T.C.
BAHÇEŞEHİR ÜNİVERSİTESİ

DETECTING PSYCHOLOGICAL PROBLEMS BY USING DATA MINING

Master Thesis

TOMRİS MUT

İSTANBUL, 2012

THE REPUCLIC OF TURKEY

BAHÇEŞEHİR UNIVERSITY

THE GRADUATE SCHOOL OF NATURAL

AND

APPLIED SCIENCES

DETECTING PSYCHOLOGICAL PROBLEMS by using Data Mining

Master Thesis

TOMRİS MUT

Thesis Supervisor: Assoc. Prof. Dr. ADEM KARAHOCA

İSTANBUL, 2012

T.C.
BAHÇEŞEHİR UNIVERSITY

INSTITUTE OF SCIENCE
INFORMATION TECHNOLOGIES
GRADUATE PROGRAM

Thesis Title : Detecting psychological problems by using data mining
Student' Name and Surname : Tomris MUT
Date of the Defense of Thesis: 30.05.2013

The thesis has been approved by the Graduate School of ___________.

Assoc. Prof.Dr. Tunç Bozbura
Graduate School Director
Signature

I certify that this thesis meets all the requirements as a thesis for the degree of Master of Information Science.

Assoc.Prof.Dr.Adem Karahoca
Program Coordinator
Signature

This is to certify that we have read this thesis and we find it fully adequate in scope, quality and content, as a thesis for the degree of Master of Information Science.

Examining Comitee Members	Signature
Thesis Supervisor Assoc. Prof. Dr. Adem Karahoca	----------------------------------
Thesis Co-supervisor Assist.Prof.Dr.M.Alper Tunga	--------------------------------
Member Assist.Prof.Dr.Yalçın Çekiç	----------------------------------

ÖNSÖZ

Bilişim çağında yaşadığımız şu günlerde, teknolojinin hızlı gelişmesiyle ortaya çıkan yeniliklerin kullanılmaya ve uygulanmaya başlanması, bu sistemlerin başarı ölçümünün değerlendirilmesi, tezin ana fikrini oluşturmaktadır. Bu tez çalışması, psikolojik rahatsızlıkların tespiti amaçlı olarak hazırlanmış ve data mining yardımıyla da güçlendirilmiştir.

Bu çalışma sürecinde yardımlarını ve desteğini esirgemeyen tez danışmanım Sayın Assoc. Prof. Dr. Adem KARAHOCA'ya, tezin içeriğini oluşturmamda yardımcı olan, bu uzun vadeli çalışma boyunca bütün sorularımı cevaplayan, desteğini esirgemeyen Sayın Tamer Uçar'a, beni hiçbir zaman yalnız bırakmayan ve her zaman desteğini hissettiğim ve sıkıntıya düştüğüm zamanlarda hep yanımda olan aileme teşekkürü bir borç bilir, sonsuz sevgilerimi sunarım.

Tomris MUT

ABSTRACT

DETECTING PSYCHOLOGICAL PROBLEMS by USING DATA MININGG

Mut, Tomris

The Institute of Sciences Information Technologies Graduate Program

Supervisor: Assoc. Prof. Dr. Adem Karahoca

August 2012, 63 pages

Because of the development of the technology, people have much more psychologıcal illnesses. To prevent the loss, the companies begin to develop new systems to ensure safe of data flow owned by the network structures. Today, data mining is a data processing technology that is used to solve many problems.

Data mining can be applied to all business areas such as financial industry, banking, telecom and biomedical fields and it is used for degree of the psychological illnesses.This study is focused on detecting psychological illnesses and their effects.Classifications are made under using information according to age, sex, marital status, income, changes of personality type, reasons to start, degree of illness, duration of illness, repeat status, and behavior to outside, changes of moods, illness situation, and using medicine. Illness affection information is taken into consideration in order to reach definite detecting psychological problems.Weka 3.7.1 (Witten, Frank, 2005), with data mining interface; Decision Trees, Logistic, Multi-Layer Perceptron, JRIP, Bayes Rule, Bayesian Networks, Part, Zeror, Oner, J48 and Rbf Networks are the classification methods that are used in this study. After studies, the performance shows that the least effective application is ZeroR (50 %). Logistic application gives best result than the others.

Key Words: Psychology, Data Mining, and Psychological Illnesses.

ÖZET

PSİKOLOJİK PROBLEMLERİN DATA MINING YÖNTEMİYLE TESPİTİ

Mut, Tomris

Fen BilimleriEnstitüsüBilgiTeknolojileriYüksekLisansProgramı

TezDanışmanı: Assoc. Prof. Dr. AdemKarahoca

Ağustos 2013, 63sayfa

Teknolojinin hızlı gelişip yayılması, kişilerin daha çok psikolojik rahatsızlıklara sahip olmalarını sağladı.

Veri madenciliği tüm iş alanlarında uygulanabilen bir yöntem olsada, sıklıkla finans sektöründe, bankacılıkta, GSM sektöründe ve biomedical alanlarda; psikolojik rahatsızlıkların seviyesinin ve neler olduğunun tespitinde kullanılır.

Bu çalışmada; Psikolojik rahatsızlıkların tespiti ve etkileri üzerinde durulmuş Kullanılan veriler olarak; yaş, cinsiyet, medeni hal, gelir, kişiliğindeki değişimler,başlama sebebi, hastalığın derecesi, hastalığın süresi, tekrarlama durumu, çevreye karşı tutumu, ruh hallerindeki değişmeler, hastalığın durumu, ilaç kullanımı bilgilerine sınıflandırma yöntemleri uygulanmıştır. Psikolojik problemlerin tespiti tanısının konulmasında ise, hastalığın etkileri göz önünde durulmuştur.

Çalışmada uygulanan sınıflandırma yöntemleri; Weka 3.7.1 (Witten & Frank, 2005) veri madenciliği ara yüzü ile; Karar Ağaçları, Logistik,Çok Katmanlı Algılayıcı, JRIP, Bayes Kuralı, Bayesian Ağları, Part, Zeror, Oner, J48 ve Rbf Ağları’ dır. Yapılan çalışmalar sonucunda, en kötü sonucu ZeroR (%50) uygulaması almıştır. Logistik uygulaması diğer uygulamalara göre daha iyi sonuç vermiştir.

Anahtar Kelimeler: Psikoloji, Verimadenciliği, PsikolojikRahatsızlıkları

TABLE OF CONTENTS

TABLES……………………………………………………………………..ix

FIGURES xi

ABBREVIATIONS xii

1. INTRODUCTION 1

1.1. PROBLEM DEFINETION 1

2. BUSINESS INTELLIGENCE 4

2.1.BUSINESS INTELLIGENCE SOLUTIONS 6

2.2. BUSINESS INTELLIGENCE COMPONENTS 7

2.2.1. Source Systems 7

2.2.2. Data Warehouse 7

2.2.3. On-line Analytical Processing (OLAP) 8

3. KNOWLEDGE DISCOVERY IN DATABASES 10

4. DATA MINING PROCESSES 13

4.1.DATA COLLECTION 13

4.2. DATA CLEANINGAND CONVERSION 13

4.3. SETTING UP THE MODEL 13

4.4. MODEL EVALUATION 14

4.5. SCORING 14

4.6. APPLICATION INTEGRATION 14

4.7. MODELMANAGEMENT 14

5. CLASSIFICATION ALGORITHMS 15

5.1. DATA PRE-PROCESSING……………………………………………….15

5.1.1. Data Cleaning………………………………………………………..15

5.1.1.1. Binning…………………………………………………………..16

5.1.1.2. Clustering……………………………………………………16

5.1.1.3. Regression…………………………………………………...16

5.2. DATA AGGREGATION…………………………………………………16

5.3. DATA TRANSFORMATION……………………………………………17

5.4. DATA REDUCTION……………………………………………………..18

6. DATA COLLECTION AND ANALYSIS 22

6.1. DATA COLLECTION 23

6.2. APPLIED CLASSIFICATION METHODS 25

6.2.1. Bayesian Networks 25

6.2.2. Multiplayer Perceptron 26

6.2.3. Ripper Algorithm 26

6.2.4. Partial Decision Trees 26

6.2.5. Bayes Rule 27

6.2.6. Oner Rule 27

6.2.7. Zeror Rule 27

6.2.8. Statistical Accuracy Metrics 27

6.2.9. Root Mean Squared Error (RMSE) 28

6.2.10. Receiver Operating Characteristic (ROC) 28

6.2.11. J48 Rule………………………………………………………….28

7. FINDINGS 30

7.1. WEKA BAYESNET APPLICATION 30

7.2. WEKA NAIVE BAYES APPLICATION 33

7.3. WEKA LOGISTIC APPLICATION 35

7.4. WEKA MULTIPLAYER PERCEPTRON APPLICATION 37

7.5. WEKA RIPPER APPLICATION 39

7.6. WEKA PART APPLICATION 41

7.7. WEKA ONER APPLICATION 43

7.8. WEKA ZEROR APPLICATION..45

7.9. WEKA RBF NETWORKS APPLICATION 47

7.10. WEKA J48 APPLICATION...49

7.11. J48 RULES' DETAILS 51

7.12. CLASSIFIER TRAINING DATA SET RESULTS………………………..53

7.13. SUMMARY OF CLASSIFICATION METHODS……………………..…56

8.DISCUSSION and CONCLUSION...57

REFERENCES...59

TABLES

Table 2.1: BI Relation to other Information Systems..…………………………………..6
Table 6.1:Variable List..24
Table 6.2:Discrete Data List 24
Table 6.3: Sorting List 25

Table 7.1: Classifying Algorithm Results………...…………………………………...53
Table 7.2: Classifying Accuracy Results 54
Table 7.3:Classifying RMSE & ROC Results 55
Table 7.4: Summary of Classification Method 56

FIGURES

Figure 2.1: Gartner Platform 5
Figure 2.2: Business Intelligence 8
Figure 2.3: An Analytical Workspace Cube 9
Figure 5.1: Confusion Matrix 20
Figure 7.1: BayesNet Statistical Values 31
Figure 7.2: BayesNet Accuracy Values 31
Figure 7.3: BayesNet ROC Curve 32
Figure 7.4:NaiveBayes Statistical Values 33
Figure 7.5: NaiveBayes Accuracy Values………………………………………….34
Figure 7.6:NaiveBayes ROC Curve 34
Figure 7.7: Logistic Statistical Values 35
Figure 7.8: Logistic Accuracy Values 36
Figure 7.9: Logistic ROC Curve 36
Figure 7.10: Mulitplayer Perceptron Statistical Values 37
Figure 7.11:Mulitplayer Perceptron Accuracy Values 38
Figure 7.12: Mulitplayer Perceptron ROC Curve 38

Figure 7.13: JRIP Statistical Values 39
Figure 7.14: JRIP Accuracy Values 40
Figure 7.15: JRIP ROC Curve 40

Figure 7.16: PART Statistical Values 41
Figure 7.17: PART Accuracy Values 42
Figure 7.18: PART ROC Curve 42
Figure 7.19: OneR Statistical Values 43
Figure 7.20: OneR Accuracy Values 44
Figure 7.21: OneR ROC Curve 44
Figure 7.22: ZeroR Statistical Values 45
Figure 7.23: ZeroR Accuracy Values 46
Figure 7.24: ZeroR ROC Curve 46

Figure 7.25: RBFStatistical Values..47
Figure 7.26: RBF Accuracy Values..48
Figure 7.27: RBF ROC Curve..48
Figure 7.28: J48Statistical Values..49
Figure 7.29: J48 Accuracy Values..50
Figure 7.30:J48 ROC Curve ………..50

LIST OF ABBREVIATION

BI	:	Business Intelligence
CRM	:	Customer Relationship Management
DB	:	Database
IS	:	Information Systems
IT	:	Information Technologies
OLAP	:	Online Analytical Process

1. INTRODUCTION

1.1 PROBLEM DEFINITION

All around the world, there are many problems that people have such as their children's problems or their jobs' problems, etc… Especially, today's busy life can cause people to avoid some problems. Psychological problems are just conclusions and we have these problems until we solve them. For this reason, we have to find the reasons why these problems happened.

There are a lot of psychological problems. When psychological problems interfere with your emotional or physical health, your relationships, work productivity, or life adjustment, you need to talk to someone who can help... a psychologist. As I said before there some major psychological problems. These are;Depression, anxiety disorders, schizophrenia, childhood disorders, impulse control disordes, personality disorders, adjustment disorders and family problems.

Psychology is the scientific study of behavior and mental process. Psychologists interest nervous system, sensation and perception, learning and memory, intelligence, language, thought, growth and development, personality, stress and health, psychological disorders, ways of treating those disorders, sexual behavior, and the behavior of people in social settings such as groups and organizations (Rathus,2005).

In this work, data mining method is used for people with psychological disorders of behavior to examine the problems giving rise to this situation. Psychological problems can have many reasons. Generally, patients with these types of problems; we look into age, sex, marital status, income, change of personality type, reason to start, degree of illness, duration of illness, repeat status, behavior to outside, changes of moods, illness situation, using medicine and effects.

Today, with the increase in the volume of digital data to new issue has fields such as; a large amount of, multidimensional and complex data processing method or develop systems; new type of method, protocol or infrastructure development; improve the use of the data and security models.

Brings to mind a large amount of the data, the first concept is “Data Mining”. Data mining, previously unknown, current and applicable information data can be defined as dynamic processes are obtained with the gas. It helps us to find with large data in hidden information of database systems. With data mining everything is simpler. Select the data mining algorithm that needs to be done, the correct one, and to determine their use of the colon.

Data mining isn’t a solution in itself. However, the decision to support the process to reach a solution, a tool which provides the information required to solve the problem. Being a new model, predictions about the record was created. The degree of the accuracy of the estimates sets out sets out the success of the model created. Data mining methods and available data classification data relationships, or links, to be grouped together, creating statistical results are generated models.

Briefly summarize the data miningworks great and complex data, it can produce all kinds of solutions by using the data, it uses some disciplines such as statistics, artificial intelligence, machine learning, and knowledge discovery in databases, computer science, and construction, it searches for previously unknown, verifiable, information can be enabled, it uses automatic or semi-automatic working solution tools, data mining is a rapidly growing sector and it's also used in many industrial. Moreover, there are tools depending on the problem issues (Tang, 2005).

There are many sectors where data mining using. The main sectors are banking where used for risk analysis, risk management, client portfolio evolution, credit cards, credit claims that detection of hidden between the existing data as well as a lot about data mining is used. Marketing where using marketing campaigns, customer relationships management, cross-sell analysis data mining are used to determination customers purchasing patterns, to keep getting existing customers, acquiring new customers, market basket analysis,sales forecasting, customer relationship management, customer value analysis. Insurance where use to find the reason for determination of customer loss, prevention of irregularities. Telecommunications where used for fraud detection, estimate the lines density. Stock exchange where used for stock price prediction, global market analysis. Health care and pharmaceutical industry where used for medical diagnosis, determining the appropriate treatment process, especially determination of

DNA in the queues. Science and Engineering where use for empirical data on model by establishing scientific and technical problems to be solved and Web Companies.

The using information data will be analyzed using WEKA, which is programmed by The University of Waikato. WEKA is a collection of different machine learning algorithms for data mining processes. In order to analyze the customer behaviors, we used pre-processing, classification and clustering methods in WEKA. By pre-processing and discretizing, the aim is to summarize the current data in the best way so that it could be understood easily just looking at the WEKA results (Dener, Dörterler & Orman, 2009).

2. BUSINESS INTELLIGENCE

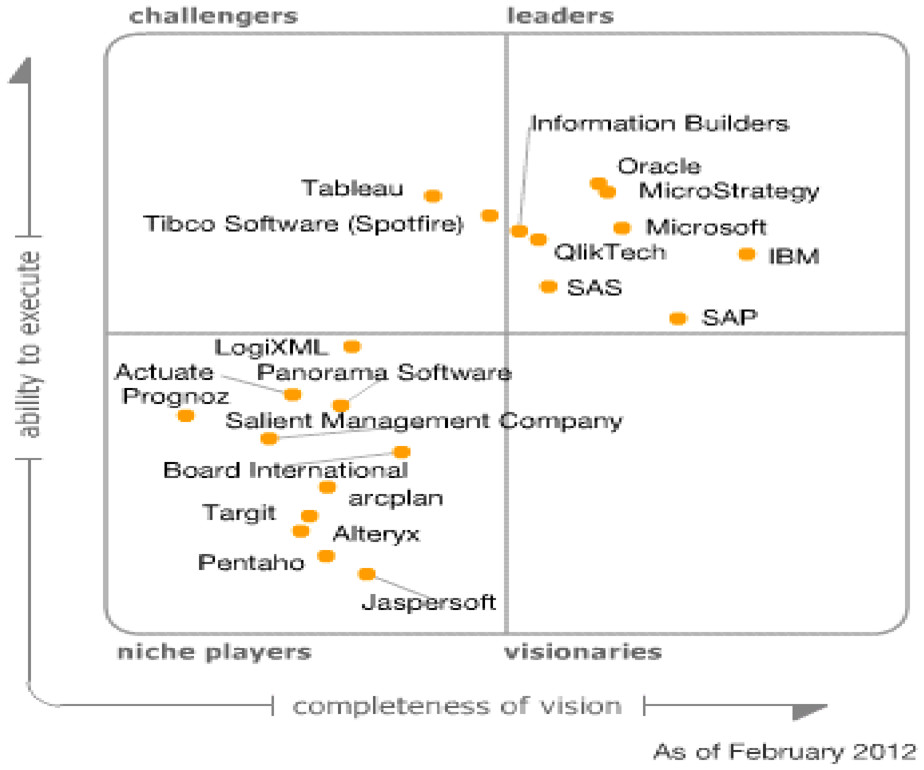
In today’s life, because of the globalization, increasing competition, developing technology helps to raise importance of information. In this situation unfortunately get busy information technology departments. Data collection, storage, processing, analyzing this information, and this information is referred to as creating business strategies with business intelligence. To analyze an organization's data and all of the various software applications used to report called business intelligence. Business intelligence solutions companies help achieve impressive ROI conversion values and in doing so have to gain profit. Business intelligence (BI) is defined as the ability for an organization to take all its capabilities and convert them into knowledge, ultimately, getting the right information to the right people, at the right time, via the right channel. This produces large amounts of information which can lead to the development of new opportunities for the organization. When these opportunities have been identified and a strategy has been effectively implemented, they can provide an organization with a competitive advantage in the market, and stability in the long run.

Business intelligence is a highly important field for organizations across all industries. A number of organizations have derived, and continue to obtain, significant benefits through the careful use of business intelligence(Sabherwal& Becerra-Fernandez, 2009).

According to the Gartner’s Business Intelligence is a user-centered process that incorporates retrieving, examining and investigating and developing insights, which leads to improved decision making (Dresner, Linden, Buytendijk, Friedman, Strange, Knox & Camm, 2002).

Business intelligence system data collection and processing, and the resulting information are responsible for its release as a user friendly.

Figure 2.1: Gartner Platform



Resource: Gartner Research Report: 2012 Business Intelligence Platform Magic Quadrant

According to figure 2.1 Gartner platform, leading business intelligence platform providers like IBM, Oracle, SAP, SAS, Micro strategy and QlikTech

According to the Gartner, leading business intelligence platform providers that force companies like Taleau and Tibco Software (Spotfire)

Generally accepted as BI is a standard & parameterized reporting, ad-hoc query and analysis environments (Relational), OLAP analysis (multidimensional), data mining and dashboard.

In all cases, use of business intelligence is viewed as being proactive. Essential component of proactive BI are real-time data warehousing, data-mining, automated anomaly and exception detection, proactive alerting with automatic recipient determination, seamless follow-trough workflow and automatic learning and refinement (Langseth and Vivatrat, 2003).

BI assists in strategic and operational decision making. A Gartner survey ranked the strategic use of BI in the following order to corporate performance management, optimizing customer relations, monitoring business activity, and traditional decision support, packaged standalone BI application for specific operations or strategies and management reporting of BI (Willen, 2002).

As seen in table 2.1, BI is related all those information systems. As a result of this situation, BI is really important to understand. If people don't understand it very carefully, their job doesn't go well.

Table 2.1: BI Relation to Other Information Systems

OLAP ↘	DATAWAREHOUSE ↓	VISULATION ↙
DATAMINING →	BUSINESS INTELLIGENCE	CRM MARKETING ←
DSS/EIS ↗	KNOWLEDGE MANAGEMENT ↑	GIS ↖

Where: OLAP = on-line data processing, CRM=customer relationship management, DSS= decision support systems, GIS = geographic information systems

2.1 BUSINESS INTELLIGENCE SOLUTIONS

Business intelligence gives you answers appropriate to specific goals. As we want to show that business intelligence solutions like integration, data transformation applications, reporting, querying, data warehousing, modeling, design, development, dashboard, BI platforms, data mining, application development, data creation and OLAP technology platforms on application development.

2.2 BUSINESS INTELLIGENCE COMPONENTS

Business intelligence is not a stand-alone software or technology. It's the combination of software and technologies. Business intelligence components, the execution of the business intelligence projects in organizations that contain data in an efficient manner (Niu & Lu & Zhang, 2009).

2.2.1 Source Systems

It provides business intelligence applications source. Business data resources are; relational databases, flat files, XML. Business intelligence applications will be analyzed to obtain the information that is the basis for the data source systems because of for this reason the starting point of the source systems business intelligence.

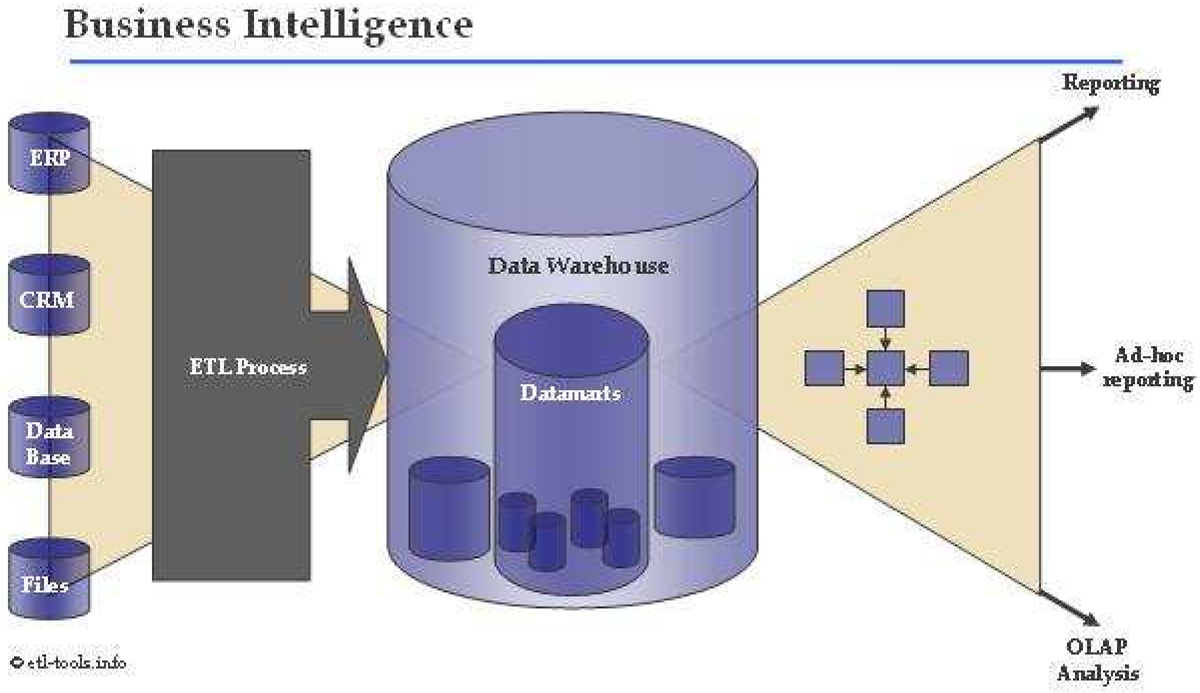
2.2.2 Data Warehouse

A data warehouse is a simple also complete and continual store of data. Moreover, it obtained from a variety of sources and it use in a business works.

Data warehouse is a process not a product. It is a technique for properly assembling and managing data from various sources for the purpose of gaining a single, detailed view of part, or all of a business.

A data warehouse is a subject oriented. Subject oriented mean that WH is organized around the major subjects of the enterprise. Moreover, it's also integrated because the source data come together from different enterprise-wide applications systems. Furthermore, a data warehouse is a time-varying.The source data in the WH is only accurate and valid at some point in time or over some time interval. The time-variance of the data warehouse is also shown in the extended time that the data is held, the implicit or explicit association of time with all data, and the fact that the data represents a series of snapshots and the last one is non-volatile. Means that data is not update in real time but is refreshing from OS on a regular basis. New data is always added as a supplement to DB, rather than replacement (Inmon, 1995).

Figure 2.2: Business Intelligence



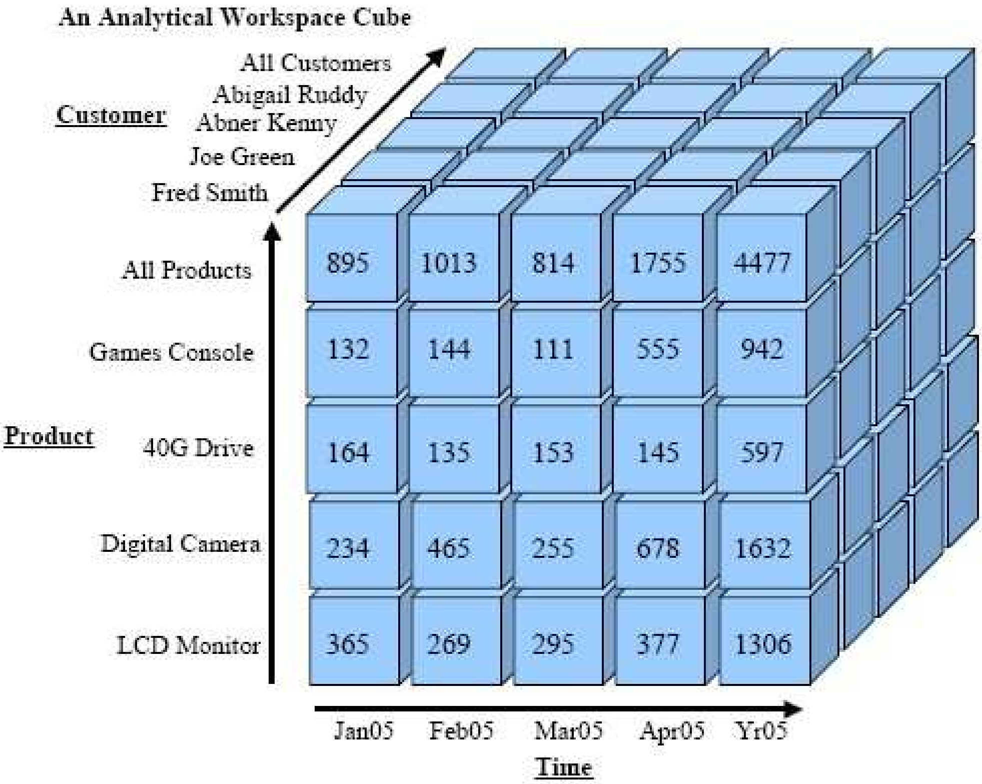
As seen in figure 2.2, BI is like a complex program. ERP, CRM, Database and files enter the ETL process to the Data warehouse and end of this there are reports, ad-hoc reporting and OLAP analysis that we get. All of these process calls business intelligence.

2.2.3 On-line Analytical Processing (OLAP)

An OLAP cube is a set of data, organized in a way that facilitates non-predetermined queries for aggregated information, or in other words, online analytical processing. OLAP is one of the computer-based techniques for analyzing business data that are collectively called business intelligence. A method that can companies look at in a lots of different ways.

There are some benefits of OLAP. Some of them are like that work with cube logics, transparency, in a server-client mode, multiple-user support, recording query from the database and solving complex calculations easily.

Figure 2.3: An Analytical Workspace Cube



	Jan05	Feb05	Mar05	Apr05	Yr05
All Products	895	1013	814	1755	4477
Games Console	132	144	111	555	942
40G Drive	164	135	153	145	597
Digital Camera	234	465	255	678	1632
LCD Monitor	365	269	295	377	1306

Resource: Comparing Materialized Views and Analytical Workspaces in Oracle Database, 2008

In the figure 2.3, the analytic workspace (AW) is used to store the multidimensional data types, e.g., the dimensions, measures and cubes.

3. KNOWLEDGE DISCOVERY IN DATABASES

Data mining the process of discovery in databases information to apply the algorithm used to occur to us. This process will need to know the properties of the data within the model will be applied to very good. Data mining algorithms to apply for transfer from pattern. In the context of large data sets from the single target low level to take a high level of information.

If we want to success in the process, first of all we should have to identify of the problem. The purpose of the business, the business must be expressed in plain language and clear focus on problem. The results obtained must be defined to measure how levels of success. Also, incorrect estimates of the costs and the benefits to be gained at this stage make accurate predictions are included in the estimates. Secondly, we have to prepare the data. For this, the establishment of this stage of the model, the problems will arise at the stage of often causes rapid restoration and reorganization of data back. This is a model for the establishment of the stages of data preparation and data analyst, and energy within the sum of the time the cause of building up discovery. There are some phases in data preparation like collection, valuation, merge and clean up, and consists of the conversion.

In the collection step, it is thought to be necessary for the defined problem, and that this data is to be collected data sources of the data. To collect data in its own data sources outside the Organization's databases are also used in organizations. Like we said in assessment, the collection of data from different sources to be used in the data mining, data can conflict as natural. These conflicts are different, and the differences in the data, the encoding for the timings of different measurements can be used. for these reasons, the best result in any good data to import models but the extent to which they are compatible with the data collected in this step are evaluated by examining. After these, there is consolidation and cleaning step. In this step, the data collected from different sources and fixed the problems identified in the previous step, and, to the extent possible, the data are collected in a single database. After that there is selection step, in this step will be done in the data selection, depending on the model. For example, a model to predict the dependent and independent variables for this step, and carries the meaning of the data set to be used in the model selected. Moreover, in the transformation step, database or data warehouse data in the summary or interconnected in a more meaningful changes to structure. For example, in the case of using an artificial neural network algorithm in an application variable is categorical yes/no; in the case of using a decision tree algorithm, for example, income variable values to be grouped as high/medium/low will increase the effectiveness of the model (Ming and the others, 1996).

After these steps, there is setting up and evaluation of the model phase. In this phase, the most appropriate model can be discovered, as defined in the problem for many is possible with a testing of the model. Supervised and unsupervised learning processes of organizational model used vary according to the models (Mohammed, 2003).

Furthermore, fourth phase is that using and updating the model. In that phase, established and can be directly accepted the validity of the model or an application. Therefore, it can be used as part of another application's sub. Established models such as fraud detection, risk analysis, credit rating business applications available or promotion planning simulation, can be integrate.

If we have to summarize these phases and steps. The prior information about the application field and with an understanding of the customer's point of view is to identify the destination of the data base information discovery process.To create the destination set of data: the discovery was to implement a subset of the variables or select a set of data, or focus on the data samples.Data cleaning and preprocessing: the collection of the necessary information for the model if appropriate noise removal for lost data that have previously been basic decision making strategies.Data reduction and projection: the task is to find the data useful to represent the bound target properties. Size reduction or conversion methods can be reduced or taken into account does not change the data representation of the number of variables can be found.The first step in the data mining method for mapping: summarizing, classification, regression, clustering is implemented as methods.Data mining algorithm and descriptive analyses, model and hypothesis selection: preferred data mining algorithms and the methods used to investigate patterns that selected data.Data mining: a special interest within the cluster form or representation is a representation of patterns; contains the classification rules and trees, regression, and clustering. Discovered information consolidation: discovered information to be grouped into another system, or the next studies simply filing, reporting is forwarded to related units.

4. DATA MINING PROCESSES

Data mining has some processing like data collection, data cleaning and conversion, setting up the model, model evaluation, scoring, application integration and model management. These processes help businesses to make their jobs easily. First of all, the most important thing is data collection.

4.1 DATA COLLECTION

Data collection is the first phase for the data mining. Data can be stored in many ways. First of all, we have to support proper data for the application from databases. After the completion of the data collection process, data is split up two parts as data testing and data test analysis (Tang, 2005).

4.2 DATA CLEANINGAND CONVERSION

Data cleaning and the conversion is the second phase of the data mining. The goal of the data conversion is to convert data source in the different formats, or values. For example, the type of the integer in the database can be converted to the type Boolean database. The reason for this is that some data mining algorithms' integer data types are more capably than Boolean data type.

The goal is the data cleaning is to clean data which in in inappropriate or incorrectly entered. In this process, the missing data is filled in automatically with the appropriate data. If the missing data is much more than usually in this point records must be deleted (Tang, 2005).

4.3 SETTING UP THE MODEL

Correctly build the model, you should understand the project' purpose very well. There is more than one algorithm for each purpose. In this case, run on-hand data on all algorithms and the most accurate result of the algorithm that is used to (Tang, 2005).

4.4 MODEL EVALUATION

Run the appropriate algorithms to the data on hand, after that there are several ways to find the most accurate result such as if there is mathematical data and we want to accurate the model then we use MAPE (Mean Absolute Percentage Error) method(Tang, 2005).

4.5 SCORING

We say that scoring instead of evaluation in the data mining.In the scoring, the main goal is to use the model for evaluation. Assessment of data that contains the new status must be for the trained model. For this reason, the trained model can be found by using the new for predictions (Tang, 2005).

4.6 APPLICATION INTEGRATION

At this point, to establish in the data mining model is that into the embedded in the application developed to run in real time.

4.7 MODEL MANAGMENT

As we know that each data mining style has a life cycle. In some applications, jobs or especially properties are static. Furthermore, there is no need to retrain the model again. Therefore, a lot of work things often changes. New data as they arrive, you must train the model again. Therefore, the data objects must be updated frequently.

5. CLASSIFICATION ALGORITHMS

5.1 DATA PRE-PROCESSING

In this step, data cleaning, data consolidation, data transformation, data reduction methods are used and in this way data analysis made available. These processes can affect the success of the model that will occur. It depends on the perspective of the person who made the application processes. Different interventions will cause different results. If you want to application to be successful, you must have the person who's knowledgeable about the topic or to work with people who are expert in this field (Oğuzlar, 2003).

5.1.1 Data Cleaning:

Data cleaning, completion of missing data, for the purposes of determining the values that the data be corrected, noise, and contrary to be reconciled, requires such as processes. There are several different ways to filling in missing data for any variable. Some of them that to disposable missing value record or records, the average of the variable can be used in place of missing values and the most appropriate value based on the existing data can be used (Oğuzlar, 2003).

The correct data may be in the next missing or incorrect data. Data cleaningis made; other phases will be used to improve the quality of the data mining model.

Data cleaning technique is used in noise data. Noise is a random variable, or that the variance as measured. Noisy data use some techniques to identify of the aim such as clustering analysis and regression. If it is a continuous variable such as price data that noisy data must be smoothed.

Some data correction techniques;

5.1.1.1 Binning:

Binning methods, low to high or high to low is used to correct the data that has been sorted. In this method, the data are sorted primarily by that separated from the amount of equal bins'. More than a bins, bin will be corrected with the help of a bin averages or medians limits.

5.1.1.2 Clustering:

Contrary to the values can be determined with clusters. Similar values can take place within the same group or cluster.

5.1.1.3 Regression:

Data can be corrected with the help of a regression function with data. Functions do not conform to the values in the wrong spots.

The other reason is the need to clean up the data in an inconsistency. Some data inconsistencies can be corrected by external functions. For example, codes can be corrected in the use of inconsistencies. A variable name can be different databases in different ways.

5.2 DATA AGGREGATION

In data mining usually data members are required merging differentdata. Different databases that the data is merged into the data warehouses. By combining different data members of the data in a single database occur in schema integration errors. For example, might be a database entries "consumer-ID", in another database it can be in the form of "customer-no". This type of schema merge to get rid of errors is used to metadata. Databases and data warehouses often have Meta data. Metadata is the data about the data. In the data combining, there is another important topic is demotion. A variable may be paid if the surplus from another table. In contradictions can cause the resulting data set redundancies (Oğuzlar, 2003).

5.3 DATA TRANSFORMATION

For using in data mining, data and data transformation are converted into suitable forms. In data transformation, there are mostly used correction, consolidation, globalization and the normalization. There are some techniques that we can show such as (Oğuzlar, 2003);

i. **Min-Max**: When we want to convert numeric values between 0 and 1 data is used. Thus the maximum and minimum numeric value is certain.

$$Y=(X-Min(X))/ (Max(X)-Min(X)) \tag{5.1}$$

In this formula;

Y: Converted values X: Observation values

Min(X): The smallest observation value

Max(X): The biggest observation value

ii. **Z-Score:** This method will take care of the average and the standard error of the data and is based on the basis of conversion to new values. In these conversions;

$$Y=(X-Mean(X))/STD(X) \tag{5.2}$$

In this formula;

Y= Conversion values X= Observation values

Mean(X) = Data’ arithmetic average

STD= the standard deviation of the data

Mean(X) = 1/n $\sum_{i=0}^{n}$(Xi) **(5.3)**

iii. **Decimal Scaling:** Taking into account the value from the decimal part diffuse daylight as a result of this situation normalization is performed. The absolute maximum number of decimal point will act depends on the value of the variable.

5.4 DATA REDUCTION

In data mining, sometimes analysis can take a long time. If you have a set of data records and removing them not cause you any problem then these data's can be reduced. Data reduction techniques are smaller volumes and data is obtained from a sample set of reduced.

There are some data reduction techniques. These are data aggregation or data cube, dimension reduction, data compression and discretization (Oğuzlar, 2003).

Data cubes are stored multivariate merged information cubes. Data cubes without making any summary information provide access to a quick calculation.

Data mining can contain hundreds of unnecessary data is sometimes done as a variable. This lowers the value of the patterns will be getting unnecessary variables and as a result of this situation data mining process becomes slow. With the aim of unnecessary variables to be disqualified-way can be done forward or back as the intuitive choices. Then each variable or variables will be included in this group is determined intuitively to the cluster. It is primarily a set of selection is performing all the way back to the intuitive.

Another method used with the aim of reducing the size decision trees. Decision trees give the best set of variables to represent the output variable, which will be discussed.

Data compression is obtained by reduced or compressed that can represent the original data, data encryption, or transformation. This will be the supported set of data is reduced data set original and this prevents data loss. These algorithms cannot be used frequently because it has limits.

Discretization, some of the data mining algorithms that handle only the category values and provides continuous data into discrete data. Categorical values are used instead of the original data values. A concept hierarchy is a decoupling of the variable given to continuous variable. Concept changes the high-level concepts instead of lower-level concept hierarchies.

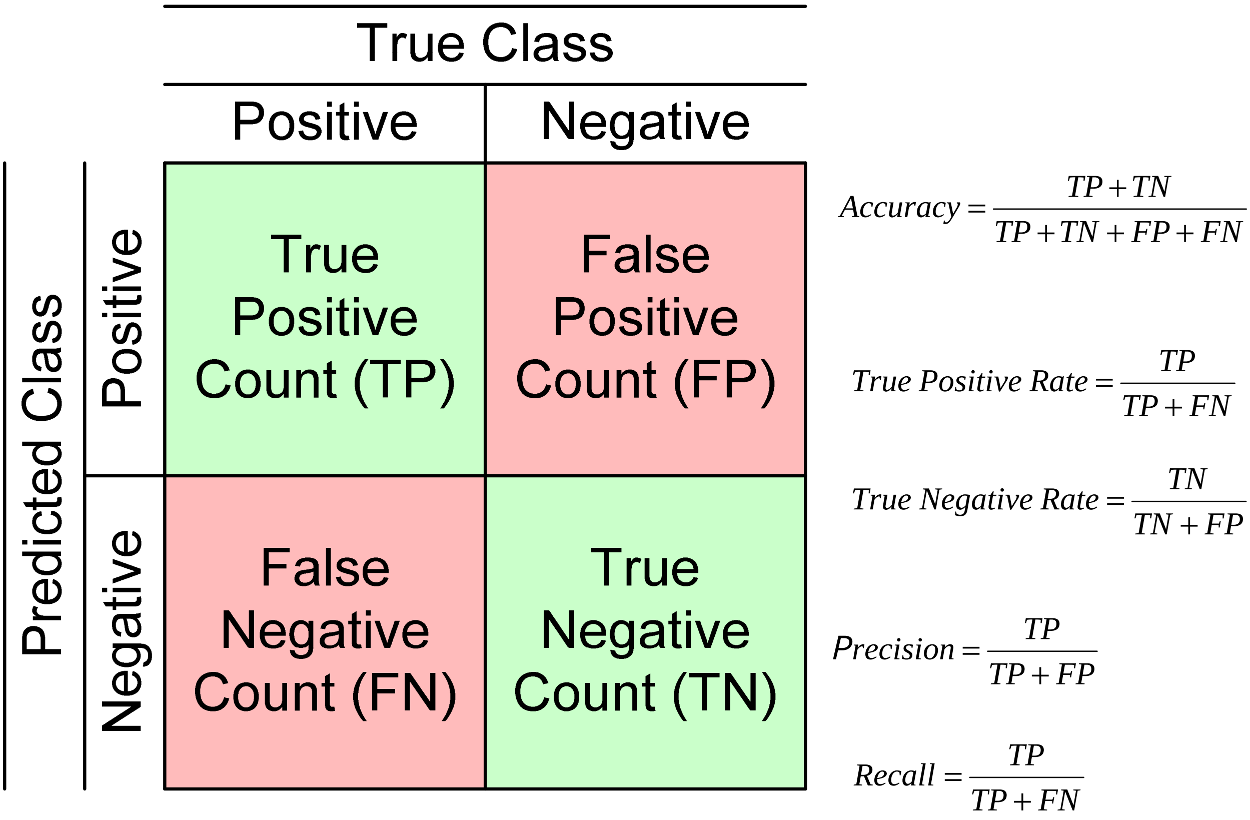
Used data mining programs is performed a data pre-processing techniques are often counted. However, special programs and related data processing or data pre-processing, there are a number of special programs in terms of strong. Especially; Bio Comp, i-Suite, KXEN, PATTERN … etc.

Different algorithms are used in the data mining, may be different for parameters. Parameters changes algorithm to algorithms or used data mining tool programs may be different. To affect the success of the model, this will consist of a selection of them.

When creating a model, the success of the model of learning and test sets that are used in determining the effects. Available as a set of test data is a set of learning and that drop is used in different methods. Learning set and test can be given different set of program files.

When evaluating the performance of the model, the basic concepts used; error rate, the precision, sensitivity and F-criteria. The success of the model, the number of right and wrong that is assigned to the class instance is related to the number of discarded class example.

Figure 5.1: Confusion Matrix



		True Class	
		Positive	Negative
Predicted Class	Positive	🟩 True Positive Count (TP)	🟥 False Positive Count (FP)
	Negative	🟥 False Negative Count (FN)	🟩 True Negative Count (TN)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$True\ Positive\ Rate = \frac{TP}{TP + FN}$$

$$True\ Negative\ Rate = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

In the figure 5.1, confusion matrix represents the success achieved as a result of the test. The real numbers in the set of rows are test samples. Columns is shows an estimate of the model.

5.5.1 Accuracy---Error Rate:

The accuracy of the model is the simplest and most popular method.Complementing this is a measure of the value-to-one ratio error.

(5.4)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

5.5.2 Precision

Certainty, which the number of True Positive example identified as a class, the class is defined as a number of samples of all.

$$Precision = \frac{TP}{TP + FP} \tag{5.5}$$

5.5.3 Sensitivity

$$Recall = \frac{TP}{TP + FN} \tag{5.6}$$

5.5.4 F--Criteria

Accuracy and precision of the criteria alone is not meaningful. You should examine each of the two criteria together. Accuracy and sensitivity of the harmonic mean of the F-criteria.

F-Criteria = (2 * Sensitivity * Precision) / (Sensitivity – Precision) **(5.7)**

6. DATA COLLECTION AND ANALYSIS

Lately, because of the technological improvements, increasing people, also environmental problems, differences between social life, family problems and education levels, there are many factors causes people' psychological life in a bad way. Moreover, there are so much effects psychological problems.

In this study, we just only talk about genetic, family, education, work, living, economic and the other psychological problems. Researches improve that because of these problems cause psychological problems.

Every psychological situation is different from people. Because of that psychological illnesses level is also different, too. These are;

Weak: This is enough to set diagnostic not too much, not too less.

Average: Symptoms are between weak and heavy.

Full Remission: In this point, there are no symptom examples or findings. But, we can indicate that this disorder can be possible by clinic way.

Partial Remission: In this point, diagnostic criteria are not enough to conquer.

These three of them are use if diagnostic criteria are not conquered. These are not psychotic but heavy, psychotic but heavy and undefined types.

If people have psychological problems, because of it his/her behavior has changed to outside. As a result of this situation, doctors can understand what the patients' illness and its degree.

At most in this study, you can see behavior problems are; behavior, anxiety, sad, crazy, offensive, exciting, concern, recession, angry, depressed and aggressive. Because of this situations are increasing by time, they should go to doctor and begin to control.

Unfortunately, psychological problems can restart again because of this using medicine can continue for life's. But sometimes it takes just less than 1 month sometime takes more than 6 years.

As we talk about before, because of the changes personality, there can be some different personality types. In this study, we talk about;

Labil: If we talk about changes of emotions.

Disinhibit Type: If we talk about people who wants to sex whatever they want and not thinking.

Aggressive Type: If we talk about changes of aggressive behavior.

Apathetic Type: If we talk about changes apatite and endiferans.

Paranoid Type: If we talk about changes behavior of skepticism and paranoid thinking.

Other Type: If we don’t know what it is?

Component Type: If patient have more than one characteristic.

Unspecified Type:

Purpose of this study is about analyzing the information discovery process and detect psychological problems by using data mining methods.

6.1 DATA COLLECTION

In this thesis, the data’s collected from a doctor who is a specialist in psychology. There are 525 data’s and 13 different variables. These variables show in table 6.1 and these variables help to detect psychological problems.

Table 6.1: Variable List

AGE	SEX
MARTIAL STATUS	INCOME
CHANGES OF PERSONALITY TYPE	REASONS TO START
DEGREE OF ILLNESS	DURATION OF ILLNESS
REPEAT STATUS	BEHAVIOUR TO OUTSIDE
CHANGES OF MOODS	ILLNESS SITUATION
USING MEDICINE	

Data’s can be numeric or the real situation. We will see it in table 6.2

Table 6.2: Discrete Time Variable List

Variable Name	Data Type	Values
Age	Integer	0-10=1,11-20=2,21-30=3,31-40=4,41-50=5,51-60=6,61-70=7
Sex	Boolean	Female=0,Male=1
Marital status	Boolean	Single=0,Married=1,Widowed=2
Income	Integer	Bad=0,Average=1,Good=2,VeryGood=3
Changes of personality type	Integer	Labil=0,Dezinhibe=1,Aggressive=2,Apatetik=3 Paranoid=4,Other=5,Component=6, Unspecified=7
Reasons to start	Integer	Genetic=0,Family=1,Education=2 Work=3,Living=4,Economic=5 Criminal=6,Psychological=7
Degree of Illness	Integer	Full remission=0,PartialRemission=1, Average=2,NotPsikoticbutheavy=3, Psikoticbutheavy=4,Undefined=5
Duration of Illness	Integer	1Year+=0,6Year+=1,6Months+=2,6Months-=3 1Months+=4,1Months-=5
Repeat Status	Integer	Yes=0,No=1,Possible=2
Behavior to Outside	Integer	Panic=0,Anxiety=1,Sad=2,Crazy=3,Offensive=4 Exciting=5,Concern=6,Recession=7, Angry=8,Depressed=9,Aggressive=10
Changes of Moods	Integer	Scared=0,Sad=1,Depressed=2,Nervous=3

		Panic=4,Cool=5,Patient=6,Peace=7
Illness Situation	Integer	Increasing=0,Decreasing=1,Same=2
Using medicine	Integer	HeavyDoze=0,Partial=1,None=2

To show data's effect, we use Weka 3.7.1(Witten & Frank, 2005) platform and use ranker method and from assessment group we use GainRatioAttributeEval and we got values as seen in table 6.3;

In this table, the biggest effect is illness situation and the lowest effect is age.

Table 6.3: Sorting List

Sort Value	Variables
0.9998	Illness Situation
0.48348	Changes of Moods
0.34812	Reasons to Start
0.31647	Changes of Personality Type
0.23136	Behavior to Outside
0.14965	Duration of Illness
0.131	Income
0.04719	Repeat Status
0.02606	Sex
0.00646	Marital Status
0	Age

6.2 APPLIED CLASSIFICATION METHODS

6.2.1 Bayesian Networks:

Bayes networks produce probability forecasts as the network output like logistic regression models.

Forecast is not produced by system. The basic purpose of the system is to estimate the probability of forecast that it is valuable for all class models for all class value. It uses the conditional and unconditional cooperation possibilities. By defining subsets of variables stand-alone classification is used between a conditional Bayes (Witten& Frank, 2005).

In this formula, we use that A, B, C for known event and we try to find X in estimating the value of presence.

$$P(X|A,B,C) = \frac{P(A|X)P(B|X)P(C|X)P(X)}{P(A,B,C)} \qquad \textbf{(6.1)}$$

6.2.2 Multiplayer Perceptron:

Artificial neural networks are simulation system and they are based on the mathematical models. These systems work as biological neural networks. Generally, there are a lot of entry and just only one output. In the entry layer, networks take values from vector values. In the privacy layer, each entry multiple own value and results adding as a result we get the new values. Later, this value feed function output (Witten & Frank, 2005).

6.2.3 Ripper Algorithm (JRIP)

In this algorithm, sets are thought like their own size. For each sets, there are separate rule (Witten& Frank, 2005).

This algorithm generated a detection model composed of resource rules that was built to detect future examples of malicious executable. This algorithm used libBFD information as features. Ripper is a rule-based learner and it builds a set of rules that identify the classes while minimizing the amount of error. The error is defined by the number of training examples misclassified by the rules (Zadok, 2001).

6.2.4 Partial Decision Trees

Decision trees are the most usable model between classification methods. They are a simple, but powerful form of multiple variable analyses. They provide unique capacities to augment andaccompany.

i. Traditional statistical forms of analysis (such as multiple linear regressions).

ii. A variety data mining tools and techniques (such as neural networks).

iii. Recently developed multidimensional forms of reporting and analysis found in the field of business intelligence.

Decision trees are produced by algorithms that recognizedifferent ways of splitting a data set into branch-like segments. These segments form aturn upside down decision tree that arises with a root node at the top of the tree. (De Ville, 2006).

Entropy is an algorithm for determine tree' branches. Entropy depends on the formulas that difference from each other data's. If we think 2 expressions which are same values and as a result of this situation their entropies' are equals zero.

$$Entropy(p1, p2, \ldots\ldots, pn) = \sum (pi\ log\ (1/pi)) \qquad \textbf{(6.2)}$$

Entropy gets different values when the branches have different modes. We called these differences between values as acquisition.

$$\text{Acquisition}(D; S) = H\ (D) - \sum P(Di)\ H\ (Di) \qquad \textbf{(6.3)}$$

6.2.5 Bayes Rule

Bayes rule depends on the probabilistic inference. It moves forward if our values and hypothesis are true. It is chosen if it has maximum probability hypothesis.

6.2.6 OneR Rule

Oner algorithm' name come from One Rule' head letters. It's easy and reliable classification model. It creates a rule for every forecast. Later, from these rules, it chose the lowest error rate part. We chose the maximum guess that is most likely to occur (Witten& Frank, 2005).

6.2.7 ZeroR Rule:

The Zeror rule classifier takes a look at the target attribute and its possible values. It always provides the output value most commonly found for the target attribute in the given dataset. Zeror as its names suggests; it does not include any rule that works on the non-target attributes (Venugopal&Patnaik, 2011).

6.2.8 Statistical Accuracy Metrics

Statistical accuracy metrics are used for measure experimental results. Widely used common metrics; mean absolute error, mean square error, root mean squared error. In this thesis, I used root mean square error to compare the methods (McClish, 1987).

6.2.9 Root Mean Squared Error

RMSE measures the quality of the fit between the actual data and the predicted model. RMSE is one of the most frequently used measures of the goodness of fit of generalized regression models. RMSE is always above zero. Its minimum occurs only when the estimate is indefinitely close to the real data(Salkind, 2010).

$$RMSE = \sqrt{E} = \sqrt{\frac{\sum_{i=1}^{N}(p_i - r_i)^2}{N}} \qquad \textbf{(6.4)}$$

6.2.10 Receiver Operating Characteristic (ROC)

A receiver operating characteristic (ROC), or simply ROC curve is a graphical plot. ROC curves are usually used in machine learning and data mining research. In the classification problems, the ROC curve is a technique for envisaging, arranging and selecting classifiers, derived from their performance(Gorunescu, 2011).

If we have to show, what is false positive rate, false negative rate, likelihood ratio positive and likelihood ratio negative, there are some algorithms that we can get them.

i. False positive rate (FP rate)=FP/(FP+TN)=1-specificity;
ii. False negative rate (FN rate)=FN/(TP+FN)=1-sensitivity;
iii. Likelihood ratio positive (LR+)sensitivity/(1-specificity);
iv. Likelihood ratio negative (LR-) = (1-sensitivity)/specificity.

6.2.5 J48 Rule

J48 is a clone of an earlier algorithm developed by J. Ross Quinlan, the very popular C4.5. Decision trees are a classic way to represent information from a machine learning algorithm, and offer a fast and powerful way to express structures in data (Witten & Frank, 2005).

J48 employs two pruning methods. The first is known as subtree replacement. This means that nodes in a decision tree may be replaced with a leaf -- basically reducing the number of tests along a certain path. This process starts from the leaves of the fully formed tree, and works backwards toward the root. The second type of pruning used in J48 is termed sub treerising. In this case, a node may be moved upwards towards the root of the tree, replacing other nodes along the way. Sub treerising often has a negligible effect on decision tree models. There is often no clear way to predict the utility of the option, though it may be advisable to try turning it off if the induction process is taking a long time. This is due to the fact that sub treerising can be somewhat computationally complex(Han &Kamber, 2006).

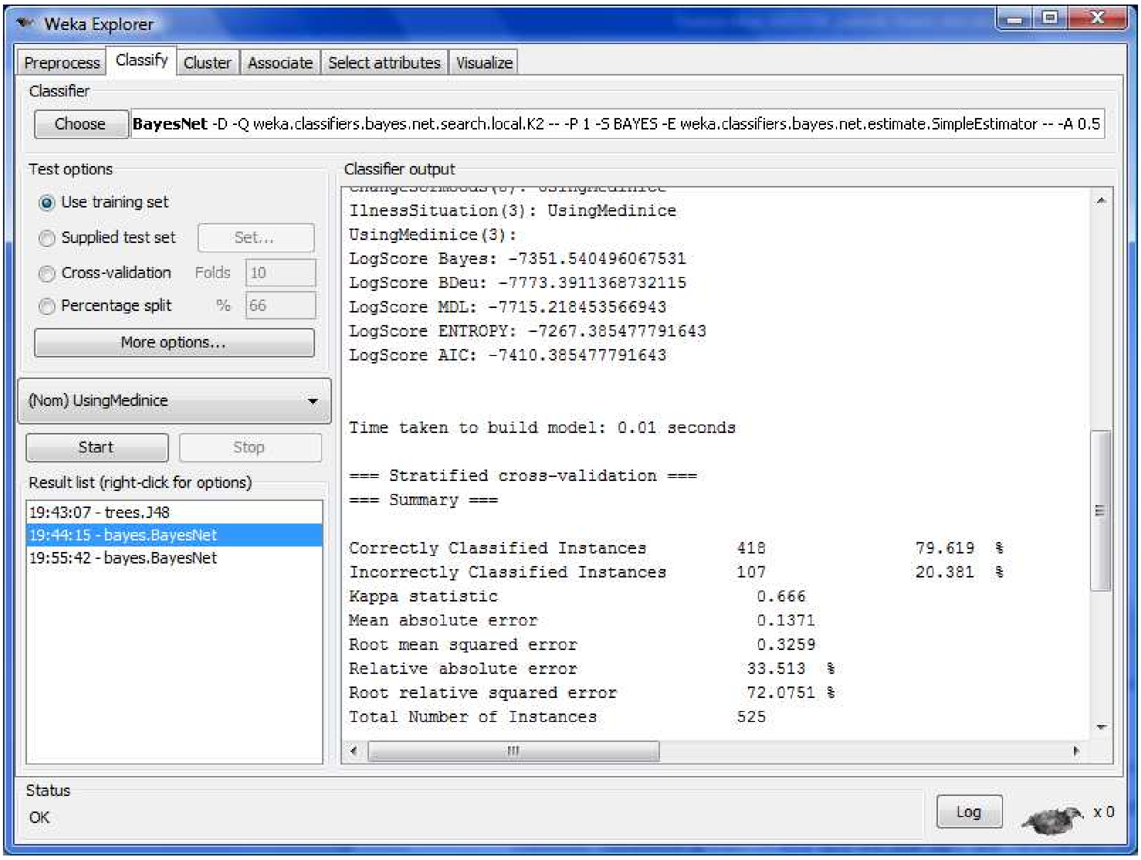
7. FINDINGS

We will use data mining classification methods for detecting and estimating psychological problems. There are many psychological problems that some of them are really important but some of them not as important as they are. There are many reasons that these psychological problems occur. In this thesis, mostly we talked about genetic, family, education, work, living, economic and other psychosocial problems.

7.1 WEKA BAYES NETWORK APPLICATIONS

We get these outputs when we use Weka 3.7.1 classification methods from Bayes Network application. In the table, there are 525 information of psychology and from this information only 107 (20.381percentage) incorrectly classified instances and the other 418 (79.619percentage) are correctly classified instances. Because RMSE value is nearly 0 and kappa statics value is nearly equal to 1. For this reason, the Bayes Network application shows that it is successful.

Figure 7.1: BayesNetStatistical Values



BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5

Test options

● Use training set

○ Supplied test set

○ Cross-validation Folds 10

○ Percentage split % 66

(Nom) UsingMedinice

Result list (right-click for options)

19:43:07 - trees.J48
19:44:15 - bayes.BayesNet
19:55:42 - bayes.BayesNet

Classifier output

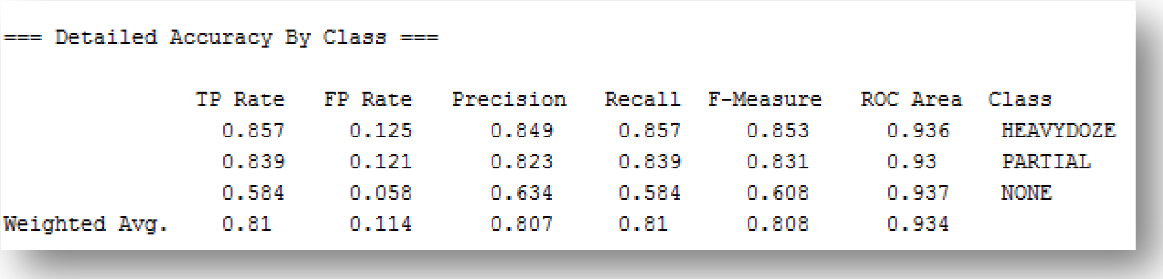
```
IlnessSituation(3): UsingMedinice
UsingMedinice(3):
LogScore Bayes: -7351.540496067531
LogScore BDeu: -7773.3911368732115
LogScore MDL: -7715.218453566943
LogScore ENTROPY: -7267.385477791643
LogScore AIC: -7410.385477791643


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         418               79.619  %
Incorrectly Classified Instances       107               20.381  %
Kappa statistic                          0.666
Mean absolute error                      0.1371
Root mean squared error                  0.3259
Relative absolute error                 33.513  %
Root relative squared error             72.0751 %
Total Number of Instances              525
```

Status

OK

In Figure 7.1, we see detailed information about Bayes Network. We use cross validation fold 10, we get detailed information about Bayes Network. However, we get correctly classified instances 418 and we just only get 79.619percentage.

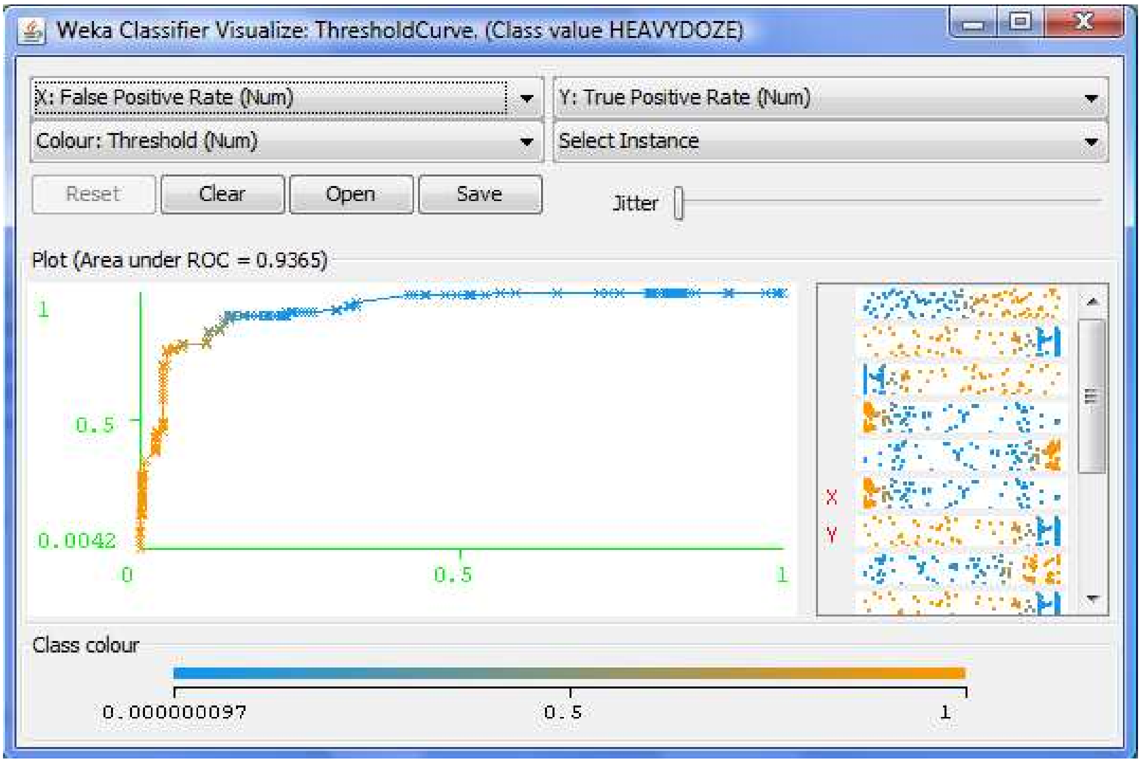
Figure 7.2: Bayes Networks Accuracy Values



```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.857     0.125      0.849     0.857     0.853      0.936    HEAVYDOZE
                 0.839     0.121      0.823     0.839     0.831      0.93     PARTIAL
                 0.584     0.058      0.634     0.584     0.608      0.937    NONE
Weighted Avg.    0.81      0.114      0.807     0.81      0.808      0.934
```

In figure 7.2, we get Bayes Network accuracy values. In this figure, the ROC areas nearly 1 as a result of this solution, we get the successful application.

Figure 7.3: Bayes Networks ROC Curve



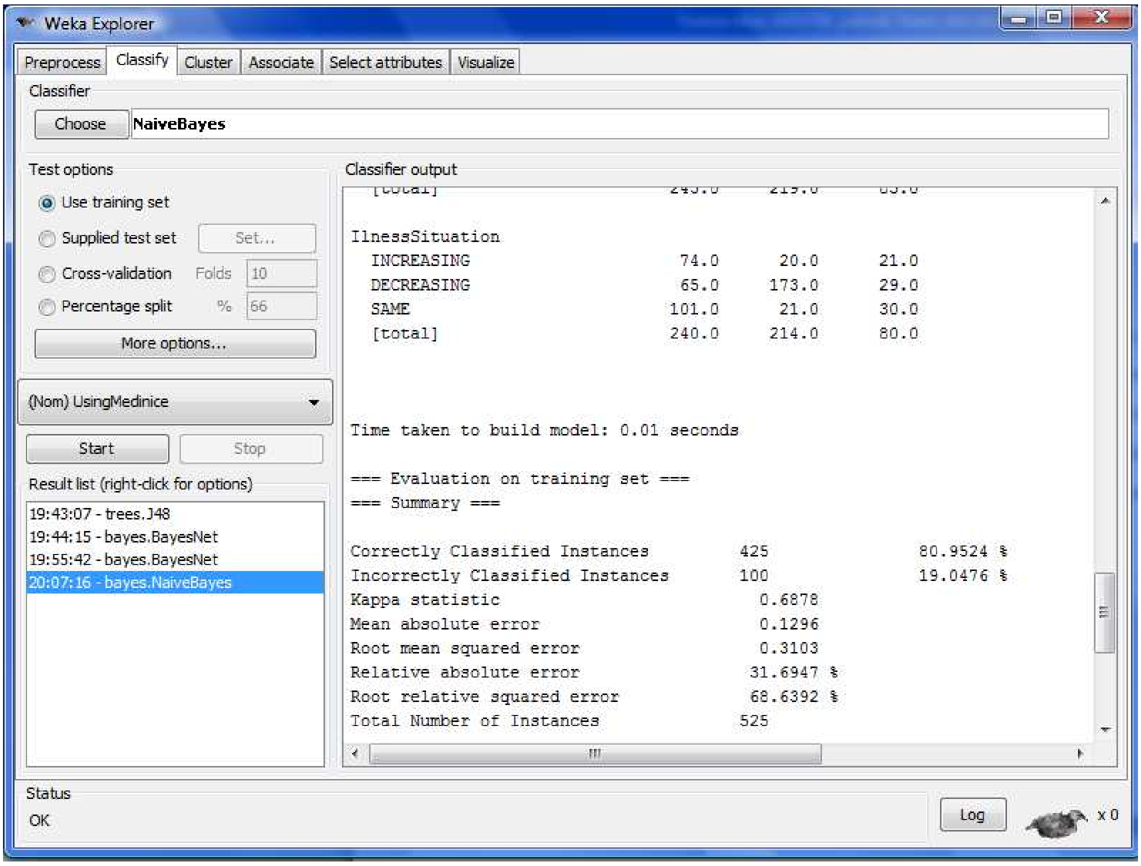
Weka Classifier Visualize: ThresholdCurve. (Class value HEAVYDOZE)
X: False Positive Rate (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot (Area under ROC = 0.9365)
1
0.5
0.0042
0
0.5
1
X
Y
Class colour
0.000000097
0.5
1

In weka classifier visualize, we chose the thresholdCurve for Heavy Doze part and we get the ROC value is 0.9365.

7.2 WEKA NAÏVE BAYES APPLICATION

After we use Weka 3.7.1 classification methods from Naïve Bayes application, we get these outputs. In the table 2.1, there are 525 information of psychology and from this information only 100 (19.0476percentage) incorrectly classified instances and the other 425 (80.9524percentage) are correctly classified instances. Because RMSE value is nearly 0, the application shows that it is successful than Bayes Network application.

Figure 7.4:Naïve Bayes Statistical Values



NaiveBayes

● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) UsingMedinice

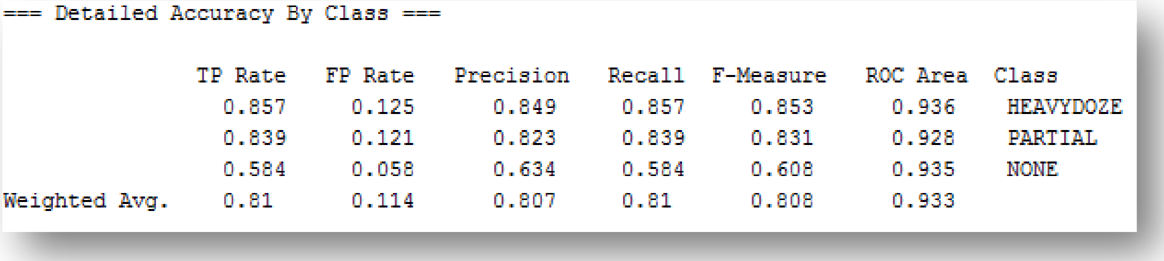
19:43:07 - trees.J48
19:44:15 - bayes.BayesNet
19:55:42 - bayes.BayesNet
20:07:16 - bayes.NaiveBayes

```
IlnessSituation
  INCREASING                        74.0       20.0       21.0
  DECREASING                        65.0      173.0       29.0
  SAME                             101.0       21.0       30.0
  [total]                          240.0      214.0       80.0



Time taken to build model: 0.01 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         425               80.9524 %
Incorrectly Classified Instances       100               19.0476 %
Kappa statistic                          0.6878
Mean absolute error                      0.1296
Root mean squared error                  0.3103
Relative absolute error                 31.6947 %
Root relative squared error             68.6392 %
Total Number of Instances              525
```

Status: OK

In Figure 7.4, we see detailed information about Naïve Bayes Network. We use Cross-validation fold 10, we get detailed information about Naïve Bayes Network. However, we get good information like correctly classified instances 425 and we just only get 80.9524percentage.

Figure 7.5: Naïve Bayes Accuracy Values

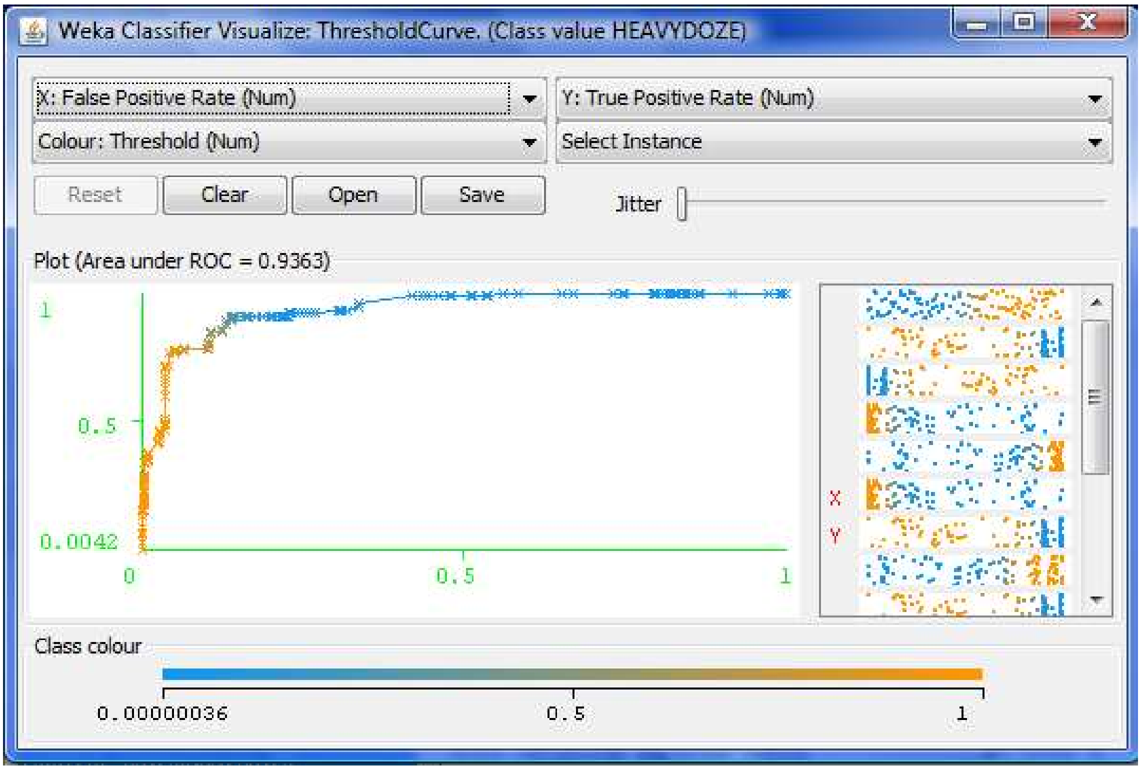


```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.857     0.125     0.849     0.857     0.853      0.936    HEAVYDOZE
                 0.839     0.121     0.823     0.839     0.831      0.928    PARTIAL
                 0.584     0.058     0.634     0.584     0.608      0.935    NONE
Weighted Avg.    0.81      0.114     0.807     0.81      0.808      0.933
```

In figure 7.5, there are ROC values. As we seen figure 8.6, under the curve value is 0.9363.

Because of the increase of FP values and decrease of the TP values to the Bayes Network, fault percentage is more less BayesNet.

Figure 7.6: Naïve Bayes ROC Curve



Weka Classifier Visualize: ThresholdCurve. (Class value HEAVYDOZE)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

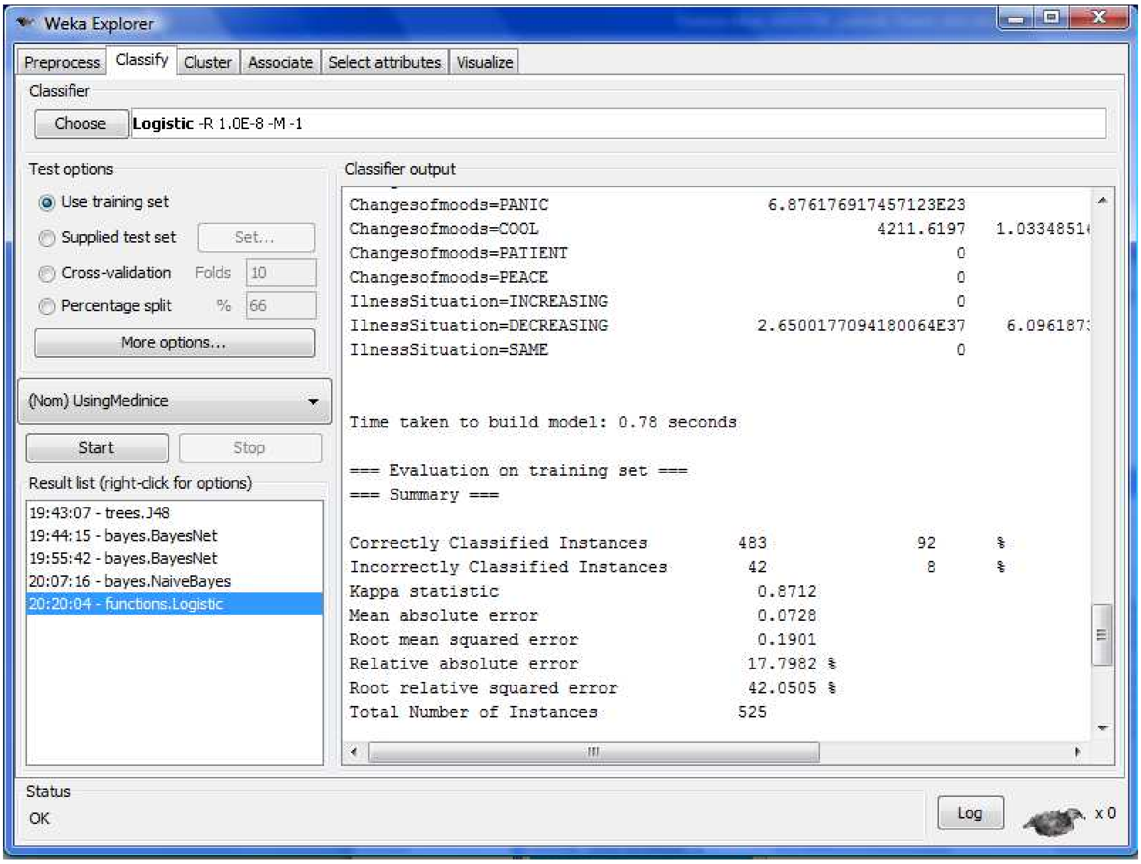
Select Instance

Plot (Area under ROC = 0.9363)

7.3 WEKA LOGISTIC APPLICATION

We get better results than Naïve Bayes and BayesNet applications. Because of the result of the RMSE is nearly 0 and the correctly classified instances 483(92percentage) and the incorrectly classified instance 42(8 percentage), we saw that these application is better than those two.

Figure 7.7: Logistic Statistical Values



Classifier: **Logistic** -R 1.0E-8 -M -1

Test options

● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) UsingMedinice

Result list (right-click for options)

```
19:43:07 - trees.J48
19:44:15 - bayes.BayesNet
19:55:42 - bayes.BayesNet
20:07:16 - bayes.NaiveBayes
20:20:04 - functions.Logistic
```

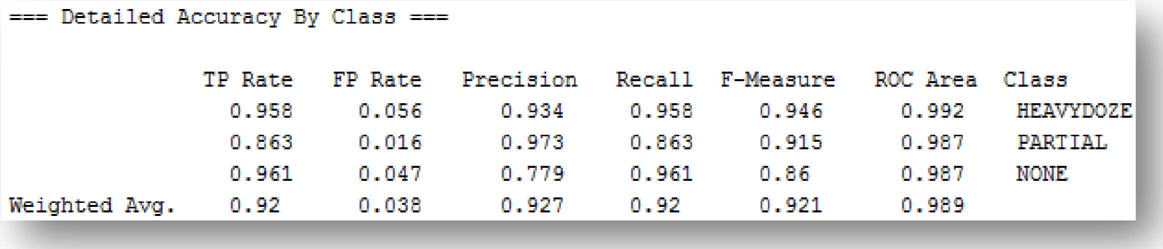
Classifier output

```
Changesofmoods=PANIC                      6.876176917457123E23
Changesofmoods=COOL                                  4211.6197    1.0334851
Changesofmoods=PATIENT                                       0
Changesofmoods=PEACE                                         0
IlnessSituation=INCREASING                                   0
IlnessSituation=DECREASING               2.6500177094180064E37     6.096187
IlnessSituation=SAME                                         0


Time taken to build model: 0.78 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         483               92      %
Incorrectly Classified Instances        42                8      %
Kappa statistic                          0.8712
Mean absolute error                      0.0728
Root mean squared error                  0.1901
Relative absolute error                 17.7982 %
Root relative squared error             42.0505 %
Total Number of Instances              525
```

Status
OK

In Figure 7.7, we see detailed information about Logistic. We use cross validation fold 10, we get detailed information about Logistic. Moreover, we get information like correctly classified instances 483 and we only get 92percentages, and just only 42 incorrectly classified instance.

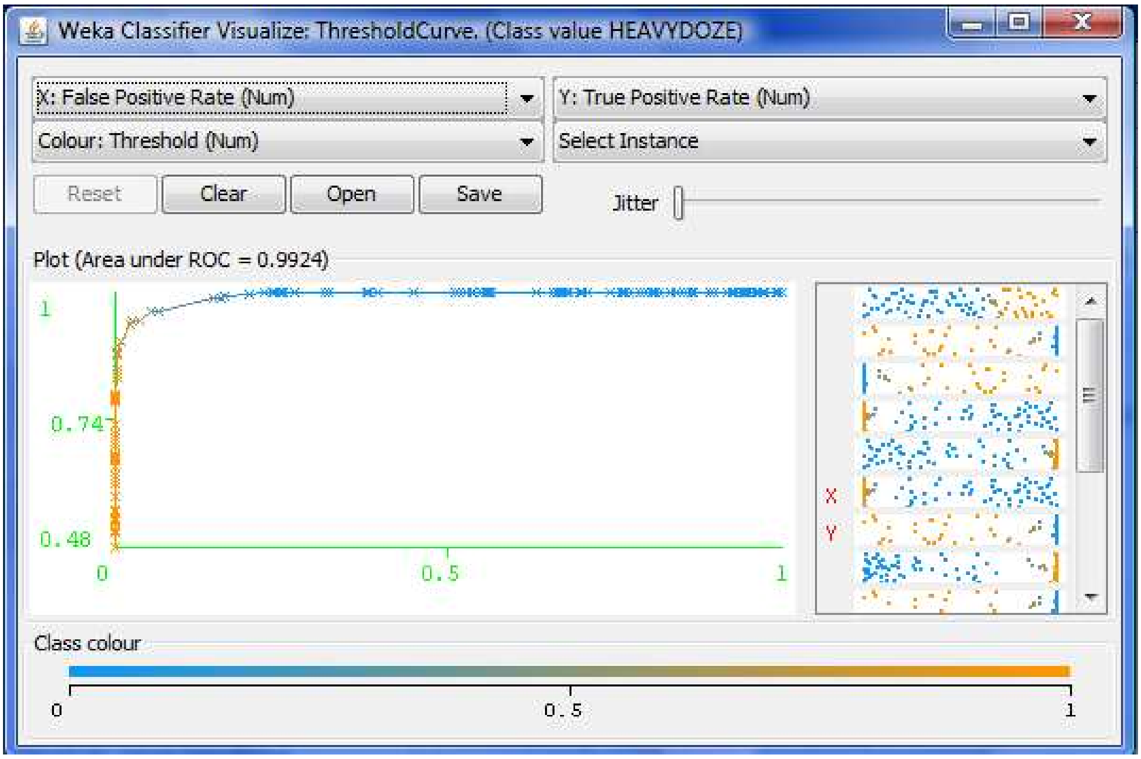
Figure 7.8: Logistic Accuracy Values



```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.958     0.056     0.934     0.958     0.946      0.992    HEAVYDOZE
                 0.863     0.016     0.973     0.863     0.915      0.987    PARTIAL
                 0.961     0.047     0.779     0.961     0.86       0.987    NONE
Weighted Avg.    0.92      0.038     0.927     0.92      0.921      0.989
```

In figure 7.8, there are ROC values. We get ROC area is nearly 1. In the ROC curve error percentage, FP values are decreasing and TP values are increasing.

Figure 7.9: Logistic ROC Curve



Weka Classifier Visualize: ThresholdCurve. (Class value HEAVYDOZE)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

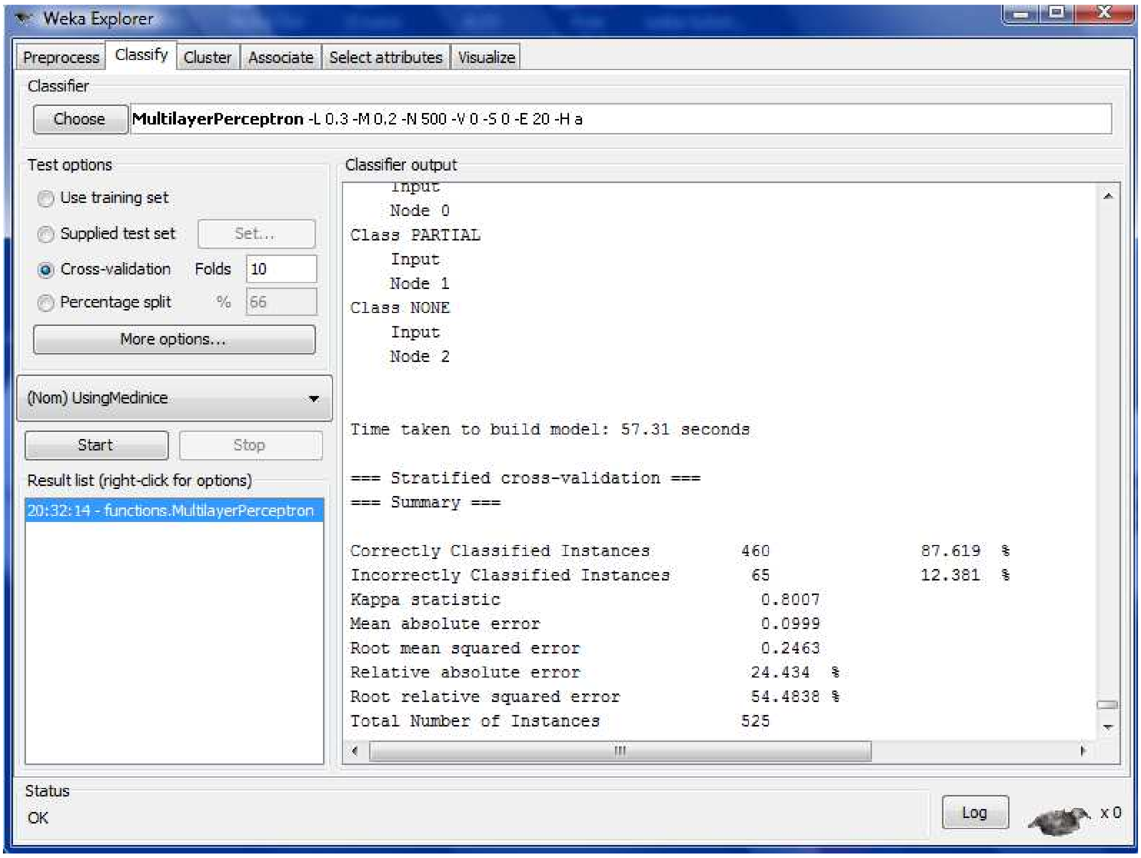
Plot (Area under ROC = 0.9924)

As we seen in figure 7.9, the successful place is under the curve 1.

7.4 WEKA MULTIPLAYER PERCEPTRON APPLICATON

In the multiplayer perceptron application when we use Weka 3.7.1 classification methods, we get these outputs. In the table, there are 525 information of psychology and from this information 460(87.619percentages) are correctly classified instances. Because RMSE value is nearly 0, the application shows that it is successful and there is no fault value.

Figure 7.10: Multiplayer Perceptron Statistical Values



Classifier: MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

Test options:
○ Use training set
○ Supplied test set Set...
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) UsingMedinice

Result list (right-click for options): 20:32:14 - functions.MultilayerPerceptron

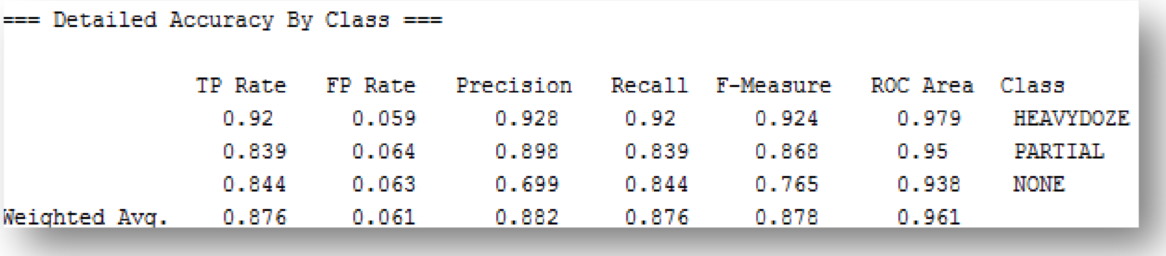
Classifier output

```
    Input
    Node 0
Class PARTIAL
    Input
    Node 1
Class NONE
    Input
    Node 2


Time taken to build model: 57.31 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         460               87.619  %
Incorrectly Classified Instances        65               12.381  %
Kappa statistic                          0.8007
Mean absolute error                      0.0999
Root mean squared error                  0.2463
Relative absolute error                 24.434  %
Root relative squared error             54.4838 %
Total Number of Instances              525
```

Status
OK

In Figure 7.10, we see detailed information about Multiplayer Perceptron. We use cross validation fold 10, we get detailed information about Multiplayer Perceptron Moreover; it gets more time to build on data (57.31 second).

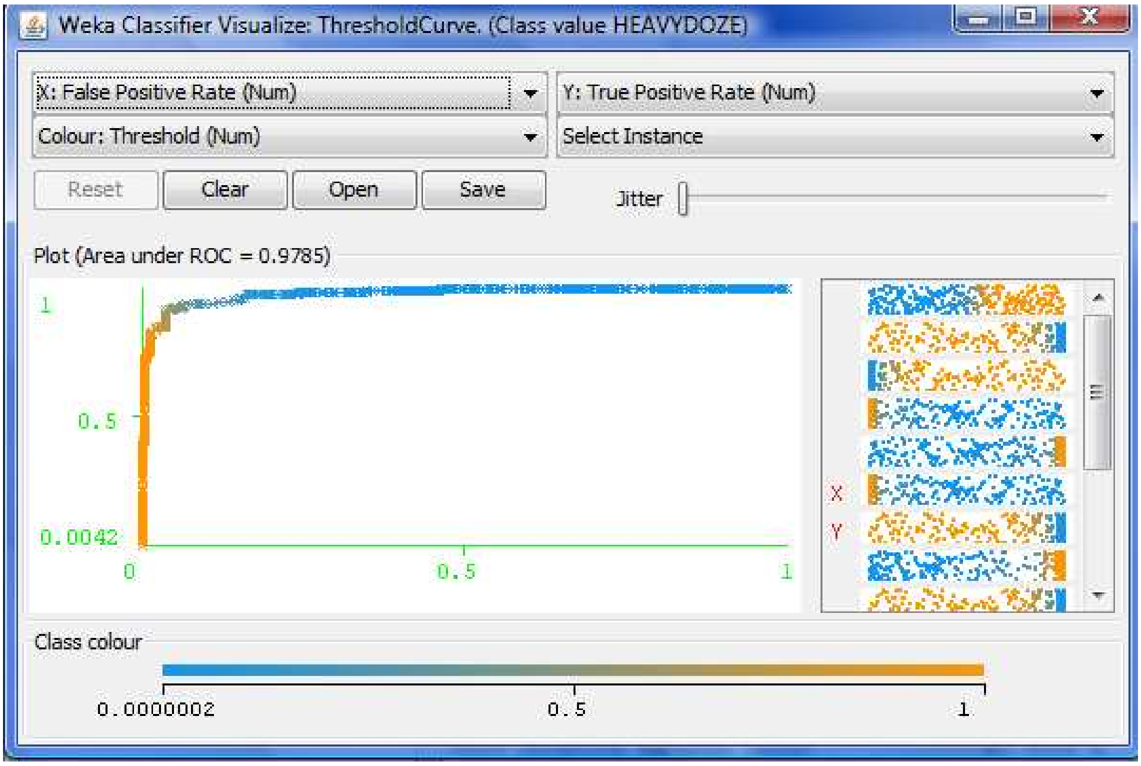
Figure 7.11: Multiplayer Perceptron Accuracy Values



```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.92      0.059     0.928      0.92      0.924      0.979    HEAVYDOZE
                 0.839     0.064     0.898      0.839     0.868      0.95     PARTIAL
                 0.844     0.063     0.699      0.844     0.765      0.938    NONE
Weighted Avg.    0.876     0.061     0.882      0.876     0.878      0.961
```

In the figure 7.11, there are ROC values.

Figure 7.12: Multiplayer Perceptron ROC Curve



Weka Classifier Visualize: ThresholdCurve. (Class value HEAVYDOZE)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

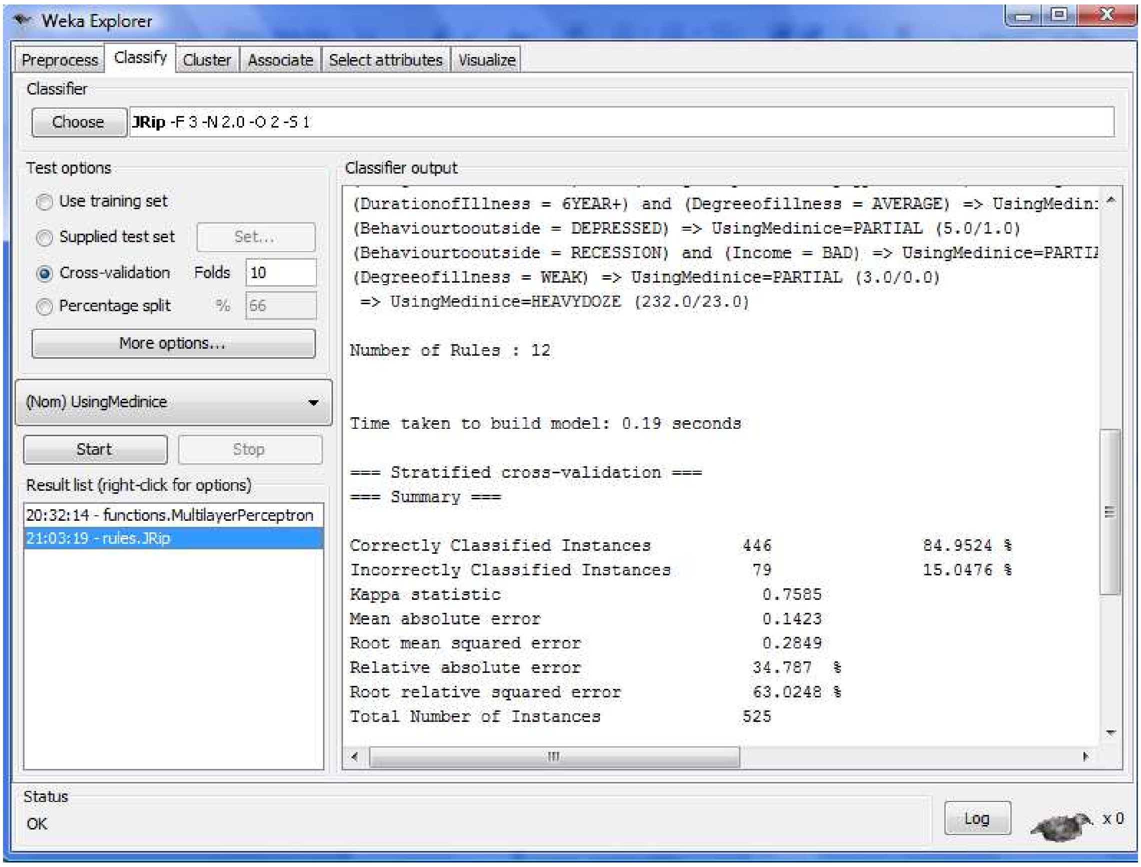
Plot (Area under ROC = 0.9785)

In figure 7.12, we see ROC curve for Heavy Doze value and as we see that the positive are is 0.9785.

7.5 WEKA JRIP APPLICATIONS

In the table, there are 525 information of psychology and from this information there is 79(15.0476 percentage) incorrectly classified instances and 446 (84.9524percentages) are correctly classified instances. Because RMSE value is nearly 0, the application shows that it is successful.

Figure 7.13: JRIP Statistical Values



JRip -F 3 -N 2.0 -O 2 -S 1

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) UsingMedinice

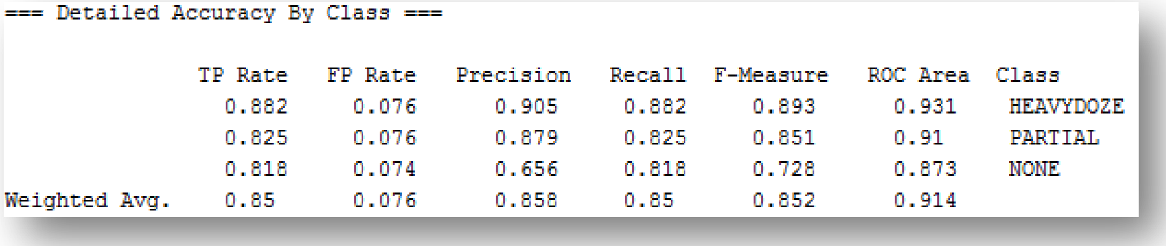
20:32:14 - functions.MultilayerPerceptron
21:03:19 - rules.JRip

```
(DurationofIllness = 6YEAR+) and (Degreeofillness = AVERAGE) => UsingMedin:
(Behaviourtooutside = DEPRESSED) => UsingMedinice=PARTIAL (5.0/1.0)
(Behaviourtooutside = RECESSION) and (Income = BAD) => UsingMedinice=PARTIA
(Degreeofillness = WEAK) => UsingMedinice=PARTIAL (3.0/0.0)
 => UsingMedinice=HEAVYDOZE (232.0/23.0)

Number of Rules : 12


Time taken to build model: 0.19 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         446               84.9524 %
Incorrectly Classified Instances        79               15.0476 %
Kappa statistic                          0.7585
Mean absolute error                      0.1423
Root mean squared error                  0.2849
Relative absolute error                 34.787  %
Root relative squared error             63.0248 %
Total Number of Instances              525
```

Status
OK

In Figure 7.13, we see detailed information about Logistic. We use cross-validation test and we get 446 correctly classified instances.

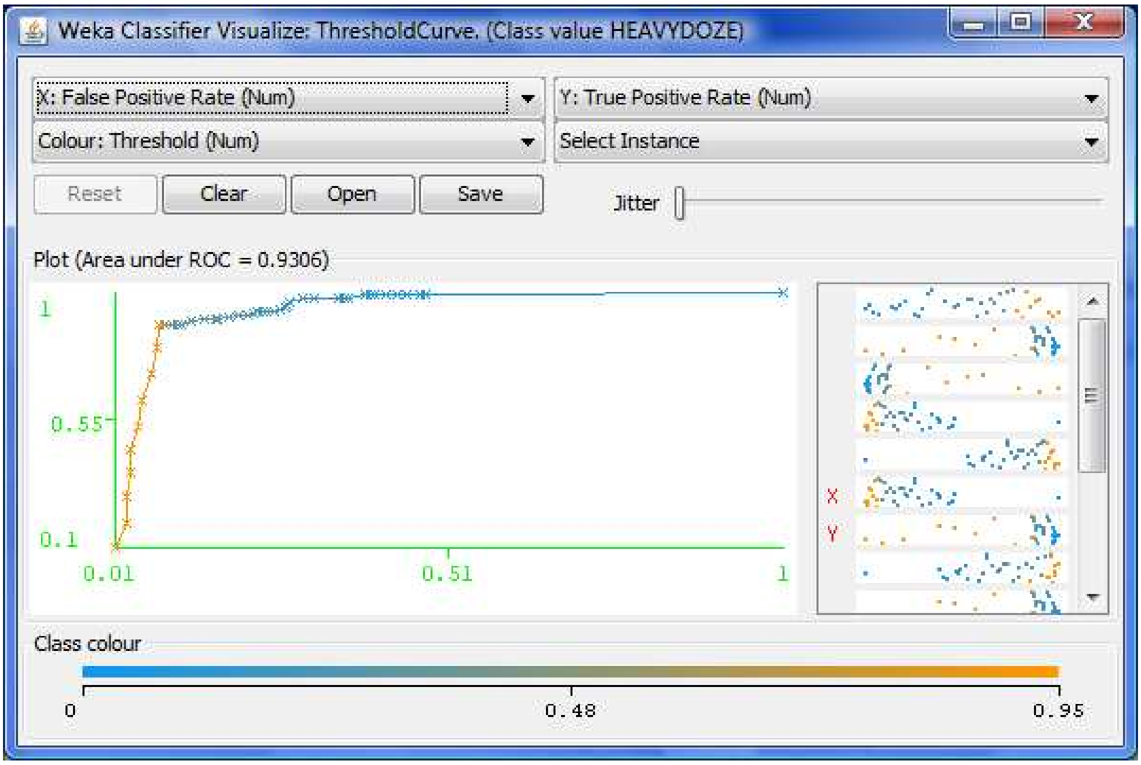
Figure 7.14: JRIP Accuracy Values



```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.882     0.076     0.905     0.882     0.893      0.931    HEAVYDOZE
                 0.825     0.076     0.879     0.825     0.851      0.91     PARTIAL
                 0.818     0.074     0.656     0.818     0.728      0.873    NONE
Weighted Avg.    0.85      0.076     0.858     0.85      0.852      0.914
```

In figure 7.14, there are ROC values.

Figure 8.15: JRIP ROC values



Weka Classifier Visualize: ThresholdCurve. (Class value HEAVYDOZE)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

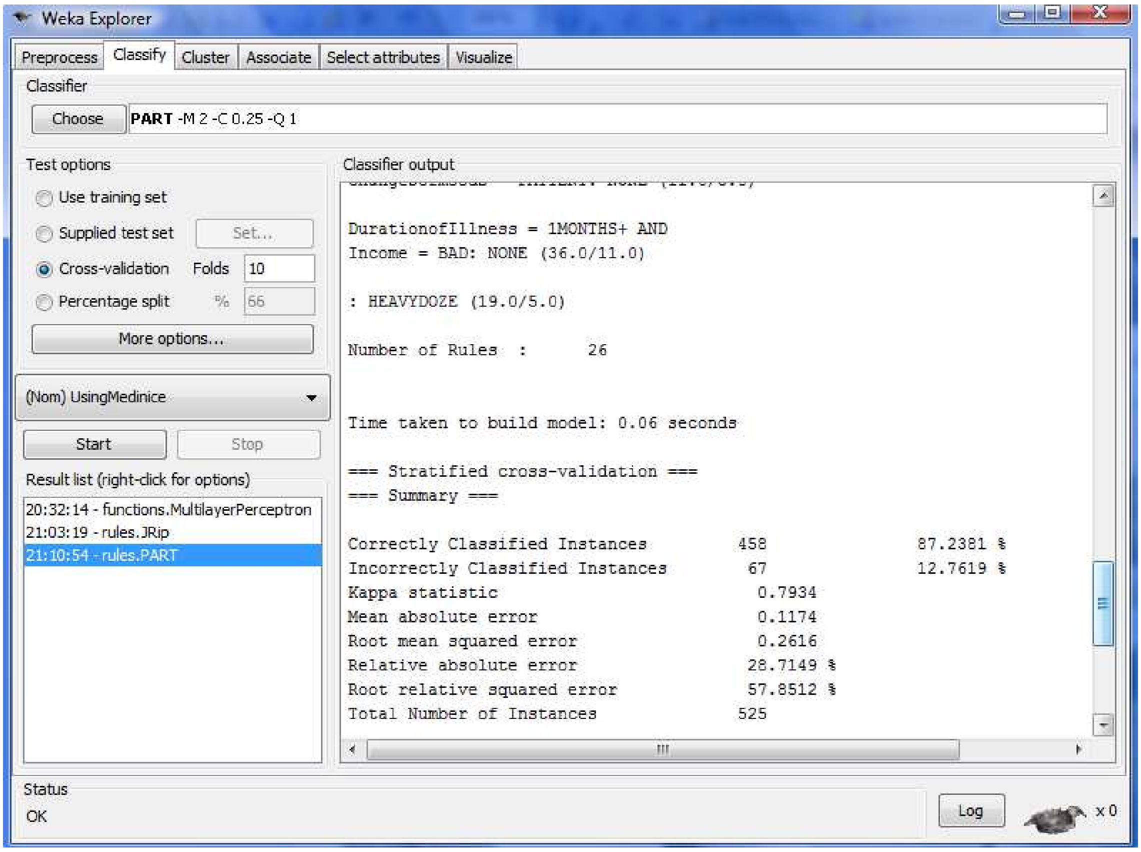
Plot (Area under ROC = 0.9306)

In weka classifier visualize, we chose the threshold Curve on the Heavy Doze part and we get the ROC value is 0.9306.

7.6 WEKA PART APPLICATION

In Part application, we use weka 3.7.1 classification methods and get these outputs. In the table, there are 525 information of psychology and from this information there is 67(12.7619 percentage) incorrectly classified instances and 458 (87.2381percentages) are correctly classified instances. Because RMSE value is nearly equal 0, the application shows that it is successful.

Figure7.16: Part Statistical Values



PART -M 2 -C 0.25 -Q 1

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) UsingMedinice

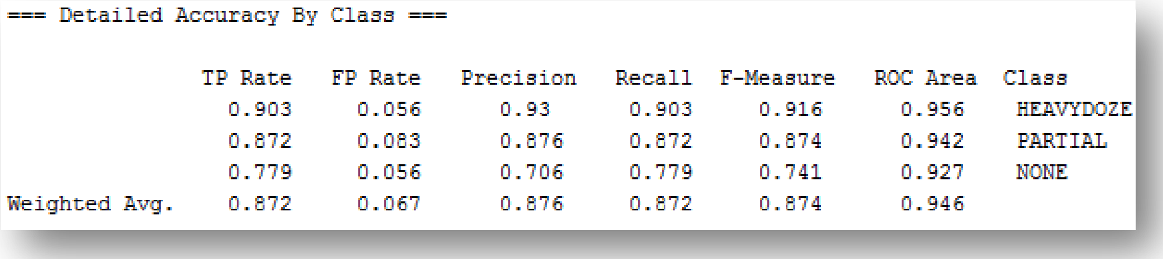
20:32:14 - functions.MultilayerPerceptron
21:03:19 - rules.JRip
21:10:54 - rules.PART

```
DurationofIllness = 1MONTHS+ AND
Income = BAD: NONE (36.0/11.0)

: HEAVYDOZE (19.0/5.0)

Number of Rules  :      26


Time taken to build model: 0.06 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         458               87.2381 %
Incorrectly Classified Instances        67               12.7619 %
Kappa statistic                          0.7934
Mean absolute error                      0.1174
Root mean squared error                  0.2616
Relative absolute error                 28.7149 %
Root relative squared error             57.8512 %
Total Number of Instances              525
```

Status
OK

In Figure 7.16, we see detailed information about PART. We use cross-validation test.

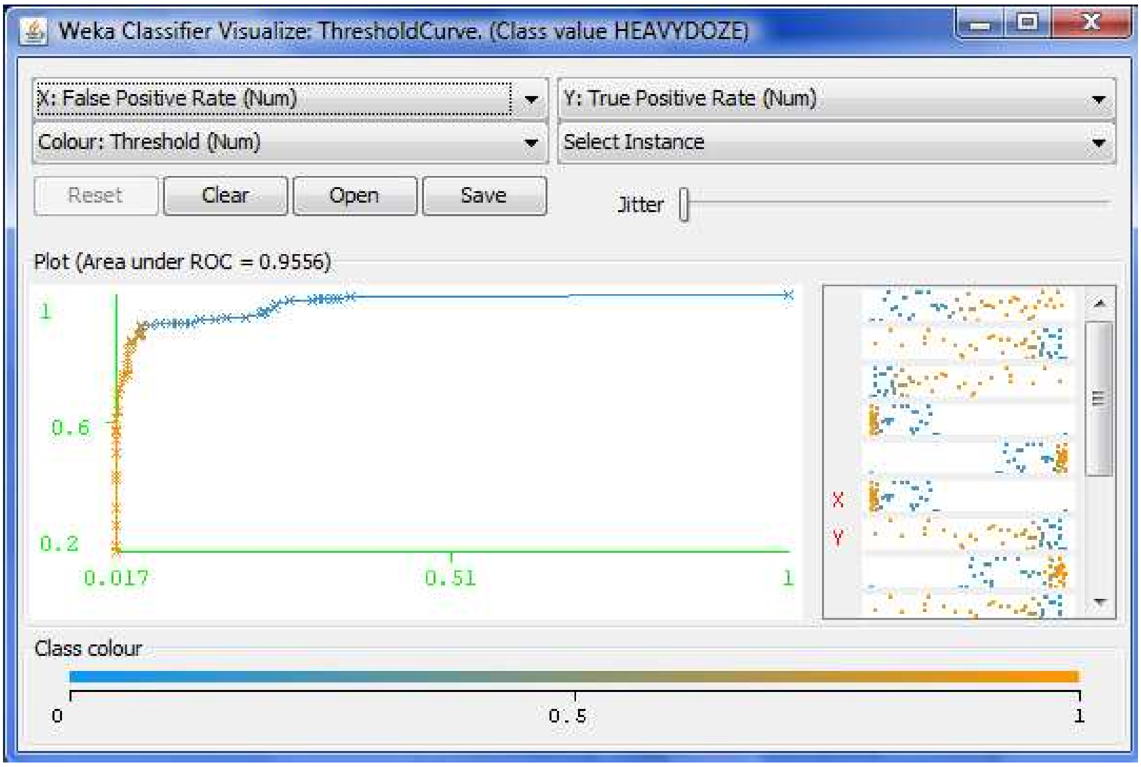
Figure 7.17: PART Accuracy Values



```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.903     0.056      0.93      0.903     0.916      0.956    HEAVYDOZE
                 0.872     0.083      0.876     0.872     0.874      0.942    PARTIAL
                 0.779     0.056      0.706     0.779     0.741      0.927    NONE
Weighted Avg.    0.872     0.067      0.876     0.872     0.874      0.946
```

In figure 7.17, there is ROC values.As a result of these situations, this is all correct.

Figure7.18: PART ROC Curve



Weka Classifier Visualize: ThresholdCurve. (Class value HEAVYDOZE)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

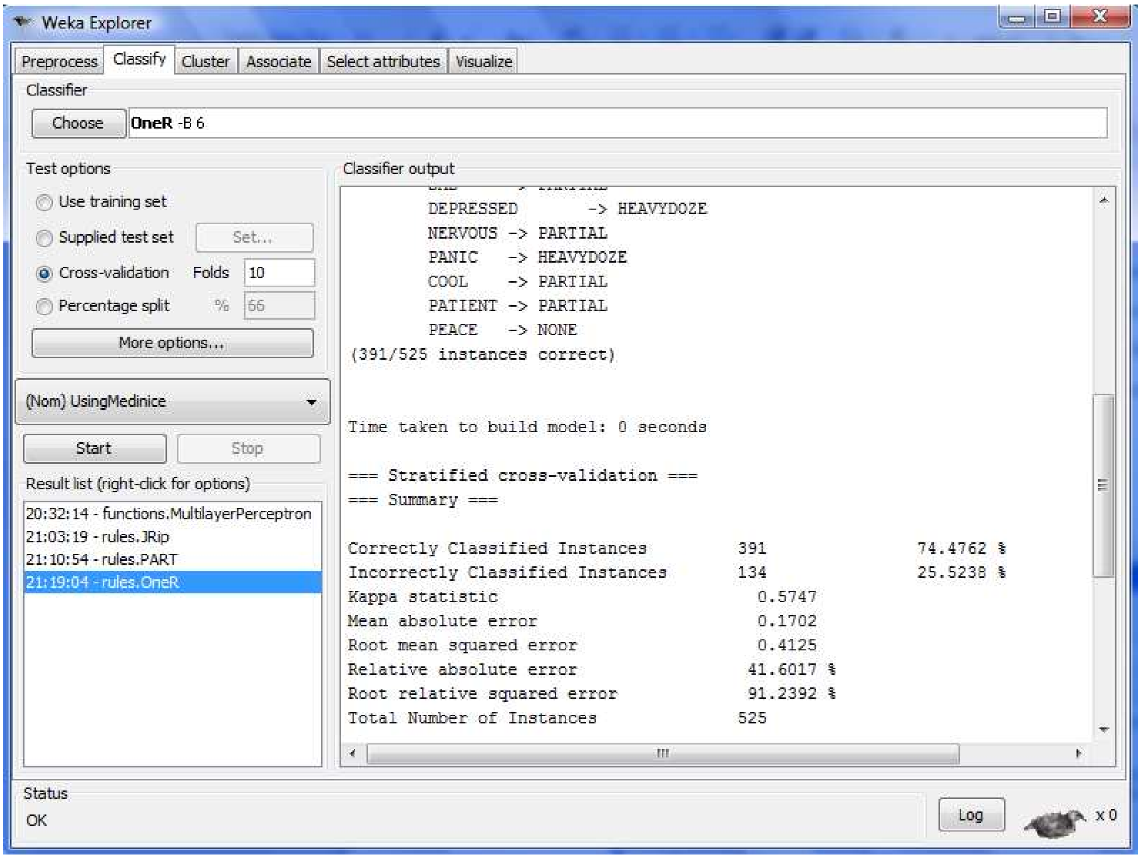
Plot (Area under ROC = 0.9556)

In weka classifier visualize, we chose the thresholdCurve on the heavy doze part and we get the ROC value is 0.9556.

7.7 WEKA ONER APPLICATION

In the OneR application, after we use Weka 3.7.1 classification methods, application gets these outputs. In the table, there are 525 information of psychology and from this information there are no incorrectly classified instances and the other 525 (100 percentages) are correctly classified instances. Because RMSE value is equals to 0, the application shows that it is successful.

Figure 7.19: OneR Statistic Values



Classifier: OneR -B 6

Test options:
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

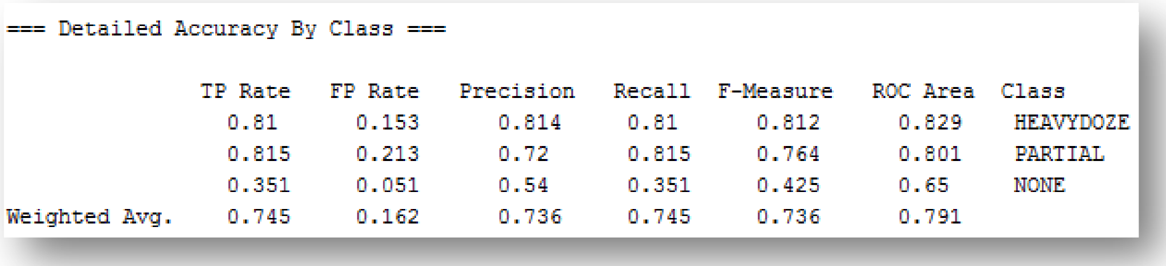
(Nom) UsingMedinice

Result list (right-click for options)
20:32:14 - functions.MultilayerPerceptron
21:03:19 - rules.JRip
21:10:54 - rules.PART
21:19:04 - rules.OneR

```
        DEPRESSED       -> HEAVYDOZE
        NERVOUS -> PARTIAL
        PANIC   -> HEAVYDOZE
        COOL    -> PARTIAL
        PATIENT -> PARTIAL
        PEACE   -> NONE
(391/525 instances correct)


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         391               74.4762 %
Incorrectly Classified Instances       134               25.5238 %
Kappa statistic                          0.5747
Mean absolute error                      0.1702
Root mean squared error                  0.4125
Relative absolute error                 41.6017 %
Root relative squared error             91.2392 %
Total Number of Instances              525
```

Status
OK

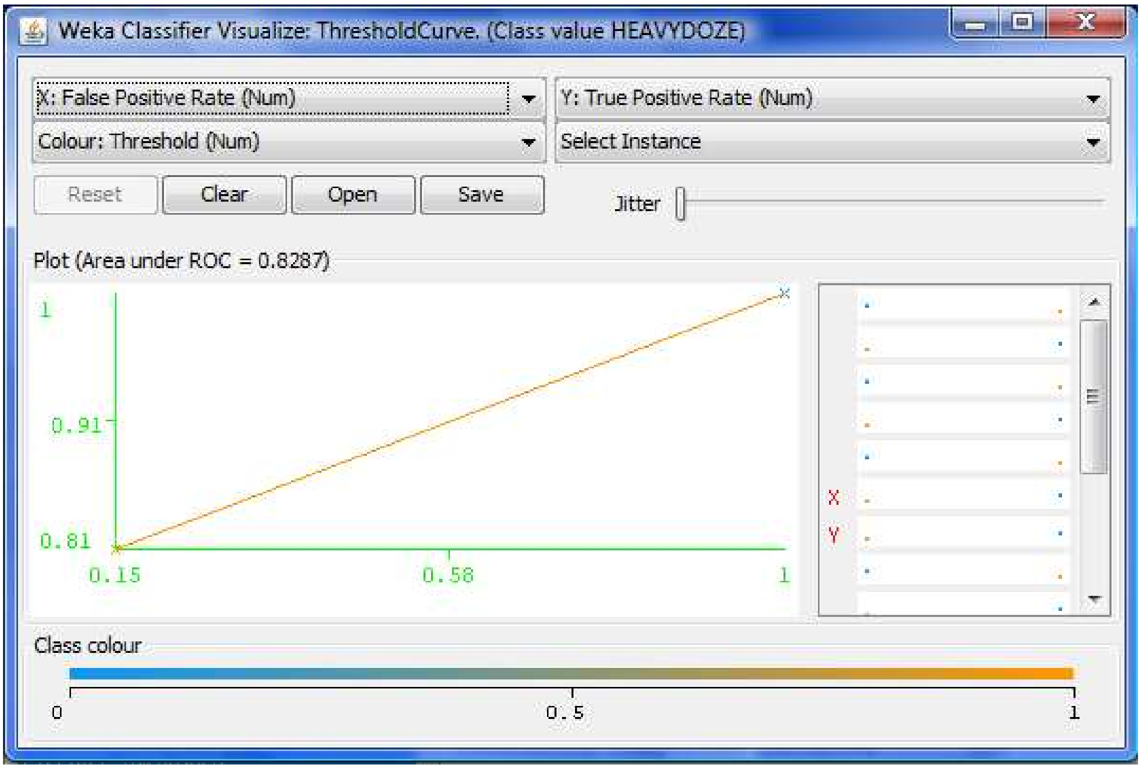
Figure7.20: OneR Accuracy Values



```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.81      0.153      0.814      0.81      0.812      0.829    HEAVYDOZE
                 0.815     0.213      0.72       0.815     0.764      0.801    PARTIAL
                 0.351     0.051      0.54       0.351     0.425      0.65     NONE
Weighted Avg.    0.745     0.162      0.736      0.745     0.736      0.791
```

As we seen in the figure 7.20, the ROC area is nearly 1. Moreover, because of this situation, it’s correct.

Figure 7.21: OneR ROC Curve



Weka Classifier Visualize: ThresholdCurve. (Class value HEAVYDOZE)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

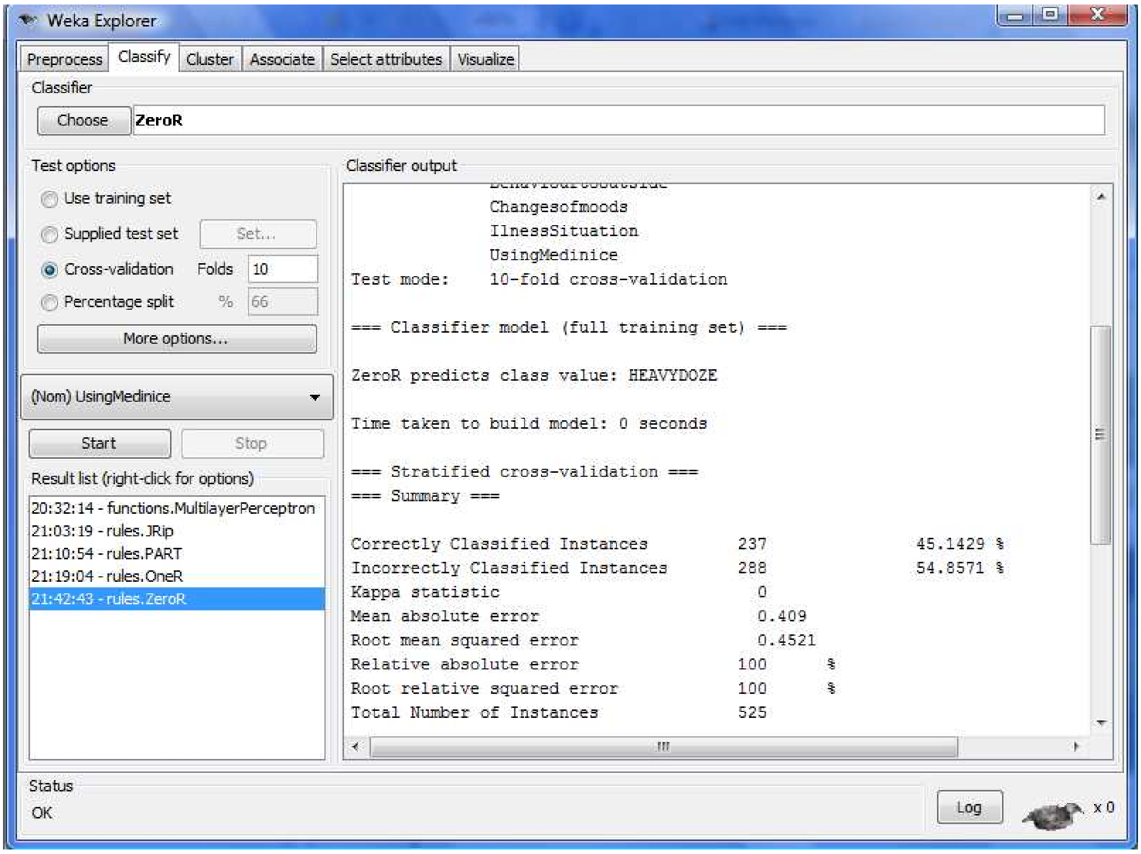
Select Instance

Plot (Area under ROC = 0.8287)

7.8 WEKA ZEROR APPLICATION

From ZeroR application in table 8.1, there are 525 information of psychology and from this information 288 (54.8571percentages) incorrectly classified instances and the other 237 (45.1429percentages) are correctly classified instances. Because kappa values are equals to 0, this application is not suitable for this application.

Figure 7.22: ZeroR Statistical Values



Classifier
Choose ZeroR

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) UsingMedinice

Result list (right-click for options)
20:32:14 - functions.MultilayerPerceptron
21:03:19 - rules.JRip
21:10:54 - rules.PART
21:19:04 - rules.OneR
21:42:43 - rules.ZeroR

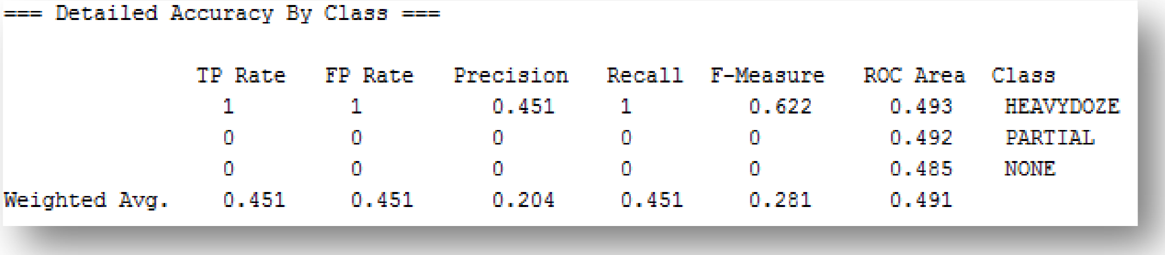
Classifier output

```
              Changesofmoods
              IlnessSituation
              UsingMedinice
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

ZeroR predicts class value: HEAVYDOZE

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         237               45.1429 %
Incorrectly Classified Instances       288               54.8571 %
Kappa statistic                          0
Mean absolute error                      0.409
Root mean squared error                  0.4521
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              525
```

Status
OK

In Figure 7.22, we see detailed information about ZeroR. We use cross-validation fold 10. This is the worst application to the other applications.

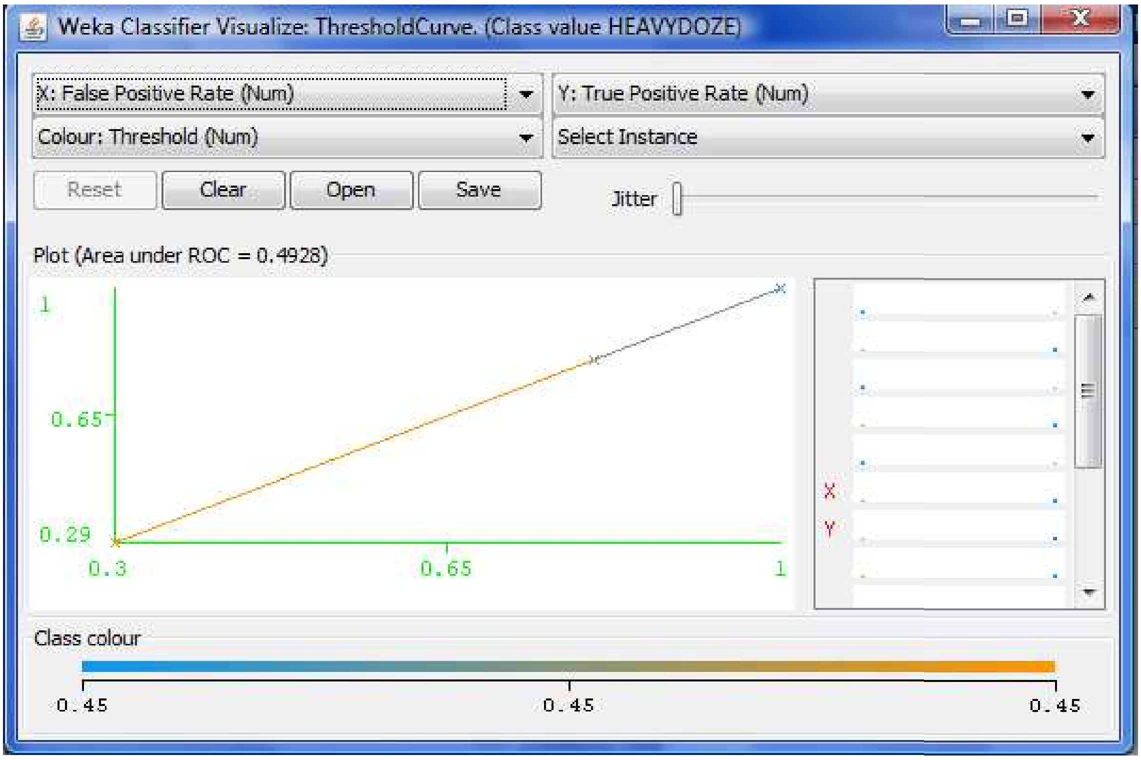
Figure 7.23: ZeroR Accuracy Values



```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         1         0.451     1         0.622      0.493    HEAVYDOZE
                 0         0         0         0         0          0.492    PARTIAL
                 0         0         0         0         0          0.485    NONE
Weighted Avg.    0.451     0.451     0.204     0.451     0.281      0.491
```

In figure 7.23, there are ROC values. As we see in table, place that under the curve is 0.5. ın the ZeroR application, FP (0, 1) and TP (0, 1) because of these values the ZeroR values is not suitable for the application.

Figure 7.24: ZeroR ROC Curve



Weka Classifier Visualize: ThresholdCurve. (Class value HEAVYDOZE)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

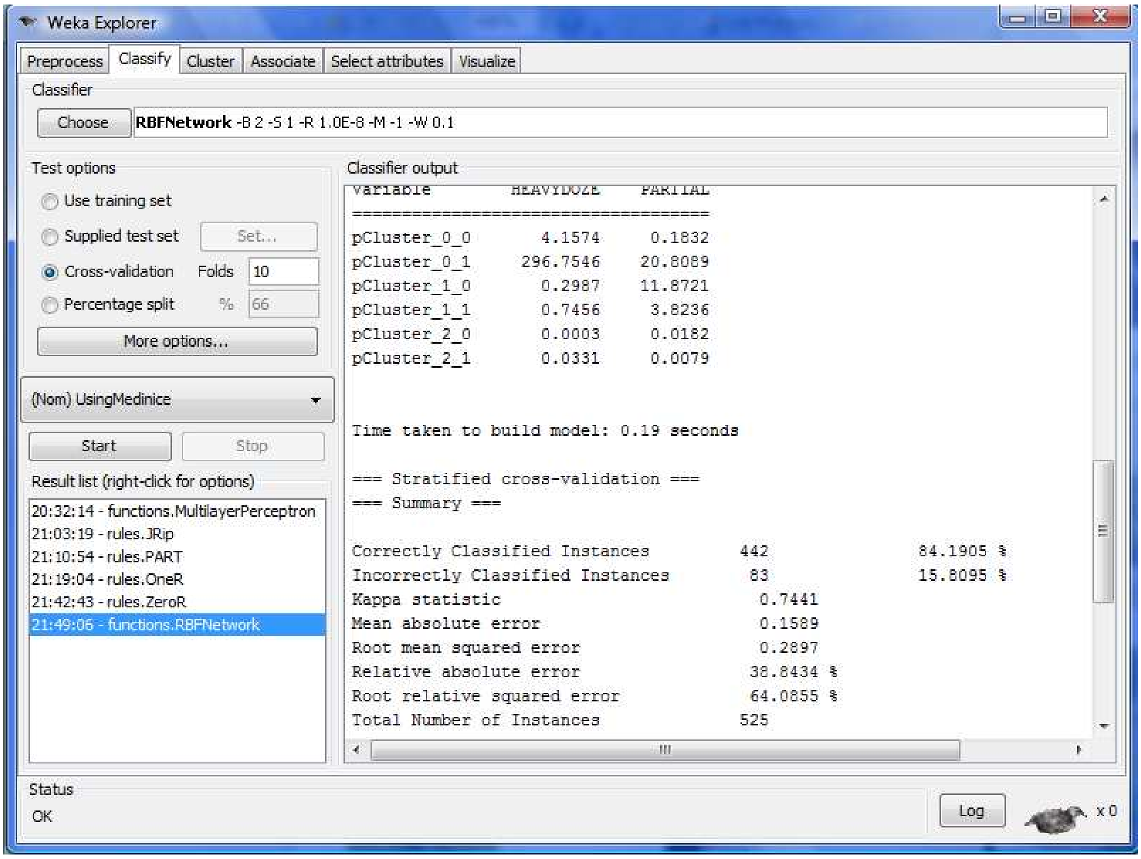
Select Instance

Plot (Area under ROC = 0.4928)

7.9 WEKA RBF NETWORKS APPLICATION

When we use Weka 3.7.1 classification methods, we get these outputs from Rbf Network application. In the table, there are 525 information of psychology and from this information just only 83 (15.8095 percentages) incorrectly classified instances and the other 442 (84.1905percentages) are correctly classified instances. Because RMSE value is nearly 0, the application shows that it is successful. There is a difference between Rbf application and the other application is that Rbfis change data values in table 7.2 discrete data values.

Figure 7.25: Rbf Networks Statistical Values



RBFNetwork -B 2 -S 1 -R 1.0E-8 -M -1 -W 0.1

○ Use training set
○ Supplied test set Set...
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) UsingMedinice

Result list (right-click for options)

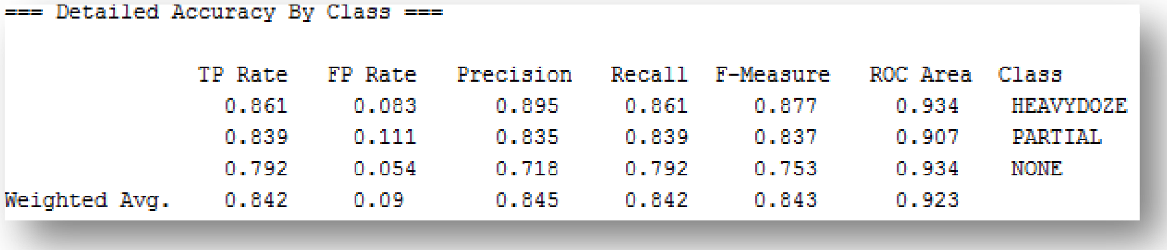
20:32:14 - functions.MultilayerPerceptron
21:03:19 - rules.JRip
21:10:54 - rules.PART
21:19:04 - rules.OneR
21:42:43 - rules.ZeroR
21:49:06 - functions.RBFNetwork

```
variable        HEAVYDOZE     PARTIAL
===================================
pCluster_0_0       4.1574      0.1832
pCluster_0_1     296.7546     20.8089
pCluster_1_0       0.2987     11.8721
pCluster_1_1       0.7456      3.8236
pCluster_2_0       0.0003      0.0182
pCluster_2_1       0.0331      0.0079


Time taken to build model: 0.19 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         442               84.1905 %
Incorrectly Classified Instances        83               15.8095 %
Kappa statistic                          0.7441
Mean absolute error                      0.1589
Root mean squared error                  0.2897
Relative absolute error                 38.8434 %
Root relative squared error             64.0855 %
Total Number of Instances              525
```

Status
OK

In Figure 7.25, we see detailed information about RBF Networks. We use cross validation fold 10, we get detailed information about RBF Network. However, we get more good information like correctly classified instances 442 and we just only get 84.1905percentages.

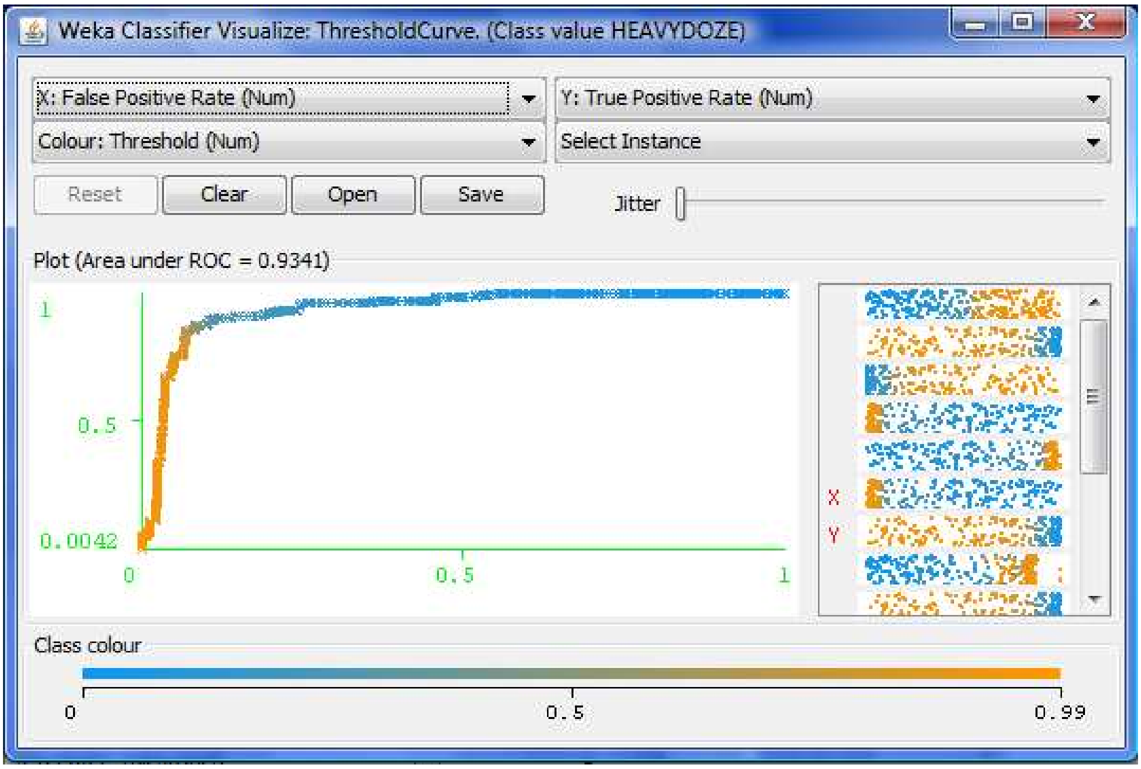
Figure 7.26: Rbf Networks Accuracy Values



```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.861     0.083      0.895     0.861     0.877      0.934    HEAVYDOZE
                 0.839     0.111      0.835     0.839     0.837      0.907    PARTIAL
                 0.792     0.054      0.718     0.792     0.753      0.934    NONE
Weighted Avg.    0.842     0.09       0.845     0.842     0.843      0.923
```

There is ROC values are in figure 7.26. System performance and outputs are successful.

Figure 7.27: Rbf Networks ROC Curve



Weka Classifier Visualize: ThresholdCurve. (Class value HEAVYDOZE)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

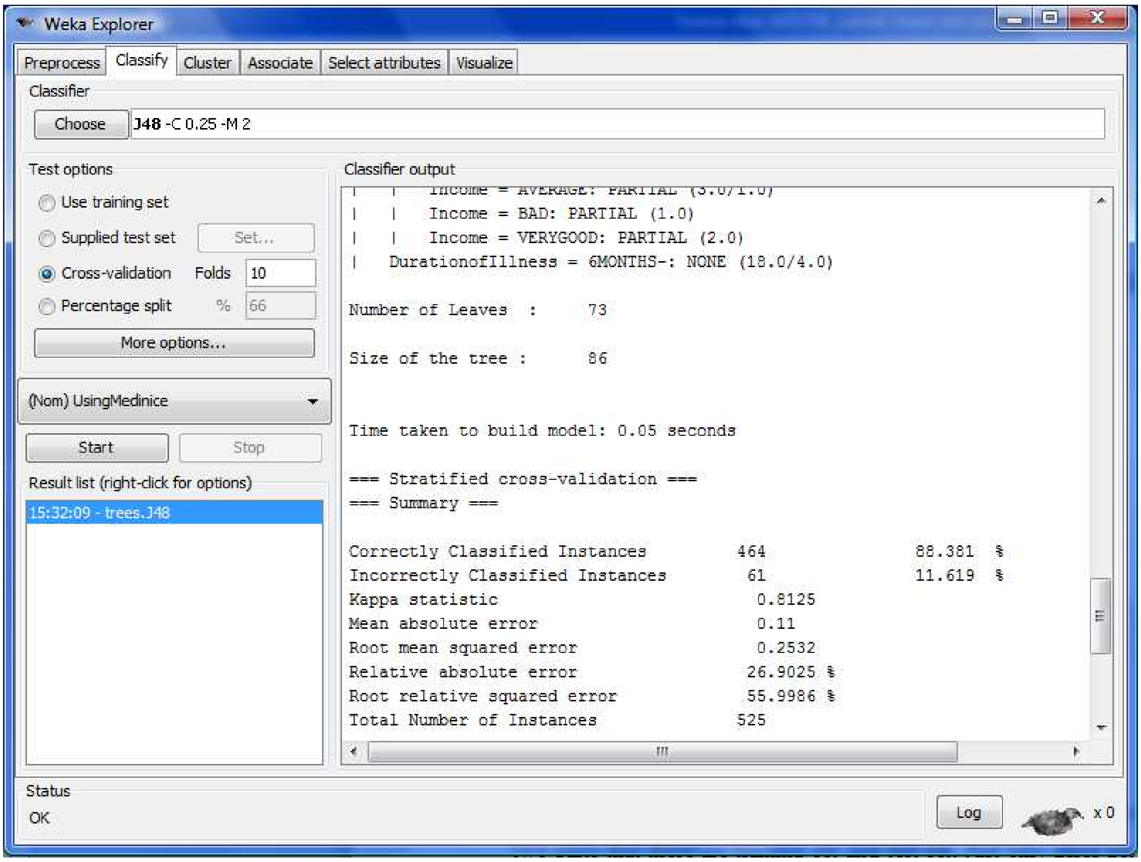
Select Instance

Plot (Area under ROC = 0.9341)

7.10 WEKA J48 APPLICATION

When we use Weka 3.7.1 classification methods, we get these outputs from J48 application. In the table, there are 525 information of psychology and from this information just only 61 (11.619 percentages) incorrectly classified instances and the other 464 (88.381percentages) are correctly classified instances. Because RMSE value is nearly 0, the application shows that it is successful.

Figure 7.28: J48 Statistical Values



J48 -C 0.25 -M 2

● Cross-validation Folds 10

(Nom) UsingMedinice

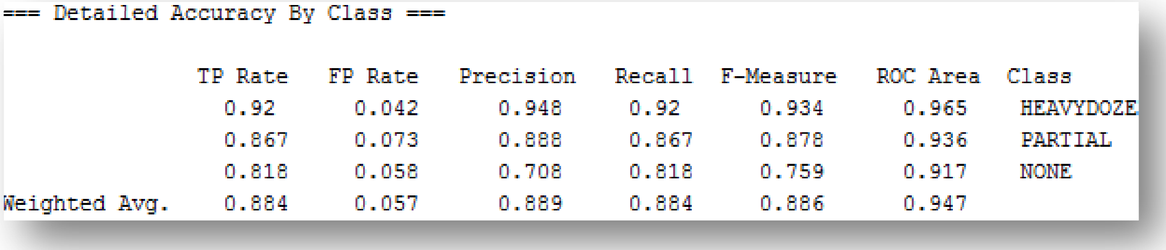
15:32:09 - trees.J48

```
|   |   Income = AVERAGE: PARTIAL (3.0/1.0)
|   |   Income = BAD: PARTIAL (1.0)
|   |   Income = VERYGOOD: PARTIAL (2.0)
|   DurationofIllness = 6MONTHS-: NONE (18.0/4.0)

Number of Leaves  :     73

Size of the tree :      86


Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         464               88.381  %
Incorrectly Classified Instances        61               11.619  %
Kappa statistic                          0.8125
Mean absolute error                      0.11
Root mean squared error                  0.2532
Relative absolute error                 26.9025 %
Root relative squared error             55.9986 %
Total Number of Instances              525
```

Status OK

In Figure 7.28, we see detailed information about J48. We use cross validation fold 10, we get detailed information about J48. Moreover; it gets time to build on data (0.05 second).

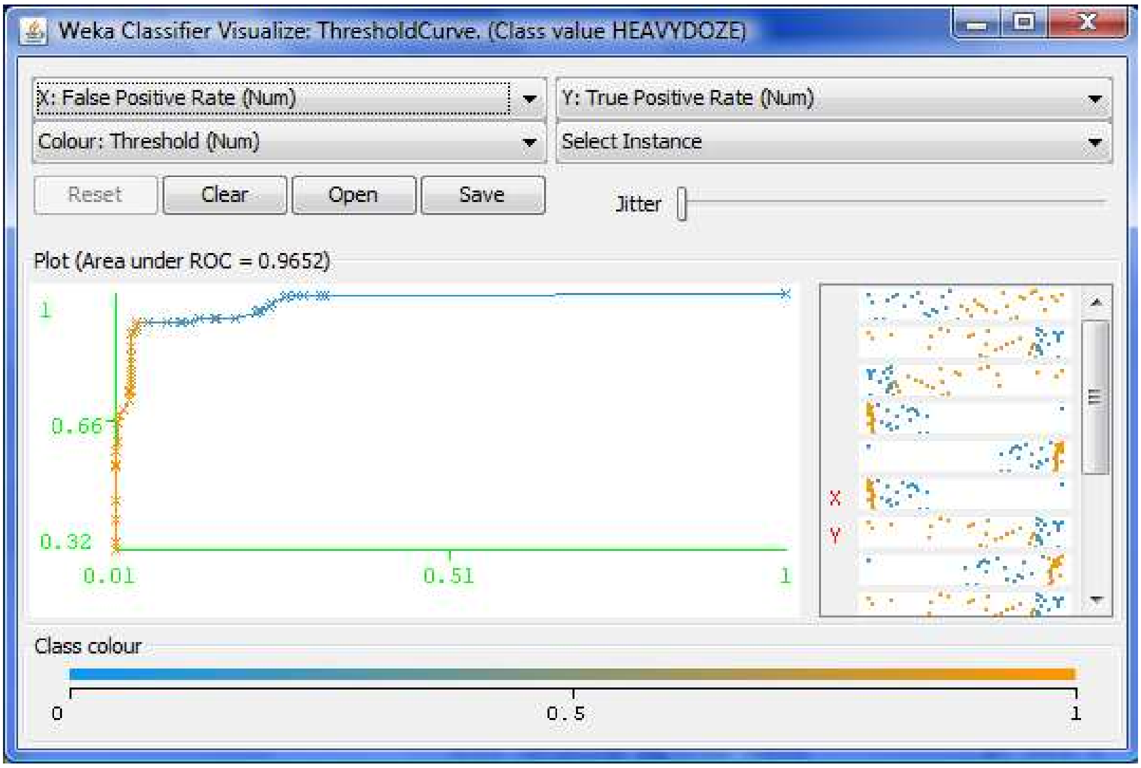
Figure 7.29: Rbf Networks Accuracy Values



```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.92      0.042      0.948     0.92      0.934      0.965    HEAVYDOZE
                 0.867     0.073      0.888     0.867     0.878      0.936    PARTIAL
                 0.818     0.058      0.708     0.818     0.759      0.917    NONE
Weighted Avg.    0.884     0.057      0.889     0.884     0.886      0.947
```

As we seen in the figure 7.29, the ROC area is nearly 1. Moreover, because of this situation, it’s correct.

Figure 7.30: J48 ROC Curve



Weka Classifier Visualize: ThresholdCurve. (Class value HEAVYDOZE)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.9652)

In weka classifier visualize, we chose the threshold Curve on the heavy doze part and we get the ROC value is 0.9652.

7.11 J48 RULES' DETAILS

If changeof mood is scared, we look into degreeofillness. If degreeofillness is fullremission or partialremission or notpsicoticbutheavy or undefined, then the using medicine is heavy doze. Moreover, if degreeofillness is average, then the using medicine is partial. If changeof mood is scared, and if degree of illness is psicoticbutheavy, we check behavior to outside. If behavior to outside is panic or sad or crazy or offensive or exciting or concern or angry or depressed or aggressive, then the using medicine is heavy doze, but if behavior to outside is recession, then the using medicine is partial.

If changeof mood is sad, we check illness situation. If illness situation is increasing, then the using medicine is heavy doze. Moreover, if illness situation is decreasing or same, then the using medicine is partial.

If changeof mood is depressed, we check duration of illness. If duration of illness is 1Year+ or 6Year+ or 6Months+ or 1Months- or 6Months-, then the using medicine is heavy doze. Moreover, if changeof moods is depressed and if duration of illness is 1Months+, then we control income. If income is good, then the using medicine is heavy doze. If income is average or bad or very good, then the using medicine is none.

If changeof mood is nervous, we also check duration of illness. If duration of illness is 1Year+, then the using medicine is heavy doze. If duration of illness is 6Year+ or 1Month+ or 1Month- or 6Months-, then the using medicine is partial. Furthermore, if changeof mood is nervous and if duration of illness is 6Months+, then we look into change of personality type. If change of personality type is component, then the using medicine is heavy doze. But, if change of personality type is labil or dezinhibe or other or paranoid or unspecified or aggressive or apatitic, then the using medicine is partial.

If changeof moods is panic, then the using medicine is heavy doze.

If changeof mood is peace, we also check duration of illness. If duration of illness is 1Year+ or 6Year+, then the using medicine is partial. However, if duration of illness is 1Month+ or 6Months+ or 6Months-, then the using medicine is none. Furthermore, if changeof mood is peace and if duration of illness is 1Month-, we look into income. If income is good, then the using medicine is none. However, if income is average or bad or very good, then the using medicine is partial.

If changeof mood is patient, we control degreeofillness. If degreeofillness is fullremission or partialremission or average or nonpsicoticbutheavy or weak, then the using medicine is partial. Moreover, if degreeofillness is psicoticbutheavy, then the using medicine is heavy doze, but if it is undefined, then the using medicine is none.

If changeof mood is cool, we control degreeofillness. If degreeofillness is fullremission, then the using medicine is none. Moreover, if degreeofillness is partial or nonpsikoticbutheavy or weak or undefined, then the using medicine is partial. Furthermore, if changeof mood is cool, and if degreeofillness is average, then we control change of personality type. If change of personality type is dezinhibe or other or paranoid or unspecified aggressive or component or apathetic, then the using medicine is partial. However, if change of personality type is labil, then the using medicine is heavy doze.

7.12 CLASSIFIER TRAINING DATA SET RESULTS

First table shows that the training sets of the applications. As discussed before there are two parts that these are training set and test set. Just only zeror get the same values as we got before. Total number of instance has changed if we use test or training set.

Table 7.1: Classifying Algorithm Results

	Bayes	**Naïve Bayes**	**Logistic**	**Multiplayer Perceptron**	**JRIP**	**PART**	**OneR**	**ZeroR**	**RBF**	**J48**
Correctly Classified Instance	76.62%	80.952%	92%	87.62%	84.952%	87.24%	74.48%	45.14%	85.20%	88.4%
Incorrectly classified Instance	20.38%	19.048%	8%	12.38%	15.048%	12.76%	25.52%	54.86%	15.81%	11.62%
Kappa Static	0.666	0.6878	0.8712	0.8007	0.759	0.793	0.574	0	0.7441	0.8125
MAE	0.1371	0.1279	0.0728	0.0999	0.142	0.117	0.170	0.409	0.1589	0.11
RMSE	0.3259	0.3103	0.1901	0.2463	0.285	0.262	0.412	0.45	0.2897	0.2532
Relative Absolute Error	33.51%	31.695%	17.7982%	24.43 %	34.79%	28.72%	41.60%	100%	38.84%	26.90%
RRSE	72.075%	68.639%	42.0505%	54.484%	63.025%	57.85%	91.24%	100%	64.09%	55.99%
Total# instance	525	525	525	525	525	525	525	525	525	525

In the table 7.2, we have detailed accuracy by class results for each application. According to ROC area result, as we seen that the best is logistic. Moreover, the worst method is zeror, again.

Table 7.2: Classifying Accuracy Results

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**
Bayes	0.81	0.114	0.807	0.81	0.808	0.934
Naïve Bayes	0.81	0.114	0.807	0.81	0.808	0.933
Logistic	0.92	0.038	0.927	0.92	0.921	0.989
Multiplayer Perceptron	0.876	0.061	0.882	0.876	0.878	0.961
JRIP	0.85	0.076	0.858	0.85	0.852	0.914
PART	0.872	0.67	0.876	0.872	0.874	0.946
OneR	0.745	0.162	0.736	0.745	0.736	0.791
ZeroR	0.451	0.451	0.204	0.451	0.281	0.481
RBF	0.842	0.39	0.845	0.842	0.843	0.923
J48	0.884	0.057	0.889	0.884	0.886	0.947

In table 7.3, we classified RMSE and ROC values for each application. As we knew that if RMSE value equals 0 or approximately 0, the result give use correct that the application is suitable. On the other hand, in the ROC value if ROC value equals 1 or approximately 1 than the application is correct and suitable. In the table, we see that for RMSE values that the best result is logistic. Furthermore, we see that for ROC values that the worst result is zeror.

Table 7.3: Classifying RMSE& ROC Results

	Bayes	**Naïve Bayes**	**Logistic**	**Multiplayer Perceptron**	**JRIP**	**PART**	**OneR**	**ZeroR**	**RBF**	**J48**
RMSE	0.3259	0.3103	0.1901	0.2463	0.2848	0.2616	0.4125	0.4521	0.2897	0.2532
ROC	0.934	0.933	0.989	0.961	0.914	0.946	0.791	0.481	0.923	0.947

7.13 SUMMARY OF CLASSIFICATION METHODS

According to table 7.4, the worst method that have highest error rate is ZeroR. If we want to see the best method, there are 5 of them. These are logistic, multilayer perceptron, jrip, part and oner. Bayes and naïve bayes also have good results but not like the others.

Table 7.4: Summary of Classification Methods

Model	Sensitivity (TPR)	Specificity (1-FPR)	RMSE
Bayes	0.82	0.8	0.3259
Naïve Bayes	0.80	0.8	0.3103
Logistic	0.92	0.91	0.1901
Multilayer Perceptron	0.88	0.85	0.2463
Jrip	0.86	0.84	0.2848
Part	0.88	0.86	0.2616
OneR	0.75	0.74	0.4125
ZeroR	0.46	0.44	0.4521
Rbf	0.843	0.841	0.2897
J48	0.884	0.056	0.2532

8. DISCUSSION AND CONCLUSION

In this work, data mining method is used for people with psychological disorders of behavior to examine the problems giving rise to this situation. Psychological problems can have many reasons. generally, patients with these types of problems; we look into age, sex, marital status, income, change of personality type, reason to start, degree of illness, duration of illness, repeat status, behavior to outside, changes of moods, illness situation, using medicine.

In my thesis, I also used J48 method to identify the result very good. As explained before, J48 is a clone of an earlier algorithm of C4.5. As a result of J48, I got that there are 525 information of psychology and from this information just only 61 (11.619 percentages) incorrectly classified instances and the other 464 (88.381percentages) are correctly classified instances.

Furthermore, we use almost every data mining' classify methods on data. On every model, we get approximately 100 percentages values.The worst value that we get is ZeroR application. We get 50 percentages.

These data are unique and reel. Moreover, there are 525 data. For the study, we work on 13 different data information to get our goal. We use GainRatioAttributeEval with Ranker because we want to see which are the most effect our data. We can say that illness situation; changes of moods and reasons to start are the best ranked attributes.

The data sets mining results are summarized in the following table. As we saw earlier the lowest result is ZeroR algorithm. ROC area results are controlled. RBF has the lowest ratio, then the other algorithms.

Best performance in representation of modeling, we can understood from RMSE values and the ROC graphics. As understanding the worst modeling value is ZeroR. As we see in table, place that under the curve is 0.5. In the ZeroR application, FP (0, 1) and TP (0, 1) because of these values the ZeroR values is not suitable for the application. There are 525 information of psychology and from this information 288 (54.8571percentages) incorrectly classified instances and the other 237 (45.1429percentages) are correctly classified instances. Because kappa values are equals to 0, this application is not suitable for this application.

REFERENCES

Books

Niu, L., Lu & Zhang G., 2009.*Cognition-driven decision support for business intelligence: Models, techniques, systems and application*. India: Springer-Verlag Berlin Heidelberg Publishing.

McClish, D.K, 1987. *Comparing the areas under more than two independent ROC curves*. Med Decis Making, pg 148-156.

Rathus, A. S., 2005. *Psychology:Concepts and connections*. 9thedn.New York: Wadsworth Publishing.

Dener, M., Dörterler M.& Orman A., 2009. *Açıkkaynakkodlu very madenciliği programları: Weka'da Örnek Uygulama,* Akademik Bilişim 2009,Harran Üniversitesi, Şanlıurfa.

Sabherwal, R. & Becerra-Fernandez, I., *Business intelligence*: *Practices, technologies, and management.* America: Wiley Publishing.

Tang, Z. &Maclennan, J.,2005. *Data mining with sql server 2005.*India: Wiley Publishing.

Mohammed, Z., 2003.*Introduction to data mining*, Springer-Verlag.

De Ville, B., 2006. *Decision trees for business intelligence and data mining.* NC: SAS Institute Inc.

Venugopal, K.R. &Patnaik, L.M., 2011.*Communications in computer and information science 157: Computer networks and intelligence computing.*India: Springer.

Witten IH, Frank E., 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Second edition. Morgan Kaufmann.

Han J, Kamber M., 2006. *Data Mining: Concepts and Techniques*. Second edition. Morgan Kaufmann.

Salkind, J. N., 2010. *Encylopedia of research design volume 3.*America : Sage Publications.

Gorunescu, F., 2011.*Intelligence systems reference library volume 12: Data mining concepts, models and technique.*Romania: Springer-Verlag Berlin Heidelberg Publishing.

Loshin, D., 2003. *Business intelligence: The savvy manager's guide getting onboard with emerging IT.* America: Morgan Kauffman.

Periodicals

Baykal, A., 2006. D.Ü. *Ziya Gökalp Eğitim Fakültesi Dergisi*, Sayı 7, sy 95-107.

Oğuzlar, A., 2003.*Erciyes Üniversitesi İktisadi ve İdari Bilimler Dergisi, sayı 21, sy.67-76.*

Other Sources

Zadok, E., 2001.*Data mining methods for detection of new malicious executables*, [online].www.fsl.cs.sunysb.edu/docs/binaryeval/node5.html [accessed 23 July 2012].

Abernethy,M.,2010.http://www.ibm.com/developerworks/opensource/library/os-weka2/index.html.

Dresner,H.,Linden,A.,Buytendijk, F.,Friedman, T.,Strange, K.,Knox, M.&Camm,M.,2002. *Strategic Analysis Report.*

Inmon, W.H, 1995. *"What is a Data Warehouse?" Prism Tech Topic, Vol. 1, No. 1, 1995.*

Russom, P., 2009.Next Generation Data Warehouse Platforms.*TDWI best practices report.* The Data Warehousing Institute:*TDWI best practices report.*

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., 1996.From data mining to knowledge discovery in databases.*AI Magazine*,[online] 27 September 1996,http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf[accessed 3 May 2012].

Sallam, R., Richardson, J., Hagerty, J., 2012. Magic quadrant for business intelligence platforms report.*Agartner research report.* Gartner Inc: *A gartner research report.*

Ming, C., Jiawei, H. & Philip, Y., 1996.*Data mining: An overview from a database perspective*.IEEE transactions on knowledge and data engineering.Vol8. No 6.[online] December 1996.

http://cs.nju.edu.cn/zhouzh/zhouzh.files/course/dm/reading/reading01/chen_tkde96.pdf [accessed 23 May 2012].