SAKARYA UNIVERSITY
INSTITUTE OF SCIENCE AND TECHNOLOGY

IMPLEMENTATION OF SOME MEDICAL DATA USING APRIORI ALGORITHM

M.Sc. THESIS

Fawad SADIQMAL

Department : COMPUTER AND INFORMATION ENGINEERING

Supervisor : Assist. Prof. Dr. Nilüfer YURTAY

June 2015

SAKARYA UNIVERSITY
INSTITUTE OF SCIENCE AND TECHNOLOGY

IMPLEMENTATION OF SOME MEDICAL DATA USING APRIORI ALGORITHM

M.Sc. THESIS

Fawad SADIQMAL

Department : **COMPUTER AND INFORMATION ENGINEERING**

Supervisor : **Assist. Prof. Dr. Nilüfer YURTAY**

This thesis has been accepted unanimously / with majority of votes by the examination committee on 16.01.2015

Assist.Prof.Dr. Nilüfer YURTAY	**Prof.Dr. Cemil ÖZ**	**Assist.Prof. Dr. Burhanettin DURMUŞ**
Head of Jury	**Jury Member**	**Jury Member**

DECLERATION

I declare that all the data in this thesis was obtained by myself in academic rules, all visual and written information and results were presented in accordance with academic and ethical rules, there is no distortion in the presented data, in case of utilizing other people's works they were refereed properly to scientific norms, the data presented in this thesis has not been used in any other thesis in this university or in any other university.

Fawad SADIQMAL

25.06.2015

PREFACE

First and foremost, a heartfelt gratitude and thank to my respected advisor (Assist. Prof.Dr. Nilüfer Yurtay) for her ample support and invaluable guidance throughout this thesis. Without the patience and support of my honorable and respected advisor this thesis was going to be simply impossible. It was hard to survive the tough times of this thesis without hearing my respected advisor words: “She used to say, no problem, you should easily solve that using this and that”. So, dear and respected Hocam “My Advisor” Your cooperation and help is never forgettable and it is highly appreciated forever.

I would also like to thank Prof.Dr. Ibrahim Çil. His spirit is strong enough to vitalize the whole Industrial Engineering Department. I took a lesson of Decision Support System from him. He is a wonderful, kind and cooperative teacher he is the one who advices me to select Nilüfer Yurtay as my advisor and I did so, Sir I really thank you for all your good advices all the time.

I would like to extend my gratitude to Assoc Prof.Dr. Celal Çeken, Assist. Prof.Dr. Seçkin Arı and Prof.Dr. Cemil Öz. I took lessons from the first two of them. Meanwhile these teachers and all my other teachers always support me and because of Allah’s blessings and their support Alhamdulillah. I come to this level. Beside of being my respected teachers they were and are always kind, cooperative and close friends to me. I would also like to take this opportunity to extend my gratitude to all the staffs of computer and information engineering department for all their unlimited cooperation and help. I am always grateful to Almighty Allah (swt), and my parents, my wife for their love and support, always encouraging me to strive for the best. Last but not the least, the big thanks goes to all my dear teachers, relatives and friends for their always unlimited support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

SUMMARY

Keywords: Medical Data Mining, Association Rules, Apriori Algorithm.

Modern medicine generates a great deal of information stored in the medical database. Extracting useful knowledge and providing scientific decision-making for the diagnosis and treatment of disease from the database increasingly becomes necessary. Data mining in medicine can deal with this problem [1].

This thesis work is based on medical data mining. I collected data from medical papers and journals, and looked for around 6000 to 7000 papers and journals then I selected 1000 papers only, and discarded the rest which were not needed or related to my thesis work. The selected papers which I used are from years 2010 to 2015. My data is, ***name or title of the paper, keywords and authors***, the focusing factor from the above data is keywords. I did an implementation on the keywords of the data. the target of the implementation is finding different relationships among these keywords. The searching keywords which I used for collecting the data are: medical data mining , medical clustering , medical classification, medical decision support system, and medical papers in fuzzy system and artificial neural network. I collected all the data manually. After I organize the data and performed the implementation using Apriori algorithm. In the result I found two things from the keywords**.** First one the most occurring words (with number of occurring), second association rules among those words.

APRIORI ALGORİTMASININ BAZI TIBBI VERİLERE UYGULANMASI

ÖZET

Anahtar kelimeler: Tıbbı Veri Madenciliği, Birliktelik Kuralları, Apriori Algoritması

Tıp alanında yapılan hastalara ait teşhis ve tedavi kayıtlarının, bilgisayar programları tarafından analiz edilmesi ve raporlanması etkili tedavi sürecini destekleyici bir unsur oluşturmaktadır. Birçok alanda olduğu gibi, tıp alanında da veri madenciliği yöntemlerinin kullanımı hızla artmaktadır. Bu yüzden tıbbi veri madenciliği başlı başına bir yöntem haline gelmiştir. Veri madenciliği yöntemlerinin kullanıldığı tıbbi karar destek sistemine yardımcı verilerin elde edilmesi ile hekimlere karar vermede yardımda bulunacak sistemin geliştirilmesi bu tez çalışmasında gösterilmiştir.

Amaç:
Modern tıp bilgileri, tıbbi veritabanında saklanan bilgilerin büyük bir kısmını oluşturmaktadır. Bu yüzden tanı ve hastalığın tedavisi için bilimsel karar vermede tıbbi veritabanından yararlı bilgilerin ayıklanması gerekli hale gelmektedir. Tıbbi veri madenciliği ile bu sorunları giderilebilir, aynı zamanda hastane bilgi yönetim düzeyini geliştirebilir ve toplumu tıp gelişimine teşvik edebilir.

Kapsam:
Bu tez çalışmasının ana konusu tıbbı veri madenciliğidir. Veriler tıbbı makale ve yayınlardan toplanmıştır. Literatürde 6000 ile 7000 arasında makale ve yayınlar incelenmiş ve bu tez çalışması için bunlardan 1000 tanesi uygulamada kullanılmıştır. İncelenen makale ve yayınlardan bu çalışma ile ilgili olmayanlar elenmiştir.

Seçilen makale ve yayınlar 2010 ile 2015 senelerinde yazılmış makalelerdir. Veri girişi olarak makalenin yazarı, makalenin ismi ve makalede geçen anahtar kelimeler kullanılmıştır. Bunlar arasından odaklanılan faktör anahtar kelimelerdir. Uygulamanın hedefi anahtar kelimeler arasındaki farklı ilişkileri bulmaktır.

Verileri toplamak için kullanılan kelimeler şunlardır; tıbbi veri madenciliği, tıbbi kümelenme, tıbbi sınıflandırma, tıbbi karar destek sistemi, bulanık sistemde ve yapay sinir ağlarda tıbbi yayınlar.

Yöntem:
Tıbbi veri madenciliği, tıbbi verilerin, farklı desen ve kaynaklardan, hızlı ve sağlam sonuçları güvenilir bir şekilde bize sunan işlem ve tekniklerden oluşmaktadır. Bu teknik ve yöntemler; bir tür yapay sinir ağı, bulanık sistem, karar destek sistemi, evrimsel algoritmalar, destek vektör makinesi gibi hesaplamalara dayalı uygulamalardır. Tıbbi veritabanlarında, hastalar ve tedavileri hakkında bilgiler büyük miktarlarda birikmiştir. Bu veriler içerisindeki ilişkiler ve desenler ile yeni tıbbi bilgiler sağlanabilir. Yeni bilgilerin üretilmesi için birkaç metodolojiler geliştirmiş ve bu gizli bilgileri keşfetmek için uygulanmıştır. Veri madenciliği teknikleri geniş bir tıbbi veritabanında veriler arasındaki ilişkileri aramak için kullanılmıştır. Tüm bu araştırmalar tıbbi veriler üzerine yapılmıştır çünkü tıp alanında birçok araştırma ve çalışmaların olması, sürekli büyüyen ve gelişen bir yapıda olması ve aynı zamanda sağlıklı yaşamanın hayatımızın en önemli kısmında yer almasıdır.

Tıbbi veri madenciliğinin önemini Stanford Üniversitesi araştırmacıları tarafından yapılan bir çalışma ile şöyle gösterebiliriz. Bu çalışma 19 Eylül 2011 tarihinde yapmıştır. Çalışmada; iki farklı tedavi için kullanılan ilaçların birlikte kullanılması halinde ortaya çıkabilecek yan etkilerin saptanmasında veri madenciliğinin kullanılması ve çalışmaya olan olumlu etkisi gösterilmiştir. Birinci ilaç antidepresan için kullanılan Paxil adlı ilaçtır ve 1992 yılında tedavi için kullanılmıştır. İkinci ilaç ise kolesterol düşürücü için kullanılan Pravachol adlı ilaçtır ve 1996 yılında kullanılmıştır. Tedavi süreçlerinde her ilaç kendi tedavisine olumlu cevap vermiştir. İlaçların birlikte alındığı süreçlerde ise oluşabilecek yan etkilerin önceden saptanabilmesi için veri madenciliği uygulanmıştır. Öncelikle tedavi sürecindeki

onbinlerce hastadan bilgileri alınarak elektronik hasta veritabanı oluşturulmuştur. Oluşturulan bu veritabanına veri madenciliği uygulanmıştır.Uygulama sonucunda her iki ilacı kullanan insanlarda yüksek kan şekeri seviyelerinin olacağı gözlemlenmiştir. İlaç etkileşimleri ile oluşabilecek yan etkileri yukarıdaki çalışmada da görüleceği gibi tıbbi veri madenciliğinin son çalışmaları ve araştırmaları ile bulunabilmektedir. Buradanda görüleceği gibi tıp alanındaki son gelişmelerin önemli parçalarından birisi de tıbbi veri madenciliğidir. Bu tıbbi veri madenciliği çalışmasında RapidMiner yazılımı kullanılmış ve Apriori Algoritması bu yazılımda çalıştırılmıştır. Apriori Algoritması' nın özellikleri ise şu şekildedir:

Bu algoritma birliktelik kurallarının oluşturulmasında yararlanılan ve yaygın olarak kullanılan bir algoritmadır. Algoritma aşağıdaki adımlardan oluşmaktadır:

- Öncelikle destek ve güven ölçülerini karşılaştırmak için eşik değerleri belirlenir.
- Her bir ürün için destek sayıları hesaplanır. Eşik değeri ile karşılaştırılan destek değerlerinin içinden eşik değerinden düşük olanlar çıkarılır.
- Kalan ürünler ikişerli gruplanarak, grup destek sayıları hesaplanır. Tekrar eşik değerleri ile karşılaştırılan destek değerlerinden eşik değerinin altında kalanlar iptal edilir.
- Daha sonra üçerli, dörderli, beşerli, vb. biçimde gruplar için aynı karşılaştırma ve eleme işlemi devam ettirilir. Eşik değerlere uygun olduğu sürece işlemler sürecektir.
- Belirlenen ürün grubunun destek ölçülerine bakarak birliktelik kuralları türetilir ve bu kurallarının her biri için güven ölçüleri belirlenir.

Apriori Algoritması ile oluşturulan birliktelik kurallarının özellikleri ise şu şekildedir;

Birliktelik kuralları (association rules), veri madenciliği alanında üzerinde çok fazla araştırma ve çalışma yapılmış olan ilgi çekici bir konudur. Birliktelik kuralları, aynı işlem içinde çoğunlukla beraber görülen nesneleri içeren kurallardır. Birliktelik kurallarının kullanıldığı en tipik örnek market sepeti uygulamasıdır. Bu işlem,

müşterilerin yaptıkları alışverişlerdeki ürünler arasındaki birliktelikleri bularak müşterilerin satın alma alışkanlıklarını çözümler. Bu tip birlikteliklerin keşfedilmesi, müşterilerin hangi ürünleri bir arada aldıkları bilgisini ortaya çıkarır ve market yöneticileri de bu bilgi ışığında raf düzenlerini belirleyerek satış oranlarını artırabilir ve etkili satış stratejileri geliştirebilirler. Market sepeti çözümlemesinin son zamanlarda çok büyük ilgi ile karşılaşmasının sebebi kullanım kolaylığı ve anlaşılabilirliğidir. Market sepet analizi ile birliktelik kuralları çıkarımı ilk olarak Agrawal ve diğerleri tarafından 1993 yılında ele alınmıştır. Kuralları oluşturabilmek için 'destek' (support) ve 'güven' (confidence) değerlerini kullanarak, kullanıcı tarafından belirlenmiş minimum destek ve minimum güven değerlerinden yaygın birlikteliklerin belirlenmesi amaçlanmıştır. Market sepet analizinde, nesneler müşteriler tarafından satın alınan ürünlerdir ve bir hareket (kayıt) birçok nesneyi içinde bulunduran tek bir satın almadır. Birliktelik kurallarının kullanışlı olması için hem konu ile ilgili hem de anlaşılabilir olması gerekir. Birliktelik kurallarında, kullanıcının kuralların tipini ve sayısını kontrol edebileceği çeşitli yollar vardır. En yaygın olarak kullanılan yöntem, eşik değerleri olarak bilinen minimum destek ve minimum güven değerlerinin belirlendiği yöntemdir. Bu yöntemde sadece kullanıcı tarafından belirlenen eşik değerlerinden büyük olan destek ve güven değerlerine sahip kurallar bulunur ve kullanılır. Diğer bir yöntemde kullanıcının sınırlanmış nesne tanımlamasıdır. Sınırlanmış nesne, kuralların içeriğinin sınırlanmasında kullanılan mantıksal bir ifadedir. Örneğin sınırlanmış nesne cips, kola ve hamburger olsun. Sadece cips, kola ve hamburger içeren kurallar ile ilgilenilir. Birliktelik kurallarındaki bir nesnenin ve bir işlemin tanımı uygulamaya bağlıdır. Market sepeti analizinde; nesneler, müşterilerin aldığı ürünler ve işlem, beraber alınan bütün nesnelerin kümesidir. Birliktelik kurallarında sıklıkla kullanılan birkaç önemli terim vardır. Bunlar; kuralın sol tarafını ifade eden önce (antecedent), kuralın sağ tarafını ifade eden sonuç (consequent), destek değeri, güven değeri, min_destek olarak gösterilen minimum destek değeri, min_güven olarak gösterilen minimum güven değeri, nesne küme, yaygın nesne kümesi ve aday nesne kümesidir.

X ürünü alan bir müşterinin Y ürününü de alma durumu (birliktelik kuralı) $X \rightarrow Y$ ile gösterilir.

Destek ölçütü: $destek(X \rightarrow Y) = \frac{sayı(X,Y)}{n}$ ile hesaplanır. A ve B ürünlerinin birlikte sayın alınma olasılığı güven değeridir.

Güven değeri: $güven(X \rightarrow Y) = \frac{sayı(X,Y)}{sayı(X)}$ ile bulunabilir. Destek ve güven ölçütlerinin yanı sıra, bu değerleri karşılaştırabilmek için eşik değerlerine de ihtiyaç duyulmaktadır. Bulunan eşik değerlerinin, hesaplanan destek ve güven değerlerinden küçük olması beklenir. Hesaplanan destek ve güven değerlerinin büyüklük derecesi birliktelik kurallarının da o kadar güçlü olduğunu ifade eder.

Örneğin 25 tane müşterinin bir defada aldığı ürün bilgilerinden yola çıkarak birliktelik kuralı şu şekilde bulunmuş olsun: $güven(Pantolon, Kazak \rightarrow Çorap)$
Burada $X = \{Pantolon, Kazak\}$ ve $Y = \{Çorap\}$ değerleri için pantolon ve kazak alan müşterilerin bunların yanında çorap da sayın alama olasılığını ifade eder. Müşterinin bu 3 ürünü birlikte satın alma sayısı 7 ve müşteri sayısı 25 ise belirttiğimiz bu kuralın destek ölçütü şöyle olacaktır.

Destek ölçütü:

$$destek(Pantolon, Kazak \rightarrow Çorap) = \frac{sayı(Pantolon, Kazak, Çorap)}{musterisayisi} = \frac{7}{25} = 0{,}28$$

Eğer pantolon ve kazak alanların sayısının 4 olduğunu farzedelim.

Güven ölçütü:

$$güven(Pantolon, Kazak \rightarrow Çorap) = \frac{sayı(Pantolon, Kazak, Çorap)}{sayi(Pantolon, Kazak)} = \frac{7}{4} = 1{,}75$$

olacaktır.

Alışveriş yerleri genel olarak müşteri bilgileri ele geçirirler. Satılan her bir hareket sepet (“basket”) olarak adlandırılır. Market–Sepet analizi, müşteri eğilimlerini tanımlayan sepet verilerini analiz eder.

RapidMiner da kullanacağımız veriler iki süreçte işlenir. İlk süreç veri toplanması, ikinci süreç veri düzenlenmesidir. Verilerin toplanması ve düzenlenmesi el ile yapılmıştır.

1. Verilerin Toplanması; Verilerin hepsi tıbbi makalelerden alınmıştır. Kullanılmış olunan makaleler sciencedirect.com' dan (%80' i) ve IEEE web sitesinden (%20' si) alınmıştır. Tıbbi makaleler sciencedirct.com' dan daha kolay bir şekilde bulunmuştur. Toplanan verilerden makale ismi, yazar ve anahtar kelimeler alınarak excel sayfasında tablolaştırılmıştır.

2.Verilerin Düzenlenmesi; Excel sayfasında toplanan verilerin direkt olarak RapidMinner' da kullanılması bize doğru sonuçlar vermez. Buyüzden toplanan verilerin öncelikle düzenlenmesi gerekmektedir. Bu düzenlemede bazı gereksiz semboller örneğin; virgül (,), noktalı virgül (;) , nokta (.), eksi işareti (-) vb. kaldırılmıştır. Birliktelik kurallarının oluşturulması, Apriori algoritmalarının çalışması için anahtar kelimeler (|) özel işareti ile karakterize edilmiştir. İyi sonuçlar elde etmek için verilerde gruplama yapılmış, yapılan gruplamayla birlikte; 'kümeleme' ve 'sınıflandırma' kelimeleri alınarak veri madenciliği oluşturulmuştur.

Sonuç:
Bu tez çalışmasında bir verimadenciliği aracı olan RapidMiner kullanarak bir uygulama yapılmış ve beklenen sonuçlar başarıyla alınmıştır.

Önceden de belirtildiği gibi anahtar kelimeler arasında farklı ilişkiler bulunmaktadır. Bu ilişkiler; birliktelik kuralları ve her birliktelik kuralları için güven ve destek değerleri ile ölçülmüştür. Bu değerlerin bu çalışma sonrası bulunması ile başta Market Sepet analizi olmak üzere bir çok karar destek sistemlerine girdi olarak kullanılması sağlanmıştır.

Uygulamada incelenen 1000 adet makalenin anahtar kelimelerinin tekrar sayısının bulunması ile Tıbbi makalelerde kullanılan yöntemler arasındaki ilişki gözlemlenmiştir.

CHAPTER 1. INTRODUCTION TO MEDICAL DATA MINING

Modern medicine generates a great deal of information stored in the medical database. Extracting useful knowledge and providing scientific decision-making for the diagnosis and treatment of disease from the database increasingly becomes necessary. Data mining in medicine can deal with this problem. It can also improve the management level of hospital information and promote the development of telemedicine and community medicine. Because the medical information is characteristic of redundancy, multi-attribution, incompletion and closely related with time, medical data mining differs from other one. Medical data mining have discussed the key techniques of medical data mining involving pretreatment of medical data, fusion of different pattern and resource, fast and robust mining algorithms and reliability of mining results. The methods and applications of medical data mining based on computation intelligence such as artificial neural network, fuzzy system, decision support system, evolutionary algorithms, rough set, support vector machine have been introduced [1].

Clinical databases have accumulated large quantities of information about patients and their medical conditions. Relationships and patterns within this data could provide new medical knowledge. Unfortunately, few methodologies have been developed and applied to discover this hidden knowledge. the techniques of data mining (also known as Knowledge Discovery in Databases) were used to search for relationships in a large clinical database. Many researches and studies have taken place in the field of medical data mining it is a really growing field of many future researches and at the same time one of the most important field to our life, because all these researches are based on medicine and we know when we hear the word of medicine then suddenly the word of Health is also coming to our mind, in-fact we can say these studies are related to our health and health is one of the most important

factor of our life and the key factor for every other living things as well. it is an accepted fact to all. Meanwhile many researches and studies took place in medical data mining for example a survey on medical data mining was done.in this survey work data accumulated on 3,902 obstetrical patients were evaluated for factors potentially contributing to preterm birth using exploratory factor analysis. Three factors were identified by the investigators for further exploration and many more examples as well [2].

An important survey was done By Neil Savage on September 19, 2011 from this study we can understand the importance of medical data mining I took a small portion of his survey it says: The antidepressant Paxil was approved for sale in 1992, the cholesterol-lowering drug Pravachol in 1996. Company studies proved that each drug, on its own, works and is safe. But what about when they are taken together? By mining tens of thousands of electronic patient records, researchers at Stanford University quickly discovered an unexpected answer: people who take both drugs have higher blood glucose levels. The effect was even greater in diabetics, for whom excess blood sugar is a health danger. The research is an example of the increasing ease with which scientists now scour digitized medical results, like glucose tests and drug prescriptions, to find hidden patterns. “You’re not constrained by the need to actually get patients lined up in a clinical trial that would be incredibly expensive,” says Russ Altman, director of Stanford’s Biomedical Informatics Training Program, whose group published the Paxil Pravachol result in the journal Clinical Pharmacology and Therapeutics this July. “We had most of this paper done probably in a month.” The spread of electronic patient records, with their computer-readable entries, is opening new possibilities for medical data mining. Instead of being limited to carefully planned studies on volunteers, scientists can increasingly carry out research virtually by sifting through troves of data collected from the unplanned experiments of real life, as preserved in medical records from scores of hospitals. Such techniques are allowing researchers to ask questions never envisioned at the time of a drug’s approval, such as how a medicine might affect particular ethnicities. They are also being used to uncover evidence of economic problems, such as overbilling and unnecessary procedures. Mining of health records “is going to build advancements in research, but also efficiencies in the health delivery system,” says

Margaret Anderson, executive director of Faster Cures, a think tank in Washington, D.C. Some large hospital systems that use electronic records now employ full-time database research teams. Laurence Meyer, associate chief of staff for research at the Salt Lake City Veterans Administration Medical Center, says he knows of more than 100 research projects using electronic records from the VA's six million patients, who are seen at 152 hospitals and 804 outpatient clinics across the country [3].

From the above survey we see people were using antidepressant Paxil and cholesterol-lowering drug Pravachol for better health, but recently it was found they have dangerous side effects, so these side effects were found because of recent studies and researches in medical data mining. It is also called recent development in medical field and we know medical data mining is also from the biggest parts in medical.

1.1. A General Overview of My Work

As I mentioned early above in the abstract that my thesis work is based on medical data mining and I performed an implementation using Rapid Miner which is a well-known data mining tool and many people use this tool for their experiments as well. In the first step of my work I collected data for my implementation the data which I collected was only and only from medical articles in data mining I wanted to find 1000 articles in medical data mining, from year 2010 to 2015 as I collect all my data manually I looked for around 6000 to 7000 articles and I selected only 1000 related articles which were needed to me and I discarded the rest. Because when I was searching for the articles, most of them were not related what I wanted, that's why it was a time consuming and challenging work to find 1000 articles manually only and only in medical data mining, however I collected my data in an excel sheet my targeted data was name or title of the article, keywords of the article and authors of the article and the main factor on which I did the implementation is Keywords from all these selected articles/papers. In our experiment I wanted to find the relationships among these keywords and perform an implementation on these keywords in Apriori algorithm using Rapid Miner tool after the experiment I found the Association rules among these keywords and also we found which keywords occur many times, so as

in these keywords there were also names of some systems which were used for implementation in the medical data mining field such as , artificial neural network,data mining, fuzzy system, decision support system , diagnosis, clustering, classification and so on… so in the implementation result I found the occurring number of these words and association rules for them as well.

1.1.1. Collecting and organizing of data

In this step of my work in fact here are two processes the first is collecting of data and the second is organizing of data first let's see how did I collect my data.

The important data for me to take it from an article was name of the article, keywords and authors, so I used to copy all these three factors and paste it in excel sheet. I collected the data from medical articles. in the beginning I was trying to collect the data from IEEE website and I looked for medical articles in data mining after sometime my advisor told me about (sciencedirect.com) website, so when I use sciencedirect.com for collecting the data it gives me much better results and I continue with (sciencedirect.com) website up to the end of my data, I can say I collected around 20% of my data from IEEE and 80% of my data from (sciencedirect.com) website, in my case sciencedirct.com was giving good results to me and to choose the needed data was easier. As I said befor I was looking just for medical articles in data mining so the searching keywords which I used for looking to articles were medical decision support system in data mining, fuzzy system in data mining, medical diagnosis, medical clustering, medical classification and so on…

Second process organizing of data: In the beginning I just used to copy and paste these factors when this session is finished I need to organize and prepare this data for implementation with RapidMiner because this data is not ready and RapidMiner can't recognize it well, because there were some unnecessary symbols to be removed like comma (,), colon(;), dot(.),minus sign(-), space and so on … so I removed all these unnecessary symbols then I put all the Keywords only in one column which were separated by a special sign (|) we use this sign for separation of the characters. after I grouped the data into some classes like I took classifcation, clustering as data

mining to get good results, grouping of data was needed, because the number of keywords are1000 so to get good results then the number of transactions should be little bit small number as in the support of an itemset we divide the number of itemset on the total number of transaction, so if it is divided by 1000 then the result will less than the min_support count which is discarded in this case. We also separate the data by characters using this sign (|) to find the number of occurrence of each character which is important in the result section, the purpose of classification is just to get good logical and understandable results and we got it.

1.1.2. Purpose of the thesis

As I mentioned earlier this thesis work is based on medical data mining and we know this field is a wide and much growing field of medical researches that's why I also try to prepare my thesis in this field.

In this study I found that many researches took place in this field and I looked for some of them and tried to analyze them and got some good results which can be used by other users as a good information source. The keywords I collected from articles then I performed and implementation on these keywords and found that there are many interesting relationships among these keywords and at the same time from my work I found which systems are used for implementation in the medical data mining field, such as fuzzy system, artificial neural network, decision support system, diagnosis, classification, clustering and so on… so I found that most of the studies are implemented through the above systems or techniques. According to my data I found that the first much used term is data mining and the second diagnosis then artificial neural network, decision support system, cancer ,fuzzy system, clustering and support_vector_machine, diabets, breast, treatment and so on ... respectively.

1.2. Data Mining Functionalities

We have observed various types of databases and information repositories on which data mining can be performed. Let us now examine the kinds of data patterns that can be mined. Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories:

Descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions.

In some cases, users may have no idea regarding what kinds of patterns in their data may be interesting, and hence may like to search for several different kinds of patterns in parallel. Thus it is important to have a data mining system that can mine multiple kinds of patterns to accommodate different user expectations or applications. Furthermore, data mining systems should be able to discover patterns at various granularities (i.e., different levels of abstraction). Data mining systems should also allow users to specify hints to guide or focus the search for interesting patterns. Because some patterns may not hold for all of the data in the database, a measure of certainty or "trustworthiness" is usually associated with each discovered pattern.

The process of mining is often controlled by the requirements of the users. The user may be a business analyst or may be a marketing manager. Different users have different need of information. Depending on the requirements we can use different data mining techniques. The different types of data mining functionalities and the patterns they discover are described below [5].

1.2.1. Association analysis

Association rule mining is an important data mining technique which is used to find out interesting patterns or associations among the data items stored in the database. Support and confidence are two measures of the interestingness for the mined patterns. These are user supplied parameters and vary from user to user. Association rule mining is mainly used in market basket analysis or retail data analysis. In market basket analysis we identify different buying habits of customers and analyze them to find associations among items those are purchased by customers. Items that are frequently purchased together by customers can be identified. Association analysis is used to help retailers to plan different types of marketing, item placement and inventory management strategies.

When we do association rule mining in relational database management systems we generally transform the database into (tid, item) format, where tid stands for transaction ID and item stands for different items purchased by the customers. There will be multiple entries for a given transaction ID, because one transaction ID indicates purchase of one particular customer and a customer can purchase as many items as he want. An association rule can look like this:

Shuaib(buys, Watch) → Shuaib (buys, Mobile) [support =1%, confidence=50%]
Where:

$$\text{Support} = \frac{\text{The number of transactions that contain Watch and Mobile}}{\text{The total number of transactions}}$$

$$\text{Confidence} = \frac{\text{The number of transactions that contain Watch and Mobile}}{\text{The number of transactions that contain Watch}}$$

The above rule will hold if its support and confidence are equal to or greater than the user specified minimum support and confidence [4], [5].

1.2.2. Clustering analysis

Cluster analysis is a major technique for classifying a ‘mountain’ of information into manageable meaningful piles. It is a data reduction tool that creates subgroups that are more manageable than individual datum. Like factor analysis, it examines the full complement of inter-relationships between variables. In cluster analysis there is no prior knowledge about which elements belong to which clusters. The grouping or clusters are defi ned through an analysis of the data. Subsequent multi-variate analyses can be performed on the clusters as groups. The focus of clustering is to maximize the intra-class similarity and minimize the interclass similarity. Clustering occurs in almost every aspect of daily life. A factory’s Health and Safety Committee may be regarded as a cluster of people. Supermarkets display items of similar nature, such as types of meat or vegetables in the same or nearby locations. Biologists have to organize the different species of animals before a meaningful description of the differences between animals is possible. In medicine, the clustering of symptoms and diseases leads to taxonomies of illnesses. In the field of business, clusters of consumer segments are often sought for successful marketing strategies [4], [5].

1.2.3. Classification analysis

In classification, by the help of the analysis of training data we develop a model which then is used to predict the class of objects whose class label is not known. The model is trained so that it can distinguish different data classes. The training data is having data objects whose class label is known in advance. There are various presentation methods for the derived model like IF-THEN rules, decision trees, neural networks, mathematical formula.

The major difference between classification and clustering is that classification is supervised and clustering is unsupervised. That means in classification the class label is known in advance, while clustering does not assume any knowledge of clusters [4],[5].

1.2.4. Deviation analysis

Deviation analysis is the differences between the current data values, and previously defined normal values. Deviation analysis is used for detecting anomalies in the datasets. It is very important for analyzing the time-related data, in which we need to identify data deviations that occur over the time. Deviation analysis tools are helpful in security systems as well, where authorities can be warned about the deviation in resource utilization by a particular user. (Deviation analysis is concerned with discovering and classifying any changes in system behavior between two identical control systems in slightly different environments) [5].

CHAPTER 2. ASSOCIATION RULES

Association rules were first introduced in [6]. It provides the results in the form of "if-then" statements. These rules are generated from the input datasets. The rules are derived from the support and confidence value given as input from the user. An association rule is, in general, an expression of the form X→Y, where X is an antecedent and Y is a consequent. Association rule shows how many times Y has occurred if X has already occurred depending on the support and confidence value. Many algorithms for generating association rules were presented over time. Some well-known algorithms are Apriori and FP-Growth [7].

Association rules are one of the most researched areas of data mining. This is useful in the marketing and retailing strategies. Association mining is to retrieval of a set of attributes shared with a large number of objects in a given database. There are many potential application areas for association rule approach which include design, layout, and customer segregation and so on. The redundancy in association rules affects the quality of the information presented. The goal of redundancy elimination is to improve the quality and usefulness of the rules. Our work aims is to remove hierarchical duplicity in multi-level, thus reducing the size of the rule set to improve the quality and usefulness without any loss [8].

2.1. Explaining Association Rules Mathematically

Let I ={I1, I2, : : : , Im} be a set of items. Let D, the task-relevant data, be a set of database transactions where each transaction T is a set of items such that $T \subseteq$ I. Each transaction is associated with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. An association rule is an implication of the form A⇒ B, where $A\subset$ I, $B\subset$ I , and $A\bigcap B=\emptyset$. The rule $A \Rightarrow B$ holds

in the transaction set D with support S, where s is the percentage of transactions in D that contain A∪B (i.e., the union of sets A and B, or say, both A and B). This is taken to be the probability, P(A∪B) The rule A ⇒ B has confidence *c* in the transaction set D, where c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability, P(B\A). That is,

$$Support(A \Rightarrow B) = P(A \cup B) \tag{5.2}$$

$$Confidence(A \Rightarrow B) = P(B \backslash A) \tag{5.3}$$

Rules that satisfy both a minimum support threshold *(min -sup)* and a minimum confidence threshold *(min- conf)* are called strong. By convention, we write support and confidence values so as to occur between 0% and 100%, rather than 0 to 1.0 .

A set of items is referred to as an itemset. An itemset that contains k items is a k-itemset. The set (*computer, antivirus- software*) is a 2-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset. This is also known, simply, as the *frequency, support count*, or count of the itemset. Note that the itemset support defined in Equation (5.2) is sometimes referred to as relative support, whereas the occurrence frequency is called the absolute support. If the relative support of an itemset *I* satisfies a prespecified minimum support threshold (i.e., the absolute support of *I* satisfies the corresponding minimum support count threshold), then I is a frequent itemset. The set of frequent k-itemsets is commonly denoted by L_k . From Equation (5.3), we have

$$Confidence(A \Rightarrow B) = P(B \backslash A) = \frac{(support\ (A \cup B)}{support\ (A)} = \frac{support\ count\ (\text{AUB})}{support - count\ (A)} \tag{5.4}$$

Equation (5.4) shows that the confidence of rule $(A \Rightarrow B)$ can be easily derived from the support counts of A and (AUB). That is, once the support counts of A, B, and (AUB) are found, it is straightforward to derive the corresponding association rules $(A \Rightarrow B)$ and $(B \Rightarrow A)$ and check whether they are strong. Thus the problem of mining association rules can be reduced to that of mining frequent itemsets.

In general, association rule mining can be viewed as a two-step process:

1. Find all frequent itemsets: By definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count, min -sup.

2. Generate strong association rules from the frequent itemsets: By definition, these rules must satisfy minimum support and minimum confidence.

Suppose, as a marketing manager of AllElectronics, you would like to determine which items are frequently purchased together within the same transactions. An example of such a rule, mined from the AllElectronics transactional database, is

buys(X, "computer") ⇒ *buys(X, "software") [support = 1%, confidence = 50%]*
where X is a variable representing a customer. A confidence, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well. A 1% support means that 1% of all of the transactions under analysis showed that computer and software were purchased together. This association rule involves a single attribute or predicate (i.e., buys) that repeats.Association rules that contain a single predicate are referred to as single-dimensional association rules. Dropping the predicate notation, the above rule can be written simply as *"Computer ⇒software [1%, 50%]"*.

Suppose, instead, that we are given the AllElectronics relational database relating to purchases. A data mining system may find association rules like:
age(*X*, "20::: 29") ∧ *income*(*X*, "20K:::29K"))*buys* ⇒ (*X*, "*CD player*")

[*Support* = 2%, *confidence* = 60%]
The rule indicates that of the *AllElectronics* customers under study, 2% are 20 to 29 years of age with an income of 20,000 to 29,000 and have purchased a CD player.

Typically, association rules are discarded as uninteresting if they do not satisfy both a minimum support threshold and a minimum confidence threshold [9].

2.1.1. Confidence and support concepts

There are two important basic measures for association rules, support(s) and confidence (c). Since the database is large and users concern about only those frequently purchased items, usually thresholds of support and confidence are predefined by users to drop those rules that are not so interesting or useful. The two thresholds are called minimal support and minimal confidence respectively. Support(s) of an association rule is defined as the percentage/fraction of records that contain X ∪ Y to the total number of records in the database. Suppose the support of an item is 0.1%, it means only 0.1 percent of the transaction contain purchasing of this item.

Support ($X \rightarrow Y$) = $\frac{\boldsymbol{Count(X,Y)}}{\boldsymbol{n}}$

Confidence of an association rule is defined as the percentage/fraction of the number of transactions that contain X ∪ Y to the total number of records that contain X.

Confidence is a measure of strength of the association rules, suppose the confidence of the association rule X⇒Y is 80%, it means that 80% of the transactions that contain X also contain Y together.

Confidence ($X \rightarrow Y$) = $\frac{\boldsymbol{Count(X,Y)}}{\boldsymbol{CountX}}$

In general, a set of items (such as the antecedent or the consequent of a rule) is called an itemset. The number of items in an itemset is called the length of an itemset [10].

2.2. Introduction To Apriori Algorithm

Apriori algorithm [11]., [12]. is data mining association rules algorithm, put forward by Imielinski, Agrawal and Swami. Apriori algorithm is a kind of data mining algorithm which is based on the horizontal data representation and breadth first search.

Apriori Algorithm is a basic algorithm that mining generates Boolean association

rules needs frequent itemsets. Apriori algorithm mainly uses a circular system to rake though one gradation in turn to complete frequent itemsets mining. The main idea of this circular system is to generate *(k*+1*)*- itemsets from the *k*- the frequent itemsets. Detailed procedure: First, find frequent 1- itemsets, called *L1*; then, use *L1* to mine *L2*, namely frequent 2- itemsets; the circular system terminates when it can't find more frequent *k*- itemsets. Mine one gradation *Lk* needs traversing the whole database once [13].

The pseudo-code for Apriori algorithm

- Pseudo-code:

C_k: Candidate itemset of size k

L_k : frequent itemset of size k

L_1 = {frequent items};
for (k = 1; L_k !=∅; k++) do begin

C_{k+1} = candidates generated from L_k;
for each transaction t in database do

increment the count of all candidates in C_{k+1}
that are contained in t

L_{k+1} = candidates in C_{k+1} with min_support
end
return $\cup_k L_k$; [14]

To make the concept much clear about Apriori Algorithm then let's see a concrete example, now we are going to see an example of Apriori Algorithm step by step then it can be much easier and understandable.

2.2.1. Explaining apriori algorithm with example

There are six transactions in the following Table 2.1., that is, |D| = 6. , $\text{Support}_{\text{threshold}}$ = %30 and $\text{Confidence}_{\text{threshold}}$= %60 Now let's see how do we proceed the Apriori algorithm for finding frequent itemsets in D.

First of all we count minimum support count:

Minimum support count is => 0.3*6 = 1.8

Table 2.1. Given itemsets

TID	NAME OF THE Diseases
1	Migraines, Anemia, Bronchitis, Asthma
2	Migraines, Anemia, Asthma
3	Asthma, Diabetes
4	Diabetes, Thyroid
5	Migraines, Anemia
6	Migraines, Anemia, Diabetes

1. In the first iteration of the algorithm, each item is a member of the set of candidate 1-itemsets, C1. The algorithm simply scans all of the transactions in order to count the number of occurrences of each item.

Table 2.2. C1 number of occurrence

Itemset	Sup.Cout
Migraines	4
Anemia	4
Bronchitis	1
Asthma	3
Diabetes	3
Thyroid	1

2. The minimum support count required is 1.8, that is, min-sup = 1.8. (Here, we are referring to absolute support because we are using a support count. The set of frequent 1-itemsets, L1, can then be determined. It consists of the candidate 1-itemsets satisfying minimum support. In our example, four of the candidates in C1

satisfy minimum support. And the two others (Bronchitis & Thyroid) are discarded.

Table 2.3. L1 prune items

Itemset	Sup.Cout
Migraines	4
Anemia	4
Asthma	3
Diabetes	3

3. To discover the set of frequent 2-itemsets, L2, the algorithm uses the join L1⋈L1 to generate a candidate set of 2-itemsets, C2 consists of $(2^{|L1|})$ 2-itemsets. Note that no candidates are removed from C2 during the prune step because each subset of the candidates is also frequent.

Table 2.4. C2 taking two itemsets

Itemset
{Migraines,Anemia}
{Migraines,Asthma}
{Migraines,Diabetes}
{ Anemia, Asthma }
{ Anemia, Diabetes}
{ Asthma , Diabetes}

4. Next, the transactions in D are scanned and the support count of each candidate itemset In C2 is accumulated, as shown below:

Table 2.5. C2 taking number of two itemsets

Itemset	Sup.Count
{Migraines,Anemia}	4
{Migraines,Asthma}	2
{Migraines,Diabetes}	1
{ Anemia, Asthma }	2
{ Anemia, Diabetes}	1
{ Asthma , Diabetes}	1

5. The set of frequent 2-itemsets, L2, is then determined, consisting of those candidate 2-itemsets in C2 having minimum support.

Table 2.6. L2 is pruning C2

Itemset	Sup.Count
{Migraines,Anemia}	4
{Migraines,Asthma}	2
{ Anemia, Asthma }	2

6. The generation of the set of candidate 3-itemsets,C3 is joining of L2 $\bowtie$ L2

Table 2.7. C3 taking three itemsets

Itemset
{ Migraines,Anemia,Asthma}

Next, the transactions in D are scanned and the support count of each candidate itemset In C3 is accumulated

Table 2.8. C3 taking number of three itemsets

Itemset	Sup.Count
{Migraines,Anemia,Asthma}	2

7. The transactions in D are scanned in order to determine L3, consisting of those candidate 3-itemsets in C3 having minimum support

Table 2.9. L3 is pruning c3

Itemset	Sup.Count
{ Migraines,Anemia,Asthma}	2

8. The algorithm uses L3 ⋈ L3 to generate a candidate set of 4-itemsets, C4

Table 2.10. C4 taking four numbers

Itemset
{Migraines, Anemia, Bronchitis, Asthma}

The transactions in D are scanned in order to determine L4, consisting of those candidate 4-itemsets in C4 having minimum support

Table 2.11. Taking the number of itemsets

Itemset	Sup.Count
{Migraines, Anemia, Bronchitis, Asthma}	1

After scanning C4 we can't proceed to L4 because the support count in C4 is 1 and the required sup-count should be ≥1.8, thus C4=Ø and the algorithm terminates . having found all of the frequent itemsets.

Looking at the support measure of the specified product groups association rules

derived and determined confidence measure for each of these rules.

Table 2.12. For association rules

Anemia	Asthma		
Migraines	Diabetes	Asthma	Bronchitis
Asthma	Migraines	Bronchitis	
Anemia			
Diabetes	Anemia	Migraines	Thyroid
Anemia	Diabetes	Migraines	

Items: { Anemia, Asthma, Migraines, Diabetes, Bronchitis, Thyroid}

13 Large Itemsets (by Apriori)

{ Anemia} (support: 66.67%)

{ Asthma} (support: 50%)

{Migraines} (support: 66.67%)

{Diabetes} (support: 50%)

{ Bronchitis} (support: 33.33%)

{Migraines, Bronchitis} (support: 33.33%)

{Migraines, Diabetes} (support: 50%)

{ Asthma, Bronchitis} (support: 33.33%)

{ Asthma, Migraines} (support: 33.33%)

{ Anemia, Diabetes} (support: 33.33%)

{ Anemia, Migraines} (support: 33.33%)

{Migraines, Diabetes, Anemia} (support: 33.33%)

{Migraines, Bronchitis, Asthma} (support: 33.33%)

In the example, Support threshold and Confidence threshold are provided which are: Support $_{threshold}$ = %30 and Confidence $_{threshold}$= 60% Looking at the Support $_{threshold}$ which is 30% we can consider all above 13 rules for support, because all of the above results are greater than Support $_{threshold}$ which is 30%.

The generated results from Table 2.12. For Support and Confidence using, the following formulas of Support and Confidence are as follow:

Support $(X \rightarrow Y) = \frac{\boldsymbol{Count(X,Y)}}{\boldsymbol{n}}$

Confidence $(X \rightarrow Y) = \frac{\boldsymbol{Count(X,Y)}}{\boldsymbol{CountX}}$

Let us see some examples how the support and confidences are calculated:

How is it calculated: {Anemia} (support: 66.67%)

Support $(X \rightarrow Y) = \frac{\boldsymbol{Count(X,Y)}}{\boldsymbol{n}}$

Support (Anemia) $= \frac{Count\ \ Anemia\ \ which\ is=4}{n=6}$, so $\frac{4}{6} = 66.67\%$ from here Support (Anemia) or {Anemia} Support=66.67%

Let's see how to get Support for two itemsets:{Migraines, Bronchitis} (support: 33.33%)

To get the above two itemsets:

Support{Migraines, Bronchitis}=$\frac{\text{o : Æ; \%KQJP / ECN=EJAO =J@ \$NK?DEPKO S DE?D =NALt}}{\text{JLx}}$

Thus $-^{t}_{x}$ L r u uu Luuuuu¤

Now let's see how to get it three itemsets: {Migraines, Bronchitis, Asthma} (support: 33.33%)

To get the above three itemsets:

Support {Migraines, Bronchitis, Asthma}=$\frac{\text{o : Æ;}}{}$

L$\frac{\text{\%KQJP / ECN=EJAO Æ\$NK?DEPKO < \#OPDI = S DE?D =NALt}}{\text{JLx}}$= Thus, $-^{t}_{x}$ L r u uu Luuuuu¤

As we saw some examples for calculating support, now let us see how to calculate confidence, we know the formula for confidence is:

Confidence ($X \rightarrow Y$) = $\frac{\boldsymbol{Count(X,Y)}}{\boldsymbol{CountX}}$

How to get { Bronchitis} => {Migraines} (confidence: 100%)

Confidence ($X \rightarrow Y$) = $\frac{\boldsymbol{Count(X,Y)}}{\boldsymbol{CountX}} = \frac{\boldsymbol{Count(Brochitis,Migraines=2)}}{\boldsymbol{Count(Bronchitis)=2}}$,Thus $\frac{\mathbf{2}}{\mathbf{2}} = \mathbf{1} = \mathbf{100}\%$

Let's see how to get confidence for three itemsets:

{Diabetes} => {Migraines, Anemia} (confidence: 66.67%)

Confidence ($X \rightarrow Y$) = [illegible] **Thus** [illegible]

Now we are going to generate association rules in the following Table 2.13., & for each rule, we will find confidence measure from Table 2.12.

16 Association Rules

Table 2.13. Generated association rules

Association	Explanation	Confidence
{ Bronchitis} => {Migraines}	If Bronchitis is occurring Then the probability of Migraines is 100% to occur after Bronchitis	$\frac{2}{2} = 100\%$
{Diabetes} => {Migraines}	If Diabetes is occurring Then the probability of Migraines is 100% to occur after Diabetes	$\frac{3}{3} = 100\%$
{Migraines} => {Diabetes}	If Migraines is occurring Then the probability of Diabetes is 75% to occur after Migraines	$\frac{3}{4} = 75\%$
{ Bronchitis} => { Asthma}	If Bronchitis is occurring Then the probability of Asthma is 100% to occur after Bronchitis	$\frac{3}{3} = 100\%$
{ Asthma} => { Bronchitis}	If Asthma is occurring Then the probability of Bronchitis is 66.67% to occur after Asthma	$\frac{2}{3} = 66.67\%$
{ Asthma} => {Migraines}	If Asthma is occurring Then the probability of Migraines is 66.67% to occur after Asthma	$\frac{2}{3} = 66.67\%$
{Diabetes} => { Anemia}	If Diabetes is occurring Then the probability of Anemia is 66.67% to occur after Diabetes	$\frac{2}{3} = 66.67\%$
{Diabetes} => {Migraines, Anemia}	If Diabetes is occurring Then the probability of Migraines and Anemia is 66.67% to occur after Diabetes	$\frac{2}{3} = 66.67\%$
{Diabetes, Anemia} => {Migraines}	If Diabetes and Anemia are occurring together Then the probability of Migraines is 100% to occur after them	$\frac{2}{2} = 100\%$

Table 2.13. Generated association rules (Continued)

{Migraines, Anemia} => {Diabetes}	If Migraines and Anemia are occurring together Then the probability of Diabetes is 100% to occur after them	$\frac{2}{2} = 100\%$
{Migraines, Diabetes} => { Anemia}	If Migraines and Diabetes are occurring together Then the probability of Anemia is 66.67% to occur after them	$\frac{2}{3} = 66.67\%$
{ Asthma} => {Migraines, Bronchitis}	If Asthma is occurring Then the probability of Migraines and Bronchitis is 66.67% to occur after it	$\frac{2}{3} = 66.67\%$
{ Bronchitis} => {Migraines, Asthma}	If Bronchitis is occurring Then the probability of Migraines and Asthma is 100% to occur after it	$\frac{2}{2} = 100\%$
{ Bronchitis, Asthma} => {Migraines}	If Bronchitis and Asthma are occurring together Then the probability of Migraines is 100% to occur after them	$\frac{2}{2} = 100\%$
{Migraines, Asthma} => { Bronchitis}	If Migraines and Asthma are occurring together Then the probability of Bronchitis is 100% to occur after them	$\frac{2}{2} = 100\%$
{Migraines, Bronchitis} => { Asthma}	If Migraines and Bronchitis are occurring together Then the probability of Asthma is 100% to occur after them	$\frac{2}{2} = 100\%$

Association rules explanation and confidence in the above Table 2.13. Are calculated from Table 2.12.

In the given example, Support threshold and Confidence threshold are provided which are:

Support $_{threshold}$ = %30 and Confidence $_{threshold}$= 60% Looking at the Confidence $_{threshold}$ which is 60% comparing this with our Table 2.13. We can find our results easily, in Table 2.13. We can see all the found Confidence values are bigger than Confidence $_{threshold}$ which is 60%, thus we got good results.

CHAPTER 3. DATA MINING IMPLEMENTATION

In this chapter, we are going to discuss the implementation part of our thesis; we use a data-mining tool for our experiment, which is called Rapid Miner. RapidMiner (formerly Yale) is an environment for machine learning and data mining processes. A modular operator concept allows the design of complex nested operator chains for a huge number of learning problems. The data handling is transparent to the operators. They do not have to cope with the actual data format or dierent data views - the Rapid Miner core takes care of the necessary transformations.

Today, Rapid Miner is the world-wide leading open-source data mining solution and is widely used by researchers and companies [15]. As we mentioned earlier that we got the results from our data. Here we are going to see the whole details of our implementation. For making it much, clear and understandable I took almost all the screen shots and tables by looking to them we can call it, self-explanatory.

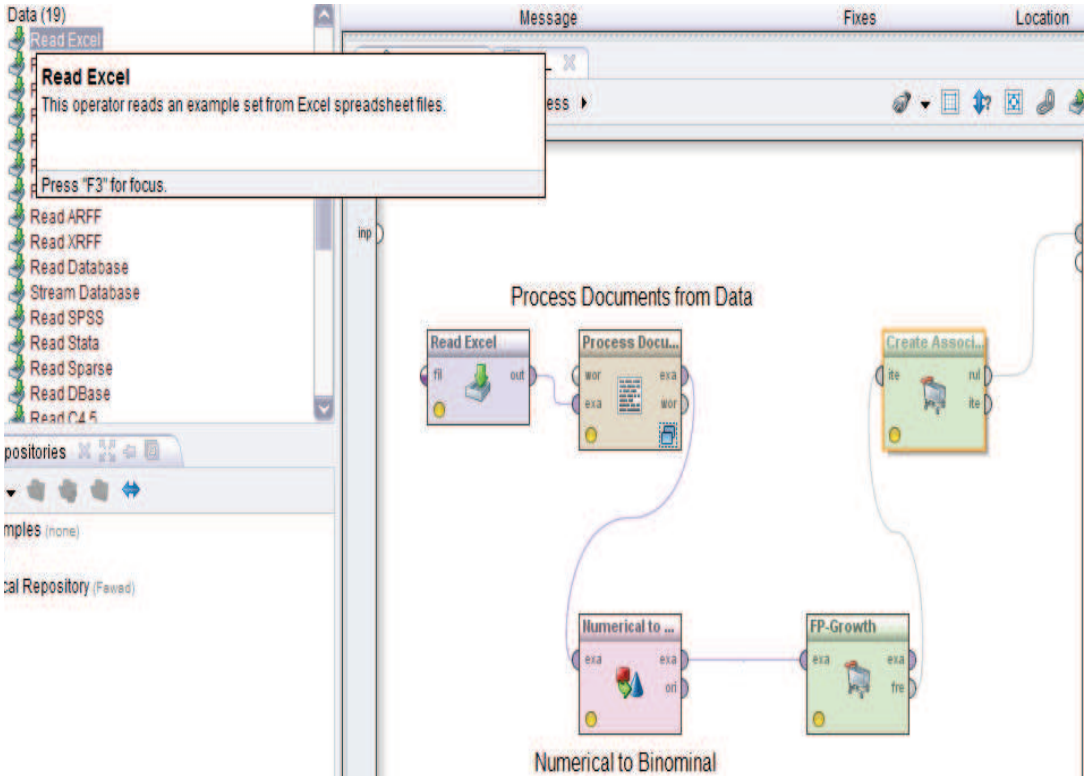


Figure 3.1. Importing excel sheet and performing text mining

In the above, interface Figure 3.1. from the left side we drag the Read Excel box and drop it here as we see, after in the same way we drag and drop Process Documents from Data from left side model (Text Processing) then drag and drop Numerical to Binominal box, so in the above process we connect the boxes by those lines the out is connected to the exa as we can see, this Process Documents from Data and Numerical to Binominal processes are important for our data, for the first process we are performing a text mining that's why we took that box (Process Documents from Data) in the second process Numerical to Binominal this is a transformation process which transforms Numerical data to Binominal as our data is not Numerical that's why we use this box.

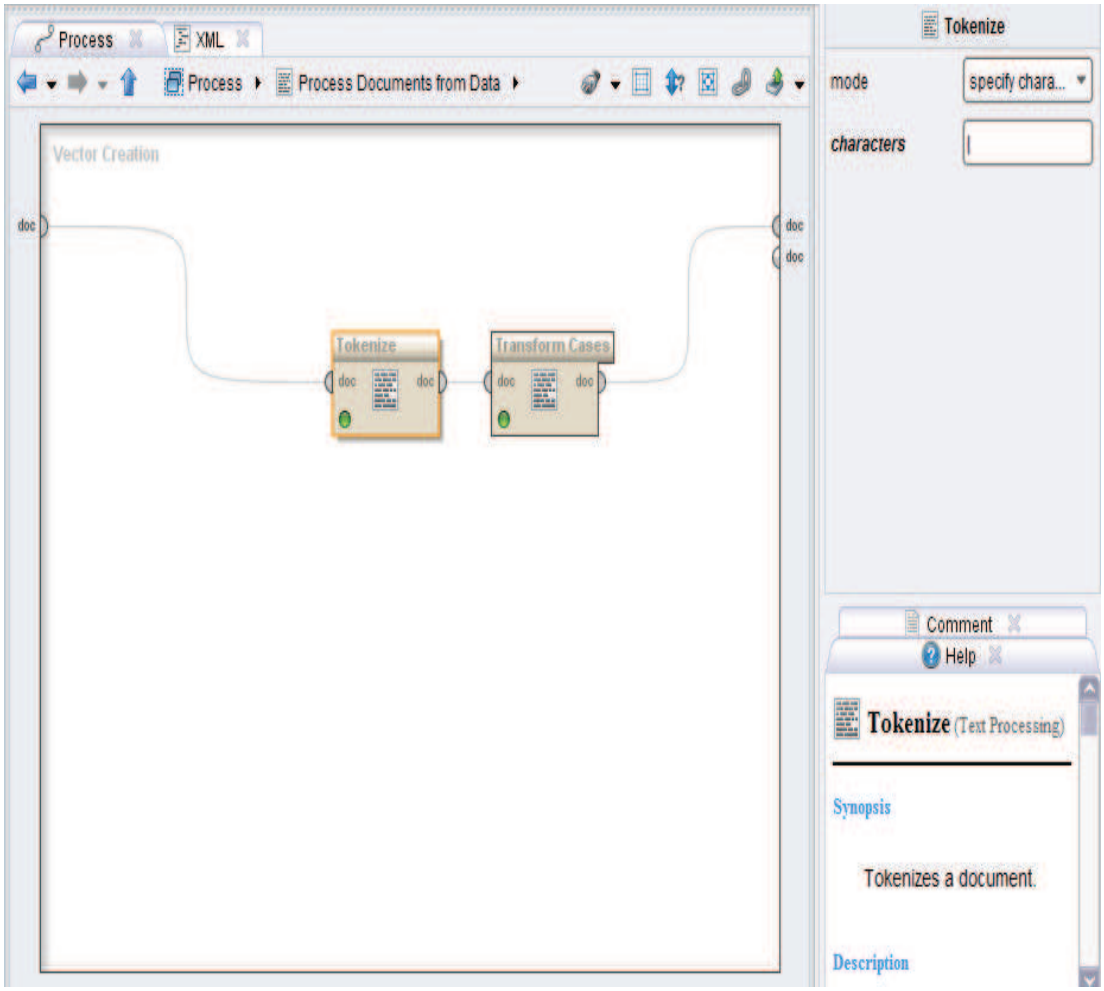


Figure 3.2. Tokenizing data and transform letters to lower case

In Figure 3.2. there are two processes one is tokenize and another is transform case, tokenize process tokenizes the data and it is needed to individualize the association rules for different itemsets, this process separates the words (itemsets) from each other and for separating the itemsets we select a specific sign (|) on the top right side by selecting Specify Character this sign (|) separates the itemsets from each other. We may use other sign as well. The second process here is called Transform Case. In this process, we select lower case letters in the results, so whenever we run our project all the results will be in lower case letters.

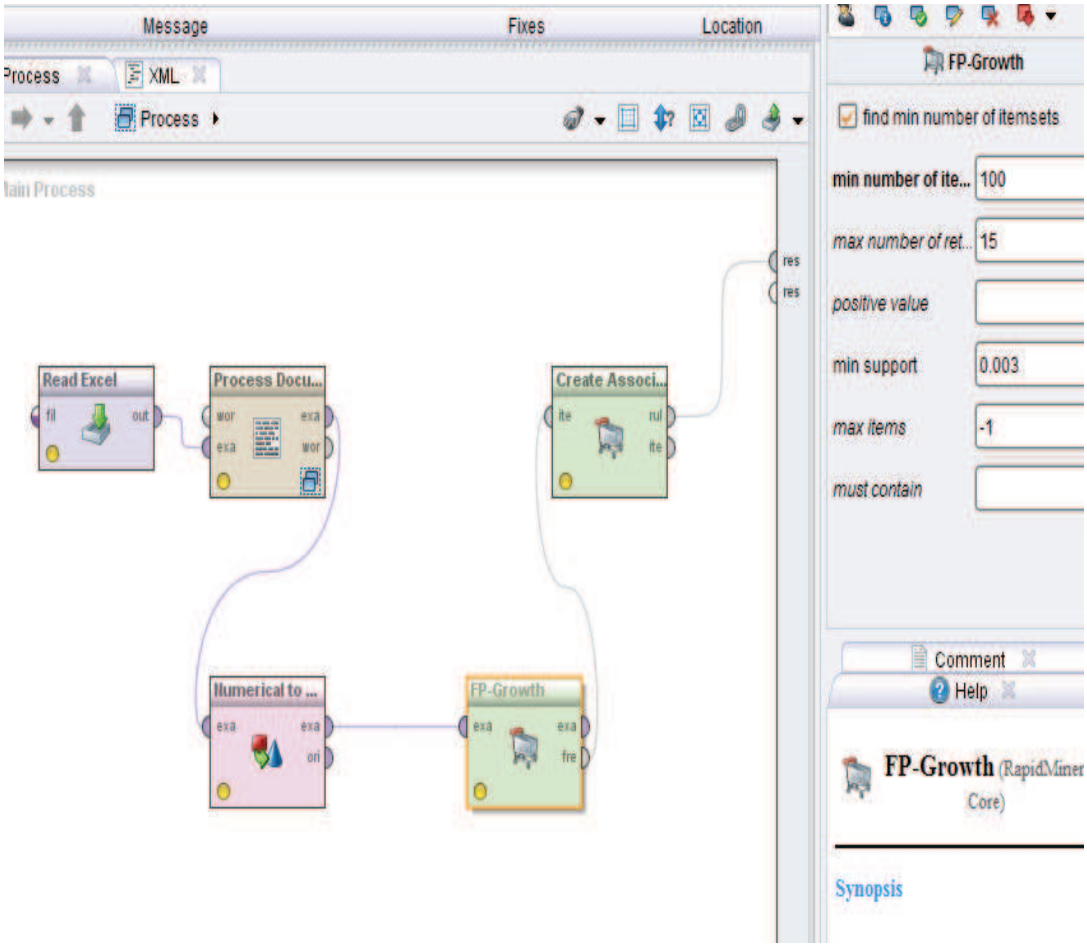


Figure 3.3. FP-Growth the box where we specify the min-sup-count

Here in FP-Growth we specify the minimum support count we can see above on the top right side we give value 0.003 for min-support. When we run our project in the result we are getting values for support and confidence, if the resulted support and confidence values are greater than or equal to the given support and confidence respectively then we take it as our successful results, otherwise discard it.

In our project all, our results are successful and we consider all of them, but the greater once are the best results and we will show it later in the table format.

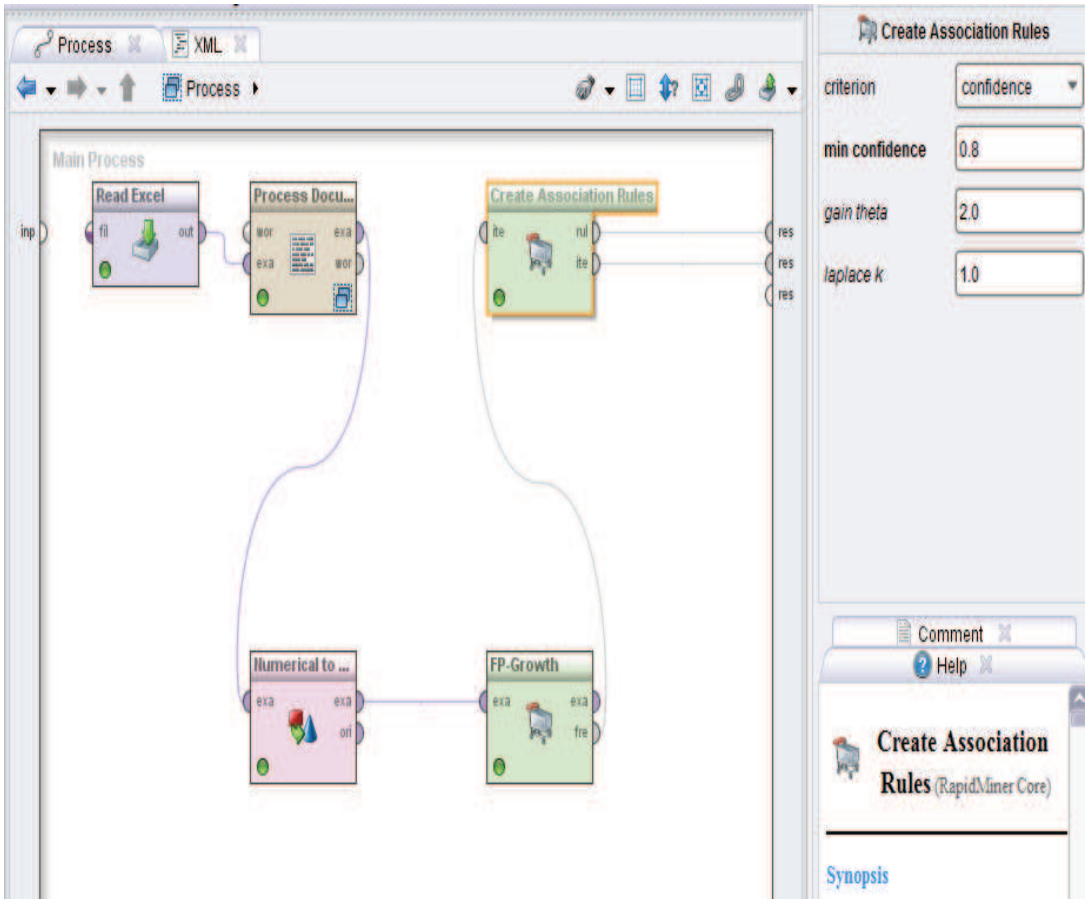


Figure 3.4. Create association rules the box where we specify the min-confidence

In Figure 3.4. There are two lines taken from creating association rules one is for association rules another is for finding length of itemsets. We specify the minimum confidence and find the length of itemsets as well, here the minimum confidence=0.8 it means after running the project all the values of confidences equal or greater than 0.8 are considered, but the smaller once are discarded. In our project all of them are considered because they are greater than the given minimum confidence value. Which will be shown later in a table format. We also found the length of itemsets, the longest one is six we can see it below Figure 3.5. In the excel screen shot.

A	B	C	D	E	F	G	H
Size	Support	item1	item2	item3	item4	item5	item6
4	0.003	fuzzy_system	accuracy	risk_prediction	attribute_selection		
4	0.003	fuzzy_system	accuracy	uci_repository	attribute_selection		
4	0.003	fuzzy_system	risk_prediction	uci_repository	attribute_selection		
4	0.003	accuracy	risk_prediction	uci_repository	attribute_selection		
5	0.003	decision_support_system	fuzzy_system	accuracy	risk_prediction	uci_repository	
5	0.003	decision_support_system	fuzzy_system	accuracy	risk_prediction	attribute_selection	
5	0.003	decision_support_system	fuzzy_system	accuracy	uci_repository	attribute_selection	
5	0.003	decision_support_system	fuzzy_system	risk_prediction	uci_repository	attribute_selection	
5	0.003	decision_support_system	accuracy	risk_prediction	uci_repository	attribute_selection	
5	0.003	fuzzy_system	accuracy	risk_prediction	uci_repository	attribute_selection	
6	0.003	decision_support_system	fuzzy_system	accuracy	risk_prediction	uci_repository	attribute_selection

Figure 3.5. Length of itemsets

The above Figure 3.5. shows the length of itemsets there is one most longest itemsets which has six items, and six other rows which are having five, five itemsets I took them as an example they are having their own meaning and logics for example the first row says in our data these six items repeated thrice and that's true we can find it by this equation 0.003 ∗ s rr Lr and same thing can be done for the other itemsets as well here 0.003 is minimum support count, 1000 is the number of our data so multiplying the min-sup-count with number of data is giving us the number of itemset in whole our data.

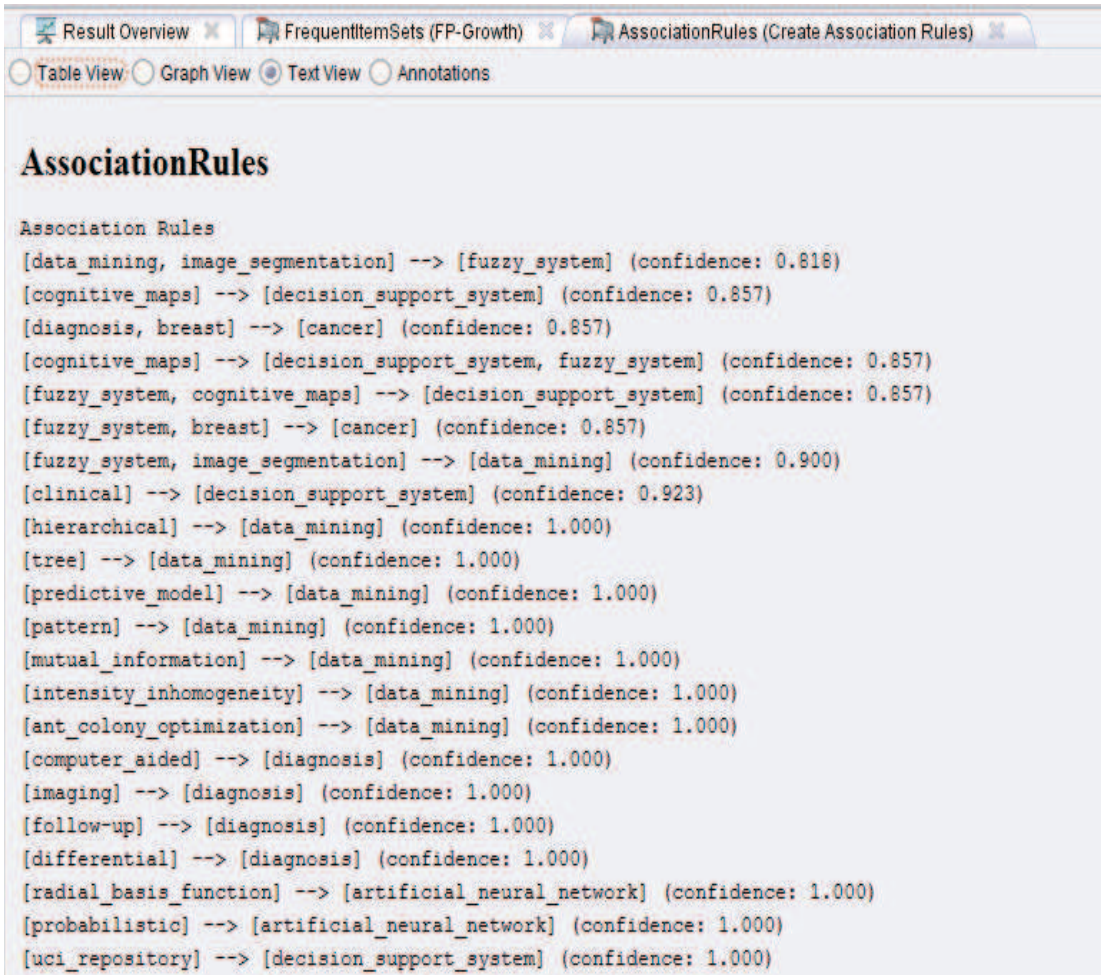
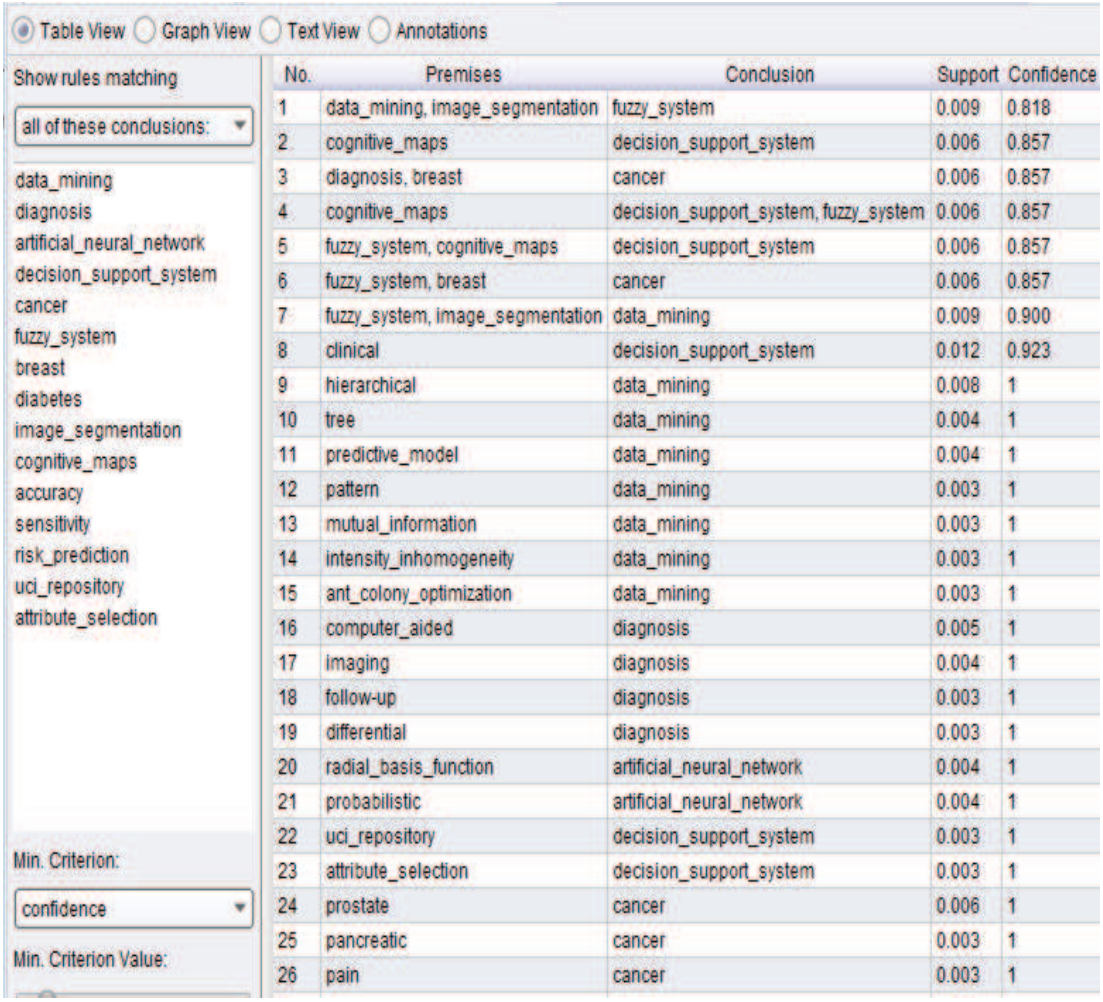


Figure 3.6. Association rules text view

Figure 3.6. is a screen shot from association rules in form of text view here we can see the confidence is greater than 80% it means we can consider all of them but when it is 100% then it is the best. The following Figure 3.7. Providing the same concept of association rules, but in a table view.



No.	Premises	Conclusion	Support	Confidence
1	data_mining, image_segmentation	fuzzy_system	0.009	0.818
2	cognitive_maps	decision_support_system	0.006	0.857
3	diagnosis, breast	cancer	0.006	0.857
4	cognitive_maps	decision_support_system, fuzzy_system	0.006	0.857
5	fuzzy_system, cognitive_maps	decision_support_system	0.006	0.857
6	fuzzy_system, breast	cancer	0.006	0.857
7	fuzzy_system, image_segmentation	data_mining	0.009	0.900
8	clinical	decision_support_system	0.012	0.923
9	hierarchical	data_mining	0.008	1
10	tree	data_mining	0.004	1
11	predictive_model	data_mining	0.004	1
12	pattern	data_mining	0.003	1
13	mutual_information	data_mining	0.003	1
14	intensity_inhomogeneity	data_mining	0.003	1
15	ant_colony_optimization	data_mining	0.003	1
16	computer_aided	diagnosis	0.005	1
17	imaging	diagnosis	0.004	1
18	follow-up	diagnosis	0.003	1
19	differential	diagnosis	0.003	1
20	radial_basis_function	artificial_neural_network	0.004	1
21	probabilistic	artificial_neural_network	0.004	1
22	uci_repository	decision_support_system	0.003	1
23	attribute_selection	decision_support_system	0.003	1
24	prostate	cancer	0.006	1
25	pancreatic	cancer	0.003	1
26	pain	cancer	0.003	1

Figure 3.7. Association rules table view

In above Figure 3.7. we can see four things at the top of the figure premises, conclusion, support and confidence as they are association rules there is a condition of IF-THEN premises stands for “IF” conclusion stands for THEN, we know support and confidence from our earlier knowledge.

Let’s see rule 10 (tree→ data_mining) support=0.004 and confidence=1.0 it means when there is tree then the occurring probability of data mining is 100% when confidence is 1 it means the probability of happening is 100% the same thing can be explained for the other rules as well. On the left side of the Figure 3.7. We can see some itemsets, so if we want to see association rules for a specific itemset then we just click on any of them we will get association rules for it. We will see some of

them in the coming figures as example.

No.	Premises	Conclusion	Support	Confidence
7	fuzzy_system, image_segmentation	data_mining	0.009	0.900
9	hierarchical	data_mining	0.008	1
10	tree	data_mining	0.004	1
11	predictive_model	data_mining	0.004	1
12	pattern	data_mining	0.003	1
13	mutual_information	data_mining	0.003	1
14	intensity_inhomogeneity	data_mining	0.003	1
15	ant_colony_optimization	data_mining	0.003	1
44	fuzzy_system, genetic_algorithm	data_mining	0.003	1
45	fuzzy_system, algorithm	data_mining	0.003	1
46	intensity_inhomogeneity	data_mining, fuzzy_system	0.003	1
48	fuzzy_system, intensity_inhomogeneity	data_mining	0.003	1
49	image_segmentation, medical	data_mining	0.003	1
50	intensity_inhomogeneity	data_mining, image_segmentation	0.003	1
52	image_segmentation, intensity_inhomogeneity	data_mining	0.003	1
137	intensity_inhomogeneity	data_mining, fuzzy_system, image_segmentation	0.003	1
139	fuzzy_system, intensity_inhomogeneity	data_mining, image_segmentation	0.003	1
141	image_segmentation, intensity_inhomogeneity	data_mining, fuzzy_system	0.003	1
143	fuzzy_system, image_segmentation, intensity_inhomogeneity	data_mining	0.003	1

Figure 3.8. Association rules table view for data mining

As we said in the preiviouse Figure 3.7. when we want to see the association rules for an itemset then we click on that after clicking all the explanation will come about that itemset here we select data mining and we can see all the explanation in the above Figure 3.8. for data mining the same thing can be done for any of the itemsets. Figure 3.8. shows the association rules for data mining as an example we can take number 10 (tree→ data_mining) support=0.004 and confidence=1.0 it means when there is tree then the occurring probability of data mining is 100% when confidence is 1 it means the probability of happening is 100% the same thing can be explained for the other rules as well. On the left side of the Figure 3.8. We can see some itemsets, so if we want to see association rules for a specific itemset then we just click on any of them we will get association rules for it. We will see one of them in

the coming figure as example.

No.	Premises	Conclusion	Support	Confidence
2	cognitive_maps	decision_support_system	0.006	0.857
4	cognitive_maps	decision_support_system, fuzzy_system	0.006	0.857
5	fuzzy_system, cognitive_maps	decision_support_system	0.006	0.857
8	clinical	decision_support_system	0.012	0.923
22	uci_repository	decision_support_system	0.003	1
23	attribute_selection	decision_support_system	0.003	1
55	artificial_neural_network, clinical	decision_support_system	0.003	1
58	fuzzy_system, knowledge_representation	decision_support_system	0.003	1
61	fuzzy_system, risk_prediction	decision_support_system	0.003	1
63	uci_repository	decision_support_system, fuzzy_system	0.003	1
65	fuzzy_system, uci_repository	decision_support_system	0.003	1
66	attribute_selection	decision_support_system, fuzzy_system	0.003	1
68	fuzzy_system, attribute_selection	decision_support_system	0.003	1
69	knowledge_representation, cognitive_maps	decision_support_system	0.003	1
71	accuracy, risk_prediction	decision_support_system	0.003	1
72	uci_repository	decision_support_system, accuracy	0.003	1
74	accuracy, uci_repository	decision_support_system	0.003	1
75	attribute_selection	decision_support_system, accuracy	0.003	1
77	accuracy, attribute_selection	decision_support_system	0.003	1
79	uci_repository	decision_support_system, risk_prediction	0.003	1
81	risk_prediction, uci_repository	decision_support_system	0.003	1
83	attribute_selection	decision_support_system, risk_prediction	0.003	1
85	risk_prediction, attribute_selection	decision_support_system	0.003	1
86	uci_repository	decision_support_system, attribute_selection	0.003	1
88	attribute_selection	decision_support_system, uci_repository	0.003	1
90	uci_repository, attribute_selection	decision_support_system	0.003	1
146	fuzzy_system, knowledge_representation	decision_support_system, cognitive_maps	0.003	1

Figure 3.9. Association rules table view for decision support system

In Figure 3.9. We took decision support system as an example and we can see the association rules for each number let's take number 8 as an example (clinical→decision_support_system) for this rule support=0.012 and confidence=0.923 it means the probability of occurring clinical with decision_support_system is around 92% to make it much clear this association rule says: If clinical is occurring then decision_support_system is also occurring with a percentage of 92.

We took some other screen shots individually for some itemsets to make the concept clearer we can have a look to the following two other once as well (fuzzy system and to diagnosis) then we will go to the graphical representation of our implementation.

No.	Premises	Conclusion	Support	Confidence
4	cognitive_maps	decision_support_system, fuzzy_system	0.006	0.857
28	cognitive_maps	fuzzy_system	0.007	1
29	rules	fuzzy_system	0.004	1
30	uci_repository	fuzzy_system	0.003	1
31	intuitionistic	fuzzy_system	0.003	1
32	intensity_inhomogeneity	fuzzy_system	0.003	1
33	attribute_selection	fuzzy_system	0.003	1
34	adaptive_neuro	fuzzy_system	0.003	1
46	intensity_inhomogeneity	data_mining, fuzzy_system	0.003	1
47	data_mining, intensity_inhomogeneity	fuzzy_system	0.003	1
59	decision_support_system, cognitive_maps	fuzzy_system	0.006	1
60	decision_support_system, risk_prediction	fuzzy_system	0.003	1
62	decision_support_system, rules	fuzzy_system	0.003	1
63	uci_repository	decision_support_system, fuzzy_system	0.003	1
64	decision_support_system, uci_repository	fuzzy_system	0.003	1
66	attribute_selection	decision_support_system, fuzzy_system	0.003	1
67	decision_support_system, attribute_selection	fuzzy_system	0.003	1
92	intensity_inhomogeneity	fuzzy_system, image_segmentation	0.003	1
94	image_segmentation, intensity_inhomogeneity	fuzzy_system	0.003	1
96	knowledge_representation, cognitive_maps	fuzzy_system	0.003	1
98	accuracy, risk_prediction	fuzzy_system	0.003	1
99	uci_repository	fuzzy_system, accuracy	0.003	1
101	accuracy, uci_repository	fuzzy_system	0.003	1
102	attribute_selection	fuzzy_system, accuracy	0.003	1
104	accuracy, attribute_selection	fuzzy_system	0.003	1
106	uci_repository	fuzzy_system, risk_prediction	0.003	1
108	risk_prediction, uci_repository	fuzzy_system	0.003	1

Figure 3.10. Association rules table view for fuzzy system

The concept and explanation is same like we did for decision support system.

No.	Premises	Conclusion	Support	Confidence
16	computer_aided	diagnosis	0.005	1
17	imaging	diagnosis	0.004	1
18	follow-up	diagnosis	0.003	1
19	differential	diagnosis	0.003	1
53	artificial_neural_network, computer_aided	diagnosis	0.003	1
54	cancer, diabetes	diagnosis	0.003	1

Figure 3.11. Association rules table view for diagnosis

Same explanation can be given for all other itemsets, which we can see on the left side of the above figure. Here We took diagnosis as an example as we took DSS and we can see the association rules for each number let's take number 16 as an example (computer_aided→diagnosis) for this rule support=0.005 and confidence=0.1 it means the probability of occurring computer_aided with diagnosis is 100% to make it much clear this association rule says: If computer_aided is happening then diagnosis is also occurring with a percentage of 100. after this, we are going to a graphical representation of our implementation. Hope it will be much clear and understandable.

In the following Figure 3.12. Whenever we want to know an association rule, we put the cursor on that rule then it will give all the explanation like premises, conclusion,

support, confidence, lift, gain, conviction, laplace and ps from all these explanation we just need the first four of them and the others are not our concern. As an example we put the cursor on rule 96 and we can see the explanation for that. Here in graphic view the concept is same as we mentioned in table view for example when we want to see the association rule for an itemset individually then we can see on the left side of Figure 3.7. There are itemsets, when we select any of them then we can see the association rule for the selected itemset. So now, we are going to see how the explanation is looking in some of the screen shots, which I took from the implementation.

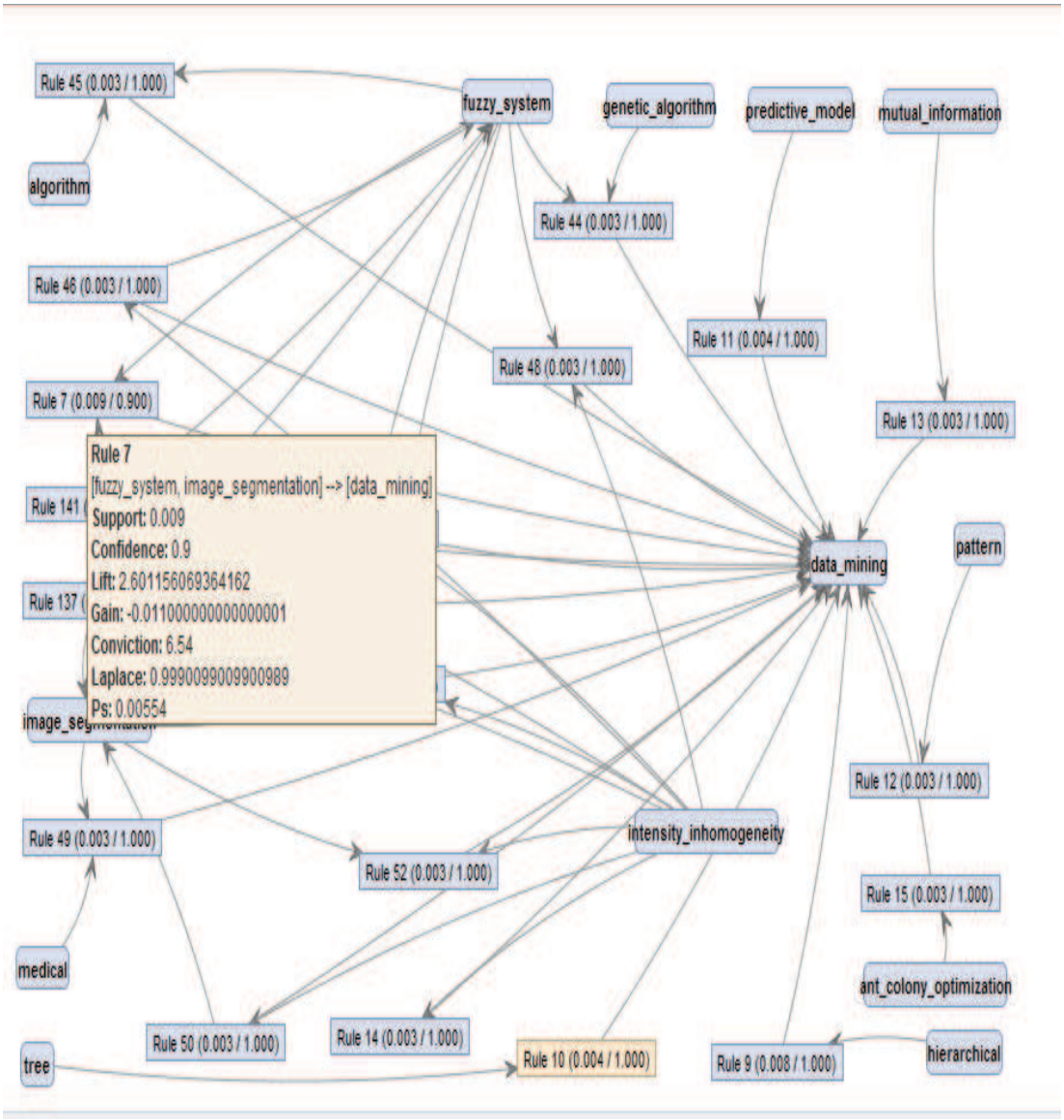


Figure 3.12. Association rules graph view for itemset data mining

Here in the above screen shot Figure 3.12., we see the graphic view for data mining itemset, we selected one rule 7 and we know how to select a rule just put your cursor on any rule then you will get all the details about it. Rule 7 (fuzzy_system, image_segmentation,) → (data_mining) support=0.009 and confidence=0.9, so this rule 7 represent that whenever fuzzy system and image segmentation occur together then the probability of happening or occurring of data mining is 90%.

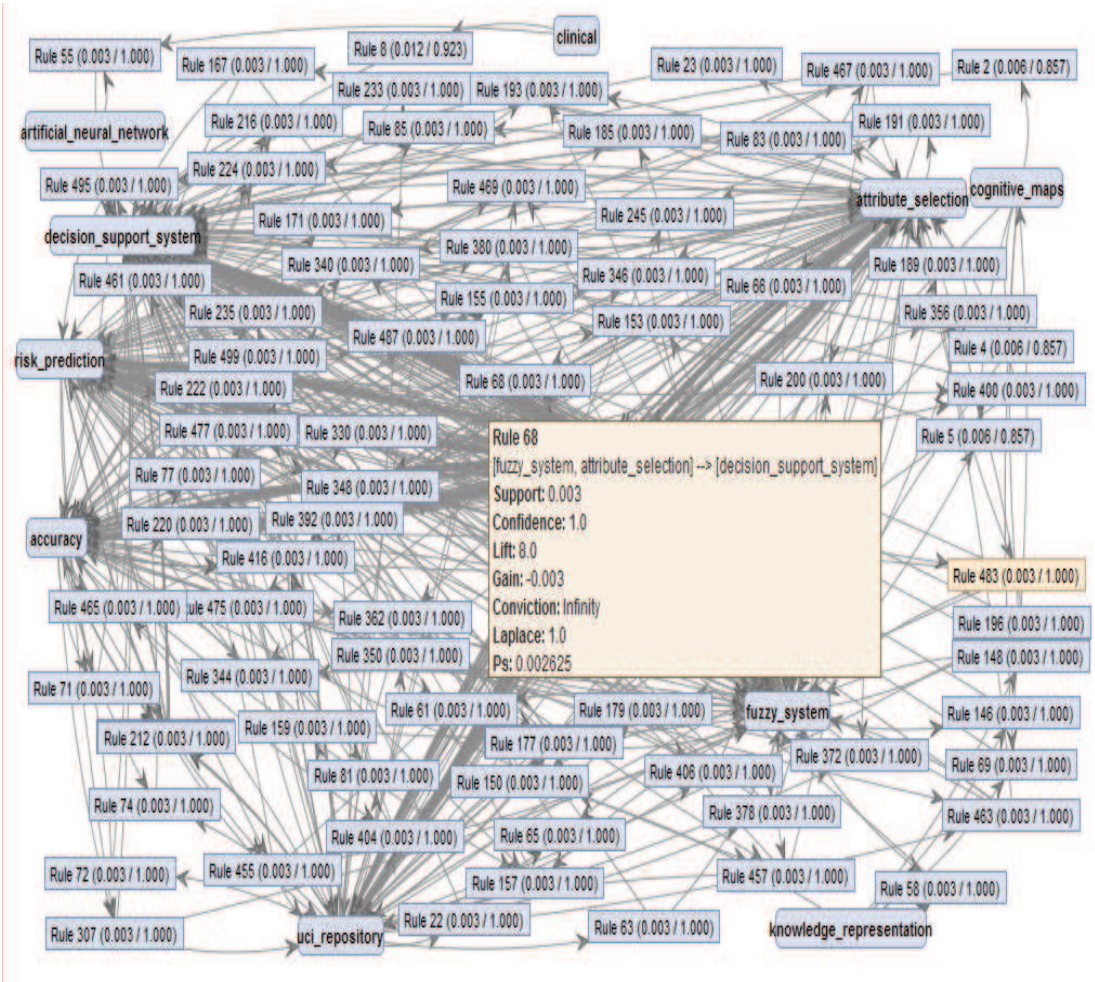


Figure 3.13. Association rules graph view for itemset decision support system

Rule 68 is (fuzzy_system, attribute_selection) → (decision_support_system) support=0.003 confidence=1.0 the explanation of the rule is: if fuzzy system and attribute selection happens together then the happening probability of decision support system after them is 100%.

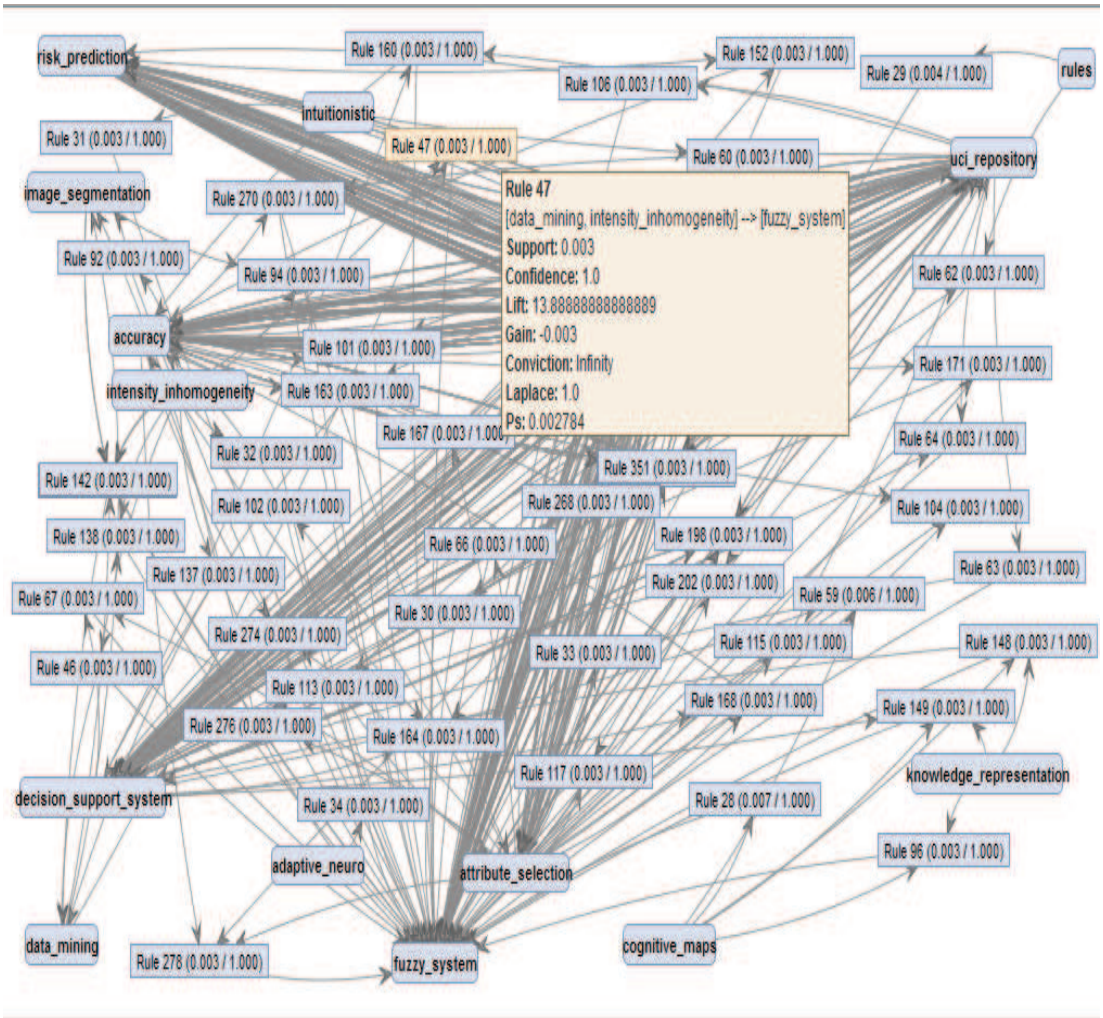


Figure 3.14. Association rules graph view for itemset fuzzy system

In Figure 3.14. the rule 47 (data_mining, intensity_inhomogeneity) → (fuzzy_system) support=0.003 confidence=1.0 here rule 47 explains whenever data mining and intensity inhomogeneity are happened together in this scenario there is 100% probability of happening or occurring fuzzy system after these two itemsets.

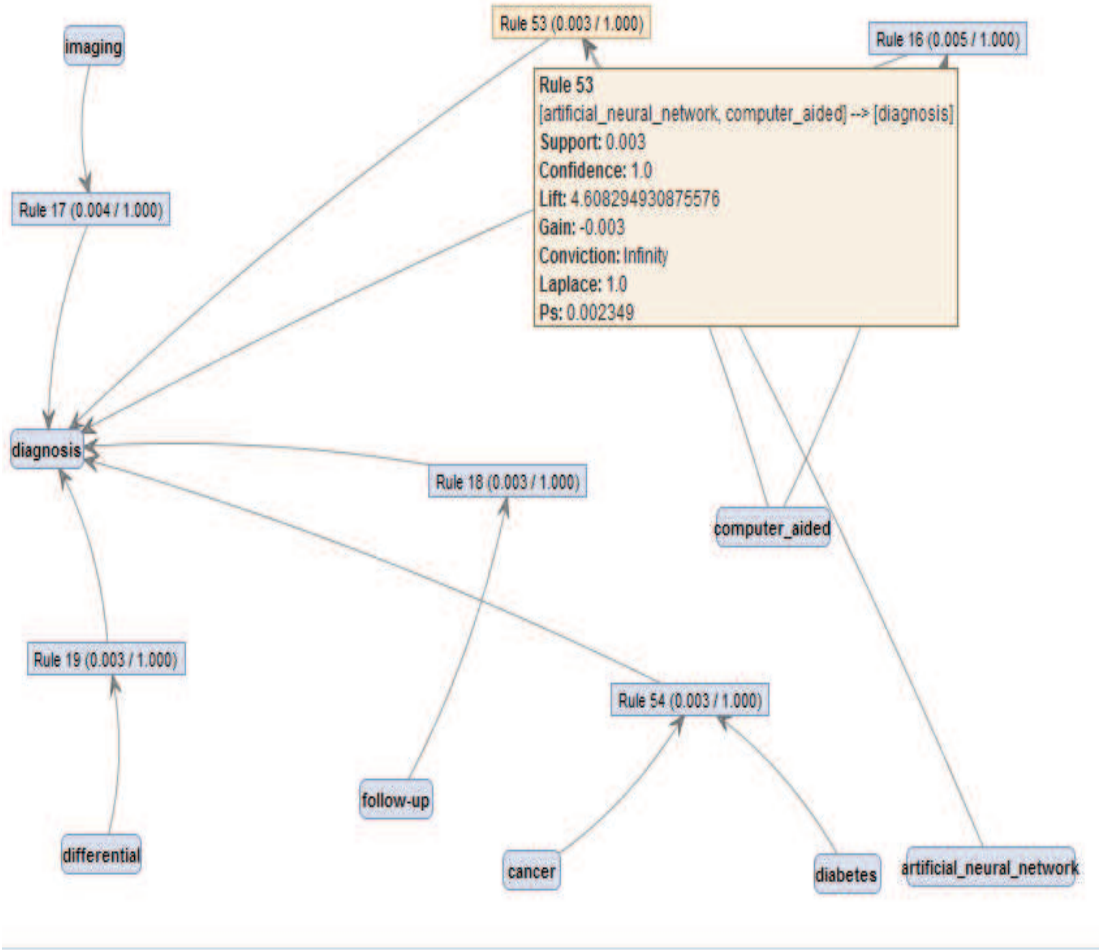


Figure 3.15. Association rules graph view for itemset diagnosis

Figure 3.15. is a graphic view of association rule for diagnosis we took here rule 53 as an example (artificial_neural_network, computer_aided) → (diagnosis) support=0.003 confidence=1.0 the explanation of rule 53 is same like other rules which we explained earlier. This rule says: if artificial neural network with computer aided occur together then there is 100% probability that diagnosis is occurring after them.

We can summarize the result of our work by looking to the following tables and graphs it will make the concept very clear because in my experiment case, tables and graphical representation is very effective and understandable, that's why I tried my best to present my work using tables and screen shot from the implementation.

All the screen shots are taken from our implementation and I took these screen shots to make my implementation understandable and clear for other users. In addition, I wanted to show all the process of my implementation.

Table 3.1. Collection of best supports

No	Premises	Conclusion	Support	Confidence
1	Clinical	decision_support_system	0.012	0.923076923
2	data_mining, image_segmentation	fuzzy_system	0.009	0.818181818
3	fuzzy_system, image_segmentation	data_mining	0.009	0.9
4	Hierarchical	data_mining	0.008	1
5	cognitive_maps	fuzzy_system	0.007	1
6	cognitive_maps	decision_support_system	0.006	0.857142857
7	diagnosis, breast	Cancer	0.006	0.857142857
8	cognitive_maps	decision_support_system, fuzzy_system	0.006	0.857142857
9	fuzzy_system, cognitive_maps	decision_support_system	0.006	0.857142857
10	fuzzy_system, breast	Cancer	0.006	0.857142857
11	Prostate	Cancer	0.006	1
12	decision_support_system, cognitive_maps	fuzzy_system	0.006	1
13	computer_aided	Diagnosis	0.005	1
14	Mellitus	Diabetes	0.005	1
15	breast, genetic_algorithm	Cancer	0.005	1
16	Tree	data_mining	0.004	1
17	predictive_model	data_mining	0.004	1
18	Imaging	Diagnosis	0.004	1
19	radial_basis_function	artificial_neural_network	0.004	1
20	Probabilistic	artificial_neural_network	0.004	1
21	Rules	fuzzy_system	0.004	1
22	Specificity	Sensitivity	0.004	1
23	artificial_neural_network, prognosis	Cancer	0.004	1
24	Pattern	data_mining	0.003	1

The above Table 3.1. Shows the best 23 supports which we collected from all 505-association rules.

We give minimum support count=0.003 so all the resulted supports are equal or greater than 0.003 in this case we can consider all the supports, but the bigger once

are much better that's why I show only the greater once on the above table, from number one to number twenty three, all the supports are greater than 0.003 but from 24 to the rest (505) all supports are 0.003 and it starts from number 24 as we can see it in Table 3.1. above. Meanwhile except one, two, three and 6-10 confidences are not one or hundred percent, but all the rest are 100%. In above Table 3.1. supports are arranged in descending order it means we took the supports from higher to the lower the higher one is 0.012 and the smaller one is 0.003.

Another important part from the conclusion is number of occurrence of each itemset. In this part, we took only those itemsets which are having the occurrence number up to five and we discard less than five. We show this part both in table and graph format for making it much understandable. The following Table 3.2. Represent the number of those itemsets that have higher number of occurrences in our data. In the following table we can see data mining is much-occurred itemset in our data, it means most of the medical applications are implemented with data mining and then others respectively and name of the sicknesses can be seen in the table as well.

Table 3.2. Number of occurrences of itemsets

No	Attribute Name	Total Occurrences	Document Occurrences
1	data_mining	435	346
2	Diagnosis	228	217
3	artificial_neural_network	181	162
4	decision_support_system	134	125
5	Cancer	100	93
6	fuzzy_system	84	72
7	support_vector_machine	39	39
8	Diabetes	33	27
9	Breast	32	32
10	Treatment	24	23
11	genetic_algorithm	19	19
12	image_segmentation	18	18
13	feature_selection	17	17
14	Algorithm	16	15
15	Ontology	16	16
16	Clinical	14	13
17	Prognosis	14	14
18	Epidemiology	13	13
19	magnetic_resonance_imaging	12	12
20	Medical	12	12
21	Biomarkers	11	11
22	Ultrasound	11	11
23	heart_disease	10	10
24	Prediction	10	10
25	Segmentation	10	10
26	heart_failure	9	9
27	medical_informatics	9	9
28	electronic_health_records	8	8
29	expert_system	8	8
30	Hierarchical	8	8
31	Hiv	8	8
32	pattern_recognition	8	8
33	Rehabilitation	8	8
34	Accuracy	7	7
35	case-based_reasoning	7	7
36	cognitive_maps	7	7
37	Depression	7	7
38	emergency_department	7	7
39	knowledge_representation	7	7
40	logistic_regression	7	7

Table 3.2. Number of occurrences of itemsets (Continued)

41	metabolic_syndrome	7	7
42	risk_factors	7	7
43	sensitivity_and_specificity	7	7
44	Biomarker	6	6
45	decision_support	6	6
46	image_processing	6	6
47	Lung	6	6
48	medical_image	6	6
49	Mortality	6	6
50	particle_swarm_optimization	6	6
51	Prevention	6	6
52	Prostate	6	6
53	Screening	6	6
54	semantic_web	6	6
55	Sensitivity	6	6
56	Survival	6	6
57	bayesian_network	5	5
58	computer_aided	5	5
59	Dementia	5	5
60	feature_extraction	5	5
61	Hypertension	5	5
62	İntuitionistic	5	3
63	Mellitus	5	5
64	Outcome	5	5
65	parkinson's_disease	5	5
66	patient_safety	5	5
67	Prevalence	5	5
68	primary_care	5	5
69	risk_prediction	5	5
70	Surgery	5	5
71	traditional_chinese_medicine	5	5
72	Validation	5	5

We can represent the above Table 3.2. In a graph form then it will be much clear and easy to understand. The occurrence number up to seven. In the following graph (Figure 3.16.) there are two terms one is document occurrence and another one is total occurrence we can see in the graph the total occurrence of data mining is 435, but 346 in the document. The total occurrence of diagnosis is 228, but 217 in the document. We can also comment the rest as we commented data mining and diagnosis.

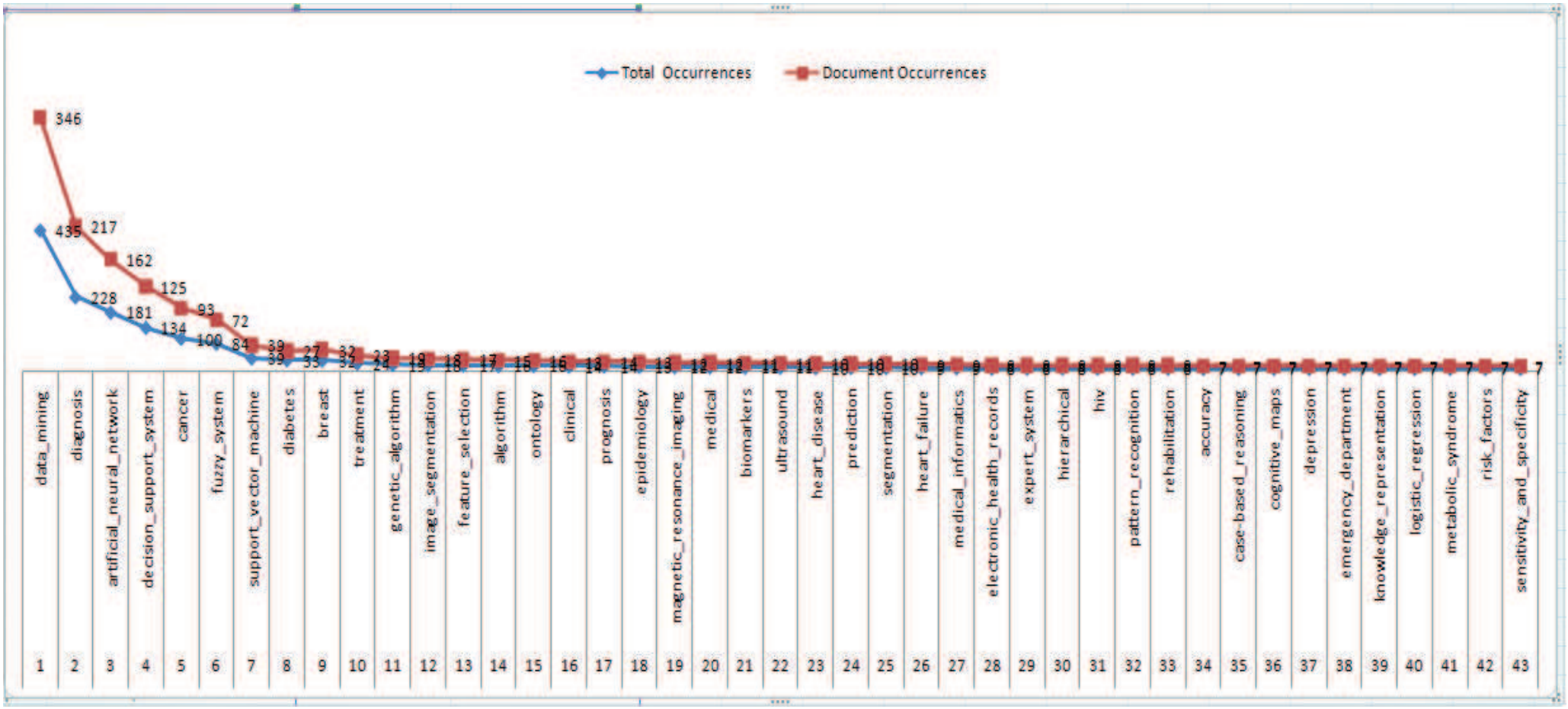


Figure 3.16. Occurrences of itemsets graphically shown

CHAPTER 4. CONCLUSION AND FUTURE WORK

Medical data mining is a broad and important area of recent researches many researches took place in this field and many other researches are going on. In this thesis work I performed an implementation using a data mining tool Rapid Miner and got the expected results successfully.

As I mentioned early the aim of this work is to find different relationships among the keywords such as: Association rules for the itemsets, support and confidence for each association rule we also found the number of occurrences of each itemset finding them are also giving us a good and meaningful result. All these resutls are very helpful for the future researchers, becasue if a researcher want to do a research in this field then from our results they can see which keywords are mentioned the most in the medical articles and they can get all the information about all the relationsips among the keywords.

We said early about our data and we said our focusing data is keywords of the articles, so in the keywords there was the name of systems, which were used for implementation in this field. For the future work, we will try to compare all the results obtained in these 1000 papers our results are also the name of some used systems like decision support system, fuzzy system, artificial neural network, data mining and so on ... the names of these system means they are the used system for implementation and we can understand which system is giving good and effective results in medical data mining from the above mentioned systems. In our case, we just collected the name of the articles, keywords and authors of the article and we performed our implementation on the keywords, but we are planning to look for all the results in these 1000 papers and then compare them, after we will find which system give the best result in medical data mining.

REFERENCES

[1] ZhU, L., WU, B., CAO, C., Introduction to Medical data mining. Sheng Wu Yi Xue Gong Cheng Xue Za Zhi. College of Automation, Chongqing University, Chonging, Sep;20(3):559-62, 2003.

[2] PRATHER, JC1., LOBACH, DF., GOODWİN, LK., HALES, JW., HAGE, ML., HAMMOND, WE., Medical data mining: knowledge discovery in a clinical data warehouse. Division of Medical Informatics, Duke University Medical Center, Durham, North Carolina, USA, 101–105, 1997.

[3] NEİL, SAVAGE., Mining Data for Better Medicine, September 19, 2011. http://www.technologyreview.com/news/425466/mining-data-for-better-medicine/, Access Date: 11.01.2015.

[4] JİAWEİ, HAN., MİCHELİNE, KAMBER., Data Mining Concepts and Techniques Second Edition 23(1):261, 2006.

[5] PANKAJ, KANDPAL., Association Rule Mining In Partitioned Databases Performance Evaluation and Analysis By M.Tech (Software Engineering), Indian Institute of Information technology, Allahabad July 2007.

[6] AGRAWAL, R., IMİELİNSKİ, T., SWAMİ, A., "Mining association rules between sets of items in large database",Proceeding of the 1993 ACM SIGMOD International Conference on Management of Data, ACM Press, pp. 207-216, 1993.

[7] MEERA, NARVEKARA., SHAFAQUE, FATMA, SYEDB., An optimized algorithm for association rule mining using FP tree, Peer-review under responsibility of scientific committee of International Conference on Advanced Computing Technologies and Applications (ICACTA), 2015.

[8] A K, CHANDANANA., M K, SHUKLAB., Removal of duplicate rules for Association Rule Mining from multilevel dataset, Peer-review under responsibility of scientific committee of International Conference on Advanced Computing Technologies and Applications (ICACTA)2015.

[9] JİAWEİ, HAN., MİCHELİNE, KAMBER., Data Mining Concepts and Techniques Second Edition , 231-233, 2006.

[10] SOTİRİS, KOTSİANTİS., DİMİTRİS, KANELLOPOULOS., Association Rules Mining: International Transactions on Computer Science and Engineering, Vol.32, A Recent Overview University of Patras, Greece, GESTS, 71-82, 2006.

[11] AGRAWA1, R., IMİELİNSKİ, T., SWAMİ, A., Mining association rules between sets of items in large databases [C] / /Proc of ACM SIGMOD Conference on Management of Data, ACM New York, NY, USA, 2072216, 1993.

[12] AGRAWAL, R., SRİKANT, R., Fast algorithm for mining association rules in large databases [C] / /Proceedings of the 20th International Conference onVery Large Data Bases, Morgan Kaufmann Publishers Inc San Francisco, CA, USA, 4872499, 1994.

[13] LI, XİANG., Simulation System of Car Crash Test in C-NCAP Analysis Based on an Improved Apriori Algorithm, International Conference on Solid State Devices and Materials Science School of Computer Engineering, Huaiyin Institute of Technology, China, Huai'an 223003, 2012.

[14] DR, BERNARD, CHEN., Ph.D.University of Central Arkansas, Data Mining Concepts and Techniques 2nd Ed Ch5 Mining Frequent Patterns, Associations, and Correlations, "slides" April 18, 2013.

[15] RAPID-I, GMBH., STOCKUMER, Str. 475 44227 Copyright by Rapid-I, Dortmund, Germany, March 14 2001-2009.

RESUME

Fawad Sadiqmal, born on 1986.06.10 in Kapisa Afghanistan. Completed primary, secondary and high school in Sarobi (Kabul) high school, finished high school at the end of 2007. İn the beggining of 2008 got a scholarship of undergraduate from OIC organization of the Islamic cooperation in Bangladesh in the department of computer science and information technology in the university of IUT (Islamic University of Technology), finished undergraduate studies at the end of 2011, worked almost a year in Afghanistan 7 months in the ministry of gas and petrollium as an IT officer and 2 months with Dyncorps as a data entry clerck. İn 2012 got a master scholarship from republic of Turkey in Sakarya University in the department of Computer and Information Engineering and still continuing in Sakarya University.