DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DATA MINING USING IN DIAGNOSIS OF THE CHRONIC RENAL DISEASE IN DIABETIC PATIENTS

by
Esin Deniz UÇAR

September, 2014
İZMİR

DATA MINING USING IN DIAGNOSIS OF THE CHRONIC RENAL DISEASE IN DIABETIC PATIENTS

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Master of Science
in Computer Engineering, Computer Engineering Program

by
Esin Deniz UÇAR

September, 2014
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"DATA MINING USING IN DIAGNOSIS OF THE CHRONIC RENAL DISEASE"** completed by **ESİN DENİZ UÇAR** under supervision of **PROF. DR. ALP KUT** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof.Dr. Alp KUT

Supervisor

Asst.Prof.Dr. Reyat Yilmaz

(Jury Member)

Asst. Prof. Dr. Derya BİRANT

(Jury Member)

Prof.Dr. Ayşe OKUR

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

I would like to thank to my thesis advisor Prof. Dr. Alp Kut for his help, suggestions and guidance.

I also thank to my husband Computer Sciences Engineer (M.Sc) Suphi Uçar for his supportive attitude and patience and also would like to thank to my director Ege University Information Technology Department Director Öğr.Gör. Ali Murat Ergin for his guidance and finally would like to thank to Doctor Soner Duman for his explanatory information descript disease.

Esin Deniz UÇAR

DATA MINING USING IN DIAGNOSIS OF THE CHRONIC RENAL DISEASE IN DIABETIC PATIENTS

ABSTRACT

Recently, data mining applications became more popular and widely used in many different sectors; also it became an important part of the health care industry. Especially, the reduction in the cost of computer hardware, computers became more accessible now, therefore computer-based solutions have been used in hospitals and then the hospitals started to collect too many data about their operations. So, hospitals have huge amount of data and making analysis on data and extract meaningful information from this data became impossible. Therefore, a software tool is required that analyzes the hospital data and provides useful data to health professionals.

Hospital Information Systems are widely used by hospitals; it offers data access without loss of time to healthcare professionals. However, it is not enough just access to data at the time. The purpose of the Hospital Information Systems, ensure that the data guidance for health professional and supports the decision making process of health professionals.

Aim of the study is to develop web-based application that supports the decision making process of health professional by determining the diabetic and hypertension patients whether under risk of renal failure or not via using data mining techniques.

This application was developed using the diabetic and hypertension patients' data of Ege University Medical School Hospital's Biochemistry Laboratory results. Naïve Bayes algorithm has used for data mining and 74322 pcs record were used to construct model and the model tested on 7296 pcs record 78% successfully.

Keywords: Data mining, hospital information system (HIS), using data mining techniques in healthcare industry.

DİYABETİK HASTALARDA KRONİK BÖBREK YETMEZLİĞİNİN VERİ MADENCİLİĞİ TEKNİKLERİ KULLANILARAK TEŞHİSİ

ÖZ

Son zamanlarda oldukça yaygınlaşan ve birçok farklı sektörde kullanılmaya başlanan Veri Madenciliği uygulamaları sağlık sektörünün de önemli bir parçası haline gelmiştir. Özellikle bilgisayar donanımlarının ucuzlaması, bilgisayarların artık daha erişebilir hale gelmesiyle birlikte bilgisayar tabanlı uygulamalar hastanelerde kullanılmaya başlanmıştır. Hastane verilerinin çok büyük miktarda olması, üzerinde analiz yapmayı ve anlamlı sonuçlar çıkarmayı imkânsız kılmaktadır. Bu sebeple, hastane verilerini analiz ederek, saklı kalan anlamlı bilgileri sağlık uzmanlarının erişimine sunacak bir yazılım aracı gerekli olmuştur.

Hastane bilgi sistemleri yaygın olarak hastaneler tarafından kullanılmakta ve sağlık uzmanlarına anında, zaman kaybı olmadan veriye erişim sunmaktadır. Ancak sadece veriye erişmek yeterli olmamaktadır. Hastane bilgi sistemlerinin amacı bu verilerin bir şekilde sağlık uzmanına yol gösterici ve sağlık uzmanın karar verme sürecini destekler nitelikte olmasını sağlamaktır.

Bu çalışmanın amacı, risk grubu içerisinde yer alan hipertansiyon ve diyabet hastalarının böbrek yetmezliği hastalığı riski taşıyıp taşımadığına karar veren sağlık uzmanının karar verme sürecini destekleyecek ve bunu yaparken de veri madenciliği tekniklerini kullanacak web tabanlı bir uygulama geliştirmektir.

Uygulama, Ege Üniversitesi Tıp Fakültesi Hastanesi Biyokimya Laboratuvarında yer alan diyabet ve hipertansiyon hastalarının verileri kullanılarak geliştirilmiştir. Veri madenciliği işlemleri için 74322 adet kayıt üzerinde Naïve Bayes algoritması kullanılmıştır ve 7296 adet kayıt üzerinde %78 başarı ile test edilmiştir.

Anahtar kelimeler: Veri madenciliği, hastane bilgi sistemi (HBS), veri madenciliği tekniklerinin sağlık sektöründe kullanımı.

CONTENTS

Page

THESIS EXAMINATION RESULT FORM ii
ACKNOWLEDGEMENTS iii
ABSTRACT iv
ÖZ v
CONTENTS vi
LIST OF FIGURES viii

CHAPTER ONE – INTRODUCTION 1

1.1 Literature Review 2
1.1.1 Oracle Data Mining 7

CHAPTER TWO – RELATED WORK 12

2.1 Data Mining Application Areas 12
2.2 Data Mining in Medicine 13

CHAPTER THREE – USED TECHNOLOGIES 15

3.1 N-Tier Programing Architecture 15
3.1.1 Evolution of Layered Architecture 17
3.1.2 N-Tier Architecture 19
3.2 Web Services 28
3.3 Visual Studio 2012 & Asp.Net 30
3.3.1 Visual Studio Features 31
3.3.2 .NET Framework 32
3.4 TOAD for Oracle 33

CHAPTER FOUR – DEVELOPMENT PROCESS 35

4.1 Preparing Data for Data Mining 35

4.2 Applying Data Mining Methods Using Oracle Data Miner 37

4.2.1 Process of Import External Data 40

4.2.2 The Process of Implementing Data Model for Data Mining 42

4.2.3 Apply Data Mining Model 47

4.3 User Interface Implementation 48

CHAPTER FIVE – CONCLUSION AND DISCUSSION 61

REFERENCES 62

LIST OF FIGURES

Page

Figure 1.1 The data mining process. 6
Figure 1.2 Evolution of database system technology..... 7
Figure 1.3 Oracle data mining algorithms for supervised functions. 11
Figure 1.4 Oracle data mining algorithms for unsupervised functions. 11
Figure 3.1 N-tier architecture of the system..... 17
Figure 3.2 Single - tier architecture..... 18
Figure 3.3 Two-tier architecture (Client/Server). 19
Figure 3.4 Three-tier architecture diagram. 20
Figure 3.5 Basic components of object-oriented programming. 21
Figure 3.6 Definition of RunSP. 22
Figure 3.7 Definition of ExecuteNonQuery method. 22
Figure 3.8 Definition of LoadSQL. 23
Figure 3.9 A constructor example for business layer class for miner application. ... 24
Figure 3.10 Business layer 'MinerBF' was added as a reference to application layer. 25
Figure 3.11 Creating instance of the business layer object. 25
Figure 3.12 Accessing the business layer's select method from application layer via business layer object. 26
Figure 3.13 Definition of Encrypt method in system framework. 27
Figure 3.14 Definition of FormatDLL method in system framework. 28
Figure 3.15 Web service process diagram. 29
Figure 3.16 Click advanced button at the bottom of the window and after that click add web reference button at the bottom of the window. 29
Figure 3.17 Enter URL of the web service then go after that name the web reference and click add reference. 30
Figure 3.18 .NET framework components. 32
Figure 3.19 Toad for oracle can be logged in with a username and password. 34
Figure 3.20 This figure represents the tables in EUTFDBA schema. 34
Figure 4.1 Definition of ProtokolDegistir function. 37

Figure 4.2 Data can be import to Oracle database. 38
Figure 4.3 Csv file imported to create table name of KAN_SONUC_KREATININ. 39
Figure 4.4 Table created successfully. 39
Figure 4.5 Formulas for females and males that determining to renal failure. 40
Figure 4.6 Contents of KAN_SONUC_KREATININ table. 41
Figure 4.7 Bobrek_Yetmezligi column added to table. 41
Figure 4.8 Bobrek_Yetmezligi column is updated depending on creatinine level in blood. 42
Figure 4.9 Naive Bayes algorithm was chosen to implement model. 43
Figure 4.10 Target table and target attributes are selected. 44
Figure 4.11 Activity completed. 45
Figure 4.12 Target class changed 0 to 1. 46
Figure 4.13 Created model applied on test data. 47
Figure 4.14 Apply output of the test data. 48
Figure 4.15 Home page of the project. 49
Figure 4.16 The selection process of the required owner. 50
Figure 4.17 Select the KAN_SONUC_KREATININ table under the EPDBA owner. 50
Figure 4.18 View of result grid according to selected value in third drop down list.51
Figure 4.19 View of grid that third drop down list remains in default mode. 51
Figure 4.20 Codes and the view of UpdateProgress. 52
Figure 4.21 Representation of the TEST_VERI_PROB_A table. 53
Figure 4.22 Codes of ASP.Net Charting component. 54
Figure 4.23 Graph represents the female and under risk patients. 55
Figure 4.24 Graph represents the male and under risk patients. 56
Figure 4.25Graph represents the under risk males and the diagnosed diabetes. 57
Figure 4.26 Graph represents the results for both female and male patients under risk of renal failure. 57
Figure 4.27 Representation of the statistical table generated after applied operation. 58
Figure 4.28 Representation of TESTVERI_EXPLAIN_A table. 60

CHAPTER ONE
INTRODUCTION

Rapidly developing technology leads increasing amounts of data on the world such as satellite data, medical data, exchange data, automation data etc. Every small organization start to collect data about their operations and more is coming every day from new data sources.

Increasing amount of data leads collection and storage problems of data. Despite these problems, a development in databases and file systems solution has been used. Hardware become cheaper than the larger amounts of data can be stored. Especially hospitals have begun to give more attention to storage of patients' data.

Healthcare industry today generates large amounts of complex data about: patients, hospital resources, disease diagnosis, electronic patient records, medical devices etc. And to use these large amounts of data in favor of patients that data need to be processed and analyzed for knowledge extraction that enables decision making.

Nowadays, although Renal Failure (RF) disorder is treatable with early diagnosis, it may result in completely lost kidney functionality. RF is the progressive loss of kidney function. Patients' renal is suffered in a way that cannot be cured. As a result due to the loss of kidney function the need of dialysis and/or transplantation may arise.

In our country which leads to RF among the top three causes of diabetes, high blood pressure and glomerulonephritis. Especially, diabetic patients are under the risk. Therefore with the scope of this thesis topic, diabetic patients' data will be analyzed and investigate that possibility of observation RF on diabetic patients. Although not mentioned in the title of the thesis, hypertension patients' data included as it includes risk groups. This correction has been proposed by the expert nephrologists.

Today, stored data has higher volume than traditional existing resources. These mountains of data include too much data but too little information. There is a need to extract useful information from the data and to interpret the data. So hospitals need to use some tools to accelerate the critic treatment process to support decision making and manage all these data. One of these tools is Data Mining.

The main purpose of this thesis, with the help of Data Mining algorithms extract hidden information that hospitals can use for decision making process to prediction of RF disease in diabetic patients. After that it provides accessing this valuable information via WEB platform.

In first chapter concept of data mining, Oracle Data Mining components, Data Mining techniques and performance analysis are described in detail.

In second chapter, widely used areas of Data mining (especially in Medicine), related works and Data Mining process will be discussed.

In the third chapter, technologies used in this application are explained in detail: N-Tier software architecture, Web Services, Visual Studio platform, TOAD and Oracle Data Miner.

In fourth chapter, the phase of data preparation for data mining, creating data model using applied activities, applying learned data mining techniques on data using Oracle data miner and screen design criteria's are discussed.

In the last part of thesis, conclusions and discussion topics about the thesis will be explained in detail.

1.1 Literature Review

In the healing process of health industry, hospital management systems become very important to finalizing the process. With software hospitals can investigate the

process and check it works well or not. Using data mining in diagnosis is based on understanding the basic symptoms of the disease, classification of these symptoms and composes a decision-maker system with tolerable mistake. Disease diagnose software has become more important due to low tolerance in these software systems.

Data mining brings a set of tools and techniques that can be applied to this processed data to discover hidden patterns that provide healthcare professionals an additional source of knowledge for making decisions. The decision rests with health care professionals. The widespread use of data mining in health care has become a need not only a luxury. Due to simplify things, to ensure the correct progress and to increase the confidence of patients against to hospitals, data mining software in hospitals could be find wide area in this market place.

Data mining is a routine of automatically searching and seeking huge amounts of data. The purpose of data mining is to discover patterns, trends and similarities. Due to the nature of data mining it uses complicated and sophisticated algorithms to divide problem into smaller data and interpret the situation and evaluate probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD).

One of the earliest and most cited definitions of the data mining process, which highlights some of its distinctive characteristics, is provided by Fayyad (1996), who define it as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." According to Weiss & Davison (2010), data mining described as "Data mining is a *process* that takes data as input and outputs knowledge."

Some of the other Data Mining definitions are as follows:

- Data mining which is define as extraction of hidden, valuable and useful knowledge from big amount of data and its use for providing strategic decision support has been giving a new perspective for usage of health data

and is becoming a method which is improving in all application domains besides as an reply for problematic domains related with huge data (Koyuncugil, & Özgülbaş, 2010, p. 21).

- Data mining has been called exploratory data analysis, among other things. Data mining refers to the analysis of the large quantities of data that are stored in computers. Masses of data generated from cash registers, from scanning, from topic specific databases throughout the company, are explored, analyzed, reduced, and reused (Olson, & Delen, 2008, p. 5).
- Data mining is a tool, not a magic wand. It won't sit in your database watching what happens and send you e-mail to get your attention when it sees an interesting pattern. It doesn't eliminate the need to know your business, to understand your data, or to understand analytical methods. Data mining assists business analysts with finding patterns and relationships in the data — it does not tell you the value of the patterns to the organization. Furthermore, the patterns uncovered by data mining must be verified in real world (Edelstein, 1999, p. 1).
- Data mining consists of the (semi-) automatic extraction of knowledge from data. This amount of stored in databases continues grow fast. With an extraction, this large amount of stored data contains valuable hidden knowledge, which could be used to improve the *decision-making* process of an organization such as hospitals. However, the number of human data analysts grows at a much smaller than the amount of stored data. Thus, there is a clear need for (semi-)automatic methods for extracting knowledge from data (Freitas, 2009, p.2).

Data mining has captivated a huge deal of attention in the information industry and in society as whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration.

This situation summarizes the emergence of need of a field called data mining and knowledge discovery. Data mining field uses sophisticated methods to extract valuable information hidden in datasets and high-level knowledge from real-world data sets. In other words data mining is exploring the patterns, relationships, changes, irregularities, rules and statistically significant structures.

Some of the data mining applications as follow (Albayrak, n.d.):

- Association - 30% of customers who buy diaper buy beer too. (Basket Analysis)
- Classification – Despite young ladies buy small cars, aged wealthy men buy luxury cars.
- Regression – Credit Scoring. (Application Scoring)
- Sequent Patterns – A customer who has paid late the first three of installments, 60% probably cannot pay back loans. (Behavioral scoring, churning)
- Similar Time Sequence – The price of shares of Company X move similar to the price of shares of Company Y.
- Anomaly – Do my customers have different behavior than usual? (Fraud Detection)
- Web Mining – Seeking for similar of a specific document on the internet.

This application can be consider in the both classification and anomaly detection. In the first stage of the development process the anomaly detection method is consider to apply by using decision trees. But in the following sections it was decided to use Naïve Bayes classification method.

Figure 1.1 illustrates the phases of data mining process below. Each part of the process represents iterative nature of a data mining project. Each phases trigger another phases.

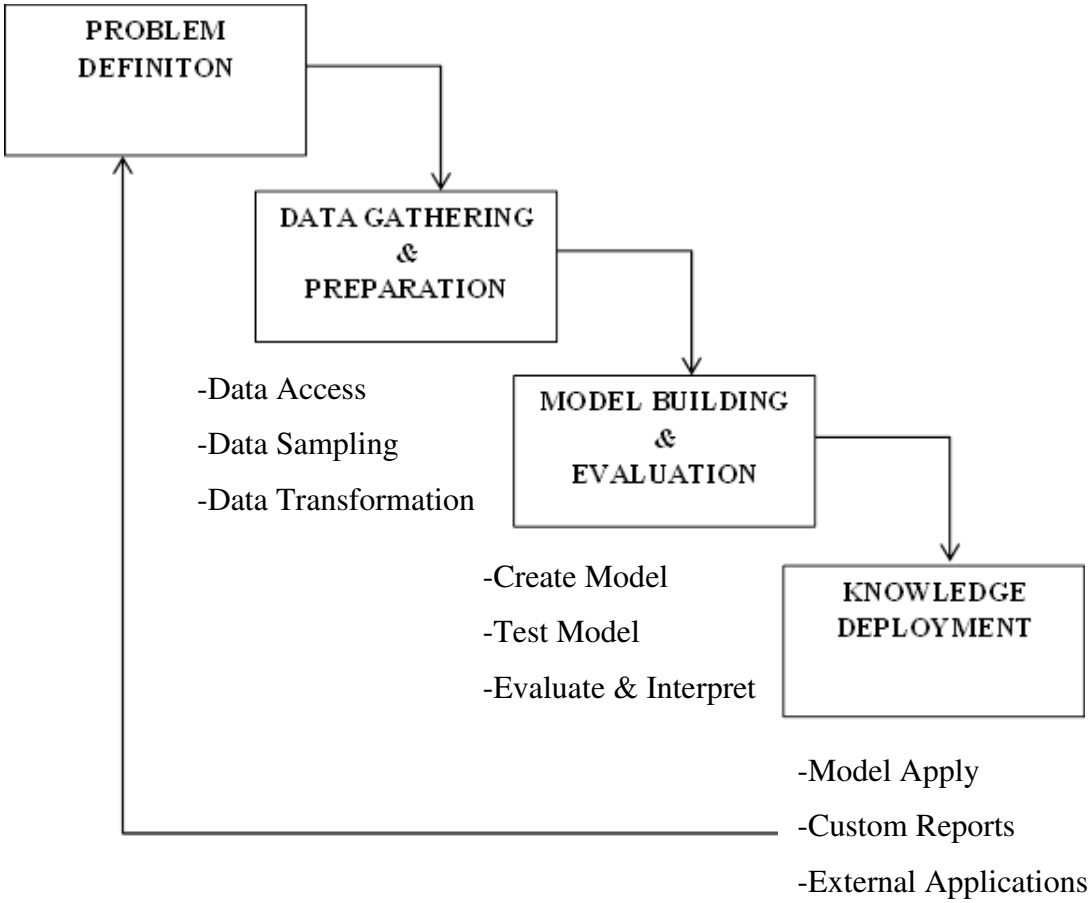


Figure 1.1 The data mining process.

Data mining can be think as a result of the natural evaluation of information technology (Han and Kamber, 2006). The database system industry has witnessed an evolutionary path in the development of the following functionalities.

As seen in Figure 1.1: *data collection and database creation, data management* (including data storage and retrieval, and database transaction processing), and *advanced data analysis* (involving data warehousing and data mining).

The figure seen on above, the first step is accessing data and prepare data for transactions, the second step is to create a model and learn model to apply it another test data, and the last step is to apply the created model in previous step and produce reports.

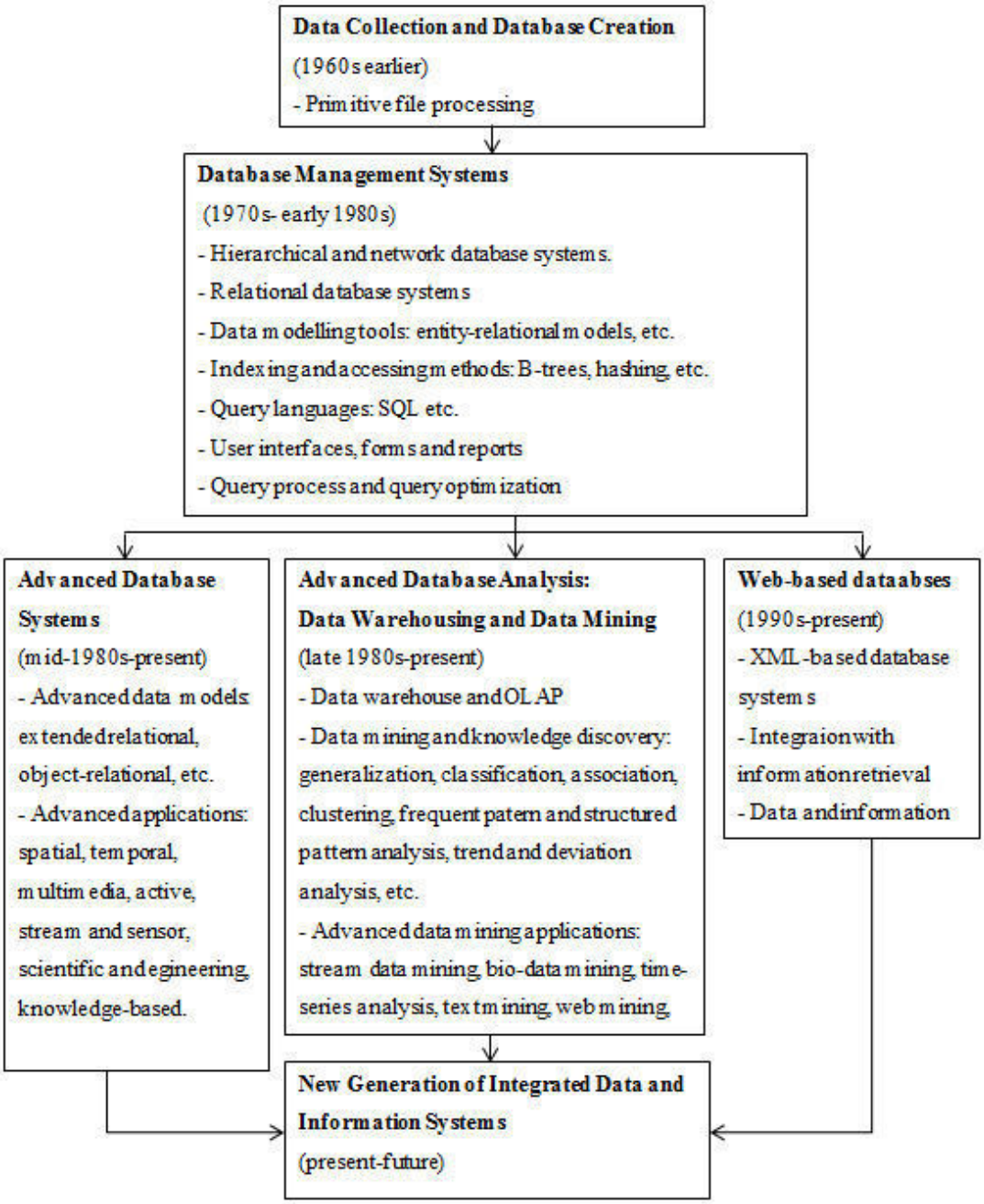


Figure 1.2 Evolution of database system technology (Han and Kamber, 2006).

1.1.1 Oracle Data Mining

A major research and development effort has been made by Oracle to add Oracle RDBMS native pattern recognition algorithms. As a result, Oracle Data Mining supports that development of (ODM) applications, integration and deployment as

part of the Oracle database environment. Data mining provides a comprehensive collection of analytics.

This data mining has infrastructure as native SQL, and PL/SQL and PL/SQL API, also it can be access from JAVA API or ODM. ODM intends to discover hidden insights. ODM is strengthened RDBMS environment with its data storage, data preparation and processing, information retrieval, security, process control, parallelism, versioning, work flow and reliability capabilities (Maimon & Rokack, 2005).

Oracle provides many advantages such as state-of-the-art data mining. Oracle Data Mining methods are implemented in the kernel of the database and it's models are first class data base objects. Oracle data mining can be used to efficiency usage of system resources.

Data mining within Oracle Database offers many advantages (Oracle, 2014):

- **No Data Movement:** Some data mining data need to transform into a specific format. With the help of Oracle data mining there is no data movement. This situation makes the process less complex, time-consuming, and error-prone.
- **Security:** In Oracle data mining data is protected by the oracle database security mechanism. Also, a specific database needs special data mining techniques. Only users that have the authority can mine.
- **Data Preparation and Administration:** Before data mined, the specific data must be filtered, sampled, normalized, cleansed and transformed into a specific format to provide a correct operation. With Oracle Data Mining data preparing process can be manage automatically. 80% of the effort is used usually to data preparation.
- **Ease of Data Refresh:** Oracle data mining has privilege to access the refreshed data. So, oracle data mining provide results according to maximizing its timelines and relevance.

- **Oracle Database Analytics:** Oracle offers many features to business intelligence. Oracle data mining can enlarge easily to other analytical features such as OLAP.
- **Oracle Technology Stack:** Oracle stack can be used in business intelligence, scientific researches and any industry.
- **Domain Environment:** Data mining models should be tested built and managed their suitable environment. Data mining operation results stored permanent so this storage provides pre and post data mining.
- **Application Programming Interfaces:** You can directly access Oracle Data Mining Functionality by using some SQL operations and Java APIs.

1.1.1.1 Data Mining Functions

The most important part of data mining operation is to decide which functionality is required. Each data mining function represents a specific domain of problem to solve it. Generally it is thought that these functions are divided into two separated area: **supervised** and **unsupervised**:

- Supervised functions are used to predict a value, require a target. It aims to provide predictive description by evaluating available data. For example Naïve Bayes algorithm is in this class.
- Unsupervised functions are used to find real relations. These functions do not need target, investigate the incidence of events. Clustering algorithms can be used to accomplish this.

Oracle Data Mining supports the following data mining functions (Oracle Data Mining Concepts 10*g* Release 2, 2005):

Supervised Data Mining:

- Classification – Dividing elements in separate classes, predict the class of and an element.

- Regression – Approximating values.
- Attribute Importance – Determine the most important features in predicting the outcome.
- Anomaly Detection – Identify items that do not satisfy the characteristics of "normal" data.

Unsupervised Data Mining:

- Clustering – Finding natural groups, similarities in data.
- Association Models – Finding relationships, connections between data.
- Feature Extraction – Crating new attributes as a combination of original features.

Data mining can also be classified as *predictive* or *descriptive*. Data mining can also be classified as *predictive* or *descriptive* (Oracle Data Mining Concepts 10*g* Release 2, 2005):

- Predictive data mining uses more than one model; these models are used to predict results of data sets. Predictive data mining functions are classification and regression. Naive Bayes is one algorithm used for predictive data mining.
- Descriptive data mining define a data set. Descriptive data mining functions are clustering, association models, and feature extraction. *K-Means* clustering is an algorithm used for descriptive data mining.

In the beginning of the development process, the anomaly detection method was considered to use but in the following sections it is decided to use Naïve Bayes instead of the decision trees. For this application Naïve Bayes algorithm which is the one of the algorithms of the predictive data mining techniques. This concept will be discussed in the next section. Here the figure in the show below represents the definition of the some data mining techniques.

Algorithm	Function	Description
Decision Tree (DT)	Classification	Decision trees extract human under-stand able rules hidden in the data and construct an if-then-else structure to predict the outcome.
Generalized Linear Models (GLM)	Classification and Regression	GLM applies logistic regression for dual objectives and continuous targets. GLM classification supports confidence limits for predicted probabilities.
Minimum Description Length (MDL)	Attribute Importance	MDL is a theoretic information model selection principle. MDL assumes that, data is the in the simplest and most compact form.
Naive Bayes (NB)	Classification	Naïve Bayes makes predictions using Bayes' theorem.
Support Vector Machine (SVM)	Classification and Regression	Different versions of SVM use different kernel functions to handle different data sets. SVM classification works to distinguish targets with the widest possible margin.

Figure 1.3 Oracle data mining algorithms for supervised functions.

Algorithm	Function	Description
Apriori (AP)	Association	Apriori performs market basket analysis by exploring co-occurring items in a set. Apriori finds rules with specified min support and greater confidence.
k- Means (KM)	Clustering	K-Means is a distance-based clustering algorithm that divides the data into a predetermined number of clusters. Each cluster has a centroid (center of gravity).
Non-Negative Matrix Factorization (NMF)	Feature Extraction	NMF constructs new features using linear combinations of the original natures. During applying of the model, NMF model matches into data array.
One Class Support Vector Machine (One–Class SVM)	Anomaly Detection	SVM creates a profile of class It aims to find the items do not compile the profile.
Orthogonal Partitioning Clustering (O-Cluster or OC)	Clustering	OC creates hierarchical, grid-based clustering model. Its algorithm creates dense clusters in space to identify areas.

Figure 1.4 Oracle data mining algorithms for unsupervised functions.

CHAPTER TWO
RELATED WORK

Nowadays, the hospitals stored huge amount of data about their patients and the patients' test results. This huge amount of data is raw data, not includes the meaningful data Therefore health professionals have some difficulties to interpret the results about the patients. The interpretation of data takes long time and sometimes human mistakes can occur. At this point to overcome all the shortcomings, there is an assistant tool is required. This assistant tool is kind of software and analyzes the data and summarize it into more understandable and human readable form.

The most popular technique to interpret data in health industry is using data mining techniques. With software uses the data mining techniques hidden information can be extracting from huge data. Not only in medicine, is data mining popular technique to simplify the data in much application area.

2.1 Data Mining Application Areas

Financial Industry: Today, finance and insurance are the fields that most in need to data mining because of their products and services. In these sectors knowledge-based management is a necessary because of the consequences of the economic crisis. The basic application of the financial industry can be summarized as risk grading, retaining existing customers, gain new customers, reduce costs, and prevent losses and leaks, creating alternative channels. A company, who can estimate customer profile, will have a big advantage of providing right offer. While company profit increasing, customer satisfaction will be increase too.

Communication Sector: The most important problem is the loss of customers in telecom sector. If they can pre-determine which customers could be losing, they can develop as cost-effective strategies for retaining customers. Sending a message with intends of ensure the retention of customer will increase the costs and make difficult

user to understand the message. For example, America's biggest wireless provider Verizon made a work purpose of not to lose their customers.

Healthcare Industry: It is important and unquestionably the impact of accurate and timely decision making on patient. The stored data in hospital database are used to identification of factors to getting successful results, testing high-risk factors of surgery, classification of patient data by age, gender, testing backwards terms of patient's health factors, developing treatment methods. Worldwide there are many successful applications. For example in San Francisco Hearth Institute started a worked that improving patient outcomes, reduction of length of stay in hospitals. And they transform medical data into information.

State Applications: Public administrators have understood the importance of data and information. Providing services tailored to the needs of the citizens is understood. The most important applications for public administrators as follows; ensure the correct use of resources, predict security problems the aim of ensure the public safety, identify the tax-related corruptions simultaneously, medical payments, implementation of programs, detection of suspicious matters and blocking the millions of dollars damage.

2.2 Data Mining in Medicine

At the present time Hospital Information Systems collect many data about patients and their medical conditions. Despite the abundance and richness of this data valuable information is hidden in hospital and clinic databases. Hospitals and medical centers need to be access useful information to improve own efficiency and future planning (Yıldırım, Uludağ, & Görür, 2008). With the help of data mining techniques hidden information discovered. So this process provides the most valuable information for researchers.

Due to it has a predictive and descriptive power data mining is widely used in healthcare industry. Calculation of healthcare cost, insurance fraud, disease

prognosis and length of stay (LOS) in a hospital can estimate easily. With the help of data mining get frequent patterns from medical databases, such as relationships between health conditions and a disease, relationships among diseases, and relationship among drugs.

Data mining brings a set of tools and techniques and use to process and interpret the data to provide healthcare professionals an additional source of knowledge for making decisions.

Today, the most important part of market area of informatics consists of health and medical fields. In last thirty years, major developments are occurred health information systems in the world. Despite health is a new field for informatics, rapid innovations has made in diagnosis tools (Kudyba, 2004).

The first usage of data mining techniques in medical information system has been in the 1970s. In the following years, especially 1990 neural networks have started to use to investigate issues such as future health status of patients in years and to cost estimates.

Consequently, diseases should be managed by considering medical terms, hospital resources and hospital costs. At this point classic queries are not sufficient to provide this management. Using data mining methods the hidden valuable information should be discovered. Discovered data must evaluate for medical research and management plan (Yıldırım, Uludağ, & Görür, 2008).

Now, to process the medical data using data mining techniques does not a luxury it becomes a need. Using data mining discovered meaningful data eases the managing medical data.

CHAPTER THREE
USED TECHNOLOGIES

In this chapter, the technologies that are used during application development process are described. Besides, this chapter associated with the used programming techniques, algorithms and methods such as n-Tier Programming Architecture, Web Services, Microsoft Visual Studio 2012 Platform, Asp.Net 4.0, and TOAD for Oracle and Oracle Data Miner.

3.1 N-Tier Programing Architecture

To perform certain tasks application Framework which forming the basis of software systems has to be developed to use with the pre-prepared libraries. Using Framework speeds up the development process and creates a standard and support to write clean understandable codes (Özbek, İnce, Turhan, & Önder, n.d.). The benefits of using Framework during development process are shown below:

- System can be test by many testers and eliminate security and performance problems.
- Speeds up and simplify the development process of the application.
- Creates a standard by providing clean and tiny code production.
- Simplifies the task-sharing between team members.
- Work easily with the database system through the database adapter that they contain.

There are important considerations while designing software. Some of these considerations that I referred from Ralph & Wand (2009) are given below:

- Compatibility – interoperability with on another product, hardware-independent.
- Extensibility – New features and abilities can be added on software

- Fault Tolerance – System continues operation in case of fault, is resistant to recover error
- Maintainability – Easily problem solving and bug fixing.
- Modularity – Divide software into well-defined independent components.
- Reliability – Accuracy and completeness of the required functions.
- Reusability – System components suitable for reuse without modification.
- Robustness – Resistance to extreme conditions.
- Security – Resistance against attacks.
- Usability – User-friendly designs for target users.
- Performance – Complete tasks in user-acceptable time, reduce CPU consuming.
- Portability – Usability of the same system on different environments.
- Scalability – Ability of adapting user number increasing.

To ensure these design specifications tiered architecture programming is necessary. Using Framework developers avoid repeated codes. Tier (Layered) Architecture is used to make the modifications easily and quick. This application was developed using C# language on .NET platform based on n-tier software architecture.

Flexibility and quality providing design pattern and technologies are used in the framework. The most important factor made system flexible and extensible is the application was developed by the n-tier software architecture (Figure 3.1).

The biggest benefit of the n-tier architecture is that it divides the program into smaller parts and provides easier control. Also, it maintains the security of data in highest level.

Using n-tier architecture reusable code capacity increases, programmers write a function once for example an encryption function and use it many times in the different parts of the project.

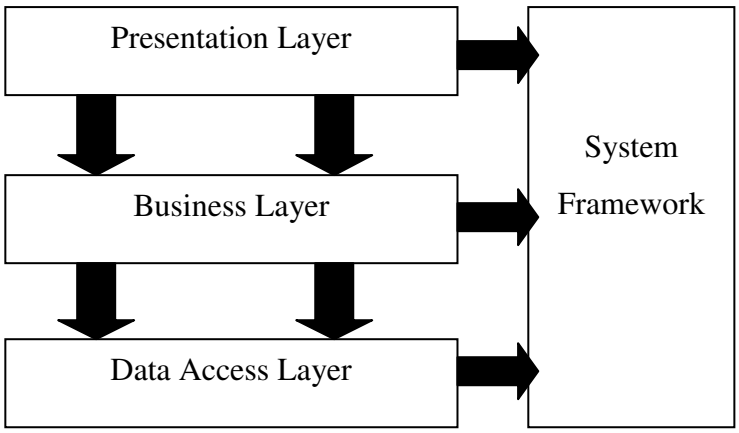


Figure 3.1 N-tier architecture of the system.

N-tier architecture in an important programming technique that gained importance with object-oriented programming and used by many programmer. It is easy to managing and controlling the program by using n-tier architecture. N-tier architecture consists of 4 separate layers: Data Access Layer, Business Layer and Presentation Layer. Also a vertical layer called System Framework.

3.1.1 Evolution of Layered Architecture

The first written programs were written as a single-layer and called 'Monolithic'. In the single-tier applications, presentation layer, business layer and data layer were located in the same computer (Figure 3.2). For example, systems such as Adabas.

Single-layer programs includes the design codes and data base connection codes in the same section. The manageability of the program is too hard.

Adabas systems are composed of a database and form that directly querying on the database (Thomas, 2005). The weaknesses of monolithic system are difficulty of repairing failed components and any change can affect all system.

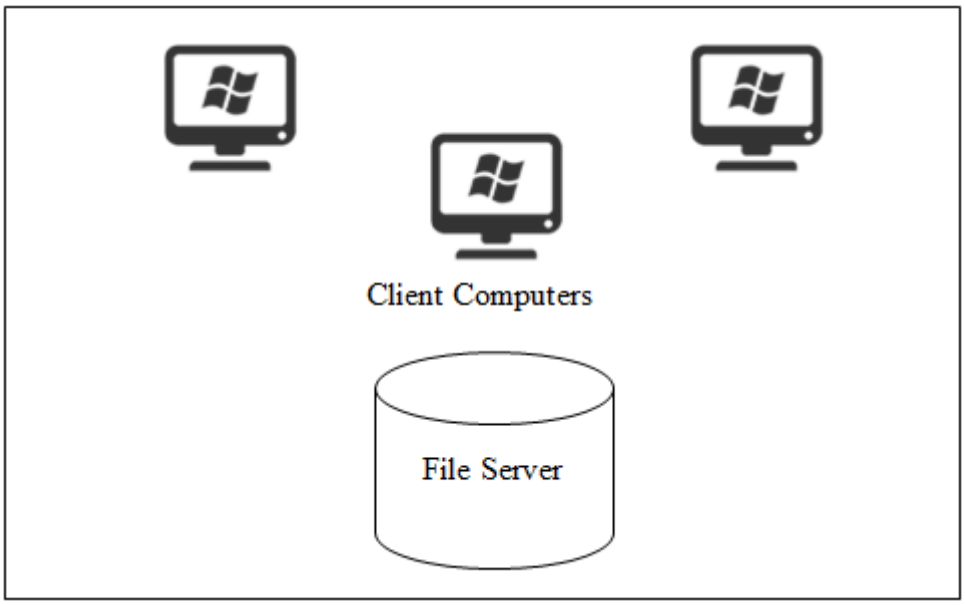


Figure 3.2 Single - tier architecture.

The two-tier architecture is like client server application. There are only two parts communicated. There is a direct communication between server and client. There are two different types of two-tier architecture:

Fat-Server, if we do transactions which belong to business layer in data layer such as controlling suitability of the procedures in data layer.

In Fat-Server technology, only related presentation graphics and their control software are included. Fat-Server is work-related software servers. It is important to send less data from the communication channel. There is no problem such as installing the software on computer. It can be used on different computers. For example: WEB client-server and a client built with J2EE/J2ME architecture.

The second type is Fat-Client, if we do transactions which belong to business layer in presentation layer such as controlling suitability of the procedures in presentation layer (Figure 3.3). In Fat-Client architecture, software needs to install on all clients from the server or any other media for each time. For example: Microsoft Outlook.

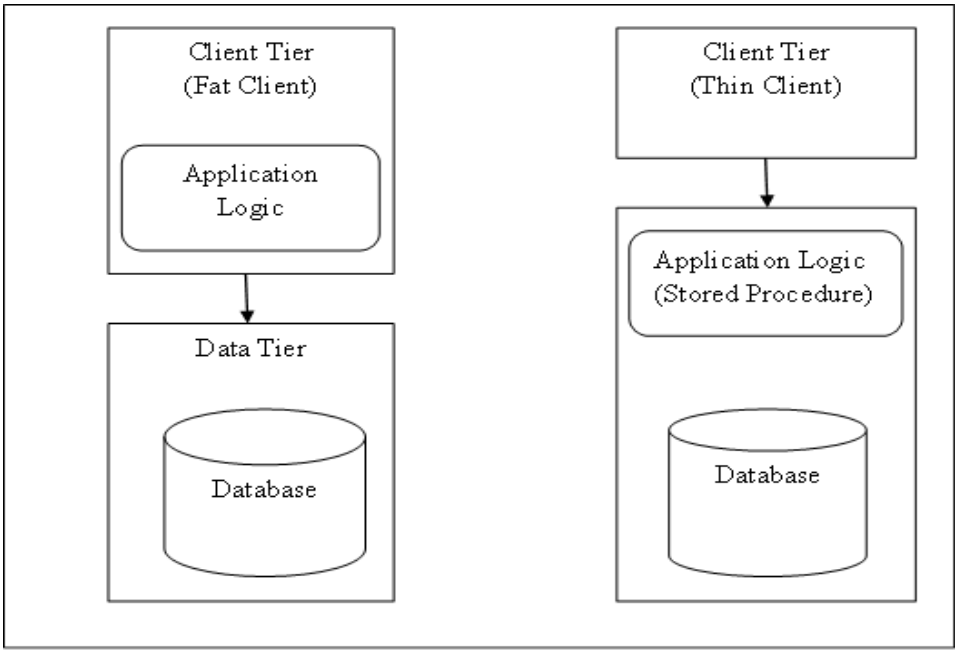


Figure 3.3 Two-tier architecture (Client/Server).

3.1.2 N-Tier Architecture

N-tier architecture is a software design pattern that data storage, data access operation, functional progress logic and user interface included in and maintained as independent modules on separate platforms. N-tier application allows update these platforms independently.

N-tier architecture is a design trend and forming the basic dynamics of database programming that becomes more popular with the object-oriented programming. The basis of object-oriented programming consists of classes. A class is an identification of an object in real time by coding. Figure 3.4 shows the basic components of the n-tier architecture using class identification.

N-tier architecture offers a structure that separates database connection, database transactions (insert, update, delete, etc.) and user interfaces in our projects. One of the most important reasons of using tiered architecture is that eliminating the number of repetitive rows. Multi-tier architecture approach provides to develop simple,

manageable applications by separating responsibilities, functionalities according to their tasks.

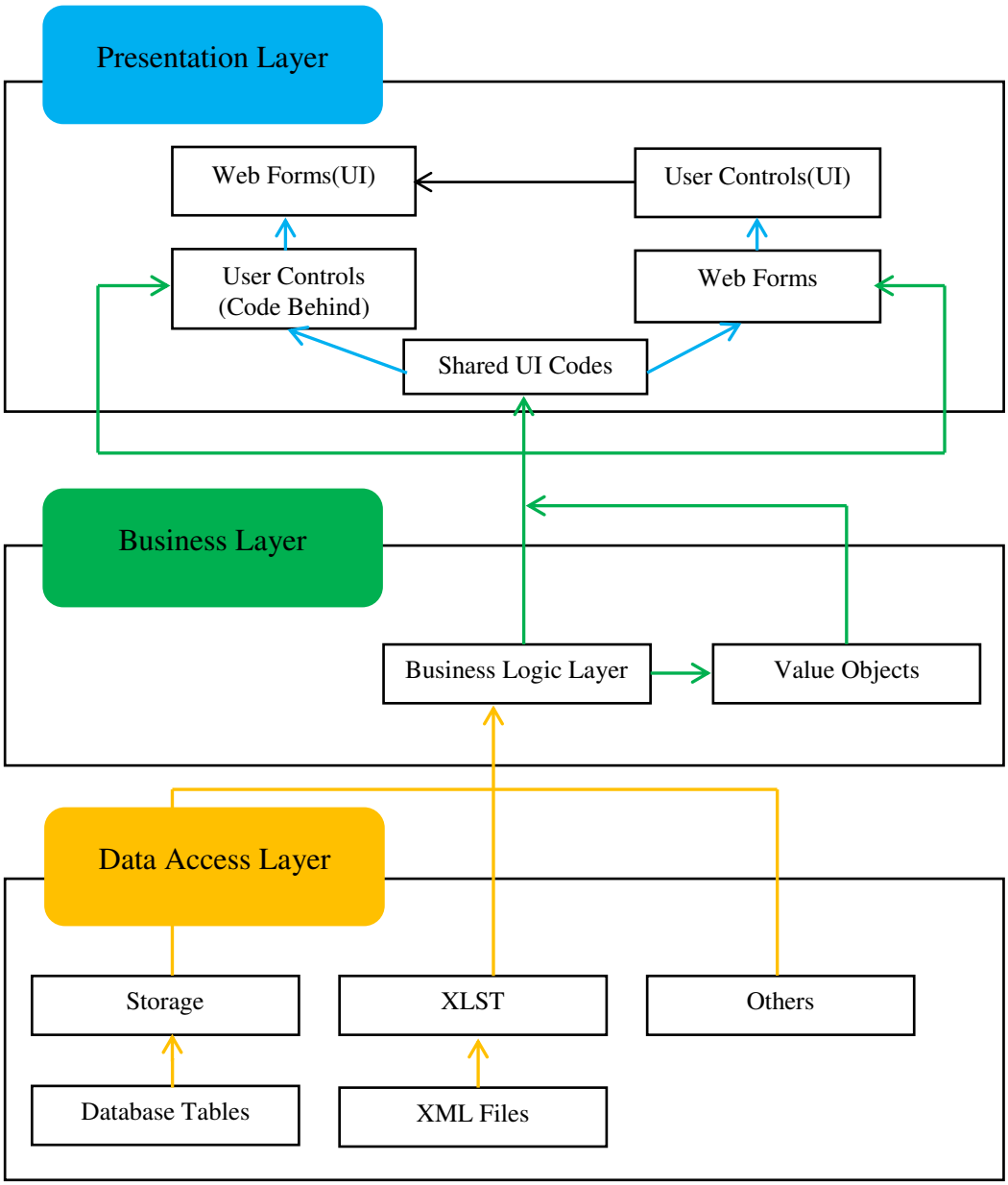


Figure 3.4 Three-tier architecture diagram.

A class generally consists of the following structures:

- Includes methods, properties, fields that each called member (Figure 3.5).
- Methods are specific code block that that executes the function and belongs to class.

- Properties are specific codes that execute the function of fields.
- Properties can be read-only or updateable, however to make the fields readable some special structures called property is needed to use instead of methods.
- Methods, fields and properties can be accessed from only in the class scope or from all classes. Therefore, the access identifiers are used such as Public, Private, Internal, Protected, and External…
- Methods can be defined with same and different parameters (method overloading).

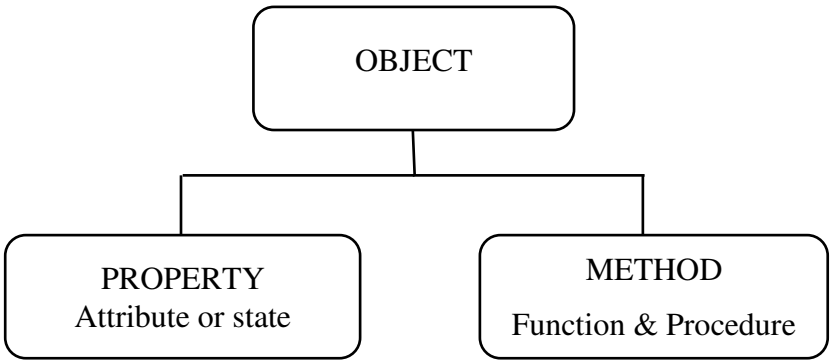


Figure 3.5 Basic components of object-oriented programming.

In n-tier architecture each tier consists of classes. These classes provide reusability. Tier describes physically separation of various installation components such as DLL, EXE on servers. Tiers can be prepared in .NET platform using classes easily and can be used in other pages. N-tier architecture consists of 4 separated tiers which are shown on Figure 3.1. Each tier has a specific function and each tier must be managed and considered separately.

Data Access Layer, only responsible for providing access to physical database. Database access methods, if necessary input/output interfaces and communication protocols are written in data access layer. In my design Data Access Layer also includes database connection string and some specific methods such as '*RunSP'* and '*LoadSQL'*. Briefly, RunSP is a method which takes two parameter *'procedurename'* and *'paramvalues'*. RunSP is a method which executes the method which name is given as parameter with the parameters which are given with an

object array (Figure 3.6). It has a simple work just takes the name of the procedure and the parameter array then executes the stored procedure on database.

```
public object RunSP(string procedurename, params object[] paramvalues)
{
    try
    {
        return OracleHelper.ExecuteNonQuery(connection, procedurename, paramvalues);
    }
    catch (OracleException ex)
    {
        throw ex;//DispatchException(ex);
    }
}
```

Figure 3.6 Definition of RunSP.

RunSP method also encapsulate another method name of '*ExecuteNonQuery*' which is a standard stored procedure executive method defined by Database companies.

When the first appearance of the n-tier architecture the examples of the big companies are like this and it is modelled again in this application (Figure 3.7).

```
public static object ExecuteNonQuery(OracleConnection connection, string spName, params object[] parameterValues)
{
    if (connection == null) throw new ArgumentNullException("connection");
    if (spName == null || spName.Length == 0) throw new ArgumentNullException("spName");

    // If we receive parameter values, we need to figure out where they go
    if ((parameterValues != null) && (parameterValues.Length > 0))
    {
        // Pull the parameters for this stored procedure from the parameter cache
        OracleParameter[] commandParameters =
            OracleHelperParameterCache.GetSpParameterSet(connection, spName);

        // Assign the provided values to these parameters based on parameter order
        AssignParameterValues(commandParameters, parameterValues);

        // Call the overload that takes an array of OracleParameters
        ExecuteNonQuery(connection, CommandType.StoredProcedure, spName, commandParameters);


        object res = 0;
        for (int i = 0; i < parameterValues.Length; i++)
            if (commandParameters[i].Direction == ParameterDirection.Output)
            {
                res = commandParameters[i].Value;
                parameterValues[i] = commandParameters[i].Value;
            }
        return res;
```

Figure 3.7 Definition of ExecuteNonQuery method.

According to recommended standard RunSP is used to execute Insert, Update, Delete procedures that written as stored procedure and LoadSQL is used to execute open Select statement.

```
public DataTable LoadSQL(string pSqlText)
{
    OracleCommand command = new OracleCommand();
    command.Connection = connection;
    command.CommandType = CommandType.Text;
    command.CommandText = pSqlText;
    DataTable dt = new DataTable();
    OracleDataAdapter da = new OracleDataAdapter(command.CommandText, connection);

    try
    {
        connection.Open();
        da.Fill(dt);
        return dt;
    }
    catch (Exception ex)
    {
        throw new Exception(ex.Message);
        //return new DataTable();
    }
    finally
    {
        connection.Close();
    }
}
```

Figure 3.8 Definition of LoadSQL.

Briefly, LoadSQL method takes and sql statement type of string then executes. According to recommended standard Select methods do not transform into stored procedures, they remain in simple sql statement form (Figure 3.8). LoadSQL method also encapsulates another method name of '*Fill*'. This method is one of the methods of '*Dataset*' object. Inserts or refreshes rows in a specified range in the '*System.Data.DataSet*' to match those in the data source using the '*System.Data.DataTable*' name. The name of the '*DataTable*' used for table mapping.

Business Layer, also called business logic, middle tier, logic tier and it controls the application functionality. Business layer works as a bridge between the presentation layer and data access layer. All data pass business layer for formatting before proceeding to presentation layer. This layer manipulates the data and prepares for presentation. Business layer formats the incoming data from database and transforms it into a suitable form that can be used in the user interface. A business layer consists of class libraries and this class must be suitable format with OOP.

Business layer is added as *Class Library* form in Visual Studio. And each class with particular tasks is added into this Class Library as a class. A business layer is constructed as a class so it has a constructor that *OracleDataHelper* is initialized and methods that execute the transactions (Figure 3.9).

In this project there is a business class name of *MinerbyBF* and a class library name of '*MinerSystem.cs*' contains methods for Select sql statements. To ensure standardization method names begins with the operation name such as 'Select' and continues with the objects' name(tables) which affected the transaction such as '*TableNames'* and followed with the parameter that identifies record which the transaction will effect.

```
public class MinerSystem
{
     ORADataHelper dh;
     public MinerSystem()
    {
        try
        {
            dh = new ORADataHelper();
        }
        catch (Exception)
        {
            dh = null;
        }
    }
```

Figure 3.9 A constructor example for business layer class for miner application.

Briefly, to associate business layer class with the web project, business layer class must be added as reference to web project. Right click on the web project and '*Add Reference'* option is selected. Then a window is open which included all business layer classes, to see the related classes on the left tab '*Solution*' option is opened and '*Projects'* is selected. Then the related classes are selected combo box one or more classes can be selected and click ok button. We can also add outsource classes to the project by choosing another tab from the left side of the window such as '*Assemblies'* or '*COM'*.

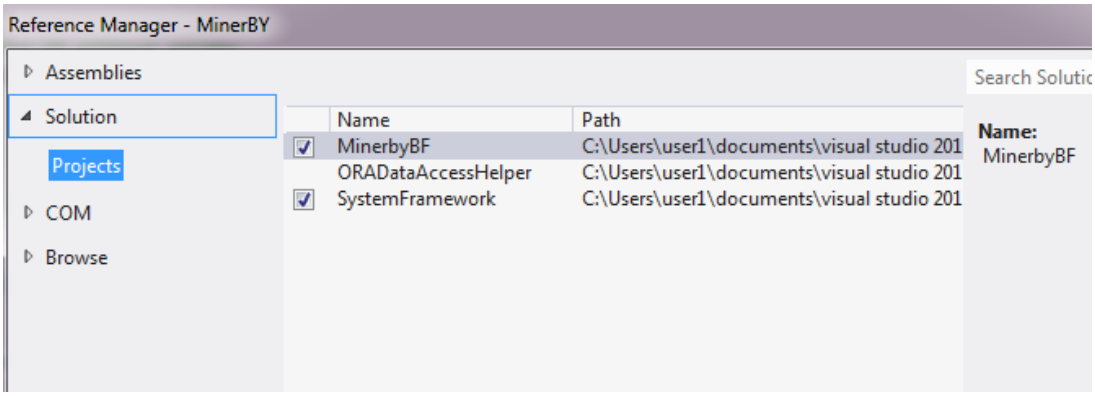


Figure 3.10 Business layer 'MinerBF' was added as a reference to application layer.

Business layer class is written once; if it is necessary some addition methods can add in it. Written class can be used in another projects without need rewrite. This situation helps to develop more effective, more consistent, faster programs by reducing the workload.

Application Layer is the top level of the three-tier architecture model. Application layer contains shared UI (User Interface) codes, code behind and design objects. The purpose of the application layer is to translate tasks and results to something the user can understand. Application layer consists of Aspx pages in ASP.NET and includes user interface controls to show data to user and getting input from the user such as buttons, textboxes, grids, labels. Also, application layer contains in code behind that includes the codes that executes the business layer class methods. Application layer use the business layer classes to retrieve data for user operations. Application layer is the layer where users interact with the application.

```
MinerbyBF.MinerSystem mnrSys = new MinerbyBF.MinerSystem();
```

Figure 3.11 Creating instance of the business layer object.

Application layer also includes the instance of the business layer object as seen in Figure 3.11 above. It is required to define instance of the business layer object to use the select methods in the business layer. The example of the using select method via business layer object is shown below in Figure 3.12.

```
MinerbyBF.MinerSystem mnrSys = new MinerbyBF.MinerSystem();
DataTable dtGraph = null;
dtGraph = mnrSys.Select_ContentOf_Table_ByGender_ByICD
    (ddlTableNames.SelectedValue.ToString(), mainTemp);
```

Figure 3.12 Accessing the business layer's select method from application layer via business layer object.

Briefly, the route of the data is as described below:

- Data layer communicates with the physical database,
- Data layer encapsulates the some specific methods to communicate the physical database (LoadSQL, RunSP),
- Data layer transfers data coming from database to business logic to format the data,
- Business logic encapsulates some specific methods that executes the user transactions on data,
- Business logic process the data coming from data layer and formats the data,
- Business logic transfers formatted data to Application Layer,
- Application Layer encapsulates the Business Logic instance and can access the its methods,
- Application Layer executes the methods to present the data,
- In Application layer, formatted data is bind to a component such as a grid or a table to make user to see the data,
- Application data presents data to user and provide user an interactive environment with data.

System Framework is the vertical layer that can interact with each layer (Figure3.1). System Framework contains anything that has universal scope across the solution; specifically the definitions of data structures used to pass information between components in the different layers (Kearns, 2010). System Framework can be accessed from each layer. Some functions need to be accessed from each layer such as an Encryption function. Consider Encryption function, in data access layer connection string parameters password and username can be encrypt by using

Encryption function (Figure 3.13). In business logic, Cryptology function can be used to encrypt a parameter to hide. In Application Layer, Cryptology function can be used to encrypt Patients' protocol number while passing as parameter between pages. System Framework offers functionality that is used by all layers; includes shared functions.

```
public static string Encrypt(string clearText, string Password)
{

    byte[] clearBytes =
        System.Text.Encoding.Unicode.GetBytes(clearText);

    PasswordDeriveBytes pdb = new PasswordDeriveBytes(Password,
        new byte[] {0x49, 0x76, 0x61, 0x6e, 0x20, 0x4d,
                        0x65, 0x64, 0x76, 0x65, 0x64, 0x65, 0x76});

    byte[] encryptedData = Encrypt(clearBytes,
        pdb.GetBytes(32), pdb.GetBytes(16));

    return Convert.ToBase64String(encryptedData);

}

public static string Encrypt(string clearText)
{
    return Encrypt(clearText, "qwerty");
}
```

Figure 3.13 Definition of Encrypt method in system framework.

Also, system framework can include some helper functions that make easier to coding. For example, frequently using code block can written here as a method and used from here many times such as FormatDLL method. In ASP.NET components the top value of a drop down list has selected in default situation.

However, in some application this field can be leaved empty. Therefore it is required to format DLL (Figure 3.14). An easy function can be written that format the DLL, but this function is needed to be access each part of the application. Therefore, this function must be written a layer which can communicate with another layer, and this layer is System Framework.

System framework includes any helper class library such as '*Enums*'. In the big companies, some classification can be required to identify the each application on the system. For example in the hospital software 26 separated applications work

together. And each application has their own ID to identify each other. This IDs are stored as enums in the '*Enums*' class library in hospital information system.

```
public static DataTable FormatDDL(DataTable dt)
{
    if (dt.Rows.Count > 0 && dt.Rows[0][0].ToString() != "-1")
    {
        DataRow dr = dt.NewRow();
        dr[0] = -1;
        dr[1] = "------";
        dt.Rows.InsertAt(dr, 0);
    }
    else if (dt.Rows.Count == 0)
    {
        DataRow dr = dt.NewRow();
        dr[0] = -1;
        dr[1] = "------";
        dt.Rows.InsertAt(dr, 0);
    }

    return dt;
}
```

Figure 3.14 Definition of FormatDLL method in system framework.

Generally System Framework includes security, operational management and communication issues. Security part can include components about authentication, authorization and validation. Operational management part includes exception handling process, configuration settings, tracing arrangements, logging and performance. Communication part includes components such as services and application to provide communication.

Finally, the most important tips to say about common layer (system framework), it must be added as reference to each layer of the project (presentation layer, business logic and data access later) to access from everywhere.

3.2 Web Services

Web services are APIs (integration methods) that based on XML messaging. The most important intended use is to provide integration and communication between the different systems to ensure messaging. Web services are platform independent, therefore they are widely used. For example, a web service written in java can be used in asp.net web applications. When you retrieve data from web service or send data to web service, your computer opens a port to transport to data.

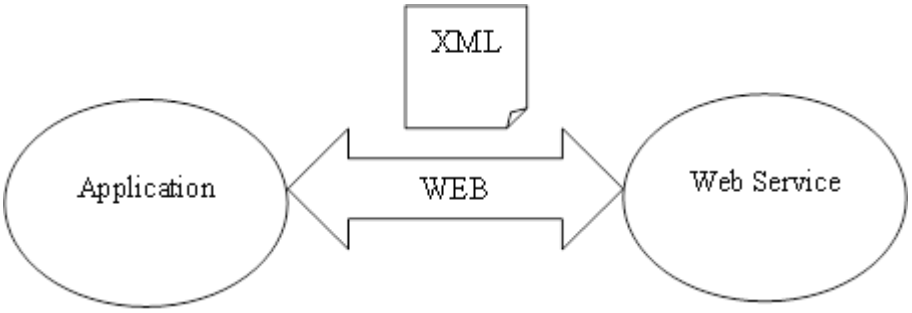


Figure 3.15 Web service process diagram.

In this project some of the patient data such as laboratory results stored in Ventura databases so my system need to be communicate Ventura database systems via web services. Current system in hospital does not contain all details about patients so it causes a problem such as the tables need to be joined with Ventura tables via web services. In this project web services are used to retrieve patients' laboratory results to examine the results to find renal failure evidence. Adding a web service to application is described the pictures on following below:

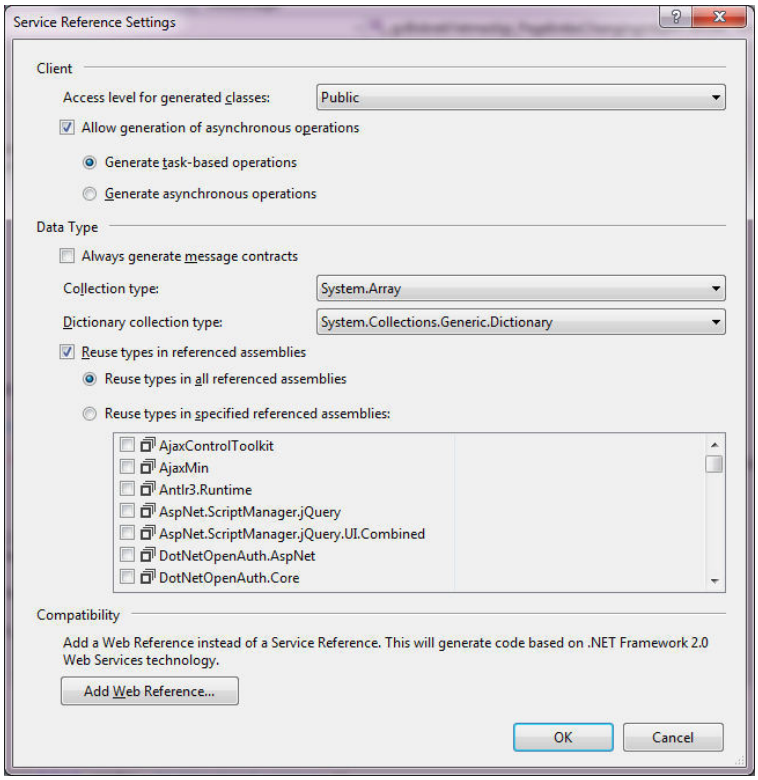


Figure 3.16 Click advanced button at the bottom of the window and after that click add web reference button at the bottom of the window.

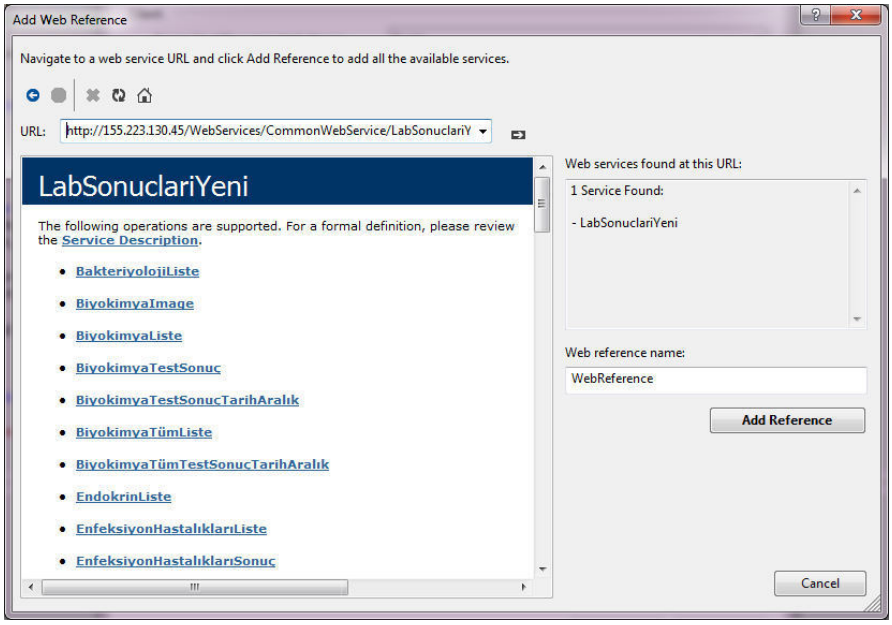


Figure 3.17 Enter URL of the web service then go after that name the web reference and click add reference.

In this project *BiyokimyaListe* and *BiyokimyaTestSonucTarihAralık* methods are used to retrieve biochemistry laboratory results of the patients'. The purpose of using web services in this operation is overcoming the barriers about platform differences retrieving data from another system and maps them with data in our database. We will discuss the contents of the methods in following chapters.

3.3 Visual Studio 2012 & Asp.Net

Visual Studio is a comprehensive collection of tools and services to help you to create variety of applications runs on Microsoft platform and another platform. Visual Studio is a program compiler that developed by Microsoft. Visual Studio support many programming languages. In addition, Visual Studio provides link between all projects, project team and participants. Current versions of Visual Studio, no matter where your team, team can operate using Eclipse or Xcode development tools with greater flexibility.

Visual Studio is used to developed windows form applications, web sites, web applications, web services, console applications and user interfaces for the platforms that supported by Microsoft windows, windows mobile, windows CE, .NET framework, .Net Compact framework and Microsoft Silverlight. Besides including IntelliSense, includes code editor supporting 'code refactoring'. Integrated debugger works both source-level and machine-level.

Visual Studio support different programming languages, its code editor and debugger provide support for all programming languages. Internal languages are C/C++ (via Visual C++), VB.NET, and C #and F # (Visual Studio 2010 onward).

3.3.1 Visual Studio Features

Code Editor, Visual Studio IDE like any other contains syntax highlighting and code completion function and methods and also includes loops, queries. It supports XML, Java Script and CSS in development of web applications. Code editor also supports bookmarks for quick navigation.

Debugger, visual studio includes both source-level and machine-level debugger. It can also use for application written in any language. In addition to this, it can observe the process. Debugger allows execution to stop temporarily at a certain position and execution progresses, keeps the track of the values of variables.

Designer provides a number of visual designers help to develop applications in visual studio. These tools are followed below*:*

- Windows Form Designer is used to create GUI applications.
- WPF Designer, was introduced in 2008 and supports drag and drop metaphor.
- Web Designer/Development, visual studio also includes a web-site editor and provides user to create applications by using drag and drop technology. Used for ASP.NET applications and also supports HTML, JavaScript and CSS. It uses code behind model to provide connection ASP.NET code.

- Class Designer, creates classes using UML models, and creates the outline classes and methods for C #and VB.NET codes.
- Data Designer, used to organize database schema and data tables.
- Mapping Designer, starting with visual studio 2008 it is used to design database schema that contained by LINQ and SQL.

Extensibility, visual studio allows developers write extensions to expand the program capabilities. These extensions are plug-into visual studio and aims to expand its functionality. Packages are created using visual studio SDK.

3.3.2 .NET Framework

.Net Framework is an application development platform developed by Microsoft and based on open internet protocols and standards. .Net framework consists of two basic components: Common language runtime (CLR) execution engine that handles running applications; and .NET framework class library which provides developers reusable codes.

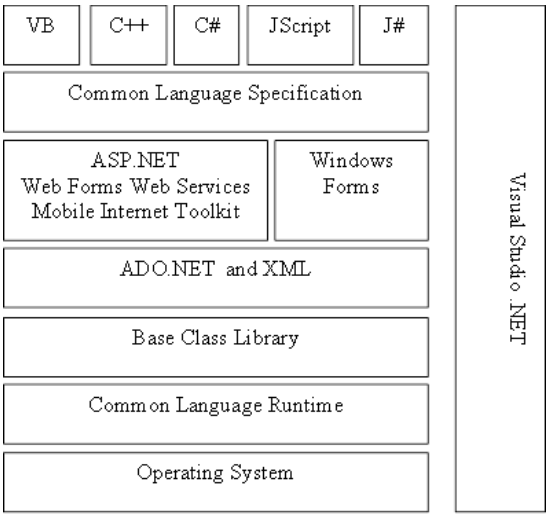


Figure 3.18 .NET framework components.

.NET framework provides following services for running applications:

- Memory management,
- Common type system,

- A comprehensive class library,
- Development frameworks and technologies,
- Cross language interoperability,
- Version compatibility and
- Multi version support.

3.4 TOAD for Oracle

Toad is an application development tool; the full name is "**t**ool for **o**racle **a**pplication **d**evelopers". Toad is a quest software product and administration/development interface of oracle. Using Toad, developers can construct PL/SQL packages, procedures, triggers and functions, test and debug. Developers can create database objects such as tables, views, indexes, constraints and users and also developers can change them.

Toad's sql editor provides the opportunity to easily write scripts and queries and tests Oracle data view and data grid. Briefly using Toad developers can create and browse or alter objects (tables, views, indexes etc.), graphically build, execute and tune queries, edit and debug PL/SQL and stored procedures (functions, packages, triggers), search objects, seek and fix the constraints, triggers, extents, indexes and database problems.

The hospital database has vital importance; therefore the data in the EUTF connection has copied in a test environment for programmers to test their application without cause a problem. Figure 3.19 illustrates the connection the EUTFZTEST database.

During the development process in this application HASTA_TANI_HAREKET and HASTA_MASTER tables are used in the EUTFDBA schema (Figure 3.20). It will be discussed in the later sections content and the structure of the tables and how they are used.

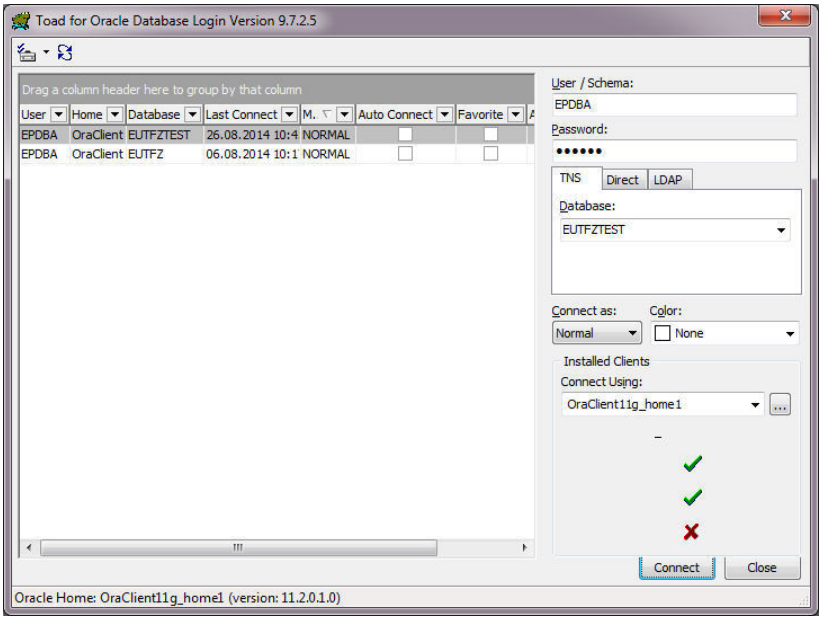


Figure 3.19 Toad for oracle can be logged in with a username and password.

In this application the related tables and views rests in EUTFZ database. And I connect to database with username and password. After the entrance is guaranteed the window will be opened. By clicking the schema browser button each table can be seen in each schema.

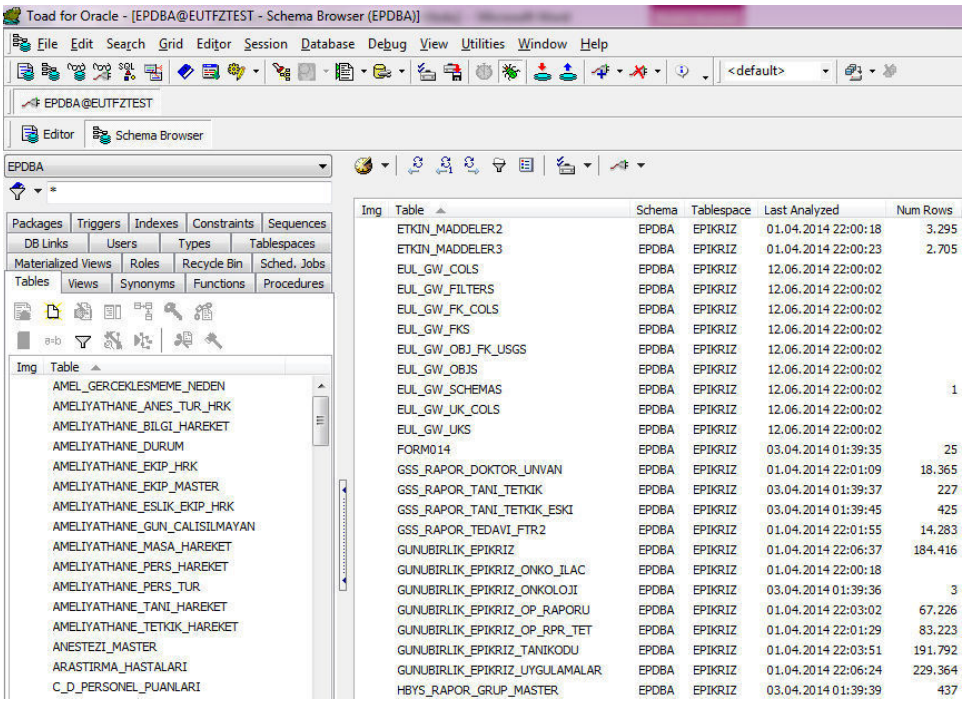


Figure 3.20 This figure represents the tables in EUTFDBA schema.

CHAPTER FOUR
DEVELOPMENT PROCESS

The first stage of the development process was to prepare data, data elimination and extract the useful data. Hospital databases contain very large amounts of data inside. Querying this crowded data is quite a difficult task.

4.1 Preparing Data for Data Mining

Data are backbone of data mining and knowledge discovery; however real-world business data usually are not available in data-mining-ready form (CRC, n.d.). Even if the data has already stored in database, it could be troublesome to retrieve the correct data. The biggest challenge for data miners, is preparing data suitable for modelling. Roughly 70% of data mining operation time is spent on preparing data obtained from different sources, therefore considerable time and effort should be spent on preparing data tables to be suitable for data mining modelling.

In this study, it is needed to obtain urinalysis and blood analysis results from the patients that diagnosed with diabetes and hypertension. Hence the structure of hospital's the database is too complicated data acquisition process was quite difficult.

The current database does not include all details of the data. It is need to join the tables in hospital database with view (EULIS.V_ALIS_SONUC) from a company name of Ventura. This view includes some determinative data such as "sender material" which specify the type of material that which will be analyzed in laboratory.

The tables and views join with patients' *protocol* which is a unique number and identifies the patient and patient's *number of repetition* which specifies how many times patients arrived. Because a patient may arrive more than once so the analysis must be related with both "protocol" and "number of repetition" to reach the exact

record. Since hospital database has huge amount of data and analyze these data would take very long time, time interval set to one week.

Also, it is need to filter EUTFDBA.HASTA_TANI_HAREKET table like according "*patient diagnosis code*" which is used to indicate the diagnosis of the patients. Such as E11, I10 etc. E11 code indicates diabetes and I10 indicates hypertension. In this study only interested in patients who were diagnosed E11 (diabetes) and I10 (hypertension) because these patients generate risk group for Chronic Renal Failure. So it is guaranteed to achieve more accurate results in this narrowing data pool.

In the first stage of data preparation part the rows usually contain information of patients. The columns describe the attributes such as '*Test_Cod*e', '*Diagnose_Code'*, '*Test_Name*', '*Date'*, '*High_Low*', '*Test_Result*', 'Unit', '*Reference_Range*' etc. Because of the first version of the analysis includes a column name of '*Test_Name*', all the tests carried out scope of the all analyzes were shown in a separate row. For instance, a patient subjected to urinalysis; "*Bilirubin, Density, Epithelial Cells, RBC, Glucose, Urine Microscopy, Ketone, Leukocytes, Nitrite, pH, Protein, Urobilinogen*" values are measured.

In the next stage of data preparation the private information about patients' identity have been hidden not to violate the privacy of the patients. This specific information was determined as patients' name, surname, date of birth, protocol, ID numbers. For the implementation of these changes queries had to be manipulated. All the information was being open via the previous queries. These queries have needed to be manipulated and run again on the data sets.

Removing properties that identifying the patients and retrieving multiple records (rows on database) for the same patient have led a new need: New identifier for each patient. This identifier must be unique for each patient but not include any protocol, ID number, name or surname of the patient. While concealing the identity of the patient, on the other hand notifying that is important if two discrete records belongs

the same patient and it is necessary to generate a new number to resolve this ambiguity.

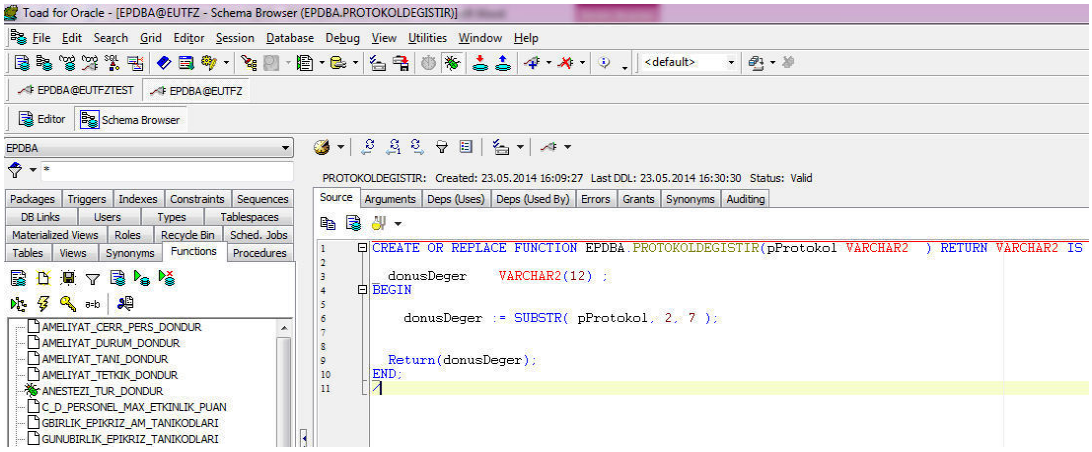


Figure 4.1 Definition of ProtokolDegistir function.

To resolve this problem a function name of “ProtokolDegistir” was created to identify each user with a number. It was necessary to provide results in output depends on patients’ protocol numbers. Therefore this function takes the patients’ protocol number as an input and apply Substring operation on the parameter. Definition of the “ProtokolDegistir” function is seen in the Figure 4.1. Thus hiding the some parts of the protocol numbers and producing results depends on these protocol numbers a new identifier number was created.

4.2 Applying Data Mining Methods Using Oracle Data Miner

The data used in the applications is located in the Oracle database. Therefore, software that can interact more conveniently with Oracle database was suitable to use.

Oracle Data Miner is and Oracle SQL Developer extension that provides applying data mining techniques on data provides data analysis, applying Oracle Data mining models.

Oracle Data miner works interact with Oracle database. Using Oracle Data Miner, data in excel format can be transferred to Oracle database. It is mentioned in the Preparing Data for Data Mining section that data was retrieved and filtered according to appropriate format.

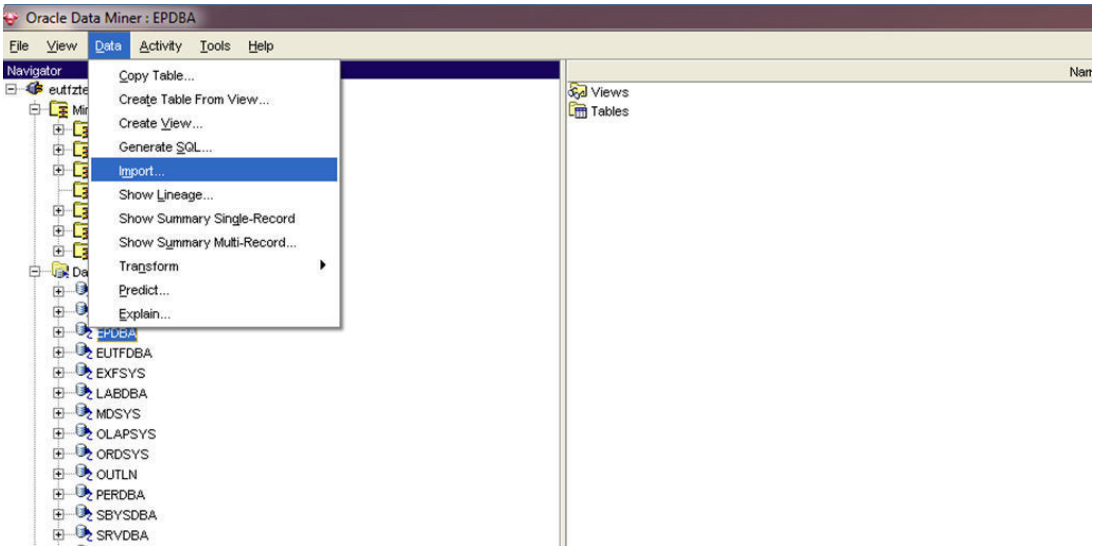


Figure 4.2 Data can be import to Oracle database.

In this study, I have data in excel format after my pre studies. The excel data consists of the diabetic and hypertension patients Creatinine results from blood test. Determining parameter is Creatinine value in blood and the excel data includes the filtered data to apply data mining techniques on.

First the excel format must be translated into csv file to apply by the oracle data miner import operation. The common separator assign as ';' for csv file. Import operation do not accrue directly on excel file, it accrues state that transformed into csv format.

After the csv format data file included by the Oracle, a new table created in Oracle database name of KAN_SONUC KREATININ. And the process statistics are given by Oracle Data Miner during the process.

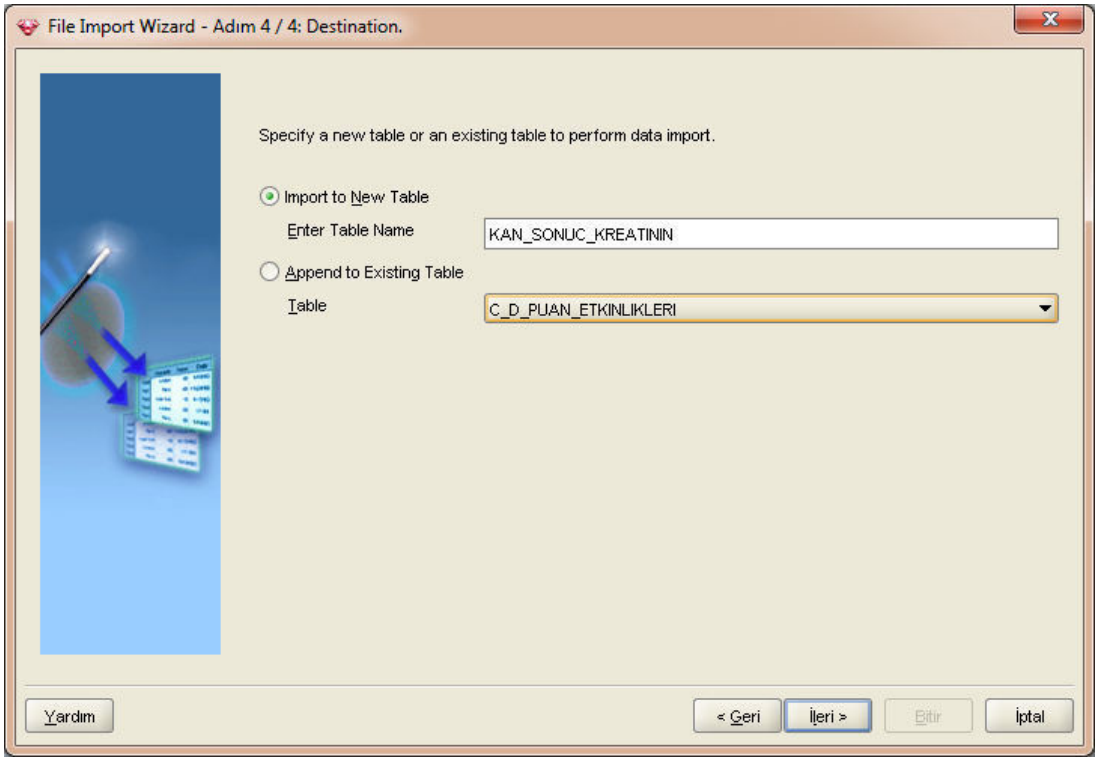


Figure 4.3 Csv file imported to create table name of KAN_SONUC_KREATININ.

As seen in the images above, the csv file selected as import file and the separator specified as ';'. The first row of csv file specifies the column names so oracle miner takes the column names and types.

A small adjustment was made at this stage: REFMIN, REFMAX and RESULT columns type changed to Number (Figure 4.3). And specify a new table name import csv data into as KAN_SONUC_KREATININ.

Name	Type	Size	Allow NULLS
CINSIYET	VARCHAR2	4000	✓
REFERANS_HASTA	NUMBER	22	✓
YAS	NUMBER	22	✓
TEKRAR	NUMBER	22	✓
TANIKODU	VARCHAR2	4000	✓
TEST_KODU	NUMBER	22	✓
TEST_ADI	VARCHAR2	4000	✓
ISTEK_TARIHI	VARCHAR2	4000	✓
SONUC_TARIHI	VARCHAR2	4000	✓
DUSUK_YUKSEK	VARCHAR2	4000	✓
SONUC	NUMBER	10	✓
BIRIM	VARCHAR2	4000	✓
REFMIN	NUMBER	10	✓
REFMAX	NUMBER	10	✓

Figure 4.4 Table created successfully.

4.2.1 Process of Import External Data

A specific formula applied during diagnosis of renal failure. This formula varies depending on gender. Further, this formula delivers results depending on such patients' creatinine levels in the blood, height and weight. However, data such as height and weight of the patients' do not store in hospital databases. Because of these limitations a new more simplified approach was used.

Here are the formulas that used for diagnosis of renal failure by doctors in hospital:

w: Weight of the patient (kg).
a: Age of the patient.
c: Creatinine level in blood of patient (mg/dl).

Formula for females	Formula for males
$[140 - a] * \frac{w}{72 * c} * 0{,}85$	$[140 - a] * \frac{w}{72 * c}$

Figure 4.5 Formulas for females and males that determining to renal failure.

In the first phase of the study it was predicted this formula used. But as the time progressed it is understood that such data cannot be obtained. Because of these limitations a new more simplified approach was used: Creatinine level in blood of patient.

Although creatinine level in blood is not the most decisive parameter for chronic diseases, least it will show the patient is under renal failure risk or not. Laboratory results of the patients are needed. This subject was mentioned in detailed in Preparing Data for Data Mining section. At Figure 4.6 table can be access from TOAD for Oracle tool.

KAN_SONUC_KREATININ: Created: 07.08.2014 11:54:50 Last DDL: 07.08.2014 11:56:14

Columns | Indexes | Constraints | Triggers | Data | Script | Grants | Synonyms | Partitions | Subpartitions | Stats/Size | Referential | Used By | Policies | Auditing

CINSIYET	REFERANS_HASTA	YAS	TEKRAR	TANIKODU	TEST_KODU	TEST_ADI	ISTEK_TARIHI	SONUC_TARIHI	DUSUK_YUKSEK	SONUC	BIRIM	REFMIN	REFMAX	BOBREK_YETMEZLIGI
E	140352	0	2	I10	667	Kreatinin	10.3.14 1:04	10.03.2014 00:57:38	D	0,1	mg/dL	0,3	0,7	1
E	140352	0	2	I10	667	Kreatinin	17.3.14 1:04	17.03.2014 00:57:31	D	0,1	mg/dL	0,3	0,7	1
K	140239	55	2	I10	667	Kreatinin	18.2.14 1:26	18.02.2014 01:20:38	D	0,18	mg/dL	0,6	1,1	1
K	140239	55	2	E11	667	Kreatinin	18.2.14 1:26	18.02.2014 01:20:38	D	0,18	mg/dL	0,6	1,1	1
K	120197	55	9	I10	667	Kreatinin	21.3.14 3:37	21.03.2014 03:31:05	D	0,19	mg/dL	0,6	1,1	1
K	120197	55	9	I10	667	Kreatinin	26.3.14 3:37	26.03.2014 03:30:57	D	0,19	mg/dL	0,6	1,1	1
K	120197	55	9	I10	667	Kreatinin	2.4.14 3:15	02.04.2014 03:08:50	D	0,19	mg/dL	0,6	1,1	1
K	130455	47	26	I10	667	Kreatinin	11.5.14 6:37	11.05.2014 06:30:57	D	0,19	mg/dL	0,6	1,1	1
K	130455	47	26	I10	667	Kreatinin	19.5.14 6:45	19.05.2014 06:38:10	D	0,19	mg/dL	0,6	1,1	1
K	140239	55	2	I10	667	Kreatinin	17.2.14 4:31	17.02.2014 04:24:38	D	0,19	mg/dL	0,6	1,1	1
K	140239	55	2	E11	667	Kreatinin	17.2.14 4:31	17.02.2014 04:24:38	D	0,19	mg/dL	0,6	1,1	1
K	50241	50	22	I10	667	Kreatinin	25.4.14 10:04	25.04.2014 09:56:44	D	0,2	mg/dL	0,6	1,1	1
K	120197	55	9	I10	667	Kreatinin	20.3.14 21:20	20.03.2014 21:14:06	D	0,2	mg/dL	0,6	1,1	1
K	120197	55	9	I10	667	Kreatinin	14.4.14 3:57	14.04.2014 03:49:56	D	0,2	mg/dL	0,6	1,1	1
K	130455	47	26	I10	667	Kreatinin	16.5.14 7:35	16.05.2014 07:28:19	D	0,2	mg/dL	0,6	1,1	1
K	130455	47	26	I10	667	Kreatinin	17.5.14 7:22	17.05.2014 07:15:03	D	0,2	mg/dL	0,6	1,1	1
K	130455	47	26	I10	667	Kreatinin	18.5.14 7:19	18.05.2014 07:12:46	D	0,2	mg/dL	0,6	1,1	1
K	130455	47	26	I10	667	Kreatinin	22.5.14 7:19	22.05.2014 07:11:48	D	0,2	mg/dL	0,6	1,1	1
K	130455	47	26	I10	667	Kreatinin	24.5.14 7:09	24.05.2014 07:02:37	D	0,2	mg/dL	0,6	1,1	1
K	130455	47	26	I10	667	Kreatinin	26.5.14 11:37	26.05.2014 11:29:10	D	0,2	mg/dL	0,6	1,1	1
K	140305	3	1	I10	667	Kreatinin	5.3.14 8:43	05.03.2014 08:36:18	D	0,2	mg/dL	0,3	0,7	1
E	140352	0	2	I10	667	Kreatinin	6.3.14 17:53	06.03.2014 17:47:16	D	0,2	mg/dL	0,3	0,7	1
E	140352	0	2	I10	667	Kreatinin	24.3.14 0:18	24.03.2014 00:11:58	D	0,2	mg/dL	0,3	0,7	1
K	130455	47	26	I10	667	Kreatinin	23.5.14 7:30	23.05.2014 07:23:42	D	0,21	mg/dL	0,6	1,1	1
E	130669	90	10	I10	667	Kreatinin	1.5.14 1:31	01.05.2014 01:24:04	D	0,21	mg/dL	0,7	1,3	1
K	120197	55	9	I10	667	Kreatinin	16.4.14 3:03	16.04.2014 02:55:53	D	0,22	mg/dL	0,6	1,1	1

Figure 4.6 Contents of KAN_SONUC_KREATININ table.

After the csv format model data included in the Oracle database as a table name of KAN_SONUC_KREATININ a specific model created that will be descriptive model of renal failure.

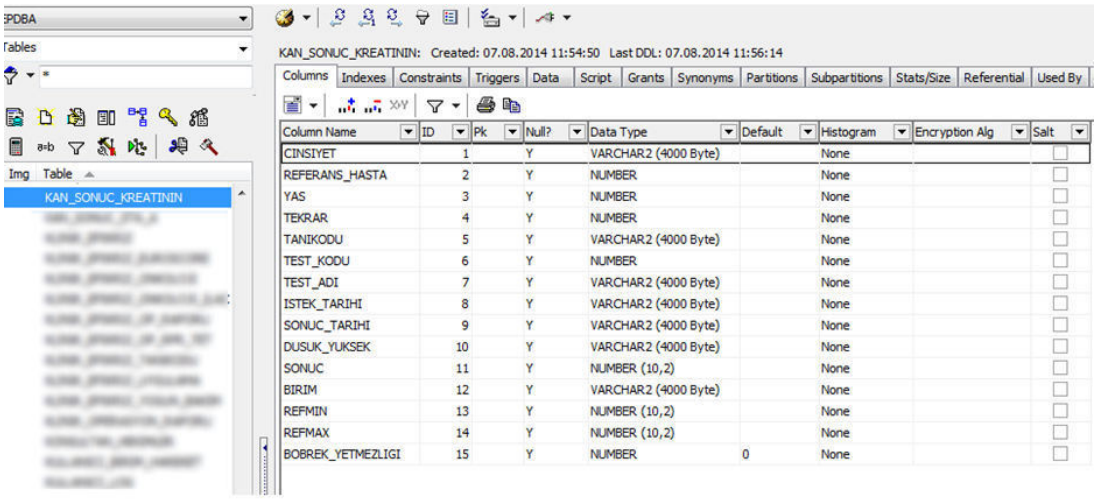
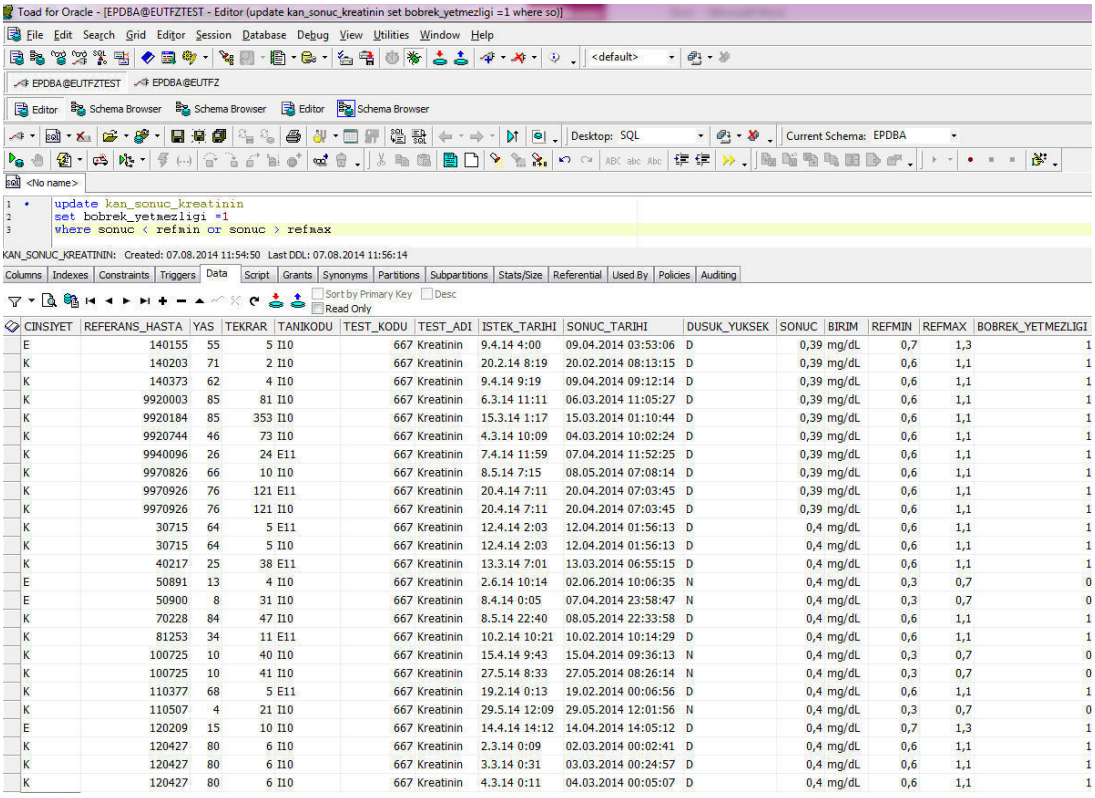


Figure 4.7 Bobrek_Yetmezligi column added to table.

This column updated depending on the creatinine level in blood of the patient. A simple query is used for this. The query compares the creatinine results with reference value boundaries. If the result is out of the reference range the

Bobrek_Yetmezligi column of the specific patient is updated as 1 otherwise updated as 0.



CINSIYET	REFERANS_HASTA	YAS	TEKRAR	TANIKODU	TEST_KODU	TEST_ADI	ISTEK_TARIHI	SONUC_TARIHI	DUSUK_YUKSEK	SONUC	BIRIM	REFMIN	REFMAX	BOBREK_YETMEZLIGI
E	140155	55	5	I10	667	Kreatinin	9.4.14 4:00	09.04.2014 03:53:06	D	0,39	mg/dL	0,7	1,3	1
K	140203	71	2	I10	667	Kreatinin	20.2.14 8:19	20.02.2014 08:13:15	D	0,39	mg/dL	0,6	1,1	1
K	140373	62	4	I10	667	Kreatinin	9.4.14 9:19	09.04.2014 09:12:14	D	0,39	mg/dL	0,6	1,1	1
K	9920003	85	81	I10	667	Kreatinin	6.3.14 11:11	06.03.2014 11:05:27	D	0,39	mg/dL	0,6	1,1	1
K	9920184	85	353	I10	667	Kreatinin	15.3.14 1:17	15.03.2014 01:10:44	D	0,39	mg/dL	0,6	1,1	1
K	9920744	46	73	I10	667	Kreatinin	4.3.14 10:09	04.03.2014 10:02:24	D	0,39	mg/dL	0,6	1,1	1
K	9940096	26	24	E11	667	Kreatinin	7.4.14 11:59	07.04.2014 11:52:25	D	0,39	mg/dL	0,6	1,1	1
K	9970826	66	10	I10	667	Kreatinin	8.5.14 7:15	08.05.2014 07:08:14	D	0,39	mg/dL	0,6	1,1	1
K	9970926	76	121	E11	667	Kreatinin	20.4.14 7:11	20.04.2014 07:03:45	D	0,39	mg/dL	0,6	1,1	1
K	9970926	76	121	I10	667	Kreatinin	20.4.14 7:11	20.04.2014 07:03:45	D	0,39	mg/dL	0,6	1,1	1
K	30715	64	5	E11	667	Kreatinin	12.4.14 2:03	12.04.2014 01:56:13	D	0,4	mg/dL	0,6	1,1	1
K	30715	64	5	I10	667	Kreatinin	12.4.14 2:03	12.04.2014 01:56:13	D	0,4	mg/dL	0,6	1,1	1
K	40217	25	38	E11	667	Kreatinin	13.3.14 7:01	13.03.2014 06:55:15	D	0,4	mg/dL	0,6	1,1	1
E	50891	13	4	I10	667	Kreatinin	2.6.14 10:14	02.06.2014 10:06:35	N	0,4	mg/dL	0,3	0,7	0
E	50900	8	31	I10	667	Kreatinin	8.4.14 0:05	07.04.2014 23:58:47	N	0,4	mg/dL	0,3	0,7	0
K	70228	84	47	I10	667	Kreatinin	8.5.14 22:40	08.05.2014 22:33:58	D	0,4	mg/dL	0,6	1,1	1
K	81253	34	11	E11	667	Kreatinin	10.2.14 10:21	10.02.2014 10:14:29	D	0,4	mg/dL	0,6	1,1	1
K	100725	10	40	I10	667	Kreatinin	15.4.14 9:43	15.04.2014 09:36:13	N	0,4	mg/dL	0,3	0,7	0
K	100725	10	41	I10	667	Kreatinin	27.5.14 8:33	27.05.2014 08:26:14	N	0,4	mg/dL	0,3	0,7	0
K	110377	68	5	E11	667	Kreatinin	19.2.14 0:13	19.02.2014 00:06:56	D	0,4	mg/dL	0,6	1,1	1
K	110507	4	21	I10	667	Kreatinin	29.5.14 12:09	29.05.2014 12:01:56	N	0,4	mg/dL	0,3	0,7	0
E	120209	15	10	I10	667	Kreatinin	14.4.14 14:12	14.04.2014 14:05:12	D	0,4	mg/dL	0,7	1,3	1
K	120427	80	6	I10	667	Kreatinin	2.3.14 0:09	02.03.2014 00:02:41	D	0,4	mg/dL	0,6	1,1	1
K	120427	80	6	I10	667	Kreatinin	3.3.14 0:31	03.03.2014 00:24:57	D	0,4	mg/dL	0,6	1,1	1
K	120427	80	6	I10	667	Kreatinin	4.3.14 0:11	04.03.2014 00:05:07	D	0,4	mg/dL	0,6	1,1	1

Figure 4.8 Bobrek_Yetmezligi column is updated depending on creatinine level in blood.

The data model is ready for learn and study by the data mining tool. The next steps include the views about learning such model and applying it.

4.2.2 The Process of Implementing Data Model for Data Mining

After creation of the model data, go back to Oracle Data Miner tool again. This model data must be learned by the system. Remember that, it was mentioned that the data import process at previous section. Now, it is time to examine this model and understand the behavior of model. It is needed to create a model to learn the behavior of this data set. To create this model Oracle Data Miner's Activity menu is used. Build option is selected in Activity menu to create a model. While implementing a model one of the Classification algorithms must be chosen. A lot of algorithm in the classification domain can be used when implementing a model.

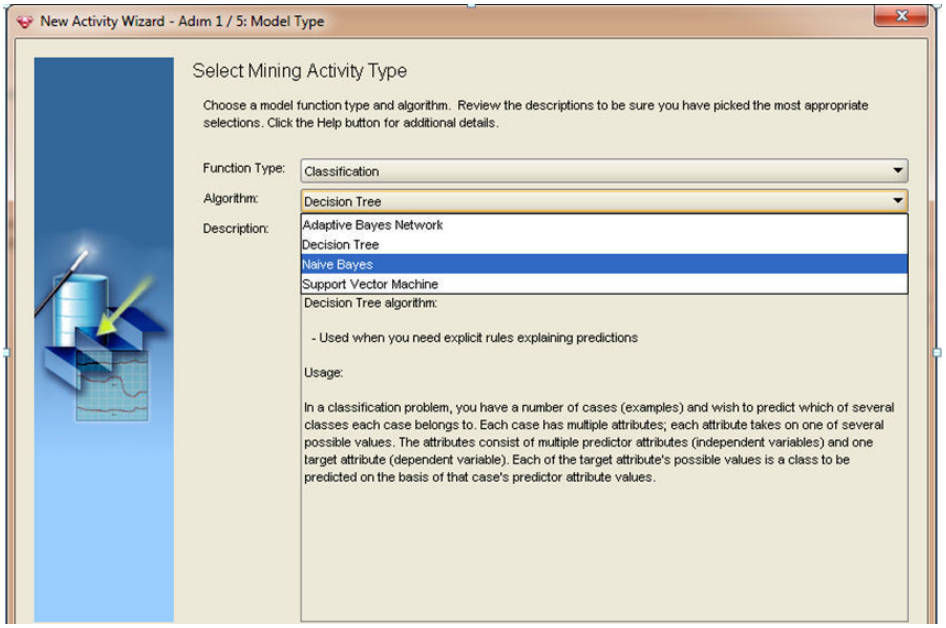


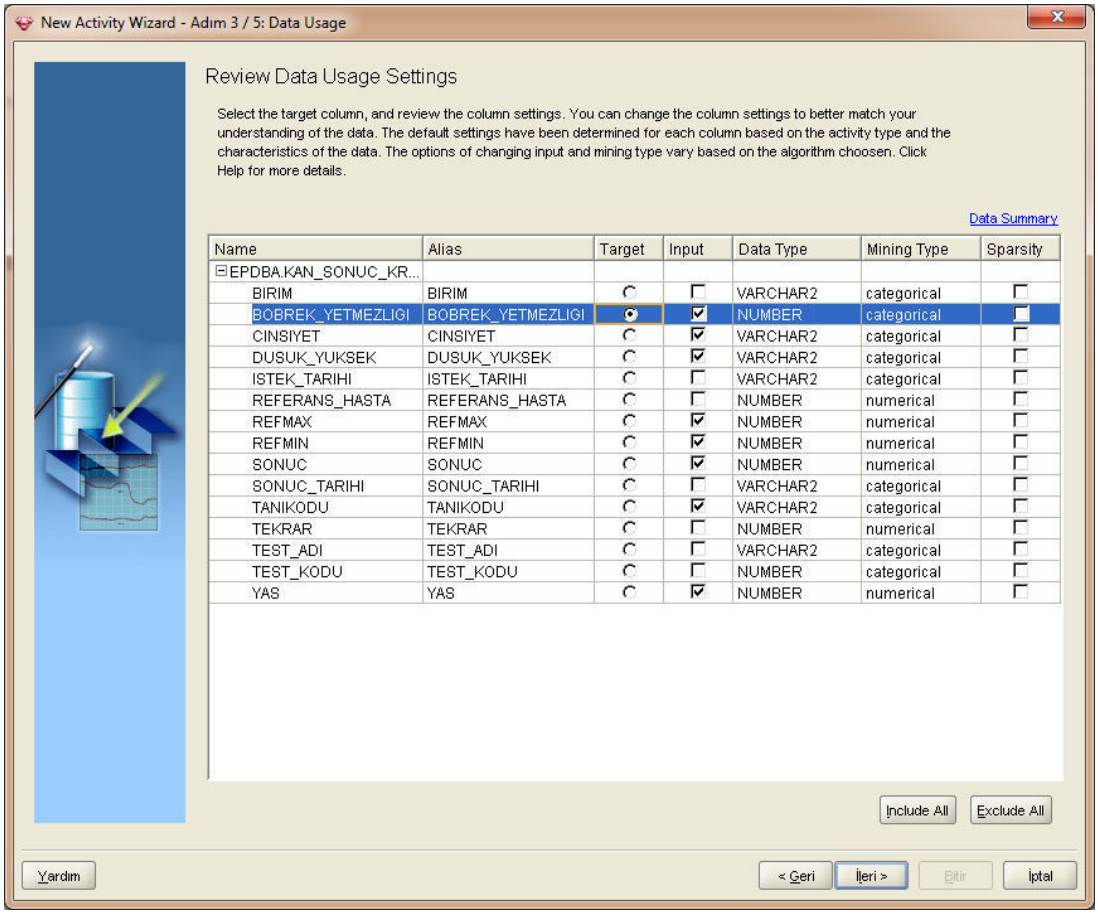
Figure 4.9 Naive Bayes algorithm was chosen to implement model.

First, because of the structure of the application, determinative parameters follows an if-then-else path and this structure is most related with the decision tree algorithm.

Therefore the first option was decision tree algorithm. But the recommended algorithm in the Oracle Data Miner User Guide is Naïve Bayes. Therefore Naïve Bayes algorithm was chosen.

Bobrek_Yetmezligi column was signed as target column that which specifies the result. As well as other decisive parameters were selected: Gender (CINSIYET), reference (DUSUK_YUKSEK), upper limit (REFMAX), lower limit (REFMIN), diagnosis code (TANI_KODU), and age (YAS).

Any different data mining algorithm can be selected from drop down list. For this application decision tree algorithm could be selected. Because of the knowledge extraction follows the if-then-else path. But Naïve Bayes was used.



Name	Alias	Target	Input	Data Type	Mining Type	Sparsity
EPDBA.KAN_SONUC_KR...						
BIRIM	BIRIM			VARCHAR2	categorical	
BOBREK_YETMEZLIGI	BOBREK_YETMEZLIGI			NUMBER	categorical	
CINSIYET	CINSIYET			VARCHAR2	categorical	
DUSUK_YUKSEK	DUSUK_YUKSEK			VARCHAR2	categorical	
ISTEK_TARIHI	ISTEK_TARIHI			VARCHAR2	categorical	
REFERANS_HASTA	REFERANS_HASTA			NUMBER	numerical	
REFMAX	REFMAX			NUMBER	numerical	
REFMIN	REFMIN			NUMBER	numerical	
SONUC	SONUC			NUMBER	numerical	
SONUC_TARIHI	SONUC_TARIHI			VARCHAR2	categorical	
TANIKODU	TANIKODU			VARCHAR2	categorical	
TEKRAR	TEKRAR			NUMBER	numerical	
TEST_ADI	TEST_ADI			VARCHAR2	categorical	
TEST_KODU	TEST_KODU			NUMBER	categorical	
YAS	YAS			NUMBER	numerical	

Figure 4.10 Target table and target attributes are selected.

Preferred target value is selected as 1, which represents the patients under the risk of renal failure. The model implemented depending on the specified target, means that the results are examined the patients which they have 1 at their Bobrek_Yetmezligi column. After that mining activity named as KAN_SONUC_KREATININ_BA.

To accomplish this it is need to add another column to table to specify the model patients under risk of renal failure or not. Bobrek_Yetmezligi column added to KAN_SONUC_KREATININ table using Toad.

Attribute Name	Value	Probability
DUSUK_YUKSEK	N	1,0000000000
TANIKODU	I10	0,6663225806
REFMAX	[,7,1,1]	0,5316129032
REFMIN	[,3,,6]	0,5316129032
CINSIYET	K	0,5294193548
CINSIYET	E	0,4705806452
REFMAX	(1,1,1,3]	0,4683870968
REFMIN	(,6,,7]	0,4683870968
YAS	(57,69]	0,3713548387
YAS	[5,57]	0,3412903226
TANIKODU	E11	0,3336774194
SONUC	(,66,,8]	0,3136774194
YAS	(69,94]	0,2873548387
SONUC	(,92,1,1]	0,2549677419
SONUC	(,8,,92]	0,2441290323
SONUC	(1,1,1,4]	0,0963870968
SONUC	[,2,,66]	0,0908387097

Figure 4.11 Activity completed.

As seen as the images above, if the target class was chosen as 0 it refers the healthy people, it can be calculate correctly depending on DUSUK_YUKSEK column. Oracle data miner statistics says that, depending on the DUSUK_YUKSEK column in KAN_SONUC_KREATININ table it can be found easily healthy people with the probability 1 (certainty) if the target class is chosen as 0.

In the picture below, the target class was assigned as 1 to refer people who under the risk of renal failure. The Oracle data miner tries to produce a model to identify the ill people and after the studies it calculated the depending on DUSUK_YUKSEK column with 0,786219 probabilities.

Also it noticed that TANIKODU column dependency has almost the same probability (0, 711465) to identify. And also it can be said according to results of Oracle data miner, gender is the determining factor too, if the patient is female according to results it can be classify under renal failure risk with 0,532365 probabilities.

Also continue to interpreting results it can be said that Hypertension (E11) patients under more risk according to Diabetic (I10) patients.

Attribute Name	Value	Probability
DUSUK_YUKSEK	Y	0,7862119013
TANIKODU	I10	0,7114658926
REFMAX	[,7,1,1]	0,5325108853
REFMIN	[,3,,6]	0,5325108853
CINSIYET	K	0,5323657475
CINSIYET	E	0,4676342525
REFMAX	(1,1,1,3]	0,4674891147
REFMIN	(,6,,7]	0,4674891147
YAS	(69,94]	0,4098693759
YAS	(57,69]	0,3336719884
SONUC	(2,09,9,84]	0,3291727141
SONUC	(1,4,2,09]	0,2955007257
TANIKODU	E11	0,2885341074
YAS	[5,57]	0,2564586357
DUSUK_YUKSEK	D	0,2137880987
SONUC	[,2,,66]	0,1933236575
SONUC	(1,1,1,4]	0,1612481858
SONUC	(,66,,8]	0,0206095791
SONUC	(,92,1,1]	0,0001451379

Figure 4.12 Target class changed 0 to 1.

Oracle data miner calculates the probability depending on the value of the attributes and generated a table contains attribute name, value and probability columns.

Activity results shows that the main decisive attribute is DUSUK_YUKSEK. Actually this shows that the result is whether range. Activity results also confirmed that being under risk of renal failure directly depend on creatinine level in blood if it is too high or too low than the upper and lower limits there is a problem.

As seen the analysis below if DUSUK_YUKSEK column includes ‘N’, it means the result value is normal, and then probability equals to 1. As because the target class was assigned as 0 (show not under risk), if DUSUK_YUKSEK is normal the probability of not being under risk equals to 1.

The target class can be set as different value after the process. But when the target class was changed the applied process should apply again.

4.2.3 Apply Data Mining Model

First, a completed build activity is selected to be used for apply activity. At the previous section, it was mentioned about how to create a build activity and its phases. Generally apply activity is the testing part of the application.

Remember that at the previous sections, the data prepared for data mining and the related record are retrieved, then the first model data implemented by assigning 1 or 0 to Bobrek_Yetmezligi column of the KAN_SONUC_KREATININ table.

After that this data examined in build activity to implement a model to learn the behavior of the data set and have ability to decide the risk groups for new data sets. The model was implemented and it was ready for the study with new data sets. This part of the study is the testing part, it is assumed that the test data satisfies the previous model rules and generates the same results.

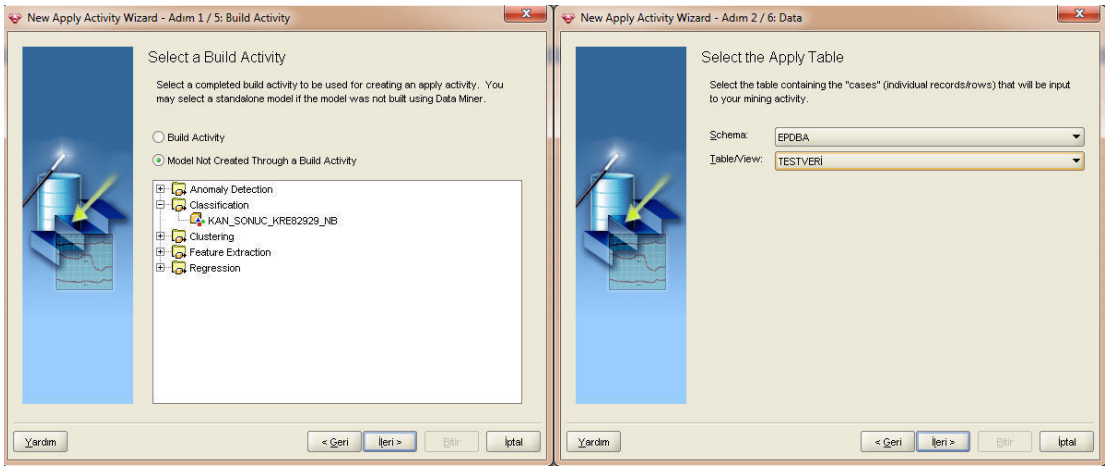


Figure 4.13 Created model applied on test data.

Firstly, a new data set that has not been used before was required. To provide that the same query was written to retrieve another data set. This new data set required for the test process to ensure the system produces the correct results (Figure 4.13).

Apply Output | Apply Settings | Task

Apply Output Table: TESTVERİ788273367_A

Fetch Size: 4000 Refresh

DMR$CASE_ID	YAS	SONUC	TANIKODU	CINSIYET	REFMIN	DUSUK_YUKS...	REFMAX	PREDICTION	PROBABILITY
577	63	0,66	E11	K	0,6	N	1,1	0	0,9999
578	35	0,66	E11	K	0,6	N	1,1	0	0,9999
579	53	0,66	I10	K	0,6	N	1,1	0	0,9999
580	38	0,66	I10	K	0,6	N	1,1	0	0,9999
581	28	0,66	I10	K	0,6	N	1,1	0	0,9999
582	77	0,66	E11	K	0,6	N	1,1	0	0,9999
583	49	0,66	I10	K	0,6	N	1,1	0	0,9999
584	71	0,66	E11	K	0,6	N	1,1	0	0,9999
585	25	0,66	E11	K	0,6	N	1,1	0	0,9999
586	71	0,66	I10	K	0,6	N	1,1	0	0,9999
587	55	0,66	I10	K	0,6	N	1,1	0	0,9999
588	68	0,66	I10	K	0,6	N	1,1	0	0,9999
589	85	0,66	I10	E	0,7	D	1,3	1	0,9999
590	65	0,66	I10	K	0,6	N	1,1	0	0,9999
591	44	0,66	I10	K	0,6	N	1,1	0	0,9999
592	45	0,66	E11	K	0,6	N	1,1	0	0,9999
593	60	0,66	I10	K	0,6	N	1,1	0	0,9999
594	52	0,66	I10	K	0,6	N	1,1	0	0,9999
595	59	0,66	E11	K	0,6	N	1,1	0	0,9999
596	41	0,66	I10	K	0,6	N	1,1	0	0,9999
597	49	0,66	I10	K	0,6	N	1,1	0	0,9999
598	70	0,66	I10	K	0,6	N	1,1	0	0,9999
599	39	0,66	E11	E	0,7	D	1,3	1	0,9999
600	63	0,66	I10	E	0,7	D	1,3	1	0,9999
601	92	0,66	I10	K	0,6	N	1,1	0	0,9999
602	68	0,66	E11	E	0,7	D	1,3	1	0,9999
603	68	0,66	I10	E	0,7	D	1,3	1	0,9999
604	77	0,66	I10	E	0,7	D	1,3	1	0,9999
605	48	0,66	E11	K	0,6	N	1,1	0	0,9999
606	84	0,66	I10	K	0,6	N	1,1	0	0,9999
607	49	0,66	E11	E	0,7	D	1,3	1	0,9999
608	64	0,66	E11	E	0,7	D	1,3	1	0,9999
609	32	0,66	I10	E	0,7	D	1,3	1	0,9999
610	63	0,66	E11	K	0,6	N	1,1	0	0,9999
611	72	0,66	I10	E	0,7	D	1,3	1	0,9999
612	71	0,66	E11	K	0,6	N	1,1	0	0,9999

Figure 4.14 Apply output of the test data.

The same operations have done to process test data and after the operation was completed another result table created by the Oracle data miner. This table can also be accessed from TOAD. Before start to testing the target value selected as 1 because of we are interested in patients that under risk of renal failure.

After the operation completed under the applied output tab the generated table can be seen. This table also includes prediction and probability columns. Prediction column updates depending on the previous model as 1 or 0 and the probability column updates depending on the result accuracy. As seen in Figure 4.14 the probability approaches almost to 1, it means results are almost accurate.

4.3 User Interface Implementation

Oracle data miner tool operates integrated with oracle database. The result tables and result analyzes stored in the oracle database. The result data can be accessed via

Toad for this application. Therefore implementation of user interface part depends on the tables that produced in the Oracle database as result of data mining operation.

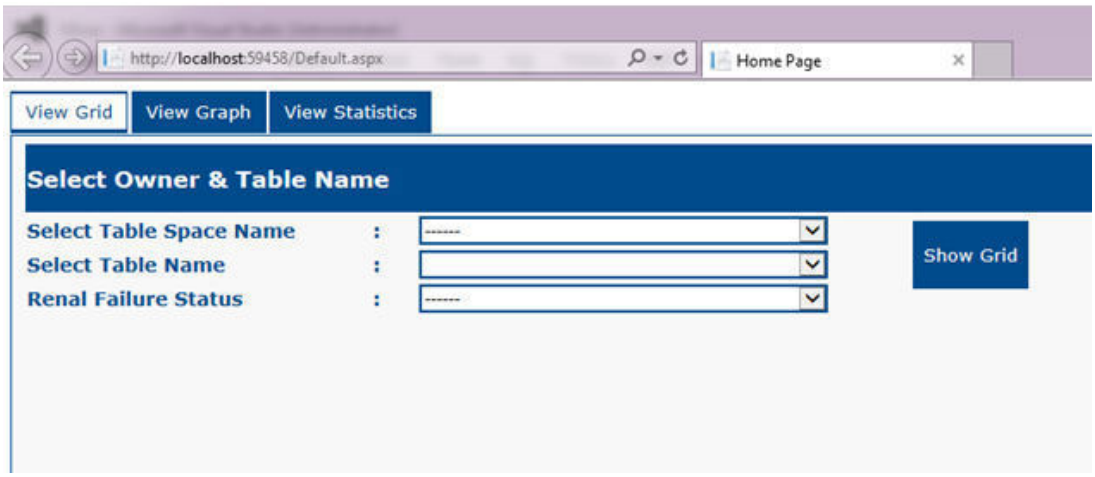


Figure 4.15 Home page of the project.

As seen above in Figure 4.15, the project consists of three sections: View Grid, View Graph, and View Statistics. And each section has its own tasks. That's why it is divided into separate sections.

Home page includes three separate drop down lists. The first drop down list includes the owner names (table space names) in Oracle database. Remember that in the data mining process the tables are created in Oracle database under given owner name (table space) and within given table name. The required owner name can be selected from the first drop down list.

Each lower drop down list's content fills up dynamically depending on the upper drop down list on the Selected_Index_Change event.

These drop down formatted with the help of the System Framework class library class, Helpers.cs. This topic was explained in detail in the used technologies section under the n-tier architecture title.

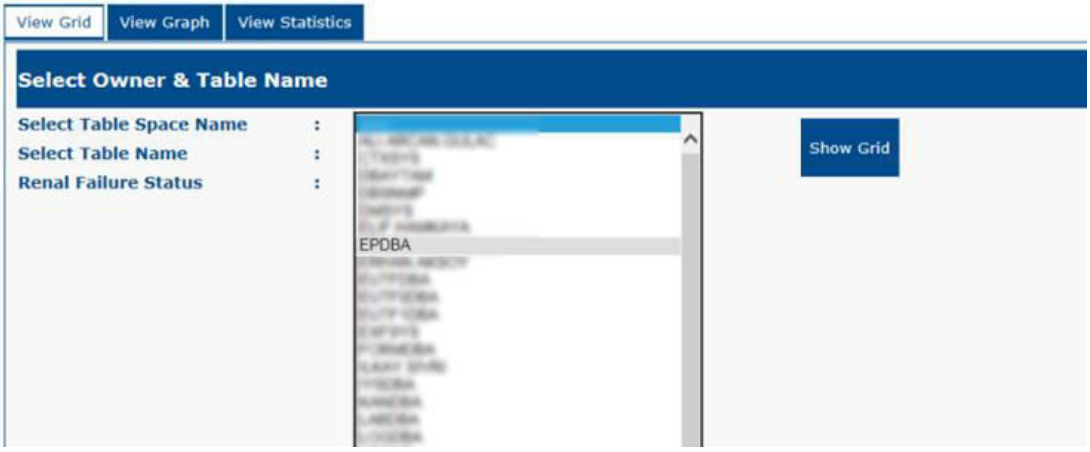


Figure 4.16 The selection process of the required owner.

In this project, tables have been created under the EPDBA table space during data mining process. This is entirely a matter of the designer.

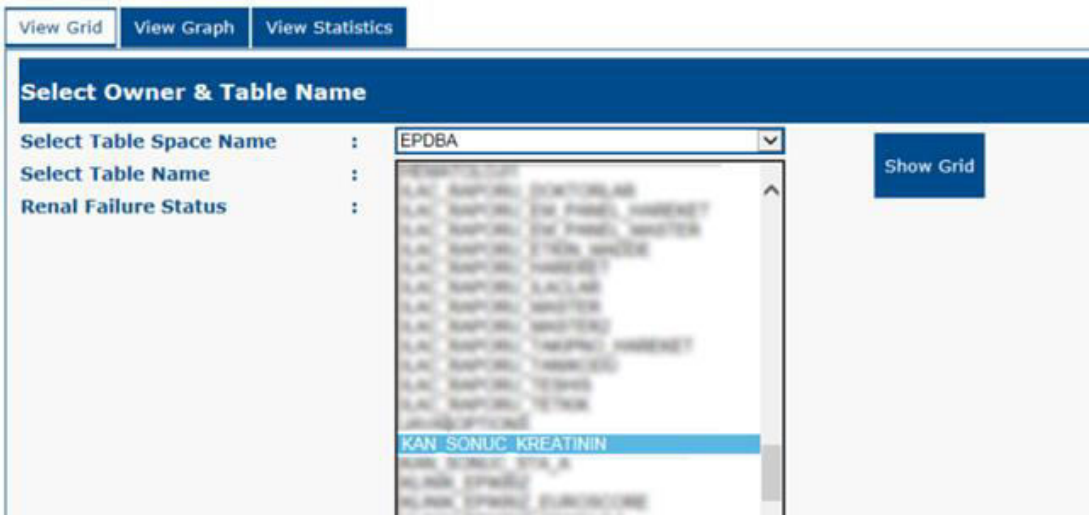


Figure 4.17 Select the KAN_SONUC_KREATININ table under the EPDBA owner.

The second drop down list's content is dynamically generated depending on the selected owner. The table names under the selected owner are listed in the second drop down list.

The third drop down list optional and it is for eliminate the results depending on their Bobrek_Yetmezligi state. Third list include only Yes and No, yes means return

the patients under the risk of renal failure, no means return the patients not under risk. As seen below, if selected value is yes that the system returns the patients with Bobrek_Yetmezligi column is 1, and if the selected value was no than the system returns Bobrek_Yetmezligi column as 0. And the determination process completed.

Renal Failure Status : Yes

1 2 3 4 5 6 7 8 9 10 ...

CINSIYET	REFERANS_HASTA	YAS	TEKRAR	TANIKODU	BOBREK_YETMEZLIGI	ISTEK_TARIHI
E	140352	0	2	I10	1	10.3.14 1:04
E	140352	0	2	I10	1	17.3.14 1:04
K	140239	55	2	I10	1	18.2.14 1:26
K	140239	55	2	E11	1	18.2.14 1:26
K	120197	55	9	I10	1	21.3.14 3:37
K	120197	55	9	I10	1	26.3.14 3:37
K	120197	55	9	I10	1	2.4.14 3:15
K	130455	47	26	I10	1	11.5.14 6:37

Renal Failure Status : No

1 2 3 4 5 6 7 8 9 10 ...

CINSIYET	REFERANS_HASTA	YAS	TEKRAR	TANIKODU	BOBREK_YETMEZLIGI	ISTEK_TARIHI
E	50891	13	3	I10	0	2.4.14 9:10
E	100546	4	17	I10	0	5.5.14 8:27
E	101463	4	3	I10	0	5.5.14 23:45
E	120373	6	3	I10	0	7.3.14 10:49
K	130254	1	23	I10	0	18.3.14 9:35
E	130498	1	24	I10	0	14.2.14 7:24

Figure 4.18 View of result grid according to selected value in third drop down list.

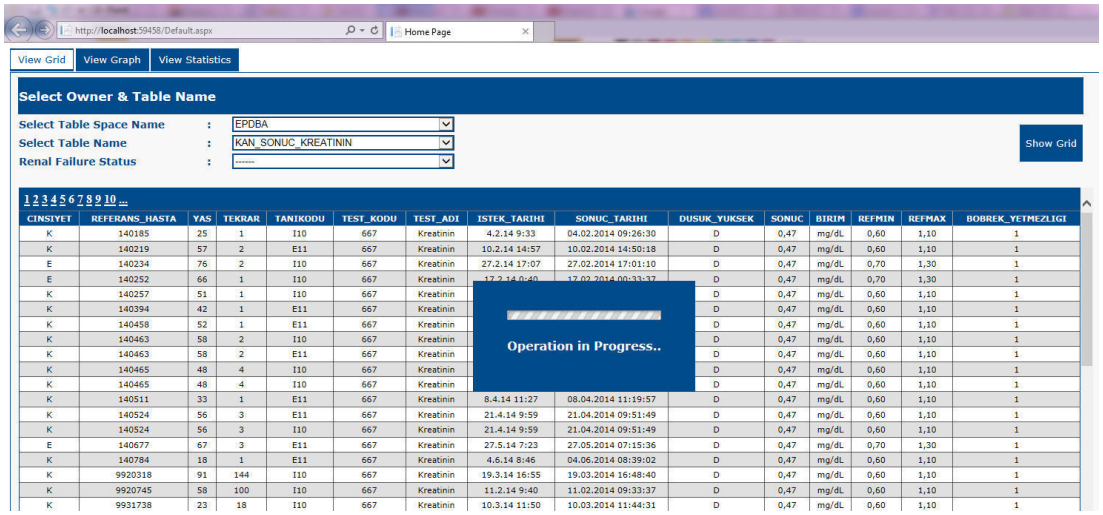


Figure 4.19 View of grid that third drop down list remains in default mode.

Because of there are many records, retrieving patients' result data form database takes long time, it is required a control that show the user that operation in progressed. For this a user control component is created name of UpdateProgress. UpdateProgress operates when the '*Show Grid*' button's event fire.

```
<%@ Control Language="C#" AutoEventWireup="true" CodeBehind="UpdateProgress.ascx.cs" Inherits="MinerBY.UpdateProgress" %>
<%@ Register Assembly="AjaxControlToolkit" Namespace="AjaxControlToolkit" TagPrefix="cc1" %>
<asp:UpdateProgress ID="updateProgress1" runat="server" DisplayAfter="1000">
    <ProgressTemplate>
        <asp:Panel ID="panelProgress" runat="server" Height="1000" Width="100%" >
        </asp:Panel>
        <asp:Panel ID="panelProgressInner" runat="server" Height="150"
            Width="300" style="top: 40%; left: 40%;position: fixed;text-align: center;vertical-align:middle;"
            BackColor="#004B8B">
            <br />
            <br />
            <asp:Image ID="imgBekleyiniz" runat="server" ImageUrl="~/Images/bekleBar.gif" />
            <br />
            <br />
            <asp:Label ID="lblBekleyiniz" runat="server" Text="Operation in Progress.."
                 Font-Names="Verdana" ForeColor="White" Font-Bold="true"></asp:Label>
        </asp:Panel>
        <cc1:AlwaysVisibleControlExtender ID="alwaysProgress" runat="server" TargetControlID="panelProgress">
        </cc1:AlwaysVisibleControlExtender>

    </ProgressTemplate>
</asp:UpdateProgress>
```

TANIKODU	TEST_KODU	TEST_ADI	ISTEK_TARIHI	SONUC_TARIHI	DUSUK_YUKSEK	SONUC	BIRIM	REFMIN
I10	667	Kreatinin	2.4.14 9:10	02.04.2014 09:03:45	N	0,30	mg/dL	0,30
I10	667	Kreatinin	5.5.14 8:27	05.05.2014 08:20:04	N	0,30	mg/dL	0,30
I10	667	Kreatinin	5.5.14 23:45	05.05.2014 23:38:26	N	0,30	mg/dL	0,30
I10	667	Kreatinin	7.3.14 10:49	07.03.2014 10:43:12	N	0,30	mg/dL	0,30
I10	667	Kreatinin			N	0,30	mg/dL	0,30
I10	667	Kreatinin			N	0,30	mg/dL	0,30
E11	667	Kreatinin			N	0,30	mg/dL	0,30
I10	667	Kreatinin			N	0,30	mg/dL	0,30
I10	667	Kreatinin			N	0,30	mg/dL	0,30
I10	667	Kreatinin			N	0,30	mg/dL	0,30
I10	667	Kreatinin			N	0,32	mg/dL	0,30

Operation in Progress..

Figure 4.20 Codes and the view of UpdateProgress.

Also in the first tab View Grid, one of the important operations is to viewing the test data results. By using Oracle data miner tool after the data mining operations the data mining model applied on test data. The applied process determine the each patient under risk of renal failure or healthy. Creating model process was discussed in the previous section. After the model applied on test data the result table named by user and I named it TEST_VERI_PROB_A. "PROB" tag indicates the table create by the result of test operation. "A" tag indicates that this is an applied operation.

View Grid | View Graph | View Statistics

Select Owner & Table Name

Select Table Space Name : EPDBA

Select Table Name : TESTVERİ_PROB_A

Renal Failure Status : ------

Show Grid

1 2 3

PREDICTION	PROBABILITY	CINSIYET	REFMIN	DUSUK_YUKSEK
1	0,999994993209838870	K	0,60	D
1	0,999993860721588130	K	0,60	D
1	0,999994993209838870	K	0,60	D
1	0,999993860721588130	K	0,60	D
1	0,999993860721588130	K	0,60	D
1	0,999993860721588130	K	0,60	D
1	0,999994993209838870	K	0,60	D
1	0,999994993209838870	K	0,60	D
1	0,999993860721588130	K	0,60	D
1	0,999994993209838870	K	0,60	D
1	0,999994993209838870	K	0,60	D
1	0,999993860721588130	K	0,60	D
1	0,999993860721588130	K	0,60	D
1	0,999994993209838870	K	0,60	D
1	0,999994993209838870	K	0,60	D
1	0,999994993209838870	K	0,60	D
1	0,999994993209838870	K	0,60	D

Figure 4.21 Representation of the TEST_VERI_PROB_A table.

This table in Figure 4.21 includes '*Prediction*' and '*Probability'* columns. During the data mining process, the '*Prediction*' column of the test data updated depending on the previously created model. System has learned the model on previous data than applied the learned model on the test data.

After doing the calculations it updates the '*Probability*' column what percentage of accurate calculation. Generally, browsing on the '*Probability'* column the values almost approach to 1 and 1 represents accuracy in probability concept.

Another important subject in this study is to observe the distribution of the results. To make easier to read of distribution for users visuals are required. The graphics are one of the visual components in software. To use the graphics in software you should have already included the specific library in your project. In this project ASP Charting component is used provide visual.

```
<asp:Chart ID="ChartResult" runat="server" Width="500px" Height="500px">
    <Series>
        <asp:Series Name="SeriesResult" ChartType="Column" ChartArea="ChartAreaResult"></asp:Series>
        <asp:Series Name="SeriesResult2" ChartType="Spline" ChartArea="ChartAreaResult" BorderWidth="2">
        </asp:Series>
        <asp:Series Name="SeriesResult3" ChartType="Point" BorderWidth="3" ShadowOffset="1" Color="#000000"
             MarkerStyle="Diamond" MarkerSize="10" ChartArea="ChartAreaResult"
            IsValueShownAsLabel="true" IsVisibleInLegend="false">
        </asp:Series>
    </Series>
    <ChartAreas>
        <asp:ChartArea Name="ChartAreaResult">
            <AxisY TitleForeColor="#004B8B">
                <StripLines>
                    <asp:StripLine TextAlignment="Near" BorderDashStyle="Solid" Interval="0" BorderColor="Red"
                        ForeColor="Red" BorderWidth="2" />
                    <asp:StripLine TextAlignment="Near" BorderDashStyle="Solid" Interval="0" BorderColor="Blue"
                        ForeColor="Blue" BorderWidth="2" />
                </StripLines>
            </AxisY>
        </asp:ChartArea>
    </ChartAreas>
double temp = 0.0;
for (int i = 0; i < 120; i = i + 10)
{
    DataRow[] arrDr = dtGraph.Select(Convert.ToInt32(i) + " < YAS AND YAS < " + Convert.ToInt32(i + 10));

    for (int j = 0; j < arrDr.Length; j++)
    {
        temp += Convert.ToDouble(((DataRow)arrDr[j]).ItemArray[10]);
    }
    temp = Convert.ToDouble(temp / arrDr.Length);

    ChartResult.Series["SeriesResult"].Points.AddXY(i + 5, temp);
    ChartResult.Series["SeriesResult2"].Points.AddXY(i + 5, temp);
    ChartResult.Series["SeriesResult3"].Points.AddXY(i + 5,temp);
    ChartResult.Series["SeriesResult3"].LabelFormat = "0.##";
    temp = 0.0;
}
ChartResult.ChartAreas["ChartAreaResult"].AxisY.StripLines[0].IntervalOffset = max;
ChartResult.ChartAreas["ChartAreaResult"].AxisY.StripLines[0].Text = "Max = " + max.ToString();
```

Figure 4.22 Codes of ASP.Net Charting component.

In the second tab of the application is View Graph, and it provides user to viewing statistical graphs about the results. The codes of the ASP.Net charting are given below. This screen offers some parameters such as Gender (Female, Male), ICD Code (I10 and E11) and health state (healthy or under risk) and generates statistical graphs depending on these parameters.

The graphs in the View Graph tab, filter the results according to selected parameters, and draw the distribution graph of results according to age range. Also, the reference range is located on the graph.

Sometimes, it is very difficult to make sense or understand something just looking at the numbers. So, these graphs were designed to solve this. It is possible to make an inference watching the falling and rising places in the chart.

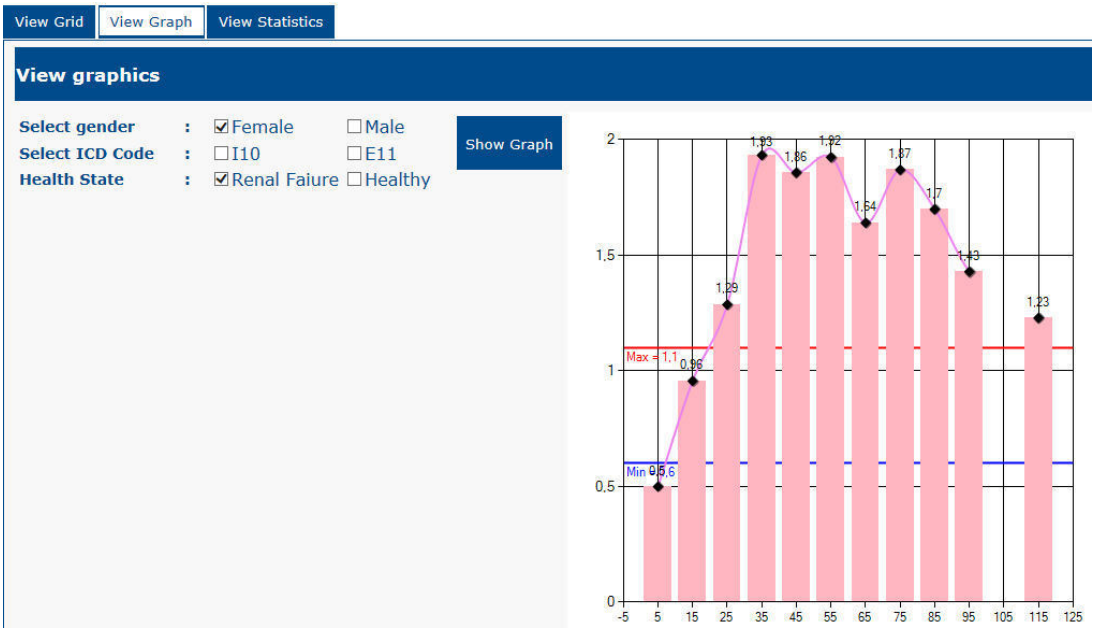


Figure 4.23 Graph represents the female and under risk patients.

The red line in the graph represents the maximum must value of the creatinine and the blue line represents the minimum must value of the creatinine. Because of the selected cluster includes the patients under risk the graphs bars are higher than the maximum must value or more less than the minimum must value.

And we can say that in the 35 age range has the highest creatinine value and equals to 1.92, and the 5 age range includes the minimum creatinine value 0.63 and it's a normal value.

The normal acceptable values are resides between blue and the red line. If the patients' value resides in the blue and red line it can be said that not under risk. Otherwise if the result is under the blue line or upper to red line it can be said that under risk of renal failure.

Female and under failure risk group were selected from the combo boxes from the left side the values exceed the must maximum value.

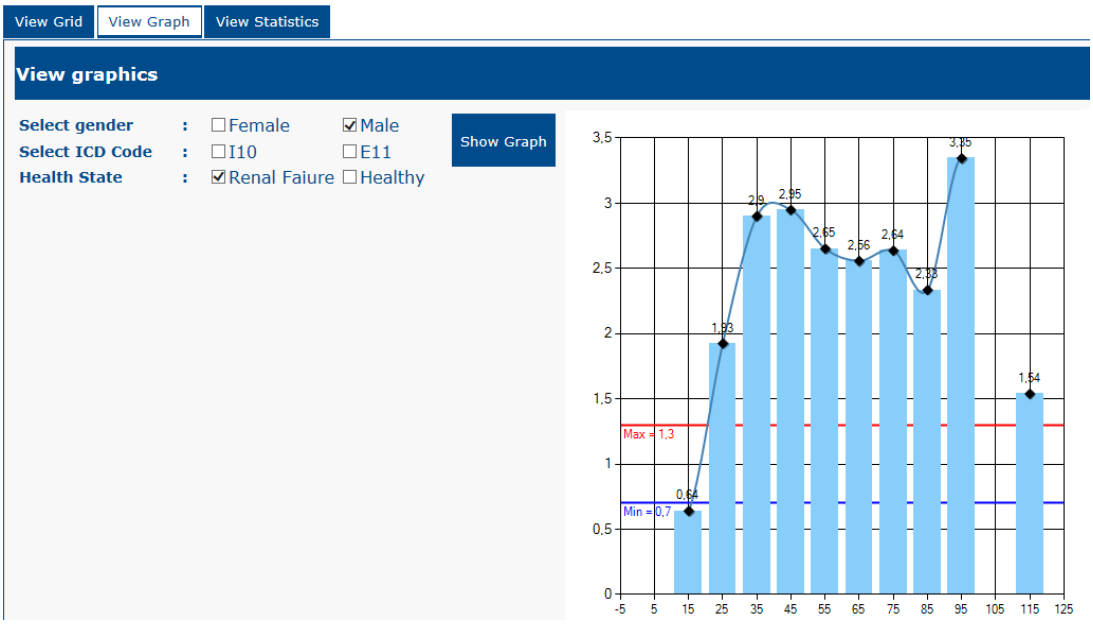


Figure 4.24 Graph represents the male and under risk patients.

If we compare two graphs, it can be said that the gender is the significant factor that identifying the risk of renal failure. Male results have higher values. As seen above the highest value of creatinine in the 95 age range and equals to 3.35. It is really a high value because the maximum value for male is 1.30. As seen in the Figure 4.24 for males the bars higher and they almost reach to 3 and 4.

It is quite difficult process to make out this analysis on the grid in first tab. But with the help of the graphs, it is easy to identify and determine which age range is under risk of renal failure and what the maximum and minimum values are reached. Also this screen offers opportunity to user to observe creatinine level in differential diagnosis codes separately.

Because of the Renal Failure check box was checked as seen in Figure 4.24 the creatinine level of the patients reaches extremely high values. It can be said that almost there is no patient who has normal creatinine level in his/her blood. If the E11 check box was check as seen in Figure 4.25 the creatinine level reaches almost 4.5 values. It shows the effect of diabetic disorder on creatinine level. To make the interfaces more attractive graphs are colored depending on the gender. To operation was done only to improve the visual appearance.

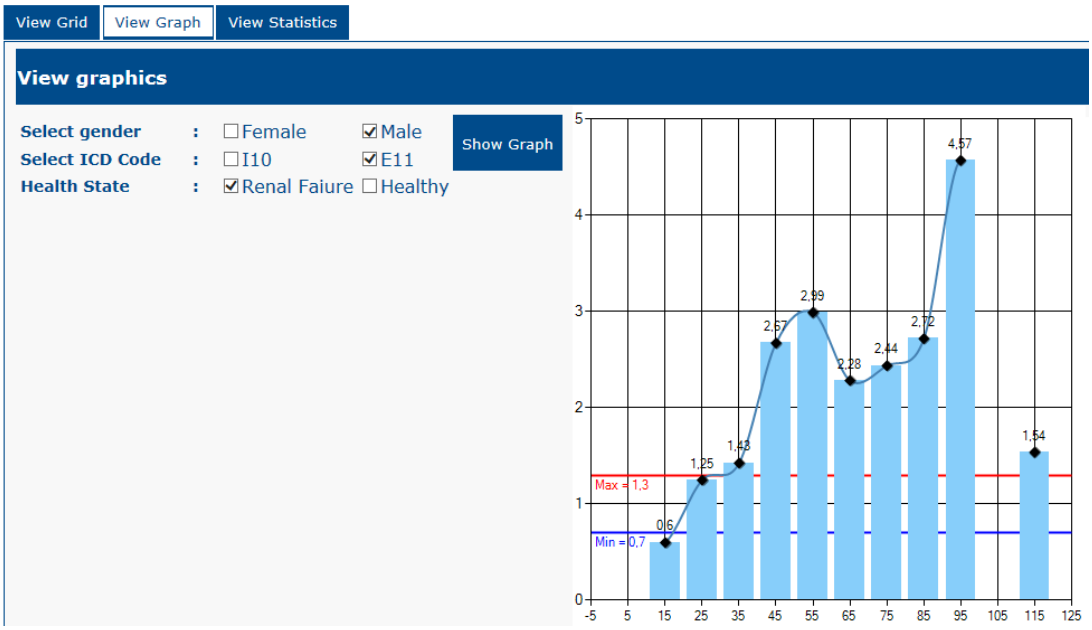


Figure 4.25Graph represents the under risk males and the diagnosed diabetes.

It is the part of the application does not concern the technical operation. When the female and the male check boxes checked together it means that return the results that belong to all patients not depending on gender. In this state the bar color is assigned as different value.

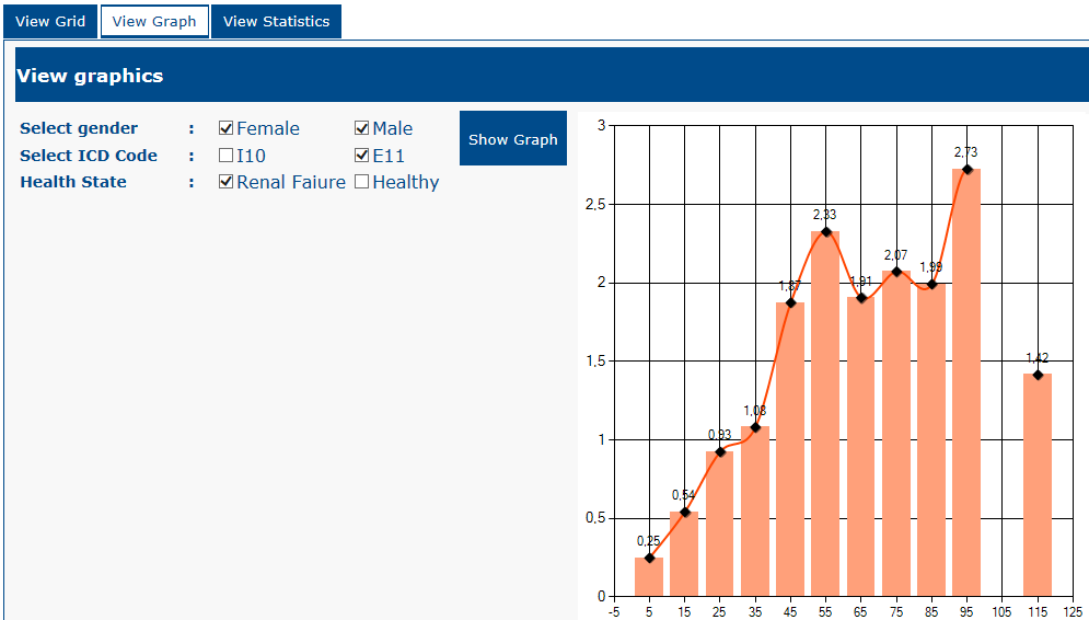


Figure 4.26 Graph represents the results for both female and male patients under risk of renal failure.

It can be generated more different graphs by marking the different combination of the checkboxes on this screen. The graph updates itself dynamically when the '*Show Graph*' button's event fire. The creatinine maximum must and minimum must value depend on the gender, therefore there is not pre-defined general maximum and minimum values for the all gender so there is no blue line and no red line on the orange graph.

Finally, the third graph includes some statistic data about the results. These operations can be conducted on more than one data table; one data table must be selected before each operation.

The name of the tables that generated during data mining process names depends of the designer of the system, so I named the table which includes the statistic data about the result as KAN_SONUC_STA_A. "STA" tag indicates the table is a statistical data table and "A" tag indicated that the table generated after applied operation.

View Grid | View Graph | View Statistics

View Results & Statistics

Select Statistic Table : KAN_SONUC_STA_A View Statistics

ATTRIBUTENAME	VALUE	PROBABILITY
DUSUK_YUKSEK	N	1,000000
TANIKODU	I10	0,666322
REFMAX	[0,7-1,1]	0,531512
REFMIN	[0,3-0,6]	0,531512
CINSIYET	K	0,529419
CINSIYET	E	0,470580
YAS	(57-69]	0,371354
YAS	[5-57]	0,341290
TANIKODU	E11	0,333677
SONUC	(0,66-0,8]	0,313677
YAS	(69-94]	0,287357
SONUC	(0,92-1,1]	0,254967
SONUC	(0,8-0,92]	0,244129
SONUC	(1,1-1,4]	0,096387
SONUC	[0,2-0,66]	0,090838

The number of Hypertension patients are 16989 and the number of Diabetic patients are 7743.
The activity applied on 13099 number of female patients and 11633 number of male patients. Total record count is 24732.
According to test results 11534 number of patients under the risk of Renal Failure and 13198 number of patients not under risk.

Figure 4.27 Representation of the statistical table generated after applied operation.

It can be inferred from this statistical table: the primary determinant parameter is DUSUK_YUKSEK attribute which indicates the level of the creatinine. If the creatinine level more than the maximum must value DUSUK_YUKSEK attribute value is “Y”. If the creatinine level less than the minimum must value DUSUK_YUKSEK attribute is “D” and if creatinine level is the acceptable range if it is between the minimum must and maximum must level than the DUSUK_YUKSEK attribute is “N”. DUSUK_YUKSEK attribute is N it can be said that the patient not under risk.

It can be also inferred from the statistical table TANIKODU attribute has the second priority identifying the risk of renal failure. If TANIKODU attribute equals to I10 (hypertension) than the prediction assigned as 1 with probability 0.666.

The third tab also includes other statistics such as the number of record and the number of female or male number of the hypertension patient and number of diabetic patients.

The table is shown above illustrates that the important attributes order by their importance to identify the renal failure disease. As seen in the table, gender attribute has importance degree of 4.

And it can be said that according to state of the gender if patient is female the probability of being health higher than the under risk.

Another table which generated as result of the data mining operation is TESTVERI_EXPLAIN_A table which can be seen in Figure 4.28 . During the data mining operation Oracle data miner tool analyzed each attribute’s importance on identifying risk of renal failure.

View Grid | View Graph | View Statistics

Select Owner & Table Name

Select Table Space Name : EPDBA

Select Table Name : TESTVERİ_EXPLAIN_A

Renal Failure Status : ------

Show Grid

ATTRIBUTE_NAME	EXPLANATORY_VALUE	RANK
DUSUK_YUKSEK	0,9930584166388830	1
SONUC	0,74562395224855	2
BIRIM	0	3
REFMIN	0	3
ISTEK_TARIHI	0	3
TANIKODU	0	3
CINSIYET	0	3
TEKRAR	0	3
REFERANS_HASTA	0	3
YAS	0	3
SONUC_TARIHI	0	3
TEST_KODU	0	3
TEST_ADI	0	3
REFMAX	0	3

Figure 4.28 Representation of TESTVERI_EXPLAIN_A table.

In the table which is shown above, the attributes ordered by the importance of the identifying renal failure. The RANK attribute shows that the attribute's importance. As seen in the table BIRIM, REFMIN, TEST_KODU_TEST_ADI, REFMAX attributes are not important in identifying process. And it can be said that the most determinative attribute is DUSUK_YUKSEK and its determinative probability approaches to 1. TESTVERI_EXPLAIN_A table can be accessed from the first tab view grid.

As a result of this process, two attributes came forward: DUSUK_YUKSEK and SONUC attributes. The most explanatory attribute was DUSUK_YUKSEK and the second explanatory attribute was SONUC. Explanatory value of DUSUK_YUKSEK equals to 0.993 and the explanatory value of SONUC equals to 0.745.

CHAPTER FIVE
CONCLUSION AND DISCUSSION

Nowadays, increasing amount of data in hospitals leads to increase need for a new tool to interpret and process this huge data. Because of the computer hardware became cheaper and more accessible, computer-based solutions have gained momentum. Within the computer became widely used in hospitals the transactions made on paper previously moved to electronic environment. Considering fallibility of the human eye, an assistant tool is required to improve decision process.

Healthcare information systems have vital importance for a well-organized healthcare system. Information technologies have a big impact on medical industry. Medical industries have to evaluate data in decision-making process to use more conscious way, have to expand service area to increase productivity. And they have to get benefit from information technologies to reduce costs, improve customer service and better management of resources.

At this point a software becomes required which have been discussed in this thesis. In this thesis data mining techniques applied on the biochemistry laboratory results of the diabetics and hypertension patients using Naïve Bayes algorithm with Oracle data miner tool. Biochemistry laboratory results retrieve via Web Services. Bayesian classification was implemented in the Oracle data miner this pre-defined version is used in the system.

Application developed on Visual Studio 2012 platform using C# and Asp.Net components. The application analyzes the data obtained from laboratory results and creates a model. After that applied the model on test data and controls that whether system has produces accurate result or not.

This application has shown the important attributes to make decision, increased the accuracy of decisions, and supported the health professionals' decision making process.

REFERENCES

Albayrak, S. (n.d.). *Veri madenciliği sınıflama ve kümeleme yöntemleri.* Retrieved May 17, 2014 from https://www.google.com.tr/url?sa=t&source=VeriMadenciliği

CRC, (n.d.). Preparing Data for data mining. Retrieved April 8, 2014 from http://www.cabnr.unr.edu/gf/dm/chap02.pdf

Edelstein, H.A., (1999). *Introduction to data mining and knowledge discovery* (3rd ed.). Potomac: Two Crows Corporation.

Fayyad, U. M. (2003). From data mining to knowledge discovery in databases. *AI Magazine*, *17* (3):37-54.

Freitas, A.A. (2009). *A survey of evolutionary algorithms for data mining and knowledge discovery* (1st ed.). Verlag NY: Springer.

Gültekin, S. (2010). *Kronik böbrek hastalığı ve tedavisi.* Retrieved March 14, 2014, from http://www.medikalakademi.com.tr/kronik-boebrekhastal-tan-ve-tedavisi/#!

Han, J., & Kamber, M. (2006). *Data mining concepts and techniques* (2nd Ed.). San Francisco: Darthmouth Publishing.

Kearns, A. (2010). *5-Layer Architecture.* Retrieved June 10, 2014, from http://morphological.files.wordpress.com/2011/08/5-layer-architecture-draft.pdf

Koyuncugil, A.S., & Özgülbaş, N. (2010). Data mining: Tıp ve sağlık hizmetlerinde kullanımı ve uygulamaları. *Bilişim Teknolojileri Dergisi, 2(2).*

Kudyba, S. (2004). Managing data mining. *CyberTech Publishing*, 3(46).

Lano, K., & Haughton, H. (1993). Integrating formal and structured methods in reverse-engineering. *Proceedings Working Conference on Reverse Engineering*(79–95).

Maimon, O., & Rokack, L. (2005). *The data mining and knowledge discovery handbook* (1315–1329). Israel: Springer US.

Olson, D.L., & Delen, D. (2008). *Advanced data mining techniques* (1st ed.). NY: Springer.

Oracle data mining concepts 10g release 2, (2005). Retrieved June 5, 2014, from http://docs.oracle.com/html/B14339_01/1intro.htm

Oracle experience, (n.d.). Retrieved May 18, 2014, from http://ora-exp.blogspot.com.tr/2007/02/veri-madenciligidata-mining-nedir-ve_24.html

Oracle, (2014). *Oracle data mining concepts.* Retrieved March 20, 2014, from http://docs.oracle.com/cd/E16655_01/datamine.121/e17692.pdf

Özbek, F., İnce, M., Turhan, M., & Önder, H. H. (n.d.). *E- Üniversite için esnek bir framework geliştirilmesi ve uygulanması.* Retrieved June 7, 2014 from http://ab.org.tr/ab14/bildiri/33.pdf

Ralph, P. & Wand, Y. (2009) *A proposal for a formal definition of the design concept* (103-136). Verlag: Springer.

Thomas, E. (2005). *Service-oriented architecture: concepts, technology, and design.* Upper Saddle River: Prentice Hall PTR

Weiss, G. M., & Davison, B.D. (2010). *Data mining.* Retrieved April 11, 2014, from http://storm.cis.fordham.edu/~gweiss/papers/data-mining-chapter-2010.pdf

What is data mining?, (n.d.). Retrieved May 5, 2014, from http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm#CHDJCFAG

Yıldırım, P., Uludağ, M., Görür A. (2008). *Hastane bilgi sistemlerinde veri madenciliği.* Retrieved May, 2014, from http://ab.org.tr/ab08/kitap/Bildiriler/Yildirim_Uludag_Gorur_AB08.pdf

Zaki, M.J., Xu Yu, J.,Ravindran, B., & Pudi, V. (2011). *Advances in knowledge discovery and data mining* (1st ed.). NY: Springer.