE has joined International Diabetic Federation (IDF) in 2000. According to World Health Organization, the percentage of diabetic patients among the national population increases with age as shown in Figure 1. There is also a slight

difference in percentages between males and females as shown in Figure 2 (Badrinath & *et al.,* 2014).

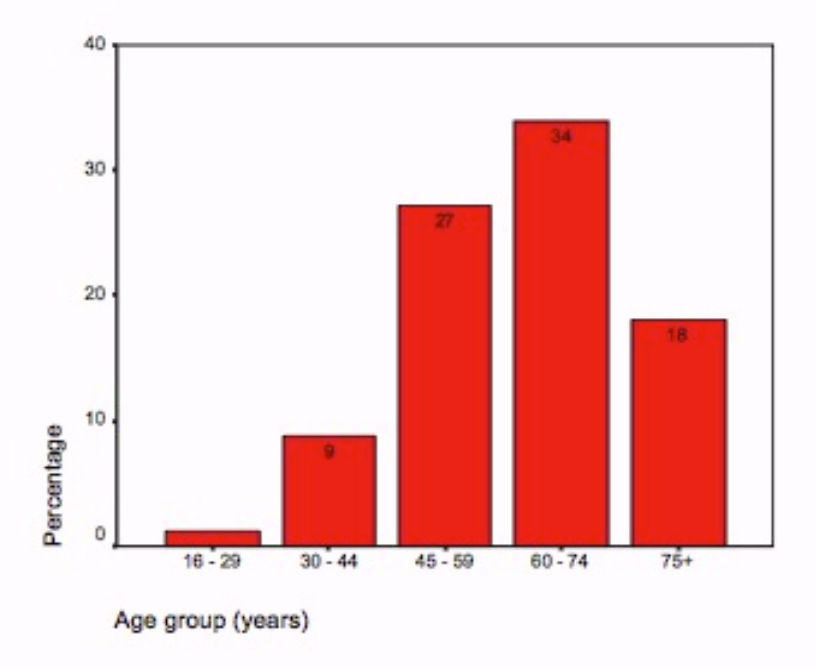


Figure 1: Proportion of Diagnosed Diabetic Patients based by age

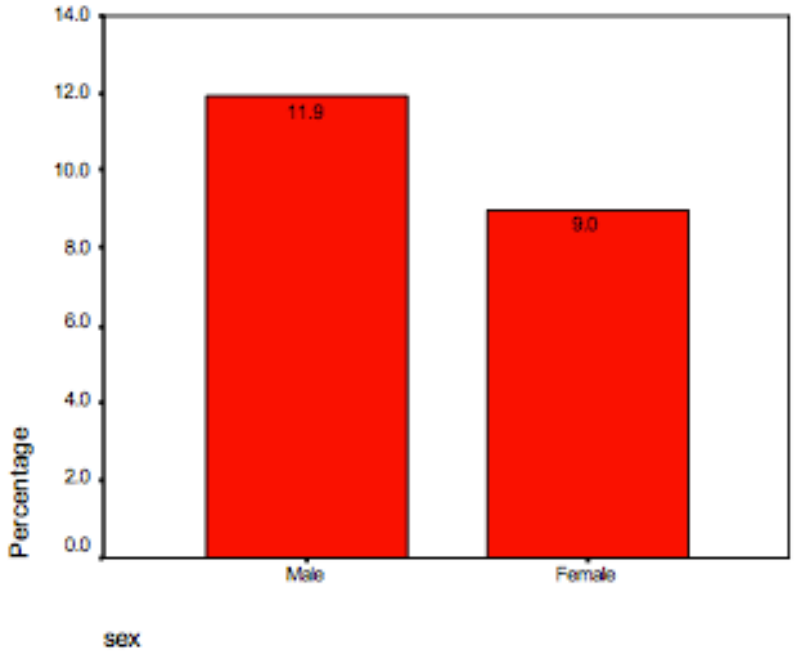


Figure 2: Proportion of people with diagnosed diabetes, by gender

The percentage of the young diagnosed diabetic citizens is extremely low and increases with age. Unfortunately, this percentage augmented enormously in 2015, and IDF and WHO reports indicate that UAE has double the proportion of diabetic people at age of 20 compared with the world prevalence of diabetics as shown in figure 3 (International Diabetic Federation, 2015).

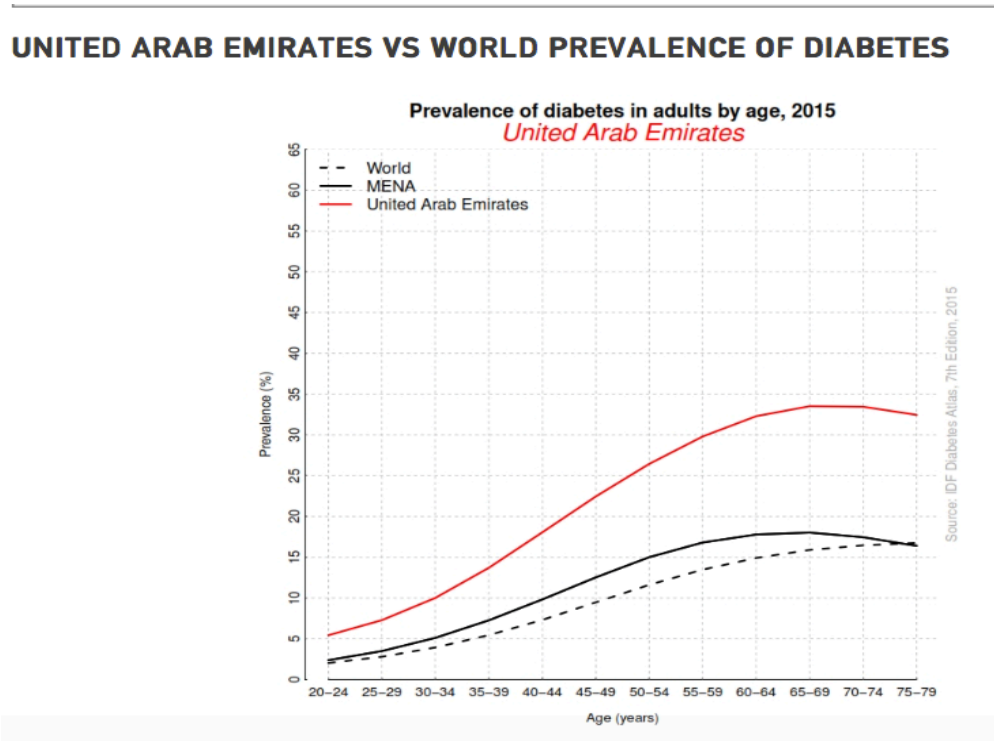


Figure 3: United Arab Emirates versus world prevalence with diabetes

Since the diagnosed disease is diabetes of type 2, which results from defect(s) in secretion of insulin, it is thought that diabetes type 2 is triggered by many risk factors. Due to the alarming increase of UAE nationals who are diagnosed as diabetic many studies are issued to study the percentage of diabetic people (diagnosed, undiagnosed, and pre-diabetic). Some researchers looked for different cultural, economical, and dietary reasons as

risk factors that may have contributed to the remarkable increase of the disease. Other researchers tried to be more specific and related diabetes type 2 to specific health problems such as obesity (Bin Zaal *et. al,* 2009), lack of vitamin D (Davis, 2011), and lack of physical activity (Baglar, 2013). Obesity, unhealthy diets, individual wealth, and cultural restrictions on physical activity for females are evolutionary sociological risk factors that are hard to change over a short period of time. Therefore, it is wise to shift attention to finding effective and supportive treatments to patients of diabetes type 2. In this study, the adolescent (age between 10 and 20 years) diabetic patients are targeted. The choice of this group is based on many reasons: 1) Adolescents are the potential pillars of the growing society. It is mandatory to help them maintain the highest possible self-esteem, especially when chronic diseases such as diabetes intervene. 2) Adolescents are more susceptible to develop comorbid-internalizing disorders such as depression and anxiety than young children (Garrison, *et al.*, 2005). 3) It is observed that patients with more comorbid internalizing disorders have higher risk of hospital readmission. This association has unclear reasons. However, the possible noted reasons can be 1) Lack of persistent adherence to treatment plan. 2) Poorer adherence to diet and exercise, 3) more missing medical visits and appointments, 4) Lack of parental control for adolescents or lack of parental intervention, as opposed to the case of young children.

1.3 Goals and Objectives

Due to the alarming increase of diabetic patients in UAE, research on diabetes is enriched. All the current research, however, has been relational, purely statistical, and documenting for percentages of diabetic and obese populations. Attention has to be given to the patients who are more frequently in need not only of medical support and closer follow-up treatments than others, but also of psychological orientation, and diabetic

educational self-management and awareness. Data mining in this case is valuable to identify those patients who need the real extra support to avoid hospital readmission and adverse complications. Some diabetic patients are more frequently readmitted to hospitals than others. Some are pre-diabetic and have symptoms that drive them to hospitalization every now and then. Others have undiagnosed diabetes because they ignore the symptoms. Due to the lack of historical patterns and behaviors among the patients, physicians treat patients evenly.

This research aims at extracting the patients who are most likely to be readmitted to the hospital and at helping the physicians, in the endocrinology department in particular, to give the more supportive medicinal treatment, and to refer the targeted patients to "Diabetic Self-Management Educational Support Program". This program should not be informative. It should be recreational to attract the patients to show more loyalty to the program and its goals. It should not only be instructional but also involve the best practices that reduce the negative effect of the common risk factors on diabetic patients. The program promotes plans of setting and serving healthy diets, physical activity programs and advising other diabetic youngsters to learn more about their body and its needs. The importance of this program is to decrease the occurrence frequency or degree of comorbidities associated with diabetes. Depression is a common comorbidity. Depression is a comorbidity that is highly associated with length of stay in hospital (LOS) (Rubin, 2015).

Economically speaking, avoiding readmission to hospital is not only supportive to the patients' self-being, but also relieves the burden of economical hospitalization costs on uninsured patients, on insurance companies, and on the government. Here lies another significant

advantageous aspect of the approach. Instead of spending financial costs on wasteful days of adolescents' lifetime in hospitals, the costs can be spared for investments in supportive projects in promoting healthy life-style activities.

Adolescents are the potential pillars in every emerging generation in every society, so it is mandatory to alleviate the individual socio-economical status by reducing the frequency of readmissions, by promoting higher self-esteem, especially in patients with chronic diseases such as diabetes, by ensuring healthy optimistic view of life, and therefore, by cutting off readmission costs.

In this dissertation, the approach is highly technical and outgoing as it 1) deals more with diagnostic medical data rather than surveys, 2) helps physicians deal with other patients more carefully, depending on their case, and 3) focuses on adolescent patients.

1.4 Research Questions

The research is designed to answer the following questions:

1) How can data mining techniques be helpful in medical field?
2) How can data mining approaches integrate different interdisciplinary domains?
3) How can data mining techniques lead to higher accuracy in predicting the patients who are more prone to hospital readmission?
4) What are the main medical features that assist in giving the best results for best practices?
5) How can data mining approaches lead to new hypotheses in the field?

1.5 Uniqueness of the research

Data mining is emerging in the medical field. Due to the severe restrictions in terms of privacy, security, and human rights, technical and pure medical data was not easy to retrieve. Most of the research was based on survey results and on various anthropology resources. Another unique feature is that the research uses clinical datasets instead of demographic statistical counts.

The third aspect is scientific. In the traditional studies performed earlier in UAE, the hypothesis is generally stated and surveys and action research are used to test the anthropological relationships. Unlike traditional approaches, data mining of clinical datasets is used to discover patterns or guess relations not observed before. Therefore, data mining in research preludes to a hypothesis. In addition, this study is unique in its objectivity in the way that not experiences or expectations pre-hypothesize, instead objective numerical data do. This research dwells beyond descriptive analysis of facts to take action of what can be done to limit the severity of the problem, especially among adolescents.

Lastly, it is evident that in UAE, despite the alarming dangers of increasing proportion of the diabetic national population and in spite of addressing the risk factors associated with this increase, all medical institutions offer mere medical, pharmaceutical, laboratory, and blunt dietary services. No suggestion has emerged earlier to enhance self-awareness, health, and psychological homeostasis of adolescents. Although there was a call for considering the latter goals (Saadi *et al.,* 2007), no suggestion aimed at deteriorating the entire medical, physical, and psychological risk factors associated with diabetes. This research acts as a prelude to an action research and project implementation that improves healthcare services, reduces costs, and elevates the socio-economical and psychological welfare of diabetic patients.

2. Literature Review

This section displays the work related to hospital readmission of diabetic patients and to the extent of effective approaches performed by data mining in the medical and healthcare field.

2.1 Hospital Readmission and Diabetes:

Hospital readmission care measure takes a modest research attention despite of its significance. Little research also focused on diabetic patients re-hospitalization although, based on studies, around 40% of the readmissions are diabetic patients (Dungan, 2012). Reducing the diabetic patient readmission is quite significant (Rubin, 2015). Different risk factors contributions to more frequent re-hospitalization should be highlighted and understood. For the 30-days readmission, the factors that increase the risk of 30-day readmission are gender (males), comorbidities, hospital stay, type of insurance, and type of discharge. Recent studies related the readmission to the level of A1C level; those patients whose A1C level is > 9% , although they look highly diabetic, are less likely to enter the hospital within 30 days, they may readmit the hospital after 30 days because of comorbidity associated with diabetes but not because of the disease itself. Those patients with A1C level $\geq$ 8% show 20% decrease in level of readmission for every 1% increase in A1C level. The patients with A1C level $\geq$ 6% are more likely revisit the hospital within 30 days because of diabetes. The risk factors overlap to include hospital readmissions to 'beyond 30 days hospitalization'. Rubin (2015) suggests two different strategies to decrease the readmission rate: 1) intensification of diabetes therapy on discharge, 2)

dedicated outpatient support, and 3) inpatient diabetes education. These strategies and risk factors recommended and mentioned by Rubin (2015) are poorly discussed in his paper. In this Literature review, research on diabetic adolescents are discussed and the data mining approaches that are done modestly till now in the medical and healthcare field is highlighted.

Most of the recent research is primarily based on anthropological approaches of the disease. Many studies, however, related the frequency increase of disease among population to different risk factors, especially among adolescents. Some related this increase to gender (Badrinath & et al., 2014), obesity (Bin Zaal et. al, 2009), physical activity (Baglar, 2013), inpatient education (Healy et al., 2013), poor healthy diets (Ali et al., 2003a). Ashraff and his co-workers (2013) in his study emphasized a poorly recognized significant risk factor, the psychological health. They studied the psychological impact of diabetes in adolescents (Ashraff, Siddiqui, and Carline, 2013). Children and adolescents are more likely to develop emotional and behavioral comorbidities. The psychological aspect of the disease is always missed to emphasize the maintenance of glucose blood. The nature of the relationship between metabolic control and psychological functioning is not clear, although studies have showed that there is some correlation between disease compliance and psychological adaption. The patients who suffer from psychological disturbances because of diabetes are more likely to re-hospitalize. Therefore, well-controlled diabetes causes less conflict among family members and poorly controlled diabetes negatively affects the child's personality, physical wellbeing, schooling, participation among the family activities, and participation in activities away from home. The more uncontrolled is the disease, the stronger is the depression and anxiety among the patients. Stress and metabolic control relation is unclear as some studies support this relation and other studies do not. Anyhow, quality of life certainly affects the diabetic adolescent's

medical control. According to Ashraff (2013), the lower the HbA1C level is the fewer are the worries and the greater is the satisfaction and the better is the health awareness of diabetic adolescents. Patients who received intensive diabetic management with a behavioral coping skills (adaptive) training have lower HbA1C percentage and show less negative impact of diabetes on their quality of life than those who received only intensive medical management alone (Ashraff (2013).

2.2 Data mining approaches in the field:

Although implementing data mining in the medical and healthcare field is modest, this field is a rich area for data mining because this field is rich in diverse huge information. Finding trends and patterns is helpful in predictions and decision-making. Due to the large amount of information, machine learning is helpful as it is tedious and almost impossible to be predicted by humans (Milovic and Vrbas, 2012).

Data mining can be descriptive as well as predictive. Descriptive data mining is unsupervised and does not need a special attribute. It declares the distribution of data and describes the nature of data. Predictive data mining is supervised and includes a special label. It demonstrates predictive trends and patterns such as classification and regression.

Data mining in the medical field has its advantages: electronic patient records are secured and collected. Information system simplifies and automates the workflow of health care institutions. Data mining can be used in many healthcare areas such as measuring clinical indicators, type of customer satisfaction, and identifying high-risk patients, etc.

Data mining in healthcare reduces subjectivity and develops more knowledge discovery. Predictive modeling can be used to determine the diseases conditions and trends, provided that medical documentation and drug prescription related to the predictive models are supported. If

knowledge is seriously considered for predictive analysis, data mining can be an ultimately useful strategy to solve quality problems such as 1) discovering new hypotheses, 2) validation of data, standards, plans, and treatments, 3) improvement and strengthening of quality indexes for data, standards, plans, and treatments.

The success of the data mining power has many obstacles: 1) privacy of documentation and ethical use of information about patients cannot be invaded; 2) raw medical data is voluminous and heterogeneous in scope and complexity, especially that the dataset may have missing, incorrect, inconsistent, or complex data; and 3) all diagnosis and treatments are not precise and not error free.

Overlooking the obstacles of data mining in the healthcare field, there are two challenges: 1) how to develop algorithms that help compare improvements, and 2) how to develop performance evaluation for algorithms (Milovic and Vrbas, 2012).

Iyer and Sumbaly (2015) implemented data mining using Decision Trees and Naïve Bayesian predictive classification techniques to find solutions to diagnose diabetes and find timely treatment for the disease. The targeted disease is non-curable but controllable. Diabetes can in its three types: diabetes 1, diabetes II, and gestational diabetes (that occurs in pregnant woman). The population of Iyer's study is pregnant woman and the target feature is whether they are diabetic or not (the variable is 0 or 1). The dataset consists of 768 instances with eight attributes (Number of times pregnant, Plasma glucose concentration, Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), 2-Hour serum insulin, Body mass index (kg/m2), Diabetes pedigree function, and Age (years). Iyer used Weka tool to replace missing values, normalize the attribute values, and run Decision Tree and Naïve Bayesian classification techniques. The classification algorithms selected the most relevant feature for the study (Plasma glucose

concentration, Body mass index (kg/m2), Diabetes pedigree function, Age (years). Two performance evaluation procedures are performed on the decision tree: data split of ratio 0.7:0.3 and cross validation, while only data split (0.7:0.3) is used for Naïve Bayesian classification (Iyer, S, and Sumbaly, 2015). The results ranged between approximately 74% accuracy in case of cross validation and 76% in case of 0.7:0.3 data split.

The same features are used by Kumari (2014) to predict whether a person is diabetic or pre-diabetic or non-diabetic. Kumari used also Weka tools and the target variable has three classes (Yes for diabetic, No for non-diabetic, Pre for pre-diabetic). Naïve Bayes classification technique is used for classification (Kumari, Vohra, and Arora, 2014).

Aljumah and his co-authors (2012) applied data mining on young and old diabetic patients. They did not work on predicting whether the patient is diabetic or may develop diabetes, nut on predicting the effective treatment based on the age and based on the training set provided by WHO in 2012 in Saudi Arabia. WHO supports the authors of this paper with data that include the following features: age group, number of patients in each group of the sample WHO has, the number of patients whose treatment was effective in controlling the disease, and therefore the percentage of the successful patients. There are five age groups of 10-years interval. These five groups are divided into two age groups: the young group p(y), ranges between 15 and 44 years and the old group p(o), ranges between 34 and 64 years; the age group from 35 to 44 is common between both sets. The treatments suggested by WHO are based on the risk factors that are commonly discussed and the values are taken based on surveys. The treatments are Drug, Diet, Weight Reduction, Smoke Cessation, Exercise, and Insulin Intake treatments. Aljumah and his co-workers implemented data mining on this data to predict the most effective treatment for diabetic patients based on age. They used support vector data mining techniques and linear

regression to perform the prediction. According to their results, a summary showing a comparative view of results is shown below (Aljumah, Ahamad, and Siddiqui, 2013).

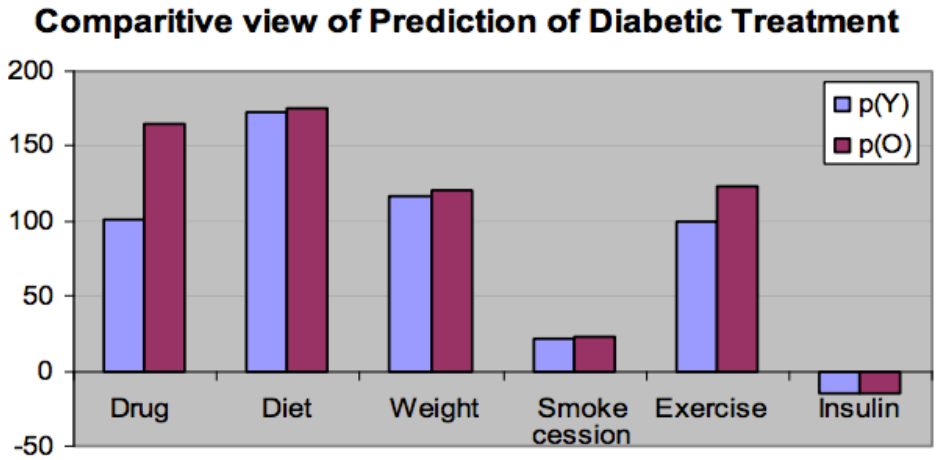


Figure 4: Comparison of p(y) and p(o).

3. Research Design

This chapter demonstrates how the research will approach data mining in the medical field and reveals how results can be beneficial not only in sciences, but also sociological and economical context. This chapter demonstrates the strategies to secure the best predictions of diabetic adolescent patients who are eligible for re-hospitalization. The aim is to assign the predicted adolescent patients to enter a recreational center that indirectly integrates treatment programs, enhances psychological and health awareness, and encourages physical activities and healthy lifestyle to reduce the disease risk factors. The research also aims at reducing the cost expenses of re-hospitalization and increasing the quality of healthcare services. As a first step, data has to be collected, preprocessed and transformed into meaningful input. In the presence of huge number of attributes, feature selection can be helpful. There are two approaches, the fast simple filtering approach, and the wrapper method approach which is shown to be more effective but more costly (Rai & Kumar, 2014). In this research, both the filter approach and wrapper-type method are used for feature selection. The latter uses optimization of attributes either by forward selection techniques or backward elimination. The attributes are discussed by experts in the medical field and weighted by computational evaluation to increase the liability of work. Different classification models are implemented and evaluated using X-validation techniques. Cross-validation

is preferred over data split because, unlike data split, the split ratio of 0.9 is reiterated 10 times with record replacement such that the records are not chosen again, causing more thorough and accurate testing. According to the cross validation results and accuracies, the best performing classification model is used for prediction. Data mining using the best classification technique for the available data may improve the medical situation and may lead to pre-hypothesis, which may be a subject for further investigation in the future (Hofmann, & Klinkenberg, 2013).

The analysis is performed using Rapid Miner as a data mining tool and Microsoft Excel as a platform for testing and comparison. Figure 5 shows the skeletal framework of this research.

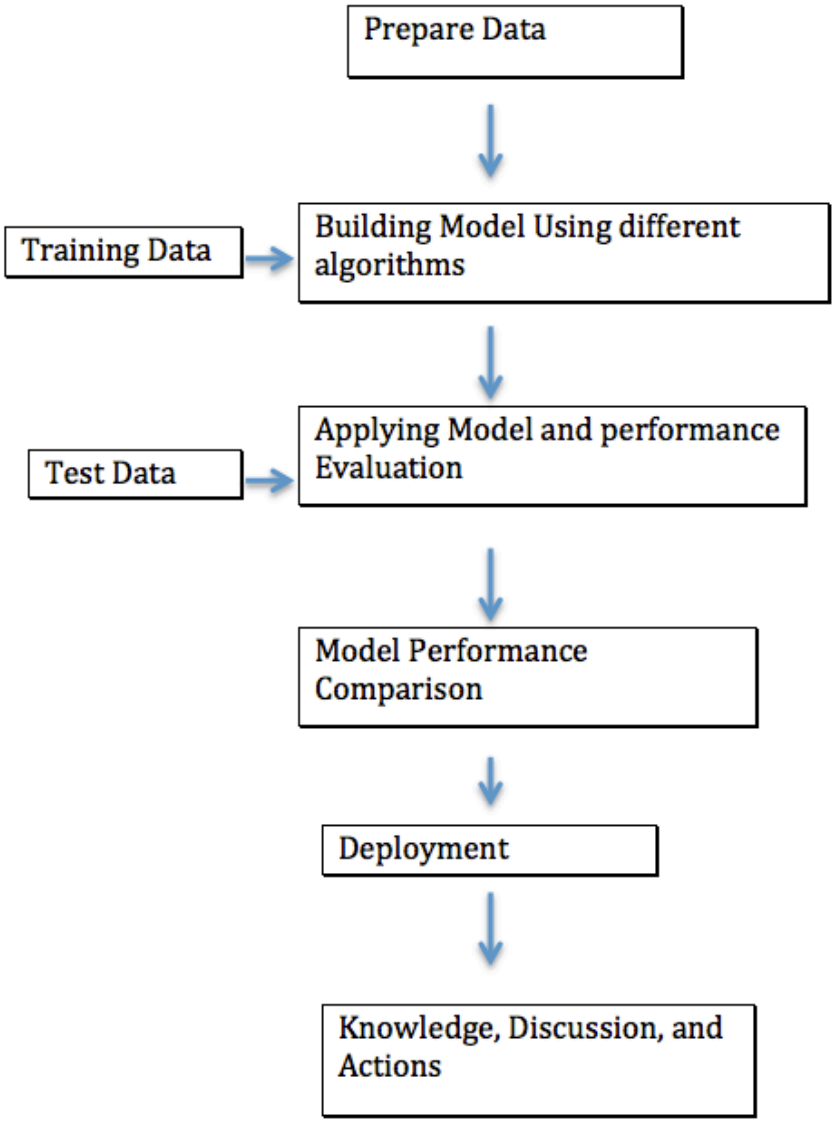


Figure 5: Data Mining Approach

3.1 Data Collection and Description

The data is collected during 2014 and 2015 from a center of diabetes in UAE whose patients vary from children to adults. Only data for adolescent patients, of age group ranging between 10 and 20 years, is extracted for study. This age group is selected for research because a sharp augmentation

of the percent prevalence of diabetes starts from this age according to WHO annual report and is much higher than that of the world as shown previously in figure 3. These patients are referred to the center either by a physician, a clinic, or an emergency room. They have diabetes of type 2, which develops due to the fact that the body does not use insulin properly and therefore, with time, cannot keep blood glucose at normal levels (Association, A.D., 1995). After treatment, they are discharged home, transferred to home with healthcare, or left against medical advice. Some have got the treatment and never got readmitted to the center, whereas some revisited the hospital within 30 days and others are readmitted after 30 days.

The frequent number of hospitalization by the young patients is distressing, substantial, and costly (Garrison, *et al.*, 2005). Reducing the potential cost of readmission to healthcare entities, with simultaneous increase of healthcare quality and guarding the psychological health of the young patients, is the concern of this research. Many features in the retrieved dataset have been recorded for the sake of deep understanding of the diabetic patients' behavior. They are good to record and may not be part of the analysis. The collected data has the following features:

1) Patient's file number which is ignored in the analysis because it acts an identity to the patient and acts only as a reference to the patient.
2) Age range (10-20years) which is a controlled variable for all the example set (all are adolescent)
3) Gender
4) Type of admission (emergency, urgent, elective, or not available)
5) Type of referral (physician referral, clinic referral, transfer from a hospital, transfer from another healthcare facility, or null)
6) Type of disposition (discharged home, discharged/transferred to another short term hospital, discharged/transferred to skilled nursing facility (SNF), discharged/transferred to another type of inpatient

care institution, or null)

7) Number of inpatient and outpatient visits
8) Number of emergencies
9) Number of medicines
10) Number of lab procedures
11) Number of diagnosis (the number of times the patient is diagnosed as diabetic)
12) Blood sugar level at the first visit (diagnosis 1 or Diag_1)
13) A1C level (Glycated Hb test) results that indicate "percentage of blood sugar attached to hemoglobin for the past two to three months" (National
14) of Diabetes and Digestive and Kidney Diseases, 2014.). If the level of A1C > 6.5 in the blood, the patient is diabetic.

    If the results are between 5.7 and 6.4, then the patient is pre-diabetic, and is the result < 5.7, then the patient is normal (National Institute of Diabetes and Digestive and Kidney Diseases, 2014.)
15) Level of insulin hormone that regulates the level of glucose in blood
16) Response of patients to 22 different types of drugs that are used as part of treatment of diabetes type 2 such as a) Biguanides (metformin is the most common drug) that decrease the level of sugar the liver makes and the intestine absorbs, increase the sensitivity to insulin level, and help, therefore, the muscles to absorb glucose, b) Meglitinides that help your body release insulin. However, in some cases, they may lower your blood sugar and cause hypoglycemia, and c) Tolazamide, one of the sulfonylureas that work by stimulating the pancreas with the help of beta cells to produce more insulin. These features agglomerate into one single feature in the original dataset, addressing whether the patients changed medicine during the course of the treatment or not

(CherneyMedically, 2015).

The values of this feature and the level of insulin are either “No” for no effect of the drug on the patient’s body, “Yes” for improvement of the medical situation because of the drug, or “down” for negative effect of the drug on the patients’ case, and “Steady” for stabilization of the health situation.

17) Diagnosed as medically diabetic. This is a categorical feature of either yes or no.
18) Readmissions. There are three attributes: No for no readmission, <30 for readmissions within 30 days, or >30 for readmission after 30 days.

Reducing the hospital readmission number to none is the ultimate goal of the research.

According to the collected dataset, 87% of the patients are medically diagnosed as diabetic. 13% have undiagnosed diabetes. The latter suffer from the disease but are not aware of their health conditions. As opposed to WHO report in 2000 as in figure 2, figure 6 shows 58% of the patients developing diabetes of type 2 are females. This increase among the females is not surprising as many recent studies in UAE show that the females suffer from diabetes, obesity, and other health risk problems more than the males (Mabry, *et al.,* 2010).

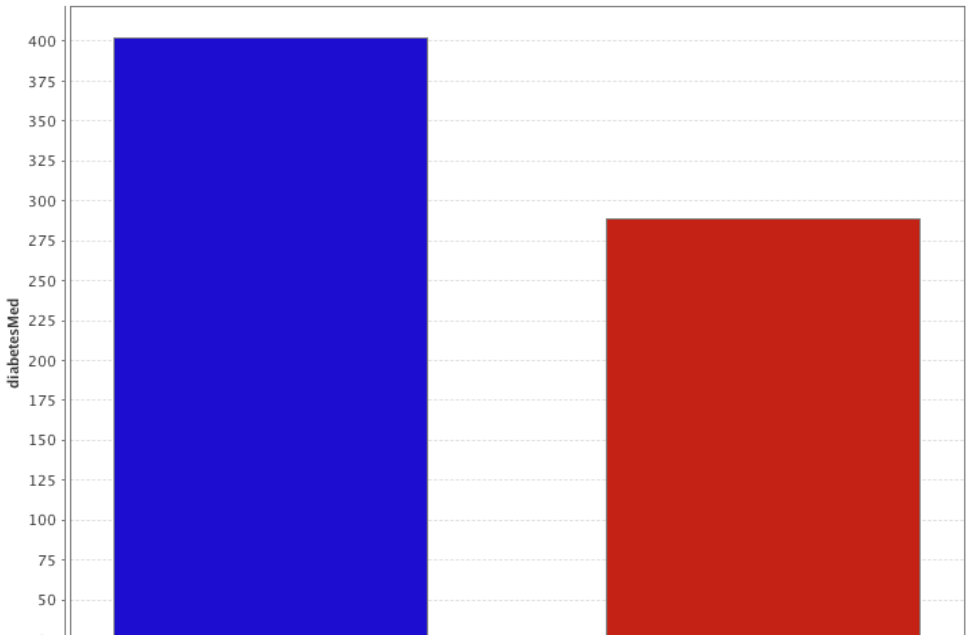


Figure 6: The proportion of diabetic patients based on gender

The proportion of the diabetic patients who are readmitted to the hospital is shown above in figure 7. 32% of the patients are readmitted after 30 days whereas 64% never got readmitted. 36% of the females and 26% of the male diabetic patients are readmitted after 30 days. In addition, the number of females that are readmitted is almost double that of the males as shown in table 1.

Gender	Readmissions		
	<30	>30	No
Females	24	147	231
Males	16	77	196

Table 1: The proportion of readmissions based on gender

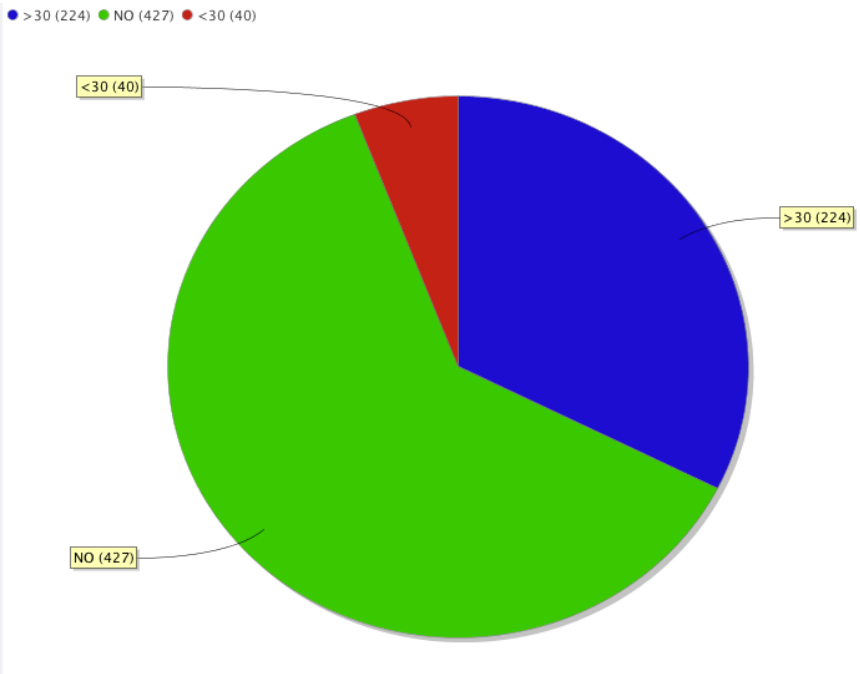


Figure7: the frequency of patients' readmissions

The goal of this research is to decrease the percent probability of readmissions and increase that of no readmissions.

3.2 Data Preparation, Transformation, and Preprocessing

The original dataset is made of 43 features. For simplicity, the dataset is divided into three main subsets as shown in the table 2:
Identity attributes, Medication attributes, and Patient characteristic attributes.

3.2.1 Description of Attributes

Identity Attributes: They are six features. Some, such as file number and encounter id, identify the patients and are not included in the analysis, but they are uploaded during the preprocessing methods for reference.

Medication Attributes: Twenty-two attributes describe the intake of certain prescribed drugs or combined drugs such as metformin, pioglitazone, rosiglitazone, metformin-rosiglitazone, and metformin-pioglitazone. One attribute describes whether the patients have changed their medication or not. These twenty-three attributes can contribute to noisy data that need to be cleaned.

Patient characteristic features: These features count for the behavior, history, medical conditions, and medical lab tests of the diabetic patients under study.

Subset of attributes	Identity Attributes	Medication attributes	Patient characteristic attributes
Number of attributes	6	23	14

Table 2: Division of Attributes

3.2.2 Possible weights assigned to attributes in each subset of features:

The attributes in the three feature subsets are weighted and discussed differently based on their characteristics.

3.2.2.1 Identity Attributes:

File number features are ignored. The age is controlled for the entire example set (10-20). Gender, type of admission, type of disposition, and admission source identify each instance. According to the preprocessing methods discussed below in details, it shows that the admission type is the most significantly relevant attribute. Weight by info-Gain information and Chi Square statistics operators are used in rapid miner to test the relevancy of the identity features with respect to the label "readmitted". Weights are normalized. All the identity features show low values of weights. However, weight by info-gain ratio show that admission type show the highest weight among the identity feature as seen in figure 8 below.

t Overview | AttributeWeights (Weight by Chi Squared Statistic)

attribute	
metformin	0
admission_source_id	0.009
gender	0.024
admission_type_id	0.228
discharge_disposition_id	0.242
number_inpatient	1

Result Overview | AttributeWeights (Weight by Information Gain Ratio)

attribute	
discharge_disposition_id	0.002
gender	0.009
metformin	0.026
admission_source_id	0.064
admission_type_id	0.379
number_inpatient	0.379

Figure 8: Weight of Identity Attributes Based on Info Gain, Chi Square, and Info Gain Ratio

3.2.2.2 Medication Attributes:

These attributes can be noisy. Three approaches are executed to remove the noisy features if any. Since these features are not numeric and are not correlated to each other, principal Component Analysis (PCA) is not used. Instead, Information-Gain feature selection, Chi-square based filtering, and forward selection and backward elimination techniques are used. The three approaches require a label, which is the time it takes the diabetic patient to get readmitted to the hospital.

For all the medication attributes, a value of +1 is given to "Up", -1 for "Down", 0 for "No" and 0.5 for "Steady".

After discretizing the nominal features into 2 bins, info-gain based filtering is executed and normalized weights of the first four features are shown in table 3.

Attribute	Info Gain Weight (Not Discretized)	Info Gain Weight (Discretized)
Metformin	1	1
Rosiglitazone	0.591	0.591
Glyburide	0.394	0.394
Pioglitazone	0.394	0.394
Glipizide	0.394	0.394

Table 3: the most relevant drug attributes based on Information Gain

After discretizing the nominal features into 2 bins, chi-square test-based filtering is executed and normalized weights of the first four features are shown in table 4. The table shows that attribute discretization has no effect on the results.

Attribute	Weight by Chi-Square (Not Discretized)	Weight by Chi-Square (Discretized)
Metformin	1	1
Rosiglitazone	0.422	0.422
Change	0.386	0.386
Pioglitazone	0.281	0.281
Glipizide	0.281	0.281

Table 4: the Most Relevant Drug Attributes Based on Weight by Chi-Square

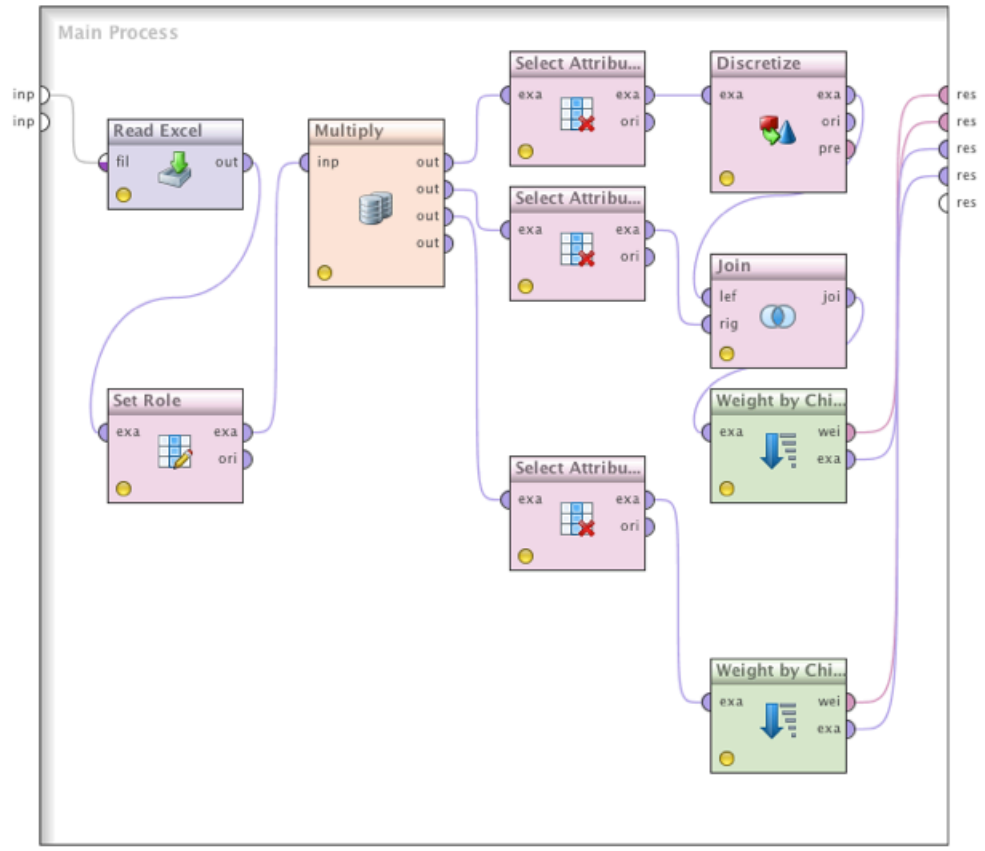


Figure 9: Weight of Drug Attributes by Chi Square Used in Rapid Miner

The rapid miner operator flow used to achieve feature selection preprocessing is shown above in figure 9. The "Discretize" operator is disabled when no discretization is to be performed.

Wrapper-type feature selection method is used to reduce the number of attributes and modify the weights to fit best the predictive models. Backward elimination and forward selection are optimization selection techniques used in a nested learning process that includes a predictive learning algorithm such as SVM, neural networks, and regression in case of numerical labels, and decision trees, random forests, Naïve Bayes, and KNN in case of categorical label attributes. The latter classification models are used and forward and backward elimination feature selection reveal that metformin is the drug that should be kept as a medication feature in the dataset, as shown in table 5.

Attribute	Attribute Weight (forward selection)	Attribute Weight (backward selection)
Metformin	1	0
Rosiglitazone	0	1
Change	0	1
Pioglitazone	0	1
Glipizide and others	0	1

Table 5: Attribute weight using forward and backward selection techniques

Asking experts and endocrinologists, metformin medication is acknowledged as the first choice as that drug is the first medication prescribed for type 2 diabetes, especially for adolescents, due to the two

effective functions: 1) Improves the sensitivity of the body tissues to insulin so that the body uses insulin more effectively. 2) Lowers glucose production in the liver. On the contrary, the other drugs either help the body secrete more insulin or stimulate the pancreas to secrete more insulin. The latter acts fast and maintains a shorter effect. They may cause low blood sugar and weight gain. Expectedly metformin, as a medication to diabetic patients, causes more homeostasis than other drugs.

3.2.2.2 Patient characteristic attributes:

They count for 14 attributes. 'Readmission to hospital' feature is categorical and acts as a label attribute. Maintained Insulin level, A1C test, medical diagnosis as diabetic or not, and gender are the three categorical characteristic features. All the others are numerical. Running different feature selection techniques and checking their performance, the weights of different attributes revealed that a) gender has a higher weight according to info-gain and wrapper-type forward and /or backward elimination techniques, b) A1C according to chi square, c) diag-1 according to weights by PCA. Executing the PCA itself with a variance of 95%, no attribute has less cumulative variance than 95%. This is explained as all are irrelevant or noisy or all are equally relevant. According to experts, these features are relevant to the patients' case.

As a conclusion, to avoid over-fitting of data, all the numerical and categorical patient characteristic attributes are considered in the analysis procedure. Figure 10 summarizes the number of attributes that are most relevant in the study.

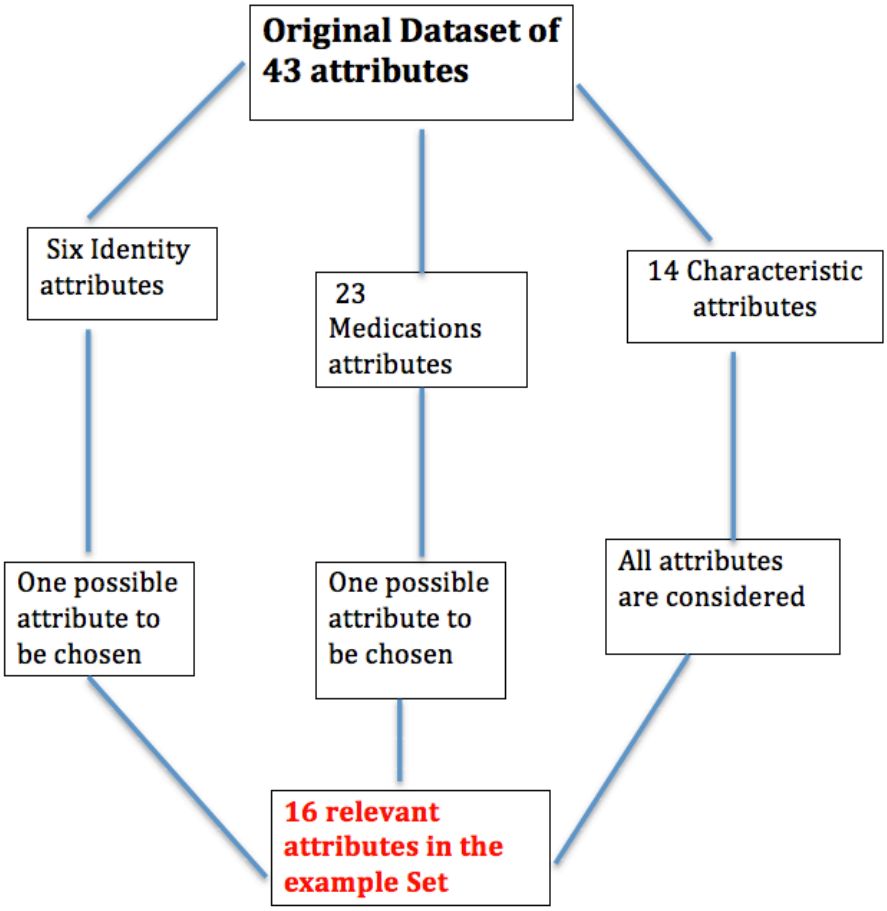


Figure 10: Summary indicating the number of the most relevant features

3.3 Features Selection

The most relevant features for study are shown in table 6. The different values of each attribute are displayed in table 7. Table 6 shows the dataset features, type of data and description for each feature as taken from the database:

Feature	Data type	Description
Gender	Binomial	Patients' gender • Male • Female
Time_ in_hospital	Integer	The greatest number of hours the patient stayed in hospital per visit
Number_lab procedures	Integer	Number of Lab procedures and medical tests done
Number_ medications	Integer	Number of medications taken or prescribed
Number _ outpatient	Integer	Number of times the patient attends a medical center without staying there overnight
Number_ inpatient	Integer	Number of times the patient slept overnight at least one night
Number_of Emergencies	Integer	Number of times the patients go to emergency room in the hospital
Diag_ 1	Integer	Glucose level detect in first diagnosis
Number_ diagnosis	Integer	The number of times the patient is diagnosed
A1C result	Polynomial	Glycated Hb test that indicates percentage of blood sugar attached to hemoglobin for the past two to three months

Metformin	Polynomial	The type of medication the triggers the secreted insulin to act on sugar and stop the liver from producing sugar
Insulin	Polynomial	The level of the hormone the regulate the amount of glucose in the blood
DiabetesMe	Binomial	Medically diagnosed as diabetic or not. All the patients in the dataset are diabetic, but some know that they are diabetic and visit the center and some others are diabetic but are not aware that they already have diabetes (this is medically called undiagnosed diabetes). Yes is given for patients are already aware of their diabetes and No for the undiagnosed diabetic patients.
Readmitted (The label)	Polynomial	The time it takes the diabetic patient to get readmitted to the hospital

Table 6: Description of Relevant Attributes

The above-mentioned features have different values as displayed in table 7

Feature	Data Type	Min Values	Max Values	Average
Gender	Binomial	1 = female	0 = male	
Time_ in_hospital	Integer	1	14	3.18

Number_lab procedures	Integer	1	103	43.158
Number_ medications	Integer	1	34	8.27
Number _ outpatient	Integer	0	7	0.17
Number_ inpatient	Integer	0	11	0.538
Number_of Emergencies	Integer	0	7	0.156
Diag_ 1	Integer	8 mg/dL	998 mg/dL	338.736
Number_ diagnosis	Integer	1	9	3.94
A1C result	Polynomial	<5.4 for low	6 for norm	>7, >8 for high
Metformin (how the patients' body react to metformin)	Polynomial	-1 for down	+1 for up	0 for No and 0.5 for steady
Insulin	Polynomial	-1 for down	+1 for up	0 for No and 0.5 for steady
DiabetesMe	Binomial	Yes = 1 for diabetic	No = 0 for undiagnosed diabetic	

Readmitted (The label)	Polynomial	No for no hospital readmission	>30 for re-hospitalization after 30 days	<30 for readmission within 30 days

Table 7: Values of Relevant Features

3.4 Size of Sample Under Study:

The original dataset is made of 691 instances where each instance stands for an adolescent patient (of age ranging between 10 and 20 years). The patient is labeled as never readmitted to the hospital, re-admitted to hospital within 30 days, or readmitted later than 30 days. For analysis and validation purposes, stratified sampling is used to select 10% of the population as a test-set and 90% as a training set. The instances are replaced in the dataset and another 10% of stratified sample test set is verified against the training set. This process is a 10-fold evaluation (X-validation) process.

3.5 Predictive Modeling: Classification

Different Data Mining techniques can be implemented to increase the popularity of knowledge discovery in databases (KDD) in many fields and crucially in medicine and in the medical healthcare, where unusual trends can lead to valuable information that alters human's destiny of life and death. To reveal some valuable hidden knowledge and hidden relationships between different parameters, many classification techniques and combination of different algorithms can be executed (Bhardwaj, & Pal, 2012). In this paper, different supervised and unsupervised predictive

techniques are implemented to predict the probability of whether the patients will be readmitted to hospital or not. In addition to clustering, Decision trees, Artificial Neural Network, k-Nearest Neighbor, Naive Bayes, and ensemble learning algorithms have been used.

3.6 Predictive Model Evaluation

Feature reduction and selection can be can be part of data preprocessing before running any model. Percent accuracy and percent recall are the evaluation measures used to assess the liability of each predictive models performance from the confusion matrix as shown below:

Predicted Class		Actual Class		
		True	False	
	Positive	True Positives	False Positives	Percent Prediction
	Negative	False Negatives	True Negatives	
		Percent Recall		

Accuracy = (TP + TN) / (TP + TN + FP + FN), where the more accurate is the prediction, the better is the model predictive power.

Therefore, a comparative analysis is conducted to compare the prediction accuracy of each classification model. The most accurate models are the most powerful models and are used to participate in ensemble learning in an

attempt to improve model prediction the most. The results may be useful to meet the goals of research, contribute to new statistical analytical approaches such as forming and testing new hypothesis, and find solutions to problems (Shillabeer, & Roddick, J.F., 2007). Not to forget, that in medical field, the minority relationships might trigger formation of some hypothesis for investigation (Cios, & Moore, 2002).

4. Classification Models And Performance Evaluation

This section covers the performance of different classification models based on the available dataset. I used various classification models and ensemble learning models to detect the best predictive model. Decision tree based on different criteria such as accuracy, Gini index, Gain ratio, and Information gain, is evaluated. Artificial Neural Network, K-nearest neighbor, Naïve Bayesian are also run on the dataset. Based on the performance of the above-mentioned classifiers, I tried to use ensemble-learning techniques that can be a combination the best performing algorithms.

4.1 Decision Tree:

Decision trees are one of the most common predictive classification models used in data mining. This section is split into two parts: 1) Description of the model and 2) Feature selection and model performance.

4.1.1 Description of the Model:

Decision trees are one of the most intuitive and frequently used data mining techniques due to their capability of capturing underlying relationships

(Bhardwaj, & Pal, 2012). They are easy to set up and easy to interpret. The decision tree takes a form of the flowchart where an attribute is tested at each node. The leaf node at the end of the tree is where the prediction is made about the target variable. Decision trees do not only help in predictive classification, but also in data visualization (Bhardwaj, & Pal, 2012). Based on the data, only post-pruning is allowed to stop over-fitting of data. Without pre-pruning and post-pruning, the decision tree is huge and complex (over-fitting is very evident). With pre-pruning only and with pre-pruning and post-pruning, the decision tree is too small indication strong under-fitting of data. The data is split into subsets based on the homogeneity of data and on reduction in data uncertainty. The split can be based on four different criteria: information gain, gain ratio, Gini index, and accuracy.

1) Information gain calculates the information before split minus the information after split. Information gain is biased towards choosing the higher values of attributes as the leaf nodes.

2) Gain ratio is a modification of information gain. The attributes with highest uncertainty tend to offer low gains upon splitting and will not be selected.

3) Gini index: measures the impurity in the categorical target variables.

4) Accuracy: promotes the selection of the attribute that maximizes the accuracy of the whole tree.

4.1.2 Feature Selection and Model Performance:

For such a dataset where there are many features per instance, it is wise to perform unsupervised clustering before performing the decision tree classification technique. Entities that belong to the same class or cluster are expected to behave more similar to each other than the entities that belong to the other classes. Figure 11 below shows how the data is clustered among

males and females based on the label categorical variable, where “>30” stands for readmission after 30 days, “<30” readmission within 30 days, or “None” for no admission. The missing values operator is used to replace missing attributes with the average; the nominal attributes are converted to numerical values to enhance clustering. “Select attribute” operator is added to perform the simple selection technique, and “optimization selection” operator is to select the most relevant attributes of the given example set. Two deterministic greedy feature selection algorithms 'forward selection' and 'backward elimination' are used for feature selection.

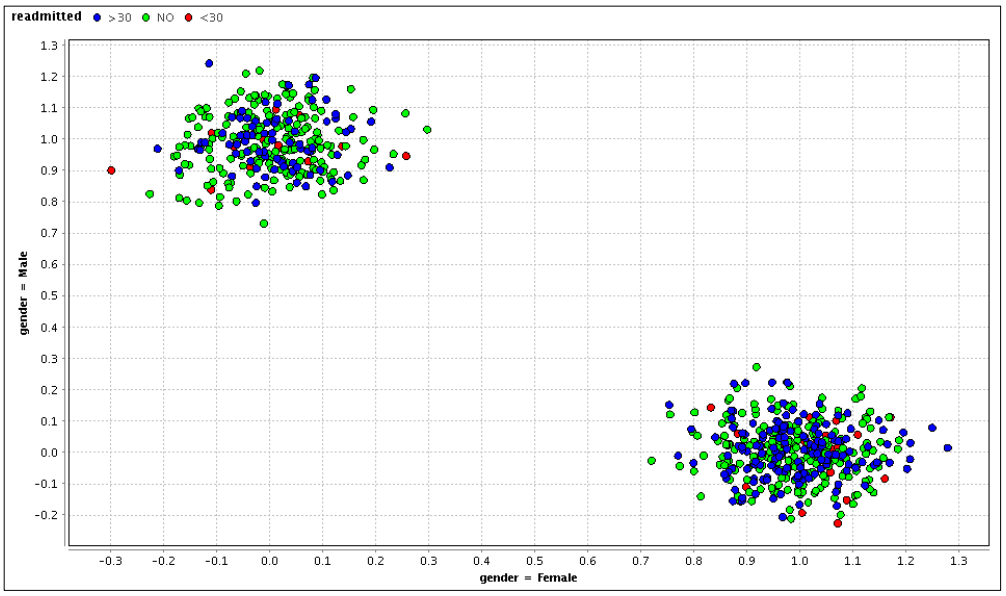


Figure 11: Clustering of diabetic patients based on readmission to hospital and gender

After clustering the data into two clusters (K=2) as shown in the above figure, decision tree classification is performed considering the four criteria previously discussed. “Select attribute” operator is added to perform the simple selection technique, and “optimization selection” operator is to select the most relevant attributes of the given example set. Two deterministic greedy feature selection algorithms 'forward selection' and 'backward elimination' are used for feature selection. The percent accuracy from the

confusion matrix of the cross validation is recorded when the example set is input with no feature selection, with simple filtered feature selection based on the previous discussion, and with optimized feature selection using forward and backward elimination. The first case is when the input data is the original dataset; the second example set is the filtered set of attributes based on their relevance as per the earlier data preprocessing discussion. “Select attribute” operator is added to perform the simple selection technique. The input data can be also optimized in an attempt to increase the percent accuracy of the model performance without enforcing any over-fitting of data. The “optimize selection” operator selects the most relevant attributes of the given example set via the nested algorithm. Two deterministic greedy feature selection algorithms 'forward selection' and 'backward elimination' are used for optimization feature selection.

Forward Selection

This process starts with an initial population with n individuals where n is the number of attributes in the input example set, is created. Each individual will use exactly one of the features, evaluate the attribute sets and select only the best k, where k means keep best parameter. For each of the k attribute sets, if there are j unused attributes, j copies of the attribute set are made and **exactly one of the previously** unused attributes is added to the attribute set to get to the next iteration if improvement occurs.

Backward Elimination

This process starts with an attribute set which uses all features. Evaluate all attribute sets and select the best k. For each of the k attribute sets, if there are j attributes used, j copies of the attribute set are made and **exactly one of**

the previously used attributes is removed from the attribute set. As long as the performance improved, the process is reiterated.

The process design of the decision tree is manifested in the figure below.

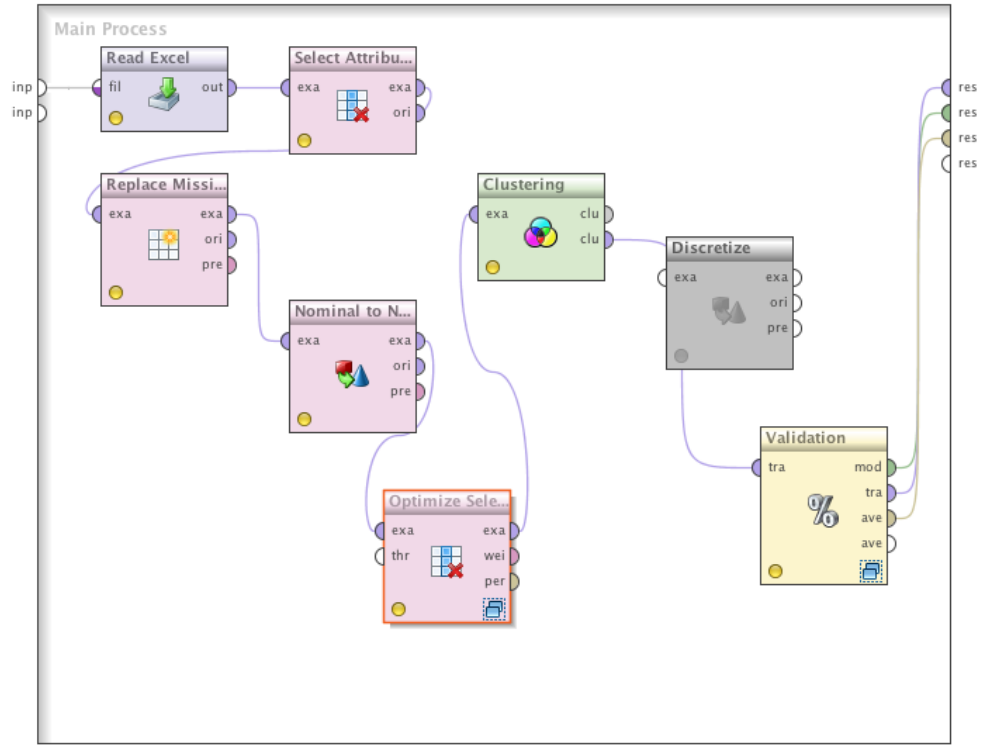


Figure 12: Decision Tree Process Design with Optimization of features

The performance of the decision tree is summarized in the table below when all the features are used in the original data set, when simple feature selection is selected, and when complex feature selection is optimized using the forward selection and backward elimination.

Decision Tree Model Performance,	Percent Accuracy (Without Discretization)		
	Simple Feature	Simple Filtered	Complex feature Selection

where the label is categorized into >30, <30, and No	Original dataset	Feature Selection	Forward Selection	Backward Elimination
Decision tree using gain ratio as a criterion	60.43%	62.75%	62.90%	59.57%
Decision tree using information gain as a criterion	62.75%	60.29%	60.00%	60.43%
Decision tree using Gini index as a criterion	62.17%	60.00%	62.75%	62.17%
Decision tree using accuracy as a criterion	63.62%	63.62%	63.48%	63.19%

Table 8: Results of the decision tree without discretizing the input data

Since diag_1 feature, that measures the level of glucose free in the blood, is numerical (continuous) and can be discretized easily into three bins, "discretization by frequency" operator is added to discretize the single attribute (diag-1) as part of preprocessing in the data mining processes. This attribute has values ranging widely from 8 to 998 mg/dL. The other numerical attributes are numerical but are not continuous. Choosing the range name type of the three bins that the values of diag_1 are discretized to as "short" not interval, there are three ranges.

Range 1 = [∞, 250.1] and its absolute count is 245, Range 2 = [250.1, 250.5] and its absolute count is 215 instances. Range 3 = [250.5, ∞]. And

its absolute count is 230 instances.

Table 9 below displays the results of the decision tree performance after discretization and using the four split criteria. Table 9 below displays the results of the decision tree performance after discretization and using the four split criteria.

Decision Tree Model Performance, where the label is categorized into >30, <30, and No	Percent Accuracy after discretization			
	Simple Feature Original dataset	Simple Filtered Feature Selection	Complex feature Selection	
			Forward Selection	Backward Elimination
Decision tree using gain ratio as a criterion	64.06%	63.48%	60.43%	61.59%
Decision tree using information gain as a criterion	58.84%	60.58%	58.99%	58.84%
Decision tree using Gini index as a criterion	58.70%	61.16%	59.98%	58.70%
Decision tree using accuracy as a criterion	61.88%	63.62%	64.20%	62.46%

Table 9: Results of the decision tree after discretizing the input data

Although discretized data in table 9 show higher accuracy percentage (64%), the results are not convincing as none of the outcomes is able to

predict the patients who are readmitted within 30 days (0% recall) as shown in the confusion matrix below (figure 13).

accuracy: 64.20% +/- 3.84% (mikro: 64.20%)

	true >30	true NO	true <30	class precision
pred. >30	64	47	14	51.20%
pred. NO	159	379	26	67.20%
pred. <30	0	1	0	0.00%
class recall	28.70%	88.76%	0.00%	

Figure 13a: Performance Evaluation of Decision Tree of Discretized Input Data.

Table View Plot View

accuracy: 63.77% +/- 4.20% (mikro: 63.77%)

	true >30	true NO	true <30	class precision
pred. >30	74	60	12	50.68%
pred. NO	149	366	28	67.40%
pred. <30	0	1	0	0.00%
class recall	33.18%	85.71%	0.00%	

Figure 13b: Performance Evaluation of Decision Tree of Non-discretized Input Data.

I had two available options: 1) down-sampling from the other two classes to have a balanced data for more experimental comparison. 2) Transformation of the data into a new dataset where each patient is labeled as either readmitted or not.

Balancing Number of Instances: The objective now is to equalize the number of instances in each class. The dataset is multiplied in a way that the each dataset is example-filtered with the same sample size. I added the append operator to merge the three samples into one sample for training purposes. The data-mining process is demonstrated in the figure 14 below.

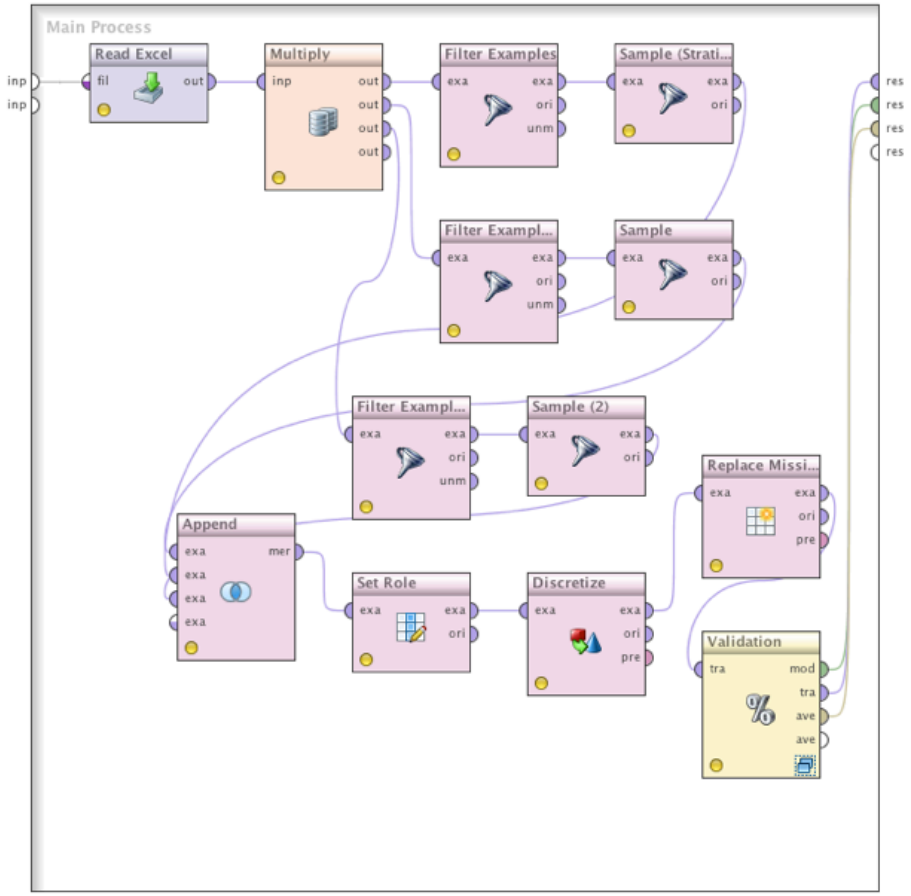


Figure 14: Rapid Miner Process of Balancing Data in a Dataset by Down-sampling

Running above process with and without discretization of data results in non-zero low percent recall for the three classes as shown in the confusion matrices below (figure 15a, 15b)

Table View Plot View

accuracy: 38.33% +/- 11.90% (mikro: 38.33%)

	true >30	true NO	true <30	class precision
pred. >30	15	11	11	40.54%
pred. NO	13	15	13	36.59%
pred. <30	12	14	16	38.10%
class recall	37.50%	37.50%	40.00%	

Figure 15a: Performance Evaluation of Decision Tree of Non-discretized Balanced Input Data

accuracy: 30.00% +/- 12.47% (mikro: 30.00%)

	true >30	true NO	true <30	class precision
pred. >30	12	16	13	29.27%
pred. NO	17	11	14	26.19%
pred. <30	11	13	13	35.14%
class recall	30.00%	27.50%	32.50%	

Figure 15b: Performance Evaluation of Decision Tree of Discretized Balanced Input Data

Transformation of Data into New Data-Set: The same previous procedure is repeated with the new data set that contains all the instances with the label attribute is categorized into two classes: Yes for readmission and No for no readmission. The results show the following:

Decision Tree Model Performance, where the label is categorized into Yes and No	Percent Accuracy (Without Discretization)			
	Simple Feature Original dataset	Simple Filtered Feature Selection	Complex feature Selection	
			Forward Selection	Backward Elimination
Decision tree using gain ratio as a criterion	67.10%	66.23%	66.09%	66.09%
Decision tree using information gain as a criterion	62.32%	64.49%	62.61%	62.32%
Decision tree using Gini index as a criterion	61.17%	64.20%	61.01%	61.16%
Decision tree using accuracy as a criterion	66.52%	66.23%	66.32%	66.65%

Table 10: Results of the decision tree without discretizing the transformed input data

Decision Tree Model Performance, where the label is categorized into Yes, and No	Percent Accuracy after discretization			
	Simple Feature Original dataset	Simple Filtered Feature Selection	Complex feature Selection	
			Forward Selection	Backward Elimination
Decision tree using gain ratio as a criterion	65.80%	66.38%	64.06%	63.16%
Decision tree using information gain as a criterion	61.59%	63.91%	61.74%	61.59%
Decision tree using Gini index as a criterion	61.45%	64.06%	62.03%	61.45%
Decision tree using accuracy as a criterion	65.80%	67.39%	68.26%	65.80%

Table 11: Results of the decision tree after discretizing the transformed input data

The decision tree using accuracy as a criterion, with discretized data and optimized feature selection shows the highest results as shown in table 11. The relationship between the different features shows interesting patterns and trends as shown below in figures 16a and 16b. According to the resulted decision tree, the featured attributes to label as readmitted or not are number_inpatient, number_outpatient, Diag_1, and A1C result.

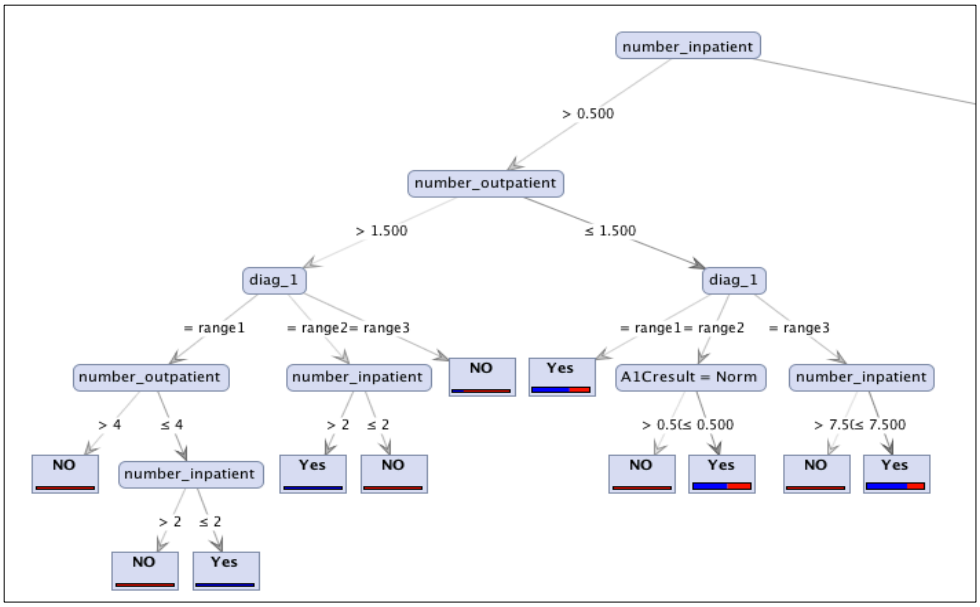


Figure16a: Part of the Best Preforming Decision Tree When Number_Patient ≥ 0.5.

The decision tree is clear to interpret and read predictions. For example according to the split f the decision tree, if a number of outpatient visits is more than 1 and when the patient first came to the center and diag_1 is in range 2([250.1-250.5], we can look at the in_patient number. If the latter is >2, then that patient will be readmitted.

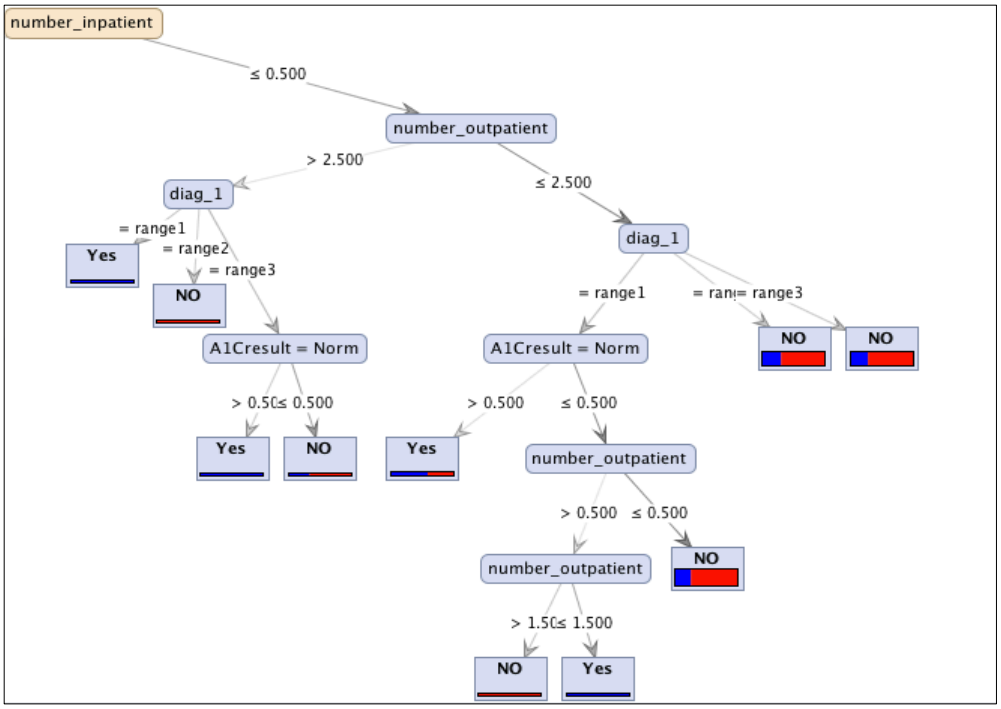


Figure16b: Part of the Best Preforming Decision Tree When Number_Patient ≤ 0.5.

Although the above model shows relatively higher performance accuracy, the percent recall as shown below for true yes is modest (figure 17).

Table View Plot View

accuracy: 68.26% +/- 3.52% (mikro: 68.26%)

	true Yes	true NO	class precision
pred. Yes	115	71	61.83%
pred. NO	148	356	70.63%
class recall	43.73%	83.37%	

Figure 17: Performance Evaluation of Decision Tree of Discretized Transformed Input Data

Therefore, it is wise to balance the data set for comparison before we try another predictive model that may increase the percent recall. Down sampling is an option, where each class has equal number of instances. The confusion matrix of the decision tree performance model before and after discretization is shown below.

Table View Plot View

accuracy: 61.77% +/- 7.34% (mikro: 61.79%)

	true NO	true Yes	class precision
pred. NO	165	103	61.57%
pred. Yes	98	160	62.02%
class recall	62.74%	60.84%	

Figure 18a: Performance Evaluation of Decision Tree of Non-discretized Balanced Transformed Input Data

accuracy: 64.27% +/- 4.51% (mikro: 64.26%)

	true NO	true Yes	class precision
pred. NO	182	107	62.98%
pred. Yes	81	156	65.82%
class recall	69.20%	59.32%	

Figure 18b: Performance Evaluation of Decision Tree of Discretized Balanced Transformed Input Data

There is small improvement in the percent of true negative (around 15% in both cases), but there is also drop in the percent recall of the true positive (around 18%). The percent accuracy of the balanced dataset is still less than that of the unbalanced transformed dataset.

4.2 Artificial Neural Network Model:

Whereas decision tree model assigns no weights to attributes, Artificial Neural Network (ANN) model does. ANN is a classification model that adjusts weights to attributes to give the best output. This section covers a detailed description of the model and a display of the model performance.

4.2.1 Description of the Model

Neural network is a useful classification technique when there is nonlinear relationship between the many different attributes that act as input data and the output label attribute. The model acts on adjusting weight to input attributes in stages to produce the output results. The model learns through adaptive adjustments between the nodes. It is a combination of nonlinear

logistic regression mathematical relationships between input and output attributes called perceptrons. The model uses back-propagation techniques for classification and the relationship between attributes is hard to explain as the process acts as a black box in the hidden layers. ANN is made of input, hidden and output nodes as shown in the topological model below.

The greater the number of hidden layers and the larger the number of attributes, the slower is the training time and therefore the performance cost.

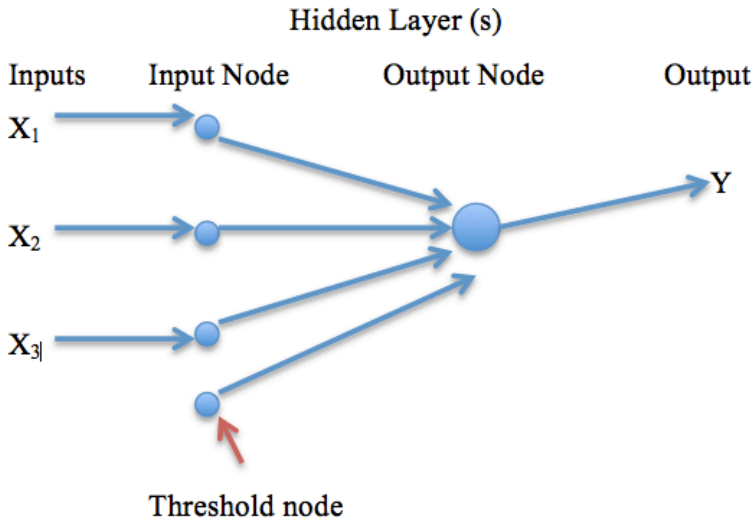


Figure 19: Model Topology of Artificial Neural Network

4.2.2 Feature Selection and Model Performance:

Artificial neural network requires strict processing as the model cannot handle categorical input data and does not function if there are any missing values. Therefore, the binomial attributes such as gender has been converted into binary values and polynomial A1C test results have been converted into numerical. ANN has showed best results of the data when the number of cycles is 1000, the learning rate is 0.3 and momentum is 0.2. The learning rate measures the extent to which the error in the previous cycle contributes

to the attribute weight in the current cycle. The closer the learning rate is to zero, the more the new weight depends on the previous weight and less on error correction. Momentum seeks to obtain globally optimized results.

ANN is run on the original dataset and on the preprocessed (filtered) data set (simple feature selection). Using ANN, optimization feature selection is redundant here as the model itself adds weights to the input attributes and adjusts weights to learn most from data. On designing the ANN data mining process, two hidden layers of six nodes each are chosen. More than two hidden layers are avoided because the execution time will be very long. The number of nodes per hidden layer is chosen based on the best model performance using cross validation techniques and in an attempt to avoid over-fitting.

The performance of the model using the original and the filtered data sets, with and without discretization, is shown in table 12.

Percent Accuracy	Original Dataset	Filtered Dataset
Discretized input data	62.61	61.59
Not discretized Input Data	61.30	59.71
Execution Time	More than 1 min	Less than 1 min

Table 12: ANN Model Performance

The accuracy results are modest ranging between 59.71% and 61.59% for the filtered dataset, and ranging between 61.30% and 62.61% for the original dataset. In spite of the low performance, only the non-discretized input dataset showed only 5% recall for patients with readmission within 30 days. Others show 0% recall. Similar trend is applied to eliminate the null recall. The label attribute is categorized into “No” for no readmission and

"yes" for readmission, irrespective of the elapsed timespan before re-hospitalization.

The neural net model is displayed in figure 20 and the percent accuracy improved with the highest is for simple feature selection without discretization as shown in the figure 21.

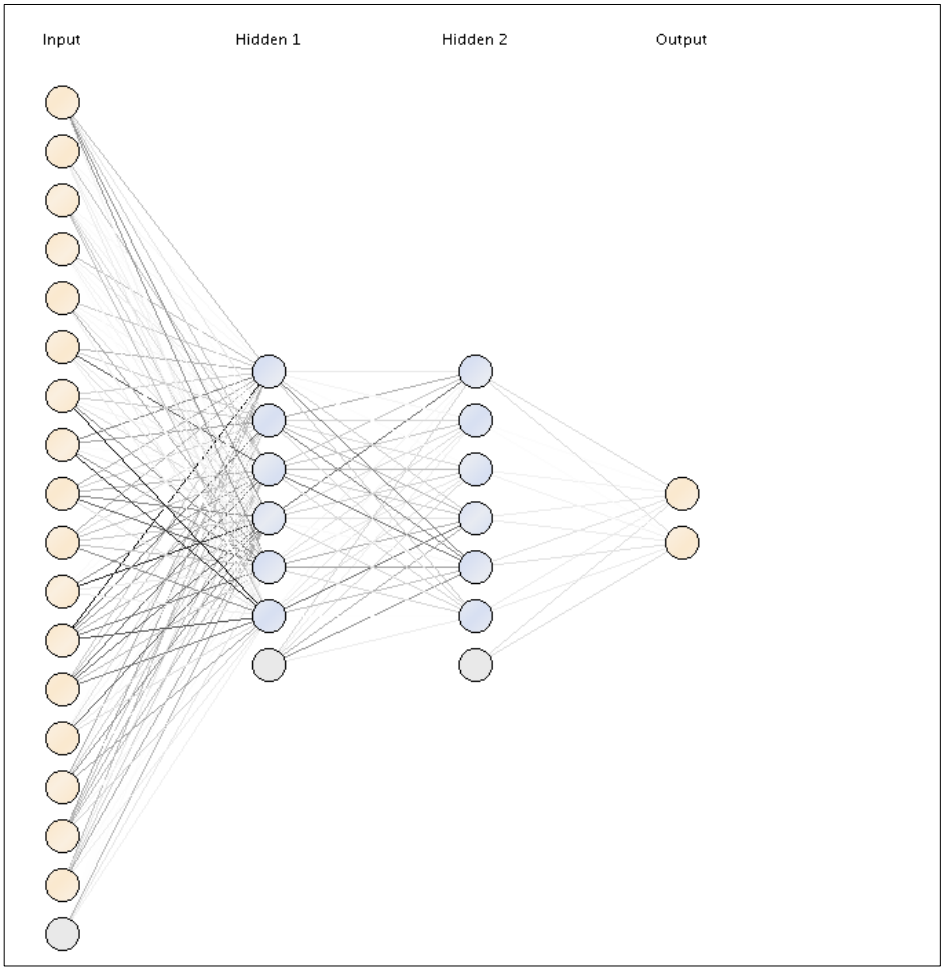


Figure 20: Artificial Neural Network Performance

accuracy: 64.49% +/- 5.35% (mikro: 64.49%)

	true Yes	true NO	class precision
pred. Yes	103	85	54.79%
pred. NO	160	342	68.13%
class recall	39.16%	80.09%	

Figure 21: Confusion Matrix of ANN Model Performance

In addition to the long execution run time, the confusion matrix above shows that 245 out of 690 examples are misclassified. The percent accuracy is less than that of the decision tree. Therefore, another predictive classification model can be tried and implemented.

4.3 K-nearest Neighbor:

K-nearest neighbor is an unsupervised learning model that can predict output based on different similarity measures among attributes. This section demonstrates the criteria on which the classification models work and the model performance evaluation.

4.3.1 Description of the Model:

KNN is a learning predictive model where the entire training set is memorized and when unlabeled test set is required to be classified, KNN compares with the training set to find a closed match (the predicted label). The key task in the KNN algorithm is the determination of the nearest record from the unlabeled test record using a measure of proximity of attributes. KNN is called a lazy learner as the relationship between input and output attributes is not explained. However, there are different techniques to find the similarity between records, such as distance, correlation, Jaccard similarity, and cosine similarity.

4.3.1.1.Distance Between Points:

For numerical data, the two dimensional distance between any two points is measured using the Euclidian distance which is defined as the square root of the sum of the squares of their coordinates. In regards with the n-

dimensional points, Minkowski distance measure is used. The Euclidian distance however is more common.

For binary attributes, Manhattan distance (Hamming distance) is used.

For categorical attributes, the distance will be either 0 or 1. If the attributes are the same, the distance is 0, otherwise 1. Therefore, converting the data into numeric gives more information.

4.3.1.2 Similarity Measures:

For Numeric data, Correlation similarity is used as the measure of linear relation between the attributes of two data points. The Pearson correlation value ranges between -1 (perfect negative correlation) and +1 (perfect positive correlations) where 0 indicates no correlation.

For binary attributes, simple matching coefficient (SMC) is applied. This similarity measure is based on the simultaneous matching occurrence (0 or 1) with respect to total occurrences.

Jaccard similarity is used when a record acts as a text document where each word acts as an attribute. Jaccard measures the similarity between two text documents by finding the ratio of the common occurrences of the same word with respect to total occurrences. Jaccard similarity is similar to SMC but nonoccurrence frequency is ignored.

Cosine similarity is also used for text attributes where attributes represent either the presence of a word or its absence (1 or 0).

In KNN implementation, the cosine similarity is the most common similarity measure for categorical attributes.

4.3.2. Feature selection and performance:

KNN works best with numerical input where Euclidian distance can be measured. In such case, normalization is required to avoid any bias by any attribute. KNN can handle categorical attributes that range between 0 for

different attributes and 1 for the same. KNN cannot handle any missing values. If the values in the test record are missing, the entire attribute is ignored in the model. Replace missing values operator is selected in processing the data. " Readmitted" is the label attribute selected. Mixed measures are used for measuring the distance between attributes.

If the missing values are not replaced, the attributes with missing values are omitted from modeling, so the dataset with missing values have fewer attributes to consider for analysis using K-nearest neighbor techniques. KNN of the raw dataset showed high accuracy because some important attributes having missing values are deleted in the algorithm because KNN cannot handle data with missing values. This percentage, in spite of its higher accuracy, is most probably misleading. The feature that has missing values is deleted; in this case the features present to predict are less and therefore the KNN performance does not have the enough and significant features or data for the reliability.

The model showed better performance when the number of nearest neighbors (K) of the unseen data is three as shown in the table below. Since K>1, weighted vote setting needs to be selected to obligate the nearest neighbor to be designated. Performance of the model before processing the data, after filtering attributes, and wrapper-type feature selection is displayed in table 13a and table 13b.

Model Performance where the readmission is categorized into >30, <30, and No	Percent Accuracy			
	Simple Feature Original dataset	Simple Filtered Feature Selection	Complex feature Selection	
			Forward Selection	Backward Elimination
K=2	54.64 %	54.35%	52.90%	54.35%
K=3	57.25%	58.26%	58.84%	59.42%

Table 13a: Model Performance Where the Readmission is Categorized into Three Labels

Model Performance where the readmission is categorized into 2 labels (Yes, or No)	Percent Accuracy			
	Simple Feature Original dataset	Simple Filtered Feature Selection	Complex feature Selection	
			Forward Selection	Backward Elimination
K=2	60.00 %	60.00 %	58.70%	60.00 %
K=3	62.90%	65.22%	60.43%	63.48%

Table 13b: Model Performance Where the Readmission is Categorized into Two Labels

KNN shows best results when K=3 with simple feature selection. The percent recall for true
Yes is 50.19% and that of true No is 74.47%. The former percentage is interesting result and higher than that in the previous predictive models performed. Compared with artificial neural network, KNN has higher predictive power for readmitted and lower for no readmission.

4.4 Naïve Bayes Classification:

Thinking of data, it is curious to know to what extent the attributes are considered independent features and can be classified independently. The well-known Naïve Bayesian classification model answers such contemplations by using probability measures. Description and performance of the model are discussed below.

4.4.1 Description of the Model:

Similar to all classification algorithms, no matter how different are their techniques; Naïve Bayesian model aims at prediction of a target variable. Naïve Bayes classification measures the posterior (conditional) probability, which is the probability of an outcome variable given the values of input variables (evidence). Since this model uses the mathematical approach of probability in its task, some assumptions need to consider. 1) The input features are independent. 2) For every example in the test dataset, there should be an example in the training dataset. Otherwise the posterior probability will be zero. 3) The latter limitation will prevail if the attributes are continuous. In such a case, discretization of continuous variables is implemented.

4.4.2 Feature selection and performance:

The dataset is split into a training set (0.9) and a test dataset (0.1) using the 10-fold cross validation technique. Stratified sampling is chosen to select the 10% test dataset and 90% the training data set. The sample is tested against the training set and this procedure is repeated for 10 times. Since the model works well with continuous and categorical attributes, the data needs no transformation. Only discretization of continuous attributes is applied.

The model shows 61.88% accuracy for the non-processed dataset and 60.58 % for the filtered input dataset. Categorizing the output data into two labels, admitted or not admitted, the model performance improved to 65.94% accuracy for the non-processed dataset, 67.25% for the filtered input dataset, and optimized forward feature selection as shown in the table below.

Naïve-Bayes Model Performance	Percent Accuracy- after discretization of continuous attributes			
	Simple Feature Original dataset	Simple Filtered Feature Selection	Complex feature Selection	
			Forward Selection	Backward Elimination
Label categorized into >30, <30, No readmission	61.88 %	60.58%	65.22%	61.59%
Label categorized into Yes, No readmission	65.94%	67.25%	67.25%	65.94%

Table 14: Percent Accuracy of Naïve Bayesian Classification Model

Rapid Miner selects, using forward selection optimization operator, eight significant attributes that apply the assumptions of the model. They are time_ hospital, number_ emergency, number_ inpatient, number_diagnosis, glucose level, and insulin.

4.5 Comparison of The Models:

Decision trees, with 4 different algorithms based on the selected criterion (Info-Gain, Accuracy, Gain Ratio, and Gini index), artificial neural networks, K-nearest neighbor, and Naïve Bayesian classification are implemented on the data set that describe the characteristics of adolescent diabetic patients. The aim is to find the most accurate predictive model of the data to successfully select the patient that may be potentially readmitted to the hospital and advise him or her to join the recreational and healthy diabetic center that works indirectly on reducing the risk factors of accentuating their disease, on augmenting the patients' self-esteem, on improving their psychological conditions.

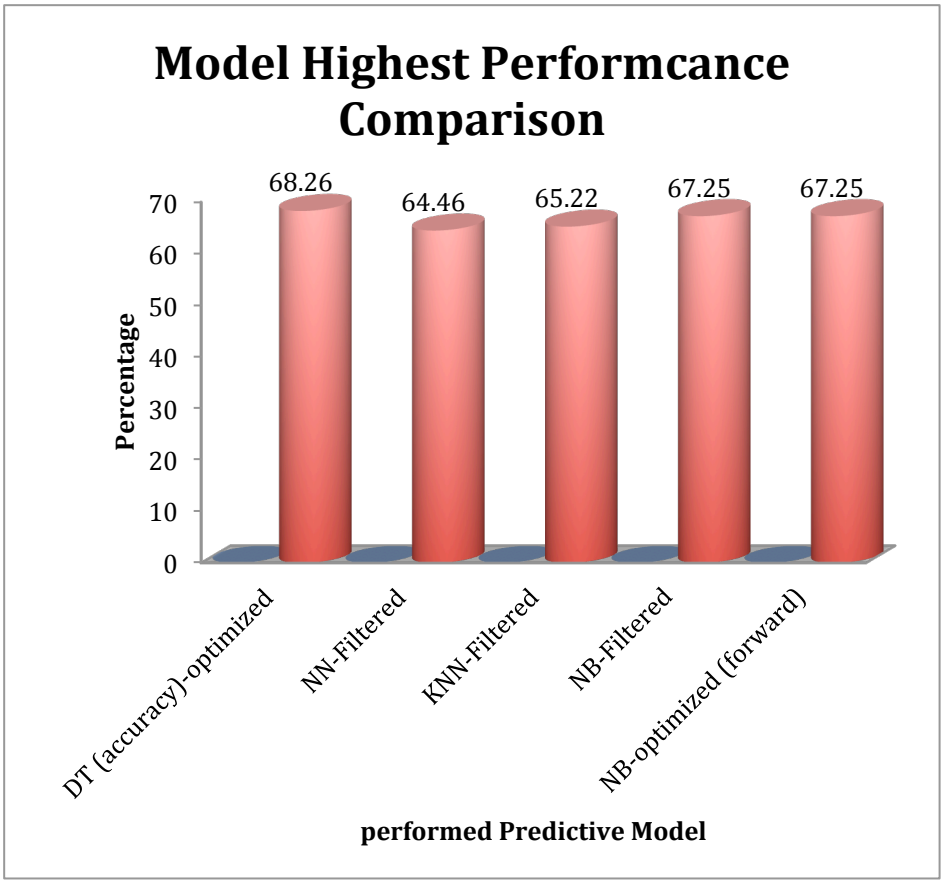


Figure 22: Comparison of Highest Performance (accuracy) of Different Models

As shown in the model comparison in figure 22, Naïve Bayesian and Decision tree with accuracy as a criterion of classification show highest results when there is feature selection, especially with the optimized forward feature selection.

A comparison of the model of highest True Negative (percent of patients who are supposed to be readmitted and not predicted to be readmitted) shows highest in KNN model with simple feature selection. Unfortunately, in this model recall of the True Positive has the lowest percentage. Therefore, it is wise to set a comparison of weighted mean recall of each model as shown in figure 23.

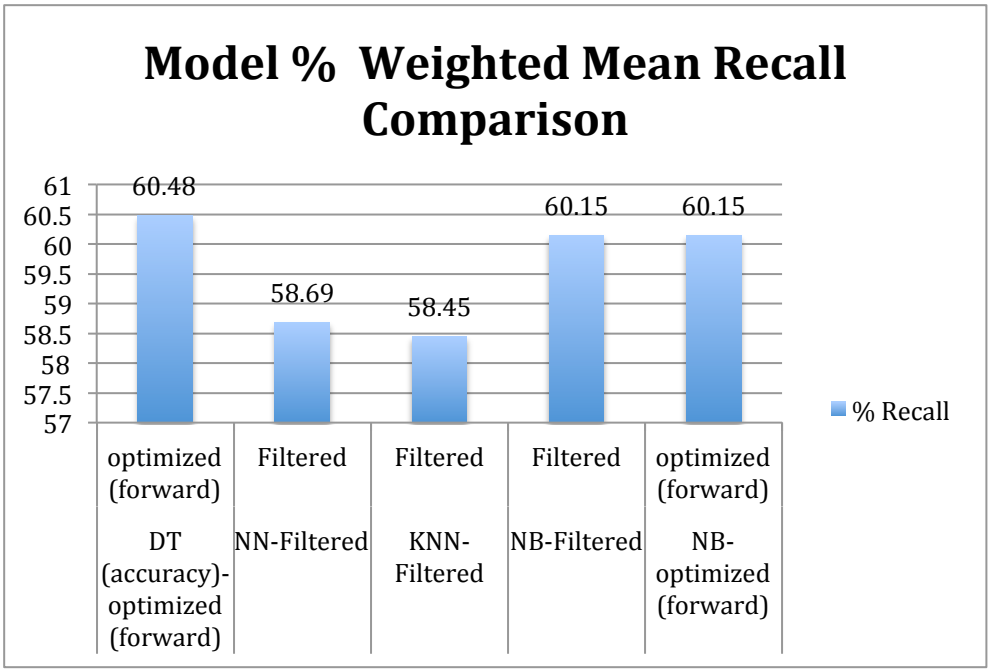


Figure 23: Comparison of Highest Performance (% Mean Recall) of Different Models

A comparison of the percent weighted mean Recall of the models also show that the best performing models are Decision tree with accuracy as criterion and optimization of features using forward selection and Naïve Bayesian with simple selection or optimized features (forward selection).

4.6 Ensemble Learning

In an attempt to improve model learning and percent accuracy of prediction, it is wise to enhance Meta-leaning, which draws together the prediction of multiple base models using the voting technique. The models use the same input data and predict their outcome individually. Their output is, then, combined, to form an ensemble output. This combined decision-making power is highly advantageous when the base models are good learners. Meta learner will be expectedly stronger. Ensemble learning does not only exceed in the predictive power its constituent individual models, but also enhances optimization of hypothesis-finding problems.

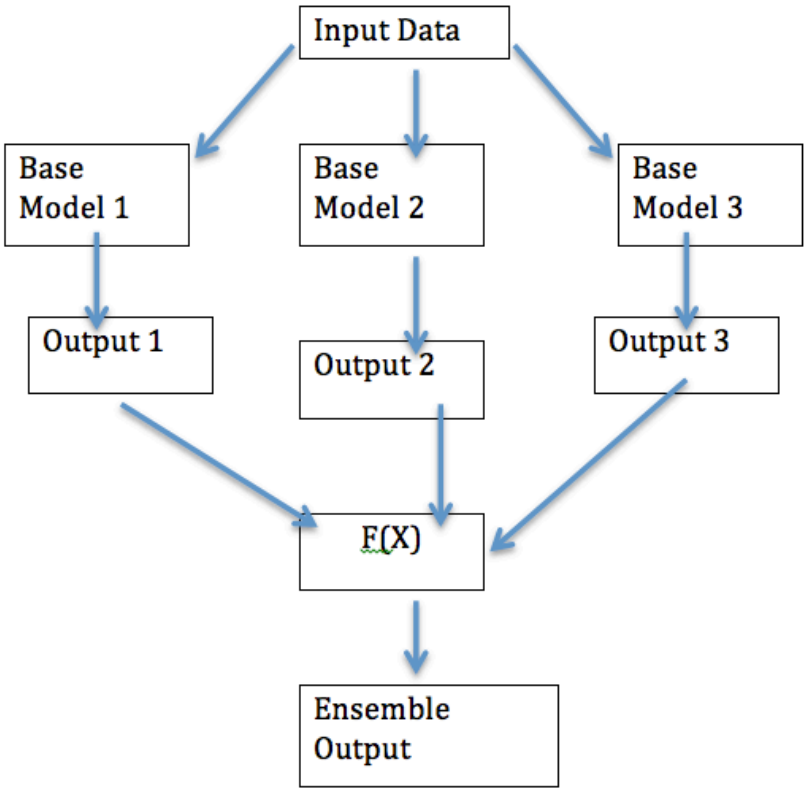


Figure 24: Ensemble Learner is a Combination of Weaker Learners

Since the best performing predictive models is Decision tree and Naïve Bayesian classification models, the following ensemble learning techniques are suggested: 1) Random Forest combining different decision trees, 2) ensemble classifier combining the two models, decision tree and Naïve Bayes, via voting, 3) ensemble classifier via bagging, and 4) ensemble classifier via boosting.

4.6.1 Random Forest:

This model creates a set of decision trees with random number of sample records automatically selected and replaced. The default number of base decision trees is 10 and the selected criterion is accuracy as this criterion has

showed highest accuracy when the individual decision tree is run separately as shown earlier in section 4.1. Once all the trees are built, they predict a target class and vote for the class of equal weights.

The performance percent accuracy of Random Forest model is 65.69%, a percentage that is lower than that of the decision tree discussed in section 4.1.2. The reason is that Random Forest algorithm selects and automatically creates different decision trees that are not only small, but also only weak leaners compared to the decision tree in section 4.1.2 (see figure 25 below).

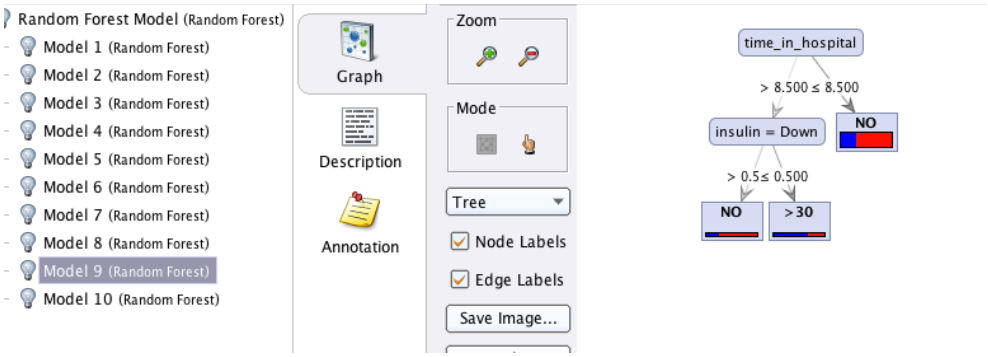


Figure 25: The Best Automatically Selected Decision Trees by Random Forest Ensemble Learner

4.6.2 Ensemble classifier combining the two models, decision tree and Naïve Bayes, via voting

Vote operator is a meta-learner and a nested operator that houses different modeling algorithms (base models). Each model uses the same input data and produces the output independently. The outputs of all the base models are voted for as a stack model whose prediction is an aggregation of all the output models. The best individually performing algorithms in figures 22 and 23 are used as base models. These are two classifiers: Decision tree and Naïve Bayes. Figure 26 below shows the operator flow to operate the vote algorithm in rapid miner.

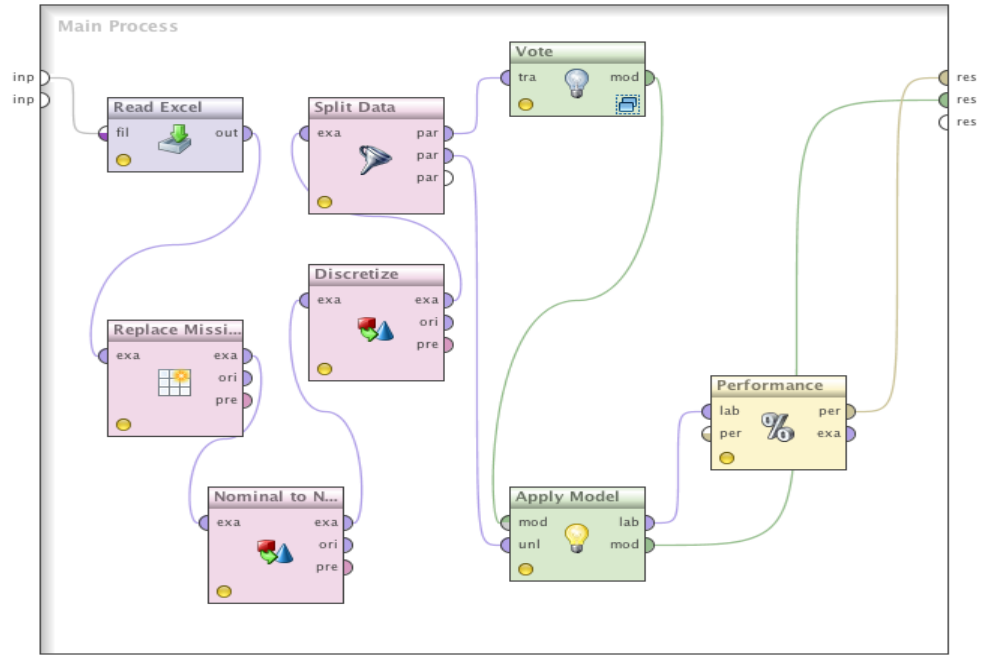


Figure 26: Rapid miner Vote Modeling Process

Ensemble algorithms via vote operator showed good improvement in percent recall of true yes; that means almost 90% of the predicted patients are eligible to readmit to hospital and they truly are. However, 80 % are predicted to re-enter the hospital, but they don't need to as shown in figure 27. This low percentage of true negative is not a good clue to adopt this ensemble learner. Other ensemble learners are performed.

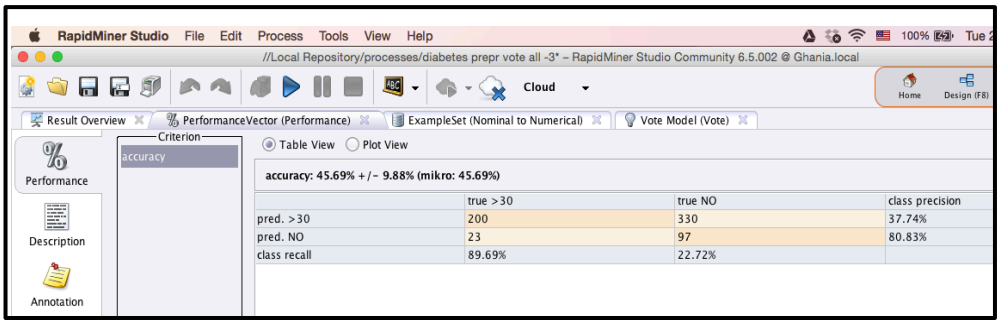


Figure 27: Cross Validation Performance of the Vote Model

4.6.3 Ensemble classifier via bagging

In this approach, the base models are not chosen as in the vote modeling, but developed by changing the training set for every model. A sample is selected from training set and replaced again, rendering the possibility of duplicates of records. Each sampled training-set is used for a base model. This process of selection is called bootstrapping. The selection is used to run each base model and the prediction of each model is aggregated for an ensemble model. Bagging is the combination of bootstrapping and aggregation.

The two models selected to create the ensemble classifier are Naïve Byes and the Decision tree. As discussed earlier, the decision tree criterion for splitting is accuracy and the model performance of the ensemble learner is 68.55% accuracy as shown in figure 28.

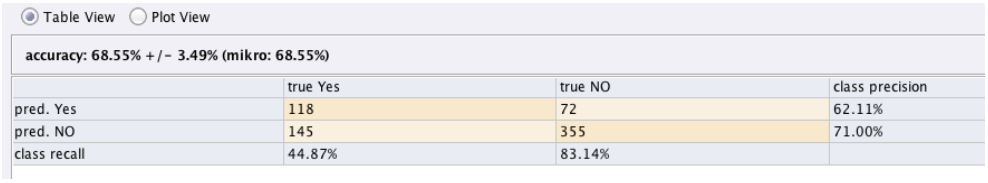


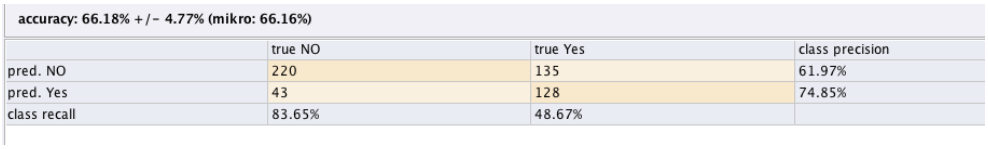
Table View Plot View

accuracy: 68.55% +/- 3.49% (mikro: 68.55%)

	true Yes	true NO	class precision
pred. Yes	118	72	62.11%
pred. NO	145	355	71.00%
class recall	44.87%	83.14%	

Figure 28: Cross Validation Performance of the Ensemble-Learning Model via Bagging

The confusion matrix of the balanced dataset shown below reveals about the small increment in percent recall of both true negative and false negative, in spite of the small drop in accuracy.


accuracy: 66.18% +/- 4.77% (mikro: 66.16%)

	true NO	true Yes	class precision
pred. NO	220	135	61.97%
pred. Yes	43	128	74.85%
class recall	83.65%	48.67%	

Figure 29: Cross Validation Performance of the Ensemble-Learning Model via Bagging Using the Balanced Dataset.

4.6.4 Ensemble classifier via boosting

This approach of ensemble learning works on training the base models in sequence one by one and assigns weights for all training records. The records that are hard to classify are reweighted. The incorrectly classified records have a higher weight and the correctly classified ones are given low weights. The classification is reiterated several times. The iterations aggregate in an ensemble learner. Adaboost operator is used in Rapid Miner to implement the boosting selection. This ensemble learner shows less improvement than the bagging-based ensemble learner. It shows 67.83% accuracy.

4.6.5 Comparison of the Four Ensemble Learners

The cross validation performance of each of the four ensemble leaners is compared with that of the best performing individual classifiers. Ensemble learning via bagging shows highest performance and is closest to decision tree with optimized feature selection. As shown in figure 30, the ensemble-learning model via bagging has the highest percent recall among all classifiers. Almost 45% are correctly predicted as true positive and 80% are predicted as false negative. Therefore, 473 out of 690 patients are correctly classified.

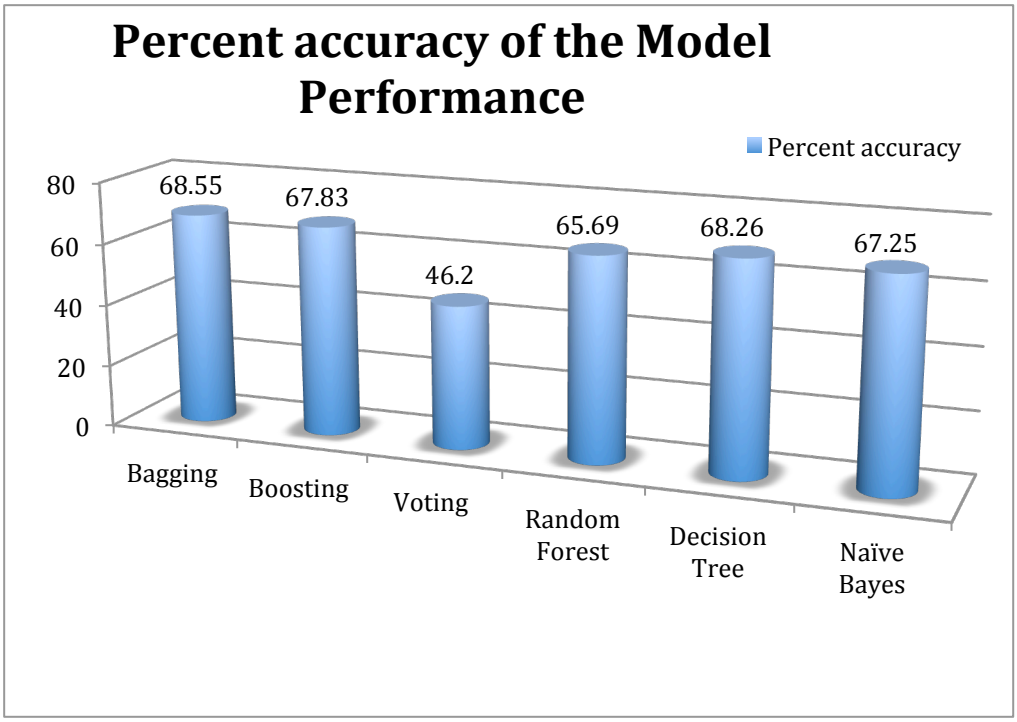


Figure 30: Comparison of the performance of the Ensemble Learners with the Best Performing Individual Classifiers

As a consequence, this learner can be used as a good predictive classifier on running new data.

In practice, if we take randomly a 10% sample of records from the dataset and run ensemble learner (bagging), the following table illustrated in figure 31 shows the resulted statistical analysis.

label **readmitted**	Polynominal	0	Least Yes (26)	Most NO (43)		Values NO (43), Yes (26)
prediction **prediction(readmitted)**	Polynominal	0	Least Yes (22)	Most NO (47)		Values NO (47), Yes (22)
confidence_Yes **confidence(Yes)**	Real	0	Min 0	Max 1	Average 0.291	Deviation 0.414
confidence_NO **confidence(NO)**	Real	0	Min 0	Max 1	Average 0.709	Deviation 0.414

Figure 31: Statistical Analysis of the Predicted Readmissions

The statistical distribution displayed as boxplot (quartile) graph (figure 32) shows the distribution range of predictions (readmissions), readmissions,

confidence (yes), and confidence (No). The default prediction is No.

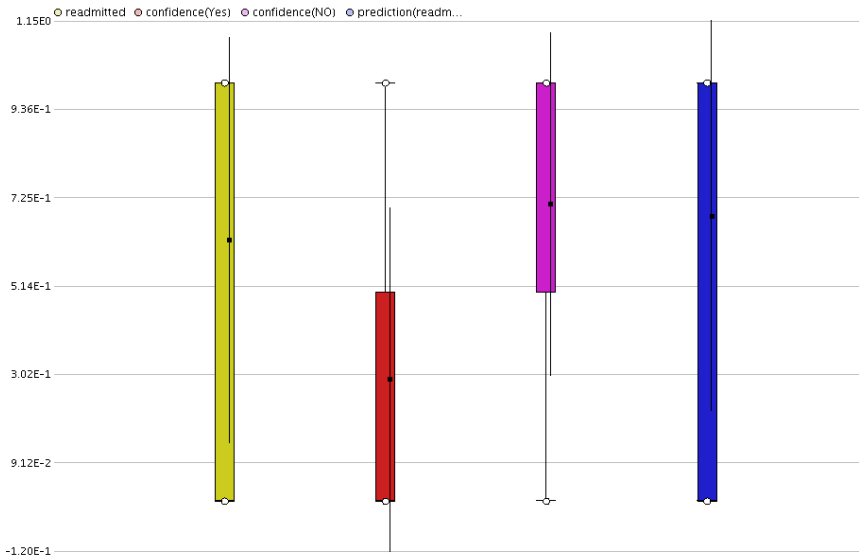


Figure 32: Box Plot (Quartile) Chart of the above distribution

Comparing the features involved in the ensemble classifier via bagging with the feature involved in the decision tree, I found out that bagging is significant not only in improving the cross validation performance (percent accuracy), but also in finding new trends and relationships that may be of high significance as a pre-hypothesis for future research. In decision tree classification technique, the algorithm selected the following feature to build the classification model such as number_inpatient, number_outpatient, Diag_1, and A1C result. In the ensemble classifier via bagging, additional feature relationships are detected such as num_lab_procedures, time_in_hospital, and num_medications as shown below in figure 33a and 33b.

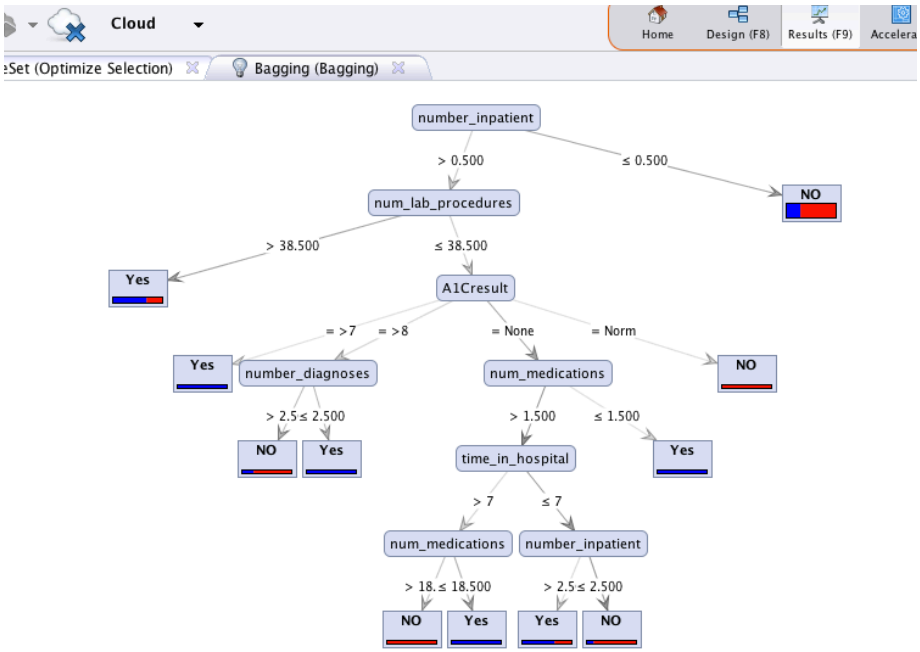


Figure 33a: One of Detected Decision Trees and Relationships via Bagging

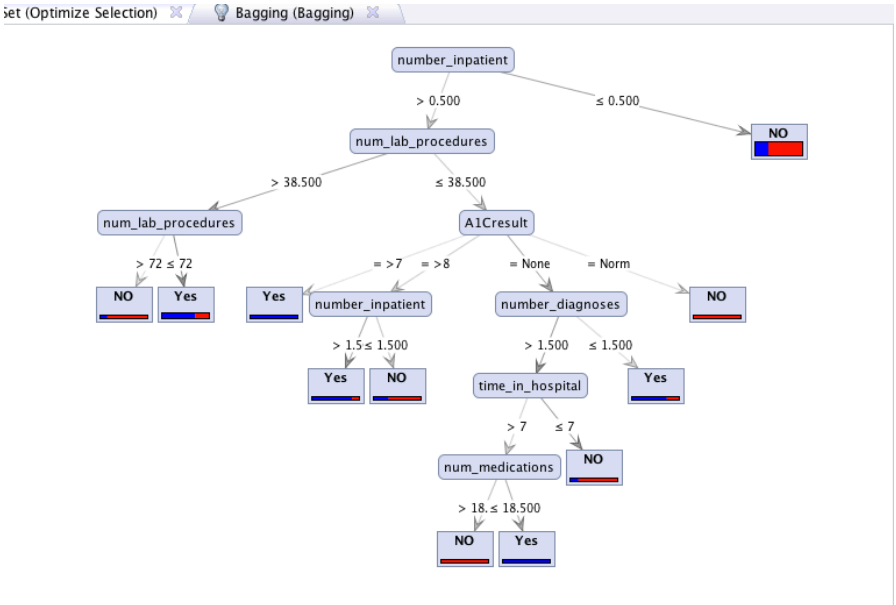


Figure 33b: Another Detected Decision Tree and Relationships via Bagging

5. Discussion

This research is unique not only in its approaches, but also in its aims. In this research, we followed thorough and deliberate data mining technical approach. Analyzing the Input data by performing different predictive classification techniques, we obtained detailed analysis of results and recommendations. The dataset consists 691 diabetic adolescent patients. As part of data preprocessing approaches, we compared the behavior of the original dataset with the data sets after conducting simple and complex feature selection. In simple feature selection, relevant features are weighted by using some statistical and other preprocessing tools in Rapid Miner, such as weight by Chi-squares, weight by Info-Gain, and weight by Principal Component Analysis (PCA) for continuous attributes. In the complex feature selection, wrapper-type feature selection is selected by optimizing the weights for attributes either by forward selection or by backward elimination. Classification processes are run on the original data, filtered data based on simple selection results, and on feature-optimized dataset in which the algorithm chooses the attributes of highest weights as input data to run the algorithm. The data set is clustered to increase the predictive results. The data is also discretized and classified using the following classifiers: Decision trees, K-nearest neighbor, Neural Network, and Naïve Bayes. They are implemented individually. Executing X-validation evaluation, the best performing individual classifiers are decision tree (with criterion as accuracy), and Naïve Bayesian. We also analyzed the confusion matrices of the models trained on balanced dataset by using the down-sampling techniques. The preprocessed dataset that rendered the best results is discretized and optimized by forward feature selection. Since ensemble classifiers are strong as they learn from their weak constituent classifiers. We suggested ensemble classification, aiming at finding a classifier of a higher predictive power. Random Forest, which is a combination of

automatically selected records and decision trees, is performed. Ensemble learners housing decision tree and Naïve Bayesian classifiers are run via voting techniques, bagging, and boosting (Adaboost) sampling methods. The results show highest prediction accuracy using bagging techniques. The second highest is the individual decision tree classifier. According to the former process, almost 70% of the diabetic adolescent diabetic patients are correctly predicted. The prediction is based on a sample of randomly selected records that are classified by the base models. The default number of iterations is 10. This creates decision trees that are more unique than the original decision tree as there are many relationships between attributes detected in the former and ignored in the individually performing decision tree as shown in figure 32. This proves the power of ensemble learning not only in improving percent cross-validation model performance, but also in spotting scarce relations between different parameters. Therefore, ensemble learner (Bagging) is run on new unlabeled dataset of diabetic patients and the classifier will be able to predict whether these patients need invitations to join the center or not.

Sharing such data mining results may lead to pre-hypotheses that can be a subject for further investigation in future. The paper is thorough and can be a base for more research in future. However, there are three limitations of this research: 1) predictive classification results may be better if more records are collected, 2) comparison of the behavior of diabetic patients of different age groups may improve results and show more trends, 3) data set collected from more than one center across UAE is preferred.

6. Conclusion

This research is unique not only in its approaches, but also in its aims. This research proves that data mining in the medical field can be as powerful as data mining in other common and well-known fields such as in marketing and ecommerce. This research proves that data mining in the medical field can contribute constructively to social, economic, and scientific domain improvement such as 1) leveraging the healthy psychological status of patients (social) (Saadi et al., 2007), 2) reducing readmission costs (economic), and 3) pre-hypothesizing (scientific) relationships between different parameters concluded from different patterns and trends predicted by machine learning techniques (Shillabeer, & Roddick, 2007). Different classification techniques are implemented, and ensemble learning via bagging of the best two classifiers, Naïve Bayes and Decision tree, shows highest percent model performance accuracy. Therefore, diabetic patients with information such as number_inpatient, number_outpatient, num_lab_procedures, time_in_ hospital, and num_medications, Diag_1, and A1C result can be predicted by the latter ensemble learner whether they are susceptible to readmission to hospital or not. If the prediction is positive, the patient is invited to register in a "recreational" program center that works indirectly on decreasing the probability of readmission by improving their healthy life-style and reducing the chronic disease risk factors. This research embeds a good number of research ideas that have not been well approached or studied before such as the efficacy of quality performance of such rehabilitation program on the diabetic patients, the relation between number of lab procedures, time in hospital, and readmissions, and the relation between num_medications, num_inpatients, and readmissions (as shown in figure 33a and 33b in section 4.6.5).

7. References:

Ali, H.I., Miller, C.J., Mohsen, A., Dunn, E.V., Ahmed, J. & Al-Nuaimie, S. (2003a). Dietary fat reducing behaviours of Arab women with type 2 diabetes mellitus. *Nutrition Research*, vol. 23(7), pp. 835–843.

Aljumah, A.A., Ahamad, M.G. & Siddiqui, M.K. (2013). Application of data mining: Diabetes health care in young and old patients. *Journal of King Saud University - Computer and Information Sciences*, vol. 25(2), pp. 127–136.

Ashraff, S., Siddiqui, M.A. & Carline, T.E. (2013). The psychosocial impact of diabetes in adolescents: A review. *Oman Medical Journal*, vol. *28*(3), pp.159-162.

Badrinath, P., Al-Shboul, Q.A., Zoubeidi, T., Gargoum, A.S., Ghubash, R. & El-Rufaie, O.E. (2002). Measuring the health of the nation: United Arab Emirates Health and Lifestyle Survey 2000. *Faculty of Medicine & Health Sciences and College of Economics Al Ain*.

Baglar, R. (2013). "Oh god, save us from Sugar": An Ethnographic exploration of diabetes Mellitus in the United Arab Emirates. *Medical Anthropology*, vol. 32(2), pp. 109–125

Bin Zaal, A.A., Musaiger, A.O. & D'Souza, R. (2009). Dietary habits associated with obesity among adolescents in Dubai, United Arab Emirates. *Nutricion Hospitalaria,* vol. *24*(4), pp.437-444.

CherneyMedically, K. (2005). *A complete list of diabetes medications* [online]. [Accessed: 6 November 2016], Available at: http://www.healthline.com/health/diabetes/medications-list#Type2diabetes3

Cios, K.J. & Moore, G.W. (2002). Uniqueness of medical data mining, *Artificial Intelligence in Medicine*, vol. *26*(1), pp.1-24.

Davis, S.V. (2011). Vitamin D deficiency and type 2 diabetes in African Americans: The common denominators. *Diabetes Spectrum*, vol. 24(3), pp. 148–153.

Dungan, K. (2012). The effect of diabetes on hospital readmissions. *Journal of Diabetes Science and Technology*, vol. 6(5), pp. 1045–52.

Garrison, M., Katon, W. & Richardson, L. (2005). The impact of psychiatric comorbidities on readmissions for diabetes in youth. *Diabetes Care*, vol. 28(9), pp. 2150–4.

Healy, S.J., Black, D., Harris, C., Lorenz, A. & Dungan, K.M. (2013). Inpatient diabetes education is associated with less frequent hospital readmission among patients with poor glycemic control. *Diabetes Care*, vol. *36*(10), pp. 2960-2967.

Hofmann, M.& Klinkenberg, R. eds. (2013). *RapidMiner: Data mining use cases and business analytics applications*. CRC Press.

International Diabetic Federation. (2015). *United Arab Emirates*. [Online] [Accessed: 26 October 2016]. Available at: http://www.idf.org/membership/mena/united-arab-emirates

Iyer, A., Jeyalatha, S. & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. *International Journal of Data Mining & Knowledge Management Process*, vol. 5(1).

Kumari, M., Vohra, R. & Arora, A. (2014). Prediction of Diabetes Using Bayesian Network. *International Journal of Computer Science & Information Technology*, vol. 5(4), pp. 5174-78.

Mabry, R.M., Reeves, M.M., Eakin, E.G. & Owen, N. (2010). Gender differences in prevalence of the metabolic syndrome in Gulf Cooperation Council Countries: a systematic review. *Diabetic Medicine*, vol. *27*(5), pp. 593-597.

Milovic, B. & Milovic, M. (2012). Prediction and decision making in health care using data mining, *Kuwait Chapter of the Arabian Journal of Business and Management Review*, vol. 1(12), p.126.

National Institute of Diabetes and Digestive and Kidney Diseases. (2014). *The A1C Test & Diabetes* [online]. [Accessed: 6 November 2016] Available at: https://www.niddk.nih.gov/health-information/diabetes/diagnosis-diabetes-prediabetes/a1c-test

Rai, A., & Kumar, D. (2014). A Comparative Study on Wrapper and Filter Methods for Feature Selection in Data Mining. *International Journal of Scientific Research and Education*, vol. 2(08), pp. 1501-11.

Rubin, D. J. (2015). Hospital readmission of patients with diabetes. *Current Diabetes Reports*, vol. 15(4), pp. 1-9.

Saadi, H., Carruthers, S.G., Nagelkerke, N., Al-Maskari, F., Afandi, B., Reed, R., Lukic, M., Nicholls, M.G., Kazam, E., Algawi, K. & Al-Kaabi, J., 2007. Prevalence of diabetes mellitus and its complications in a population-based sample in Al Ain, United Arab Emirates. *Diabetes Research and Clinical Practice*, vol.*78*(3), pp.369-377.

Shillabeer, A. & Roddick, J.F. (2007, December). Establishing a lineage for medical knowledge discovery, *In Proceedings of the Sixth Australasian Conference on Data Mining and Analytics*. Australian Computer Society, Inc. Volume 70, pp. 29-37.