KADIR HAS UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

Importance of Data Preprocessing For Improving Classification Performance on CAC Data Set

KAMRAN EMRE SAYIN

January 2013

KADIR HAS UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

Importance of Data Preprocessing For Improving Classification Performance on CAC Data Set

KAMRAN EMRE SAYIN

Prof Dr. Hasan Dağ (Thesis Supervisor)	Kadir Has University
Asst. Dr. Songül Albayrak	Yıldız Teknik University
Asst. Dr. Öznur Yaşar	Kadir Has University

January 2013

Abstract

Data Mining usage in Health Sector increased much in this decade because of the need for efficient treatment. From cost-cutting in medical expenses to acting as a Decision Support System for patient diagnosis, Data Mining nowadays is a strong companion in Health Sector. The dataset used in this thesis belongs to Dr. Nurhan Seyahi. Dr. Nurhan Seyahi and his colleagues made a research about Coronary Artery Calcification in 178 patients having renal transplantation recently. They used conventional statistical methods in their research. By using the power of data mining, this thesis shows the importance of feature selection and discretization used with classification methods for acting as a decision support system in patient diagnosis for CAC Dataset. Just by looking at seven important attributes, which are; age, time of transplantation, diabetes mellitus, phosphor, rose angina test, donor type and patient history, doctors can decide whether the patient has coronary artery calcification or not with approximately 70% accuracy. After the discretization process this accuracy approximately increases to 75% in some algorithms. Thus becoming a strong decision support system for doctors working in this area.

Özet

Veri Madenciliği'nin sağlık alanında kullanımı son 10 yılda verimli tedavi ihtiyacı dolayısıyla artmıştır. Veri Madenciliği günümüzde sağlık alanında güçlü bir yardımcıdır. Sağlık harcamalarının kesilmesinden, hasta teşhisinde karar destek sistemi olarak rol almasına kadar uzanır. Bu tezde kullanılan verisetinin sahibi Dr. Nurhan Seyahi'dir. Dr. Nurhan Seyahi ve meslektaşları 178 tane böbrek nakli geçirmiş hastalarda Koroner Arterlerde Kalsifikasyon üzerine araştırma yapmışlardır. Onlar araştırmalarında geleneksel istatistik metodlarını kullanmışlardır. Bu tez, veri madenciliğinin gücünü kullanarak, öznitelik seçme ve ayrıklaştırma metodlarıyla beraber sınıflandırma algoritmalarının kullanılmasıyla Koroner Arterlerde Kalsifikasyon olup olmadığınının incelenmesinin önemini göstermektedir. Sadece yedi özniteliğe bakarak, ki bunlar; yaş, nakil süresi, diabet, fosfor, rose anjina testi, verici tipi ve hastanın hastalık geçmişi olmak üzere, doktorlar hastada koroner arterlerde kalsifikasyon olup olmadığına yaklaşık 70% doğrulukta karar verebilirler. Veri ayrıklaştırma işleminden sonra bu başarı oranı bazı algoritmalarda 75% civarlarına yükselir. Bu nedenle bu alanda çalışan doktorlar için kuvvetli bir karar destek sistemi olur.

Acknowledgements

I'm really grateful to the most respected minister of Turkey, **Kamran İnan** for his support during my educational career and for his guidance to improve my vision. His advices acted like a light during my self improvement, about learning different languages, following the world issues and so on. I have always worked hard to fulfill his expectations and will continue to do so.

Special thanks to my mother, Hasibe Halise Sayın who acted as a light to guide me through my life. She is my friend, my co-worker, my all.

Thanks to my father Adnan Sayın and my grandmother Cevheret Çeltikcilioğlu, their sprits were all around me and they will always guide me. Praying for them all the time.

In the end I would like to thank Prof. Dr. Hasan Dağ and Işıl Yenidoğan. They are the only reason for me to study the field of Data Mining. They let me in to their Data Mining Group and gave great contributions and also thanks to Assistant Professor Songül Albayrak for teaching me the Introduction to Data Mining course and for all her help in Data Mining conference papers.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CAC	Coronary Artery Calcification
DM	Data Mining
KL	Kullback-Leibler
MLP	Multi Layer Perceptron
NB	Naive Bayes
P	Phosphorous
SMO	Sequential Minimal Optimization
SVM	Support Vector Machine
ToT	Time on Transplantation

Chapter 1

Introduction

Improvements in data storage capacity and the introduction of large databases paved the way for another concept which is data warehouses. These data warehouses contain enormous amounts of data but huge amounts of data do not necessarily mean valuable information by itself. Data Mining is the extraction of valuable information from the patterns of data and turning it into useful knowledge.

Data mining uses the power of conventional statistical methods, decision trees, neural networks and other areas. Most important factor in data mining is field expert supported decision making. Data mining methods are only used as a decision support system, so at least one expert at the field of the data set used is required for an accurate work.

There are several application areas of Data Mining, one of them is healthcare. This thesis uses a medical data set which was formed by Dr. Nurhan Seyahi and his colleagues. They examined coronary artery calcification and the important features that contribute to this disease in renal transplant patients.

Aim of this thesis is to build an efficient decision support system for doctors working on CAC Presence in renal transplant patients. Doctors will spend their time less during diagnosis phase and cost of tests will be reduced by reducing the number of necessary tests. First step is to apply classification rules and see the classification accuracy performance.

Second step is to find the most valuable attributes to decrease the number of tests made and increase the classification accuracy performances. Third step is to use discretization for easier interpretation of numerical attributes turning into nominal (categorical) ones and increasing the classification accuracy of tree and rule based algorithms that depend on nominal data. So our methodology is to learn about the dataset, apply classification algorithms, apply feature selection algorithms then apply classification algorithms and in the end use discretization algorithms on the dataset before applying one last classification algorithms to see the benchmarks and improvements.

The second chapter explains the aspects of Data Mining and its application areas. Answers what differs Data Mining from ordinary Statistics and other conventional methods.

Third chapter goes into detail of Data Mining, used in Health Sector nowadays and how it all started.

Fourth chapter explains Dr. Nurhan Seyahi's Coronary Artery Calcification Dataset, how it is formed and the attributes inside the dataset.

Fifth chapter dives into Data Mining Algorithms used in this thesis. Starting from feature selection algorithms, going to classification algorithms and ending with the discretization process.

Sixth chapter shows the test results and discusses them at the end of each experiment.

The final chapter is the conclusion, explains what is gained from the work and what can be done as a future work.

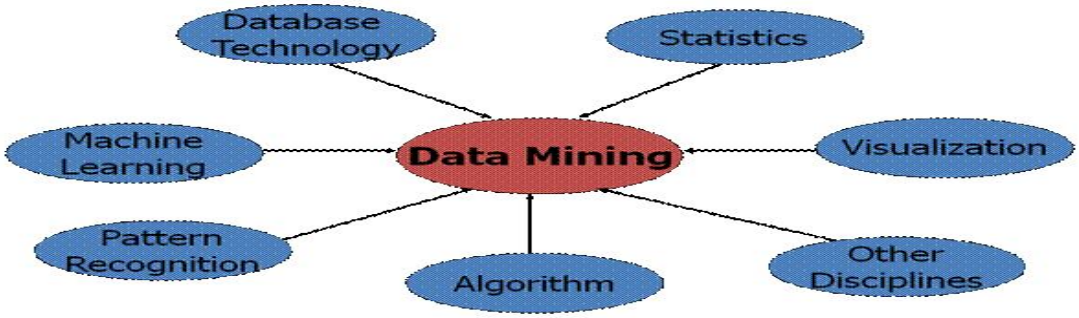
Chapter 2

Data Mining

Data Mining is the process of gathering implicit information from a dataset and extracting this hidden information to provide useful explicit information in the end [1]. It is a collection of statistical methods, mathematical algorithms, decision trees, artificial neural networks, support vector machines, clustering methods. On top of that it directly stands on the vision of a field expert. That is why it is generally used as a Decision Support System [2].

Combination of many fields that form and contribute to the Data Mining is shown in figure 2.1. Statistics, artifical intelligence, visualization, machine learning, pattern recognition and other fields unite to form data mining. They are not individually Data Mining alone.

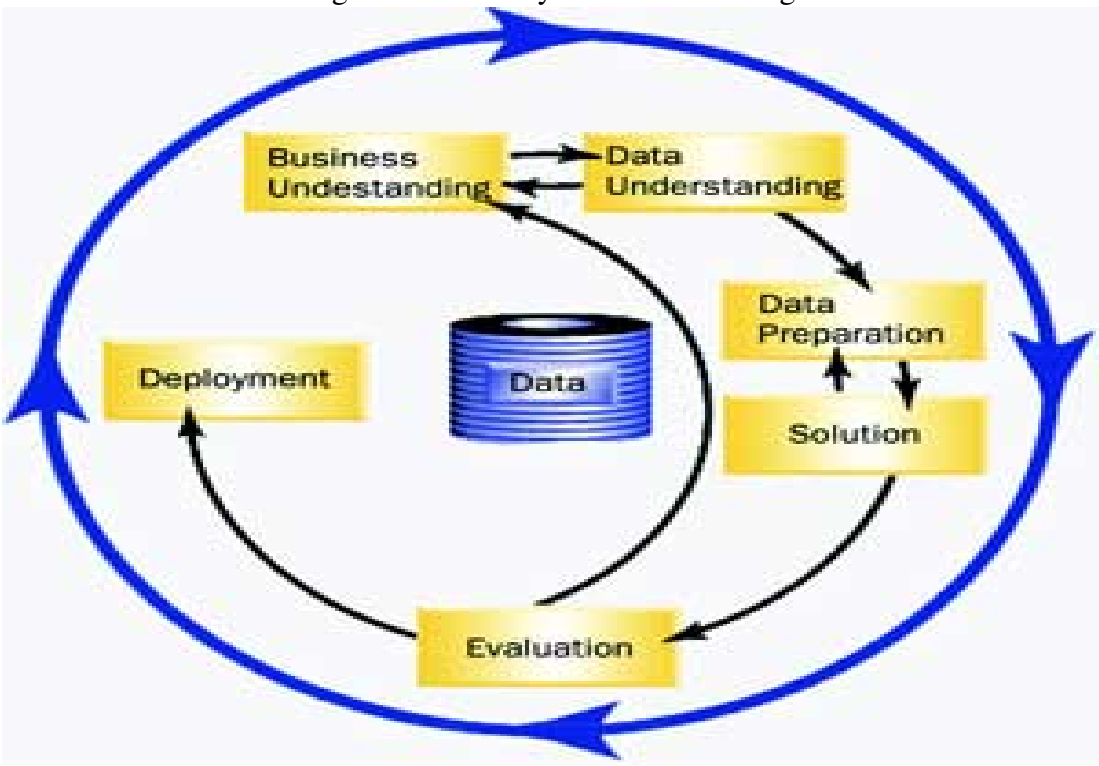
Figure 2.1: Data Mining is the Combination of Many Fields



Data Mining is not Data Warehousing, simple SQL, Software Agents, Online Analytical Processing (OLAP), Statistical Analysis Tool or Data visualization.

Life cycle of Data Mining consisting of business understanding (objectives), data understanding, data preparation, modeling (solution), evaluation and deployment can be seen in figure 2.2 [3].

Figure 2.2: Life Cycle of Data Mining



Without one phase to be completed perfectly other phases cannot be initiated correctly.

2.1 Application Areas of Data Mining

Today, Data Mining is used in different areas with different purposes.

- Insurance Companies for calculating the risk factors.
- Banks for managing loan ratios and fraud detection.
- Marketing Companies for deciding which item is sold with which item.
- Weather prediction and other climate related future events.

- Health sector for reducing costs and making efficient diagnosis.

2.2 Tasks of Data Mining

Data Mining tasks can be divided into four main categories.

- Classification and Prediction.
- Association Rules.
- Clustering Analysis.
- Outlier Detection.

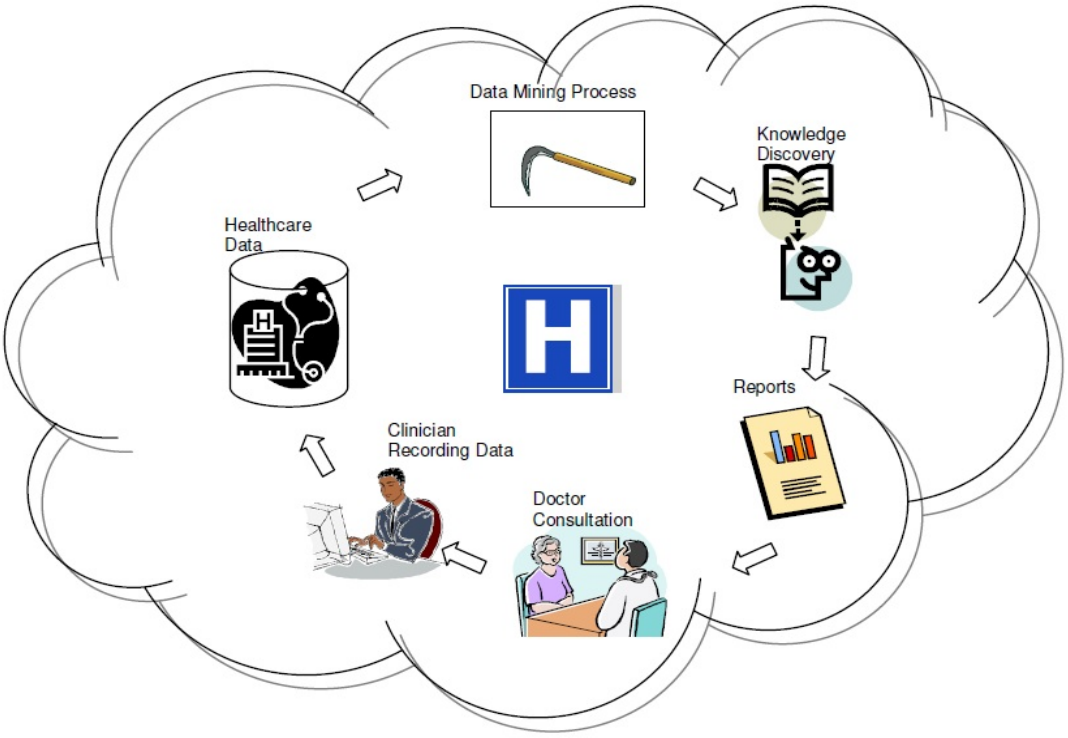
This thesis focuses on Classification methodology.

Chapter 3

Data Mining in Health Sector

Just like data mining used in other fields, data mining in health has a similar methodology for a life cycle shown in figure 3.1[4].

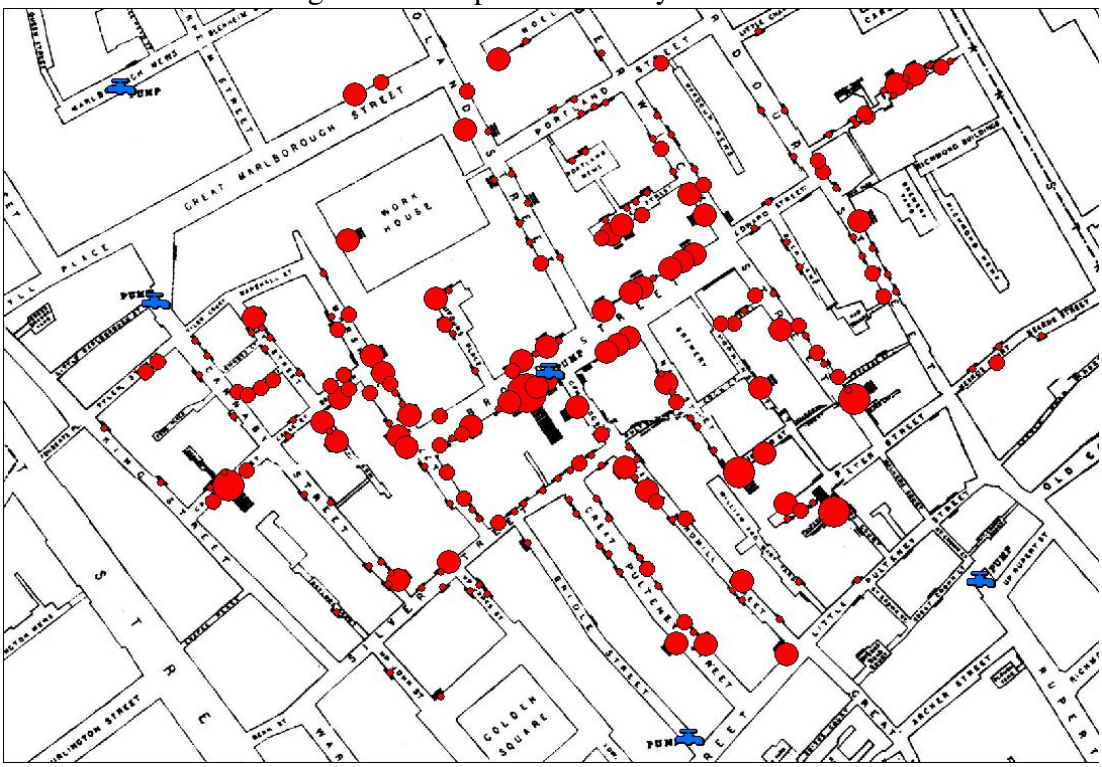
Figure 3.1: Data Mining Life Cycle in Health



The only difference is the importance of the field. Always remembering that data mining is not a final verdict just a decision support system to ease the life of doctors. Understanding the objective is the most important part of the lifecycle.

In 1854 John Snow saved people from a massive cholera outbreak in Soho, London by mapping the locations of the deaths on a map and found they clustered around a pump. This was one of the ealiest usage of Data Mining in Health. His map is shown in figure 3.2 [5].

Figure 3.2: Map of London by John Snow



Karaolis and his friends applied data mining method C4.5 to 528 patients having coronary heart diseases to show that the important factors for myocardial infarction (MI); is age, smoking, family history, and for percutaneous coronary intervention (PCI); is the family story, hypertension and diabetes, for coronary large veins bypass (CABG); is age, hypertension and smoking [6].

Karaolis and his friends worked on the case with 620 patients using C4.5 decision tree algorithm finding the most important risk factors are; gender, age, smoking, blood pressure and cholesterol [7].

The team of Srinivas made a research about a coal mine in India named Singareni to compare the cardiovascular diseases in this region with other regions, taking the other attributes into account [8].

In another work, Rothaus and his friends made a research about heart rate disturbances using K-means algorithm to assess which attributes are useful in classifying heart rate order [9].

In China, Xing and his colleagues worked on the survival rates of patients having coronary heart diseases, using 1000 patients and monitoring them for 6 months. 502 of the patients formed a dataset and they applied 3 different Data Mining algorithms on this dataset. [10]

Also in China, Z. Lin and his colleagues studied to produce an association rule depending on the support confidence. They saw that the correlation of the model before and after the rule applied did not match each other [11].

Chapter 4

CAC Dataset

The 178 records from the dataset which are used in this thesis were gathered from outpatients in İstanbul University Cerrahpaşa Medical Faculty between March 2006 and December 2007. This dataset is published in the work [12] by Dr. Nurhan Seyahi and his colleagues.

Dataset consists of 26 attributes, having 8 categorical and 18 numerical attributes. Class information can take two values which are; ”CAC Present” and ”CAC Absent”

4.1 Attributes of the Data Set

Details of the attributes in the dataset containing 178 patients having renal transplantation is given in table 4.1

Table 4.1: Attributes of the CAC Dataset

No.	Attribute Name	Description
1	Age	Age of the patient
2	Sex	Gender of the patient (Woman, Man)
3	Time on Transplantation (month)	Time after the transplantation until the CAC measurement (70.6 ± 59.5)
4	Donor Type	Type of the Donor (Living or Cadaver)
5	Dialysis Vintage (month)	Dialysis duration after renal transplantation (24.5 ± 23.5)
6	hs CRP (mg/L)	High-sensitivity C-reactive Protein test (3.1 ± 3.7)
7	Rose Angina Test	Rose Angina Questionnarie under doctor's control (Yes, No)
8	Cigare Ever	Smoking status (Yes, No)
9	Cigare Duration Period (box/year)	Smoking frequency (4.8 ± 9.6)
10	Past Cardiac Disease	Any Cardiovascular diseases encountered in the past (Yes, No)
11	Family History	Cardiovascular disease presence in 1st degree relatives (Yes, No)
12	BMI (kg/m^2)	Body Mass Index (25.7 ± 4.3)
13	Diastolic (mmHg)	Blood pressure on the vessels and arteries when the heart is relaxed (79.9 ± 11.1)
14	Systolic (mmHg)	Blood pressure on the vessels and arteries while the heart is beating (122.6 ± 16.3)
15	Hypertension	If Systolic is more than 140 mm Hg and if Diasystolic is more than 90 mm Hg and if the patient is using medicine than this is Yes, otherwise No (Yes, No)
16	T kol (mg/dL)	Total Cholesterol (LDL+HDL+VLDL) (188.8 ± 41.6)
17	LDL-Cholesterol (mg/dL)	Low Density Lipoprotein or Bad Cholesterol (111.8 ± 33.2)
18	HDL-Cholesterol (mg/dL)	High Density Lipoprotein or Good Cholesterol (49.1 ± 12.3)
19	Triglyceride (mg/dL)	Tryglyceride (151,6±76,1)
20	Albuminuri (mg/day)	Albumin amount in urine measured in 24 hours (250.9 ± 586.7)
21	Ca (mg/dL)	Calcium concentration in blood (9.6 ± 0.5)
22	P (mg/dL)	Phosphor concentration in blood (3.4 ± 0.7)
23	Ca P product (mg^2/dL^2)	Multiplication of Phosphate and Calcium concentrations in blood (32.2 ± 6.3)
24	PTH (pg/dL)	Parathyroid Hormone Concentration in blood (114.6 ± 113.6)
25	Diabetus Mellitus	Diabetes Presence Information (Yes, No)
26	MDRD ($mL/min/1.73m^2$)	Renal function depending on age and creatine taking

Chapter 5

Algorithms Used For Data Mining

Data mining algorithms are used to process the data into valuable information. Based on the objective of mining they are basically divided into four categories.

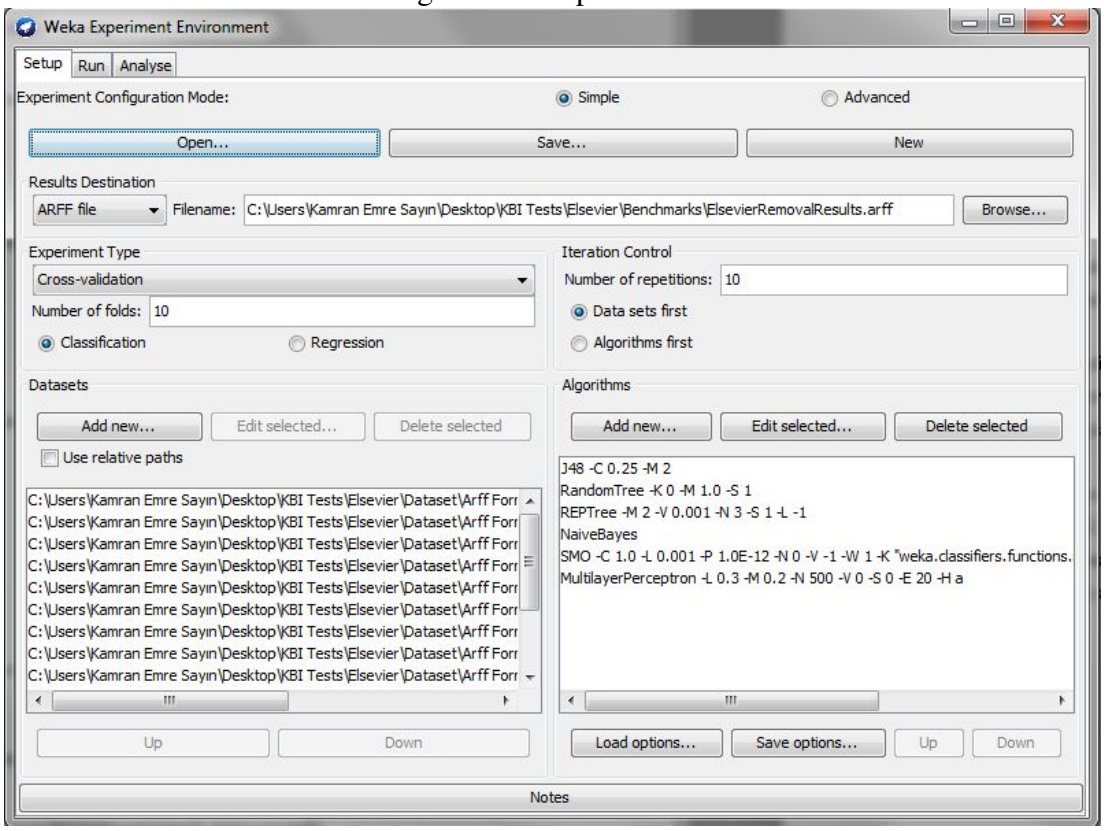
The algorithms used in this thesis are from Weka [13]. Weka is an open-source software developed at the University of Waikato and the programming language is based on Java.

Weka has 4 different applications, Explorer, Experimenter, KnowledgeFlow and Simple CLI. Knowledge Flow is a node and linked based interface and Simple CLI is the command line prompt version where each algorithm is run by hand. Below in figure 5.1 is a screenshot of the program.

Figure 5.1: Weka GUI

In this thesis only Explorer and Experiment applications of the Weka are used. Below in figure 5.2 the Experimenter interface is shown.

Figure 5.2: Experimenter

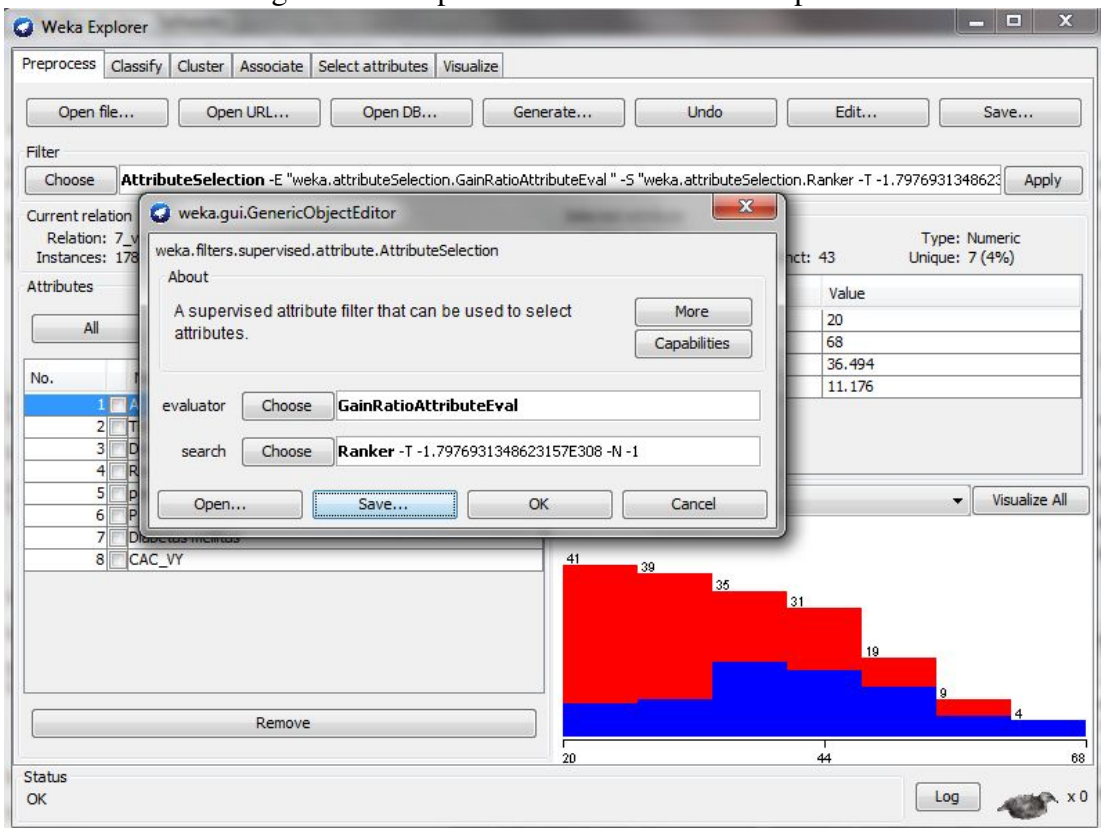


5.1 Feature Selection Algorithms

Feature Selection Algorithms choose the best set of features that contribute to the class value, that is attributes that are more important are considered in the selection. Three types of Feature Selection Algorithms are used in this thesis, Information Gain, Gain Ratio and Correlation Based Feature Selection (CFS) algorithms.

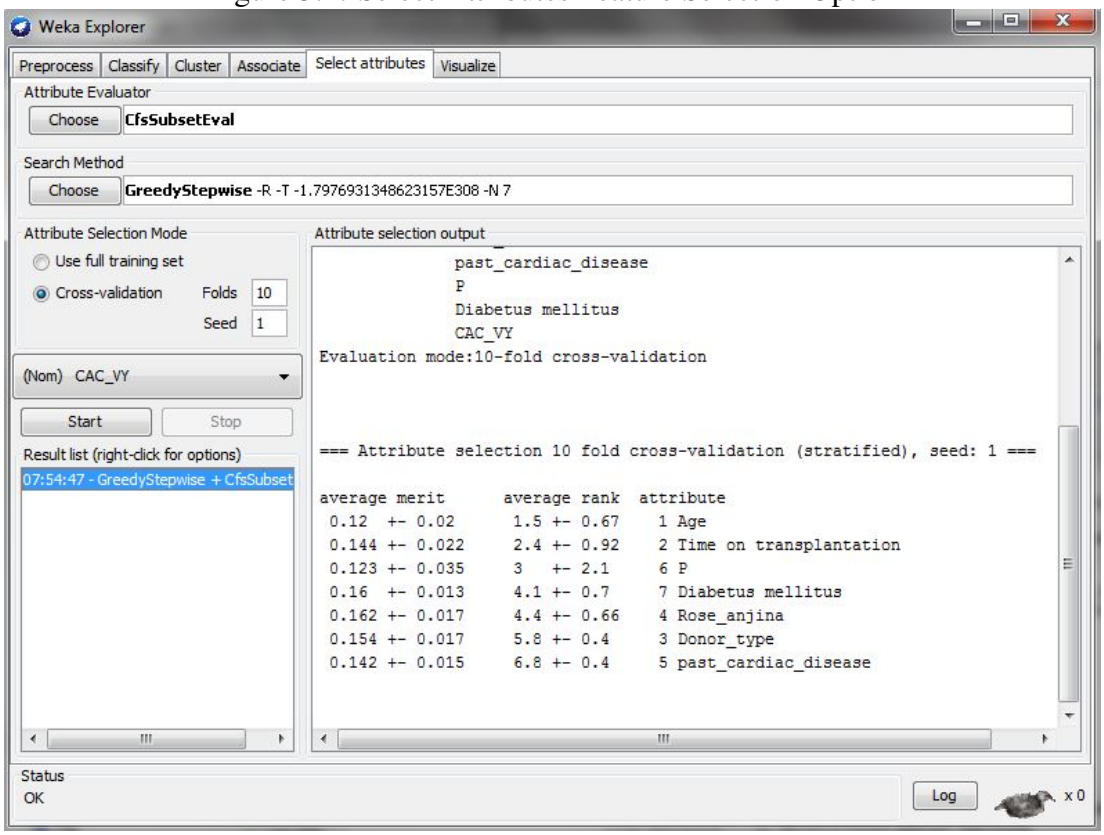
Feature Selection can only be done from the Explorer interface. It can be done either from the Preprocess tab which is shown in figure 5.3.

Figure 5.3: Preprocess Feature Selection Option



or from the Select Attributes tab shown in figure 5.4. Here average merit and average rank for each attribute is shown also unlike the preprocess tab's usage for feature selection algorithms.

Figure 5.4: Select Attributes Feature Selection Option



5.1.1 Information Gain

First proposed by Kullback and Leibler [14], information gain calculates the information gathered from entropy measurement of the given attribute with respect to the class.

$$Info\ Gain(Class,\ Attribute) = H(\ Class\ ) - H(\ Class \mid Attribute\ ). \tag{5.1}$$

here H is the entropy. It is defined as;

$$H(x) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i). \tag{5.2}$$

here p shows the probability of a particular value occuring and x is the sample (record) space.

5.1.2 Gain Ratio

Gain Ratio is simply the normalized version of Information Gain algorithm. It reduces the bias of Information Gain, dividing by the entropy of the given attribute.

$$Gain\ Ratio(Class,\ Attribute) = \frac{H(\ Class\ ) - H(\ Class \mid Attribute\ )}{H(Attribute)}. \tag{5.3}$$

Ranker

Information Gain and Gain Ratio algorithms both use the ranker method as searching type. The attributes are ranked with respect to the average merit and average rank. An option to assign a cut off point (by strictly assigning number or a threshold value) to select spesific number of attributes is available [15].

5.1.3 Correlation based Feature Selection

CFS is a feature subset selection algorithm found by Mark Hall in 1999 [16] in his PhD study. This feature selection algorithm looks for the best subset of attributes holding the highest correlation with the class attribute but lowest correlation between each attribute.

Greedy Stepwise Search Algorithm

This search algorithm can be used with any feature subset selection algorithm. I used Greedy Stepwise Search Algorithm in conjunction with CFS algorithm to search for the best subset of attributes [17]. ”It performs greedy forward search from an empty set of attributes or greedy backward search from full set of attributes. It adds/deletes attributes to the set until no longer any other attributes changes the evaluation performance”[18]. There is also an option to rank the attributes and use a cutoff point (by number of attributes or by a threshold value) to select the attributes, which I used in this thesis.

5.2 Classification Algorithms

Classification is the process assigning an appropriate class label to an instance (record) in the dataset [19]. Classification is generally used in supervised datasets where there is a class label for each instance.

5.2.1 J48

It is the Weka specific version of C4.5 algorithm developed by Quinlan [20]. Uses the normalized version of Information Gain which is Gain Ratio for building trees as

the splitting criteria. Has both reduced error pruning and normal C4.5 pruning option. As a default it comes with C4.5 pruning option, which is used in this thesis.

5.2.2 Rep Tree

Builds a decision tree (if the class is nominal) or regression tree (if the class is numerical) using information gain and variance respectively and prunes it using reduced-error pruning [21]. Reduced Error Pruning is the pruning of a node from bottom to top, looking for replacements with the most frequent class, if the misclassification rises than the node is unpruned otherwise if there is no rise in misclassification errors, then the node is pruned. In the REPTree numerical values are sorted only once for each numerical attribute The sorting process is for determining split points on numeric attributes.The tree is built in a greedy fashion, with the best attribute chosen at each point according to info-gain. [18]

5.2.3 Random Tree

Constructs a decision tree that regards K randomly chosen features at each node. No type of pruning is done. Random Tree can only work with nominal classes and binary classes [18]. Generally used as a base for building Random Forests. [22]. K value is left as 0 in our work which sets the number of randomly chosen attributes. which

$$\log_2(number\ of\ attributes) = K \tag{5.4}$$

so the number of features to be considered becomes 1.

5.2.4 Naive Bayes

Naive Bayes is a probabilistic approach to classification problem[23]. The difference from the conventional Bayes theorem is that, it assumes all atributes are independent and not affected by each other that is where the ”Naive” word comes.

$$P(C,\ X) = \frac{P(\ X|\ C\ ) - P(\ C\ )}{P(X)}. \tag{5.5}$$

Here X is the sample containing the attributes and C is the value of the Class attribute.

5.2.5 Sequential Minimal Optimization

Sequential Minimal Optimization is a Support Vector Machine based algorithm, developed by John Platt [24] at Microsoft technologies. It is generally used in solving optimization problems during the training process of Support Vector Machines. SMO replaces all missing values and turns categorical (nominal) attributes into binary attributes. It also normalizes all attributes by default. In this thesis it is options are used in default mode, with PolyKernel as the kernel selection and filter type as Normalize training data. SMO works better on numerical attributes.

5.2.6 Multilayer Perceptron

MLP is a type of Artifical Neural Network depending on neurons (nodes) [25]. MLP is a feedforward and backpropagation algorithm. Consists of multilevels with each layer fully connected to the next one. . MLP utilizes a supervised learning technique called backpropagation for classifying instances. In Weka it uses a sigmoid activation function and has options for an independent GUI usage to monitor the training phase [18]. MLP is introduced to solve the limitations of perceptrons which could solve only linearly seperable problems [26].

5.3 Discretization

Discretization is simply the transformation of numerical values to nominal (categorical) values. This process is done by dividing a continuous range into subgroups. Suppose there are 200 people in a group that want to apply for a bank loan and their ages are between 20 and 80. If the bank workers want to categorize them, they have to put them into some groups. For example one can categorize people between 20 and 40 as young, people between 40 and 65 as middle aged and 65 to 80 as old. So there will be three subgroups, which are; young, middle-aged and old. This subgroups can be increased depending on the choice of the field expert. This makes it easy to understand and easy to standardise.

Discretization can be grouped into two categories, Unsupervised Discretization and Supervised Discretization. As the name implies Unsupervised Discretization is generally applied to datasets having no class information. The types of Unsupervised Discretization are;

- Equal Width Binning
- Equal Frequency Binning

mainly but more complex ones are based on clustering methods [27]

Supervised Discretization techniques as the name suggests takes the class information into account before making subgroups. Supervised methods are mainly based on Fayyad-Irani [28] or Kononenko [29] algorithms.

Weka uses Fayyad-Irani method as default, so in my thesis I used Fayyad-Irani Discretization method with better encoding option.

Chapter 6

Preprocessing and Test Results

As implied before, one of the goals of this work is to apply classification algorithms to CAC dataset for classifying future instances and make a set of rules. Raw dataset had 80 attributes but the original dataset that Nurhan Seyahi and his colleagues used in their work have 26 attributes.

6.1 Benchmark of Classification Algorithms Applied to the Original Dataset

The accuracy performance of six different classification algorithms applied on the original data set with 26 attributes is shown in table 6.1.

Table 6.1: Accuracy Performance of Classification Algorithms on Original Dataset

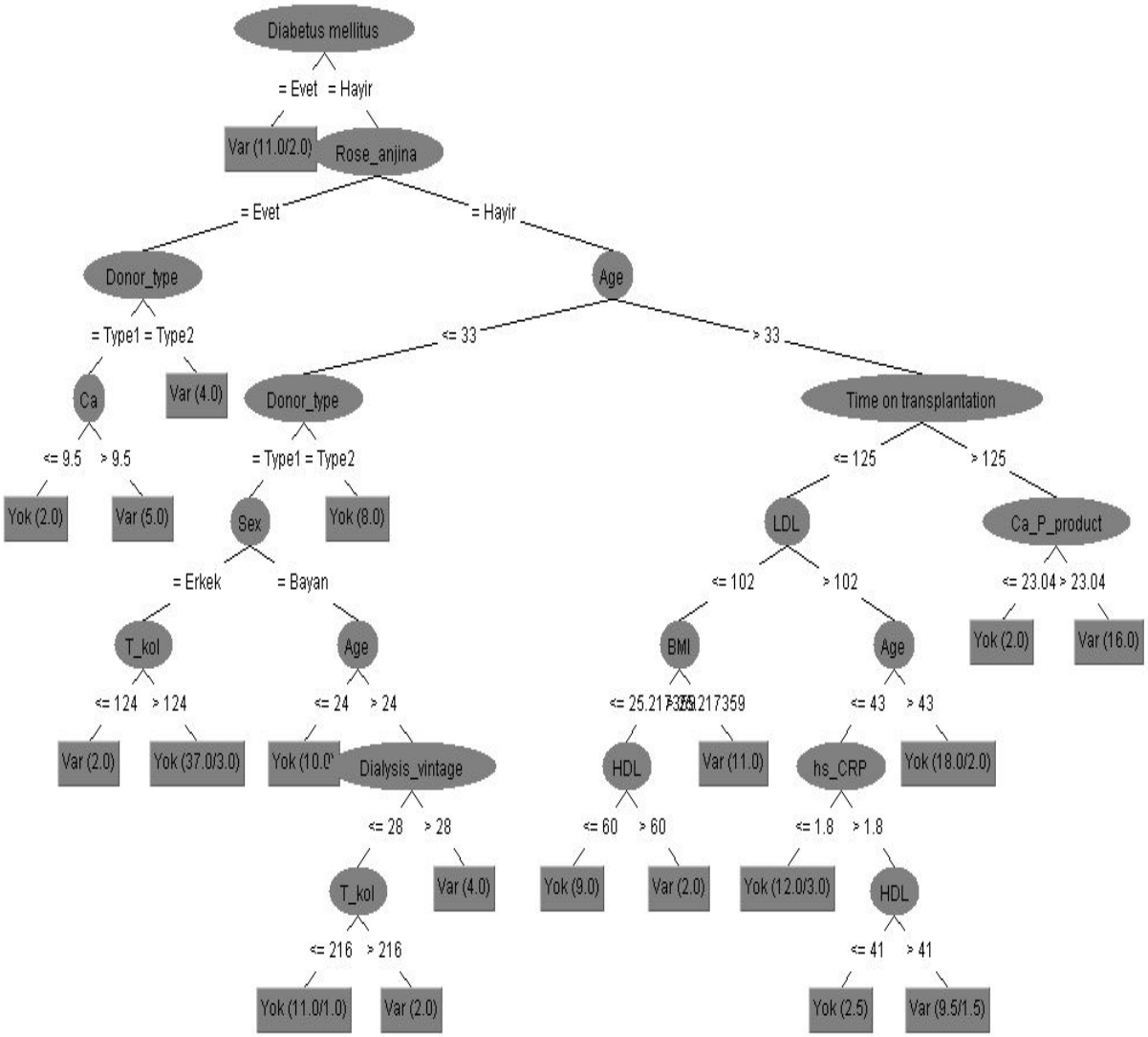
Weka Experimenter Test Results	J48	Random Tree	REPTree	Naive Bayes	SMO	MLP
26 Attributes	58.9	59.5	60.73	65.7	68.2	60.1

As seen in the benchmark table, it can be seen that SMO with 68.2% accuracy comes on top and J48 with 58.9% accuracy becomes the worst.

As shown in figure 6.1 having 56.29% classification accuracy, Diabetus Mellitus is the most important factor in Coronary Artery Calcification as it is put in to the root. The second important factor is the Rose Angina Test. Numbers in the parantheses represent coverage in the set and errors in the set respectively. Only left side of the parantheses, covered instances are shown, while on the right side error of misclassified

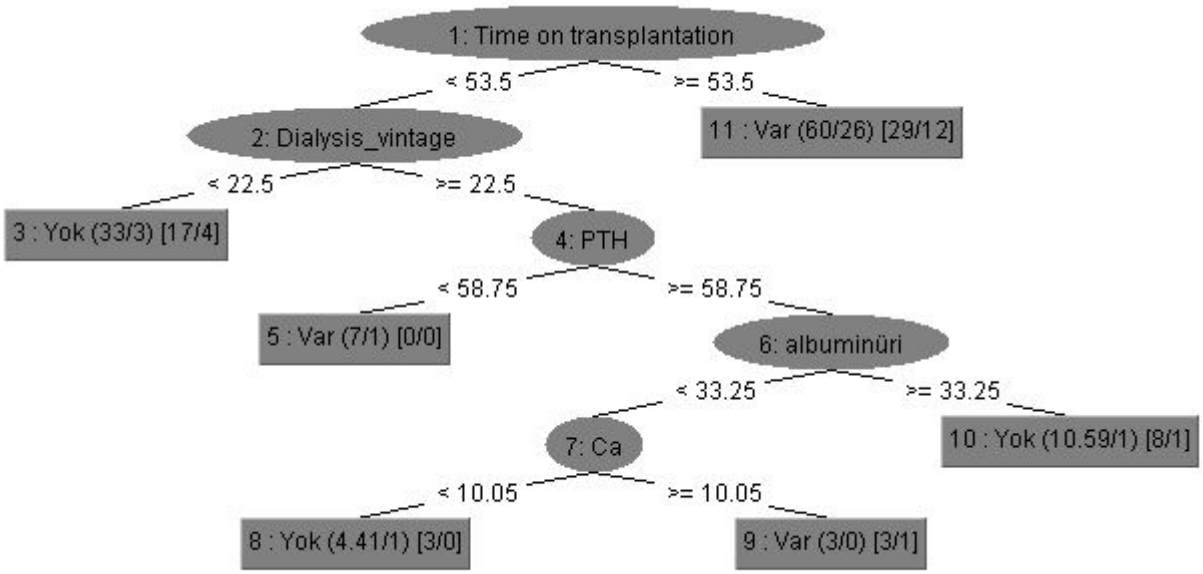
instances for the set is shown. Here HDL and Total Cholesterol are declared to be least important for J48 tree.

Figure 6.1: J48 Tree of 7 Attributes Dataset



Another tree example is the REPTree which is shown in figure 6.2 having 58.98% classification accuracy, Time on Transplantation is the most important factor in Coronary Artery Calcification as it is put in to the root. The second important factor is the Dialysis Vintage. Numbers in the parantheses represent coverage in the set and errors in the set respectively. Only left side of the parantheses, covered instances are shown, while on the right side error of misclassified instances for the set is shown. In addition to the J48, REPTree has reduced error pruning so the numbers shown in brackets represent coverage in the pruning set and errors in the pruning set respectively. Only left side of the brackets, covered instances in the pruning set are shown, while on the right side error of misclassified instances for the pruning set is shown. Here Ca and Albimunuri are declared to be least important for REPTree.

Figure 6.2: REPTree of 26 Attributes Dataset



6.2 Feature Selection Algorithms Applied

Classification accuracy performances on 26 attributes dataset was not satisfactory, but other than that as the methodology of this work, 3 feature selection algorithms should be applied for finding the most important attributes in the dataset. Also making diagnosis phase easier for the doctors and removing cost/time waste.

Our aim is to remove the attributes that contribute less merit to the class information determination and preserve the valuable attributes that directly have an effect on the class value. We are going to apply the Information Gain, Gain Ratio (which is a

normalized version of Information Gain) and CFS feature selection algorithms respectively.

7 selected attributes using Information Gain Algorithm is given in table 6.2

Table 6.2: Attributes Selected with Information Gain

Attribute Selection with Information Gain
Age
Time on Transplantation
Diabetes Mellitus
Rose Angina
P
Donor Type
Past Cardiac Disease

Table 6.3 shows the attributes selected by Gain Ratio.

Table 6.3: Attributes Selected with Gain Ratio

Attribute Selection with Gain Ratio
Age
Diabetes Mellitus
Time on Transplantation
P
Rose Angina
Past Cardiac Disease
Donor Type

Table 6.4 shows the 7 attributes selected by CFS algorithm.

Table 6.4: Attributes Selected with CFS

Attribute Selection with CFS
Age
Time on Transplantation
Diabetes Mellitus
Rose Angina
Donor Type
P
Hypertension

After finding the selected attributes for all 3 feature selection algorithms, we need to compare their selection style and see how many of the attributes out of 5, 6 and 7 attributes are same to the other algorithms' selected attributes. Their similarities can be seen in table 6.5.

Table 6.5: Similarity of Selected Attributes among Feature Selection Algorithms (IG, GR, CFS)

Data Set	**Ratio of**	**Number of Attributes Selected**		
		5	6	7
CAC	IG/GR	5/5	5/6	7/7
	IG/CFS	4/5	6/6	6/7
	GR/CFS	4/5	5/6	6/7

6.2.1 Benchmark of Classification Algorithms after Feature Selection

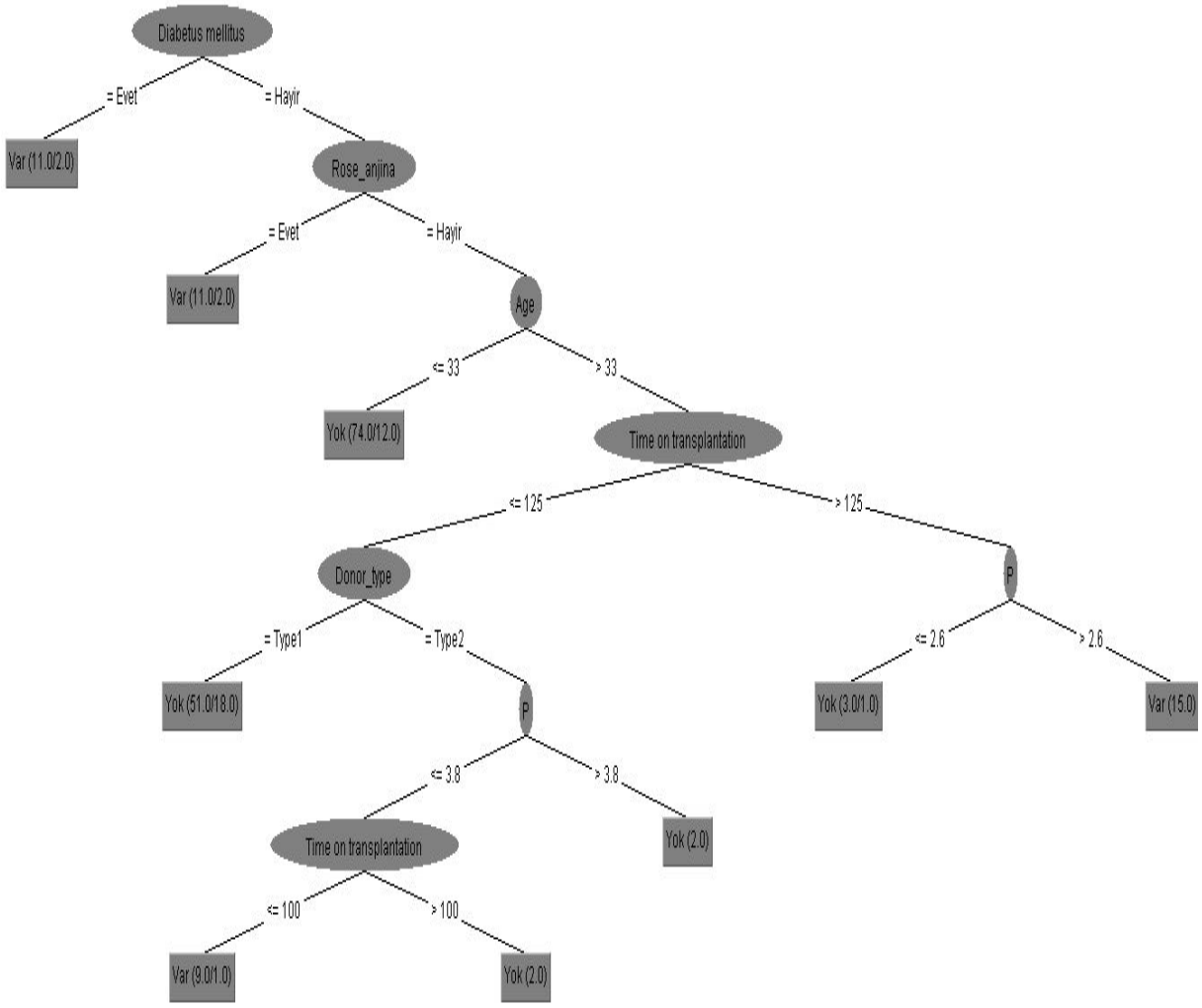
The accuracy performance of six different classification algorithms applied the reduced data set with 7 attributes (after the feature selection) is shown in table 6.6. The selected 7 attributes after the agreement are age, time of transplantation, diabetes mellitus, phosphor, rose angina test, donor type and patient history (past cardiac disease). Every classification algorithm have been used in Weka Experimenter environment with 10 fold cross validation and 10 times repetition. It is clearly seen that the classfiication performance increases fairly after the feature selection process.

Table 6.6: Accuracy Performance of Classification Algorithms on 7 Attributes Dataset

Weka Experimenter Test Results	J48	Random Tree	REPTree	Naive Bayes	SMO	MLP
7 Attributes	68.5	61.4	69.3	65.29	70.9	71.6

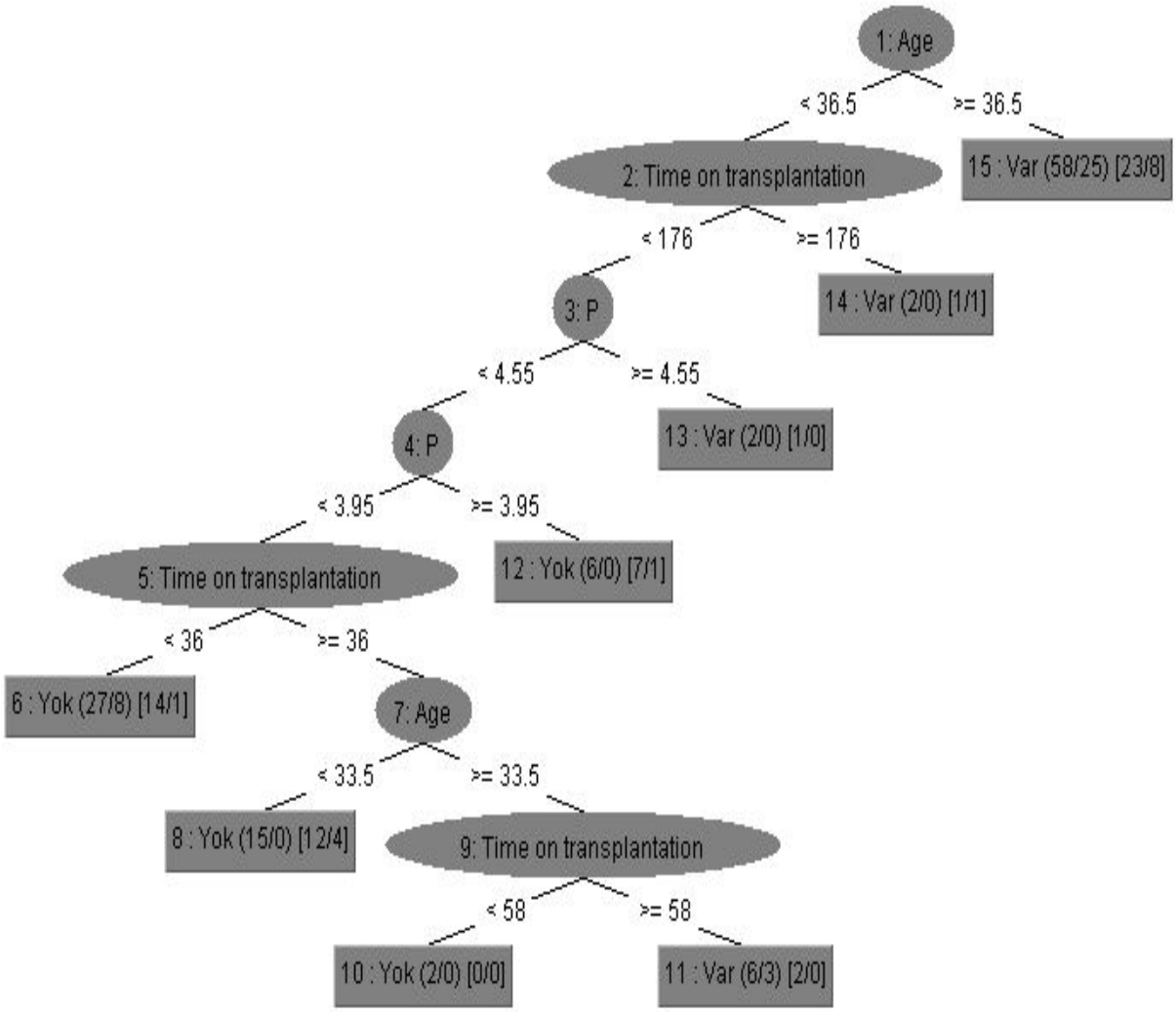
As shown in figure 6.3 having 66.29% classification accuracy, Diabetus Mellitus is the most important factor in Coronary Artery Calcification. The second important factor is the Rose Angina Test. If it's result is "yes" then this shows the patient has "CAC present" as the diagnosis result.

Figure 6.3: J48 Tree of 7 Attributes Dataset



The REPTree of the 7 attributes dataset is shown in figure 6.4 having 65.73% classification accuracy.

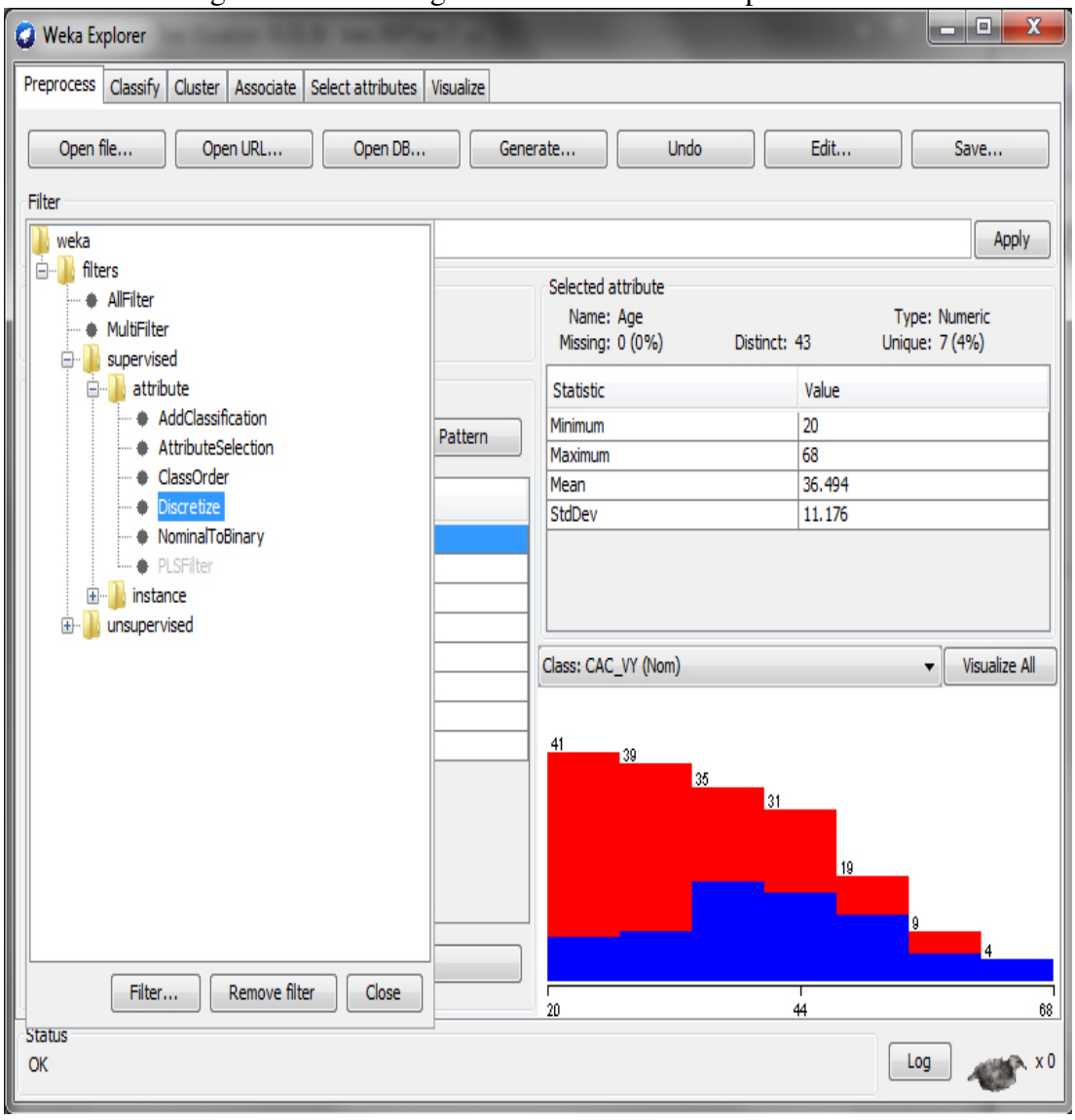
Figure 6.4: REPTree of 7 Attributes Dataset



6.3 Discretization Applied

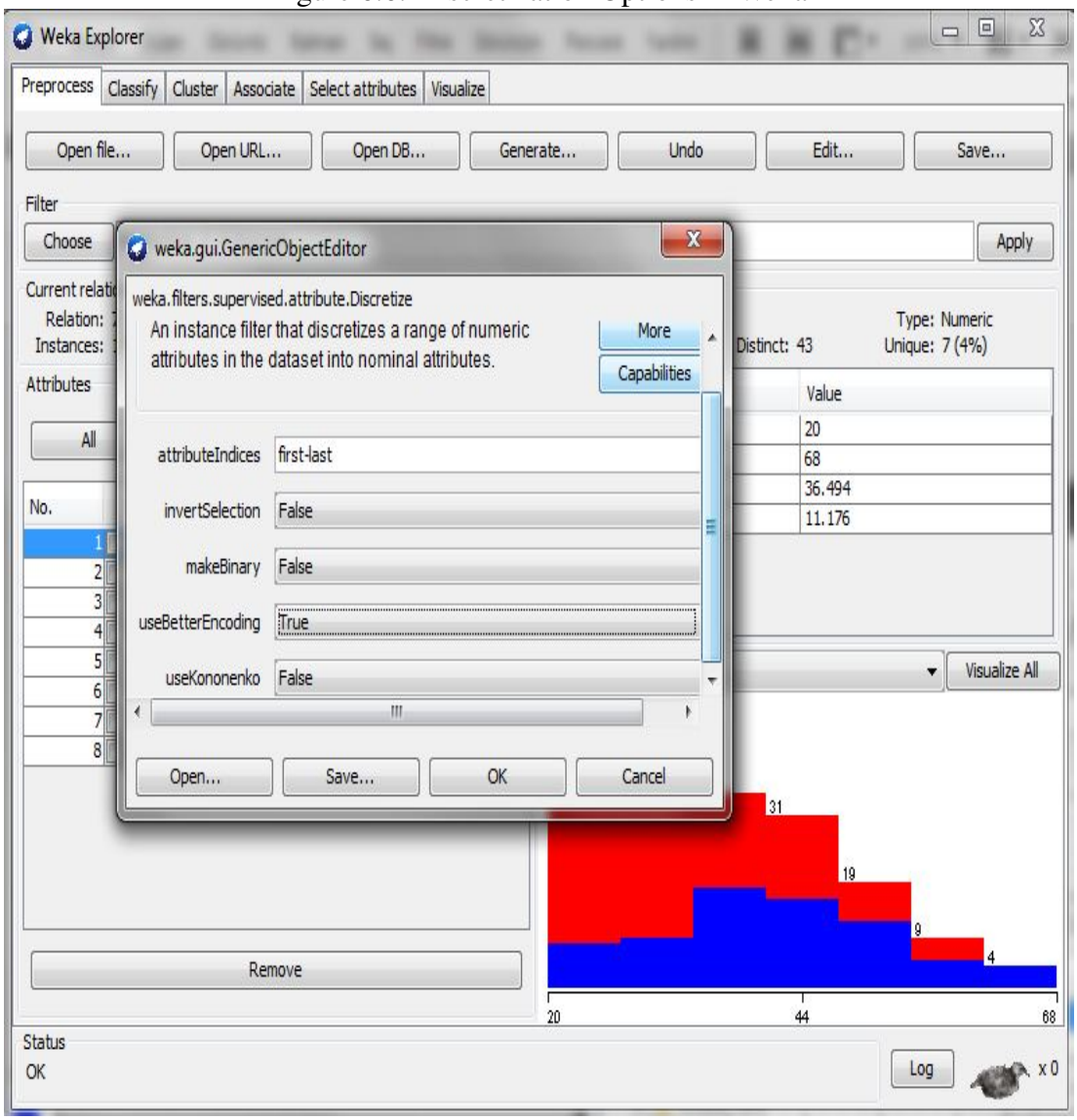
Discretization is the transformation of numerical values into categorical values as discussed before. Weka has the Discretization algorithm under the preprocessing tab. As shown in figure 6.5 it is embedded right under supervised and attribute options.

Figure 6.5: Selecting Discretization from Preprocess Tab



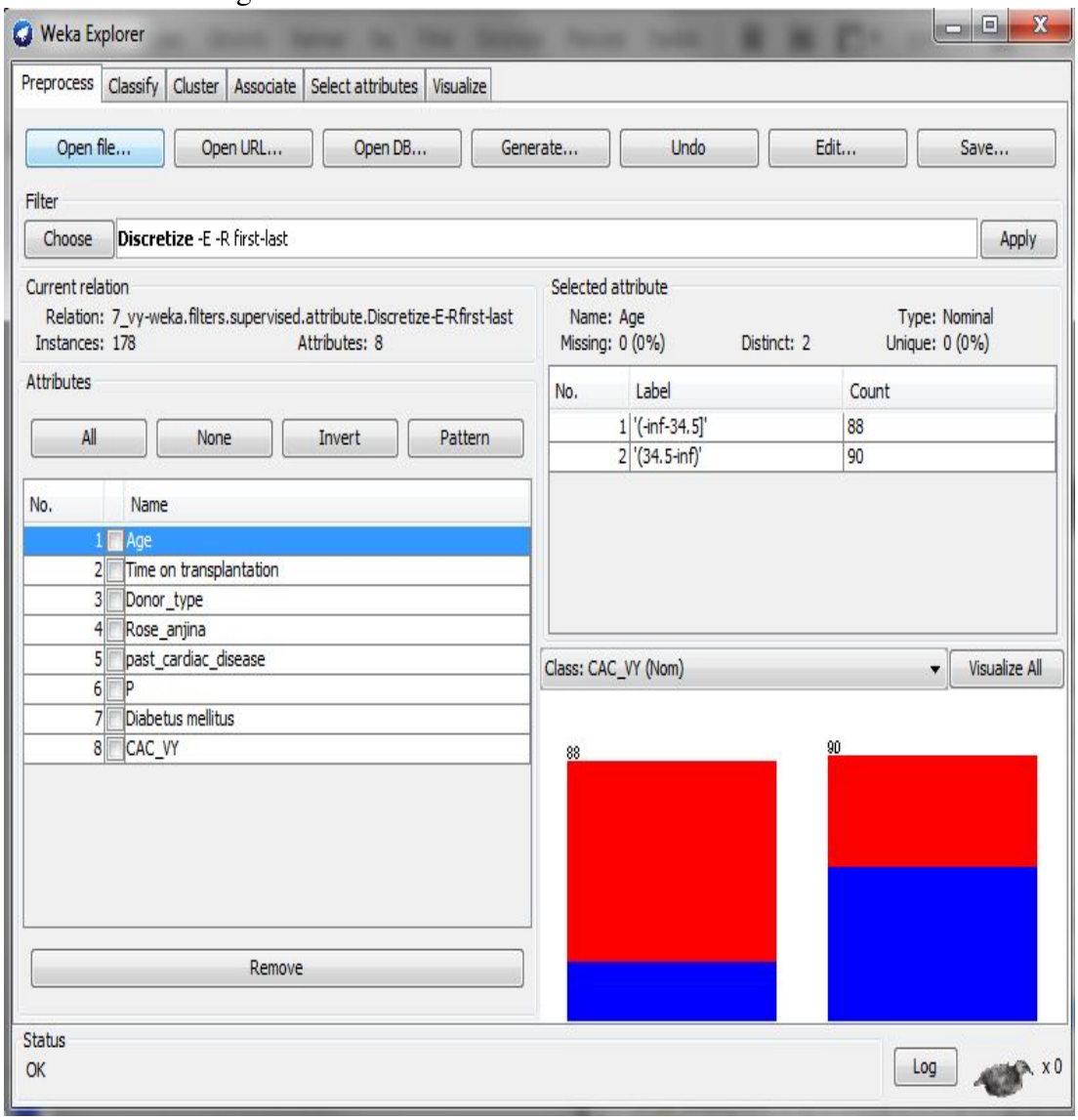
Options in the discretization algorithm are shown in figure 6.6. Weka uses Fayyad-Irani method as default, in this thesis better encoding is enabled only for forcing cut-off points to be found.

Figure 6.6: Discretization Options in Weka



An example of discretized age attribute is shown in figure 6.7. It can easily be seen that the cut-off point is 34.5 for the age attribute.

Figure 6.7: View of The Dataset After Discretization



6.3.1 Benchmarks of Classification Algorithms after Discretization

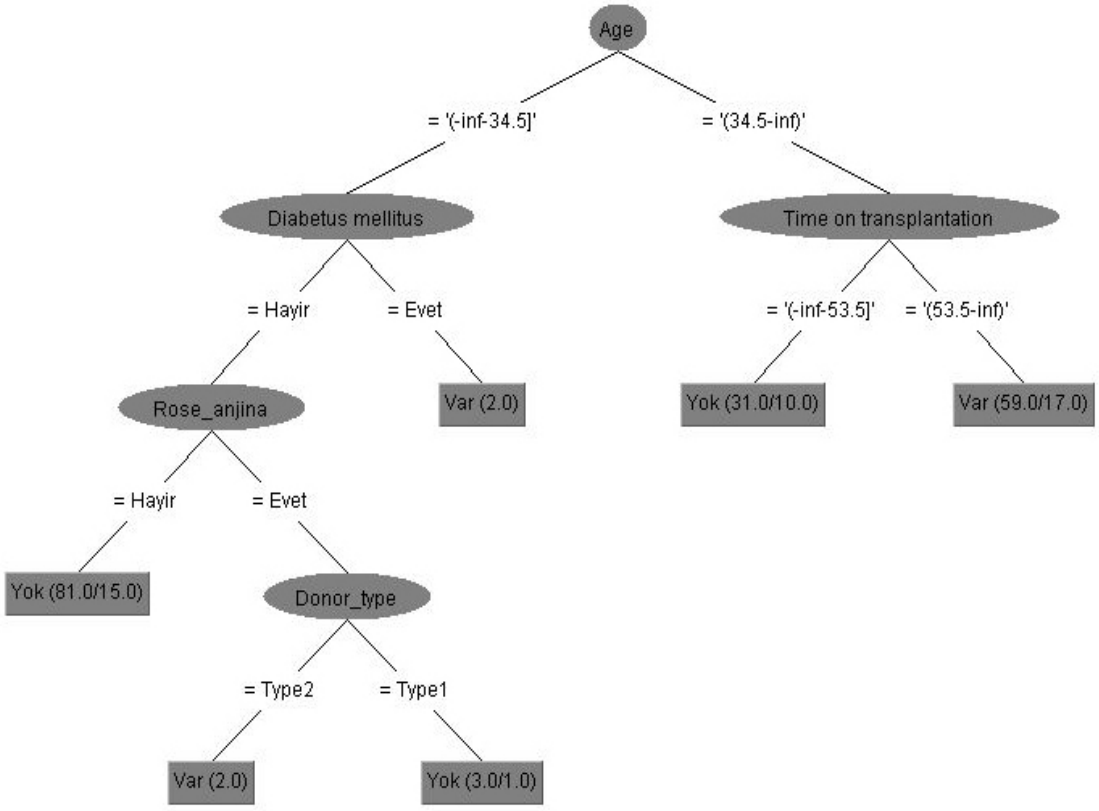
After applying discretization datasets (26 and 7) are put in Weka Experimenter for the Classification Accuracy Benchmarks. Table 6.7 shows the classification accuracy performances after the discretization process.

Table 6.7: Accuracy Performance of Classification Algorithms After Discretization Process

Weka Experimenter Test Results	J48	Random Tree	REPTree	Naive Bayes	SMO	MLP
Discretized 26 Attributes	73.16	68.12	72.21	73.89	69.04	70.99
Discretized 7 Attributes	73.73	72.5	72.04	73.78	68.62	72.67

As shown in figure 6.8 Age attribute becomes the most important factor in J48 tree after discretization process. 34.5 is the cut-off point for the age attribute.

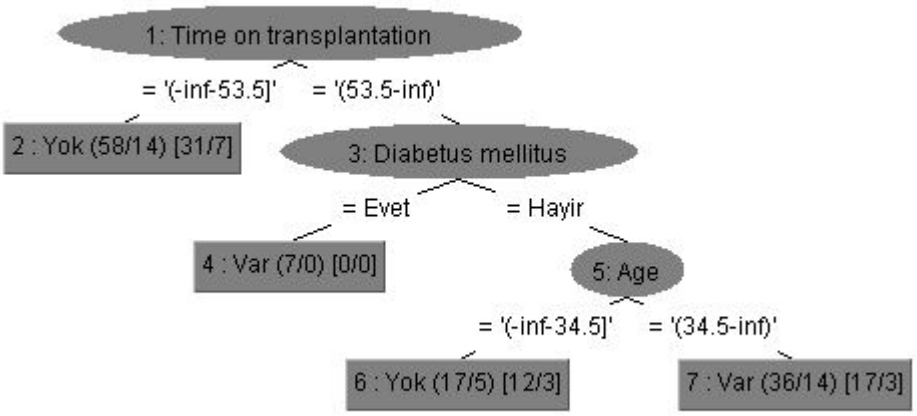
Figure 6.8: Discretized J48 Tree of Reduced Dataset with 7 Attributes



As shown in figure 6.9, ToT attribute becomes the most important factor in REPtree after discretization process and is put in the root of the tree. 53.5 is the cut-off point

for the ToT attribute.

Figure 6.9: Discretized REPTree of Reduced Dataset with 7 Attributes



Chapter 7

Conclusion

The first step of Data Mining, preprocessing process showed its benefits during the classification accuracy performance tests. Also besides the IT side of the work, there are some contributions to the medical side. Reducing the number of attributes with feature selection algorithms decreased the burden of doctors, applying many medical tests to patients before a correct diagnosis can be made. Saving the cost of the tests and also sparing some time for the doctors and the patients. By looking at only 7 important factors, which are; age, time of transplantation, diabetes mellitus, phosphor, rose angina test, donor type and patient history; CAC Presence can be determined by 70% - 75% accuracy. Importance of phosphor and is correlation to CAC Presence was also important and not discovered explicitly before by means of data mining. One important thing is to remember that this is only a decision support system to ease the diagnosis phase for doctors working in this field, not a final verdict for the health status of the patients.

Feature Selection methods significantly improved our classification performance results. One can conclude that there is not much difference between Information Gain and Gain Ratio, however their order of importance varies for each attribute during selection process. Feature subset of CFS algorithm differed from the other two but they had 66.7% similarity between them at worst so it is not wrong to say that any of these algorithms may be selected during feature selection in the preprocessing level.

Discretization of the numerical attributes increased the performance of Naive Bayes and J48 by approximately 5% but did not change the performance of SMO algorithm.

SMO strongly depends on numerical attributes, so changing them to nominal (categorical) ones did not give any advantage. Other benefit of discretization came after the visualization of J48, making the tree easy to interpret for the doctors, because of the cutting-points it assigned after the discretization of numerical attributes.

There are two limitations of this thesis, one is the number of records in the CAC dataset. 178 records cannot be considered large enough when talking about datasets if a generalization for classification can be made precisely. Having more records(instances) means that the training process will be better, resulting as the increase in classification performance. The second limitation is the use of only one software; Weka, which is an open souce software.

As a future work, using more software like SPSS Modeler, SAS Enterprise Miner and Rapid Miner will make it more solid to talk deterministic and in a bold manner about the selected features and performance of classification algorithms. Because other commercial data mining softwares, especially like SPSS Modeler and SAS Enterprise Miner, have their own specific algorithms especially for classification. Applying other feature selection algorithms also to see if there would be any difference in classification accuracy performance will be a double check. Above all, using all the steps that is applied for this thesis to the new CAC Progress Dataset of Nurhan Seyahi and his colleagues, may turn more information to valuable knowledge about the dependency of Coronary Artery Calcification in patients having renal transplantation.

s

References

[1] M. Kamber J. Han and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kauffmann Publishers, 2001.

[2] J. W. Seifert. *Data mining: An overview*. The Library of Congress, 2004.

[3] http://msquaresystems.com/wp-content/uploads/2011/07/data-mining.jpg.

[4] http://www-users.cs.umn.edu/ desikan/pakdd2012/dmhm.html.

[5] http://www.r-bloggers.com/john-snow

[6] M. Karaolis, J. A. Moutiris, and C. S. Pattichis. Assessment of the risk of coronary heart event based on data mining. In *IEEE International Conference on BioInformatics and BioEngineering*, pages 1–5, 2008.

[7] M. A. Karaolis, J. A. Moutiris, D. Hadjipanayi, and C. S. Pattichis. Assessment of the risk factors of coronary heart events based on data mining with decision trees. *IEEE Transactions On Information Technology In Biomedicine*, 14:559–566, 2010.

[8] K. K. Srinivas, G. R. Rao, and A. Govardhan. Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques. In *The 5th International Conference on Computer Science & Education*, pages 1344–1349, 2010.

[9] K. Rothaus, X. Jiang, T. Waldeyer, L Faritz, M. Vogel, and P. Kirchhof. Data mining for detecting disturbances in hearth rhythm. In *Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming*, pages 3211–3216, 2008.

[10] Y. Xing, J. Wang, and Z. Zhao. Combination data mining methods with new medical data to predicting outcome of coronary heart disease. In *International Conference on Convergence Information Technology*, pages 868–872, 2007.

[11] Z. Lin, W. Yi, M. Lu, Z. Liu, and H. Xu. Correlation research of association rules and application in the data about coronary heart disease. In *International Conference of Soft Computing and Pattern Recognition*, pages 143–148, 2009.

[12] N. Seyahi and et.al. Coronary artery calcification and coronary ischaemia in renal transplant recipients. *Nephrol Dial Transplant*, 26:720–726, 2011.

[13] *Weka: Data Mining Software in Java*.

[14] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[15] http://weka.sourceforge.net/doc.dev/weka/attributeselection/ranker.html.

[16] Mark A. Hall. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, University of Waikato, 1999.

[17] http://weka.sourceforge.net/doc.dev/weka/attributeselection/greedystepwise.html.

[18] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.

[19] G. Silahtaroğlu. *Kavram ve algoritmalarıyla temel veri madenciliği*. Papatya Yayıncılık, 2008.

[20] Steven L. Salzberg. C4.5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning*, 16:235–240, 1994. 10.1007/BF00993309.

[21] J. R. Quinlan. Simplifying decision trees, 1986.

[22] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[23] George John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345. Morgan Kaufmann, 1995.

[24] J. Platt. *Fast training of support vector machines using sequential minimal optimization.* MIT Press, 1998.

[25] F. F. Moralı, A. I. and Aygün. Çok katmanlı algılayıcı ve geriye yayılım algoritması ile konuşmacı ayırt etme. In *Akademik Bilişim 07 - IX. Akademik Bilişim Konferansı Bildirileri*, 2007.

[26] Leonardo Noriega. Multilayer perceptron tutorial. Technical report, School of Computing Staffordshire University, 2005.

[27] João Gama and Carlos Pinto. Discretization from data streams: applications to histograms and data mining. In *Proceedings of the 2006 ACM symposium on Applied computing*, SAC '06, pages 662–667, New York, NY, USA, 2006. ACM.

[28] Usama M. Fayyad and Keki B. Irani. Multi-interval discretization of continuousvalued attributes for classification learning. In *Thirteenth International Joint Conference on Articial Intelligence*, volume 2, pages 1022–1027. Morgan Kaufmann Publishers, 1993.

[29] Igor Kononenko. On biases in estimating multi-valued attributes. In *14th International Joint Conference on Articial Intelligence*, pages 1034–1040, 1995.

Curriculum Vitae

I was born in 17/03/1985. I finished Medeni Berk Elementary School in 1996 then attended Yeşilköy Anatolian Highschool.After finishing highschool, I took the ÖSS examination and qualified for full scholarship in Physics at Işık University. While studying Physics, I also studied Information Technologies as a minor degree. I finished Işık University in 2007.

I worked as a teaching assistant and organized labs in Operating Systems, Logic Design, Internet and Web Programming and Introduction to Computing.

I attended Kadir Has University in 2010 and studied Information Technologies in Institute of Science and Engineering.

Publications

[1] I. Yenidoğan Tiryakiler, N. Seyahi, S. Albayrak, K. E. Sayın, E. Ergin, ve H. Dağ, "Veri Madenciliği Teknikleri ile Böbrek Nakli Geçirmiş Hastalarda Koroner Arter Kalsifikasyonunun İncelenmesi", TIPTEKNO2011, Antalya, 2011.
[2] H. Dağ, K. E. Sayın, I. Yenidoğan, S. Albayrak, and Ç. Acar, "Comparison of Feature Selection Algorithms for Medical Data", INISTA2012, Trabzon, 2012