T.C.

BEYKENT ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

EKSİK DEĞERLERİ EN OLASI DEĞER İLE DOLDURMANIN SINIFLANDIRMA ALGORİTMALARI ÜZERİNDEN KARŞILAŞTIRILMASI

(Yüksek Lisans Tezi)

Tezi Hazırlayan: Çağdaş KEKLİK

İSTANBUL,2017

T.C.

BEYKENT ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

BİLGİSAYAR MÜHENDİSLİĞİ BİLİM DALI

EKSİK DEĞERLERİ EN OLASI DEĞER İLE DOLDURMANIN SINIFLANDIRMA ALGORİTMALARI ÜZERİNDEN KARŞILAŞTIRILMASI

(Yüksek Lisans Tezi)

Tezi Hazırlayan:

Çağdaş KEKLİK

Öğrenci No:

130820001

Danışman:

Yrd.Doç.Dr.Cengiz ÖRENCİK

İSTANBUL,2017

YEMİN METNİ

Yüksek lisans tezi olarak sunduğum "**Eksik Değerleri En Olası Değer İle Doldurmanın Sınıflandırma Algoritmaları Üzerinden Karşılaştırılması.**" başlıklı çalışmanın, bilimsel ahlak ve geleneklere uygun bir şekilde tarafımdan yazıldığını, yararlandığım eserlerin tamamının kaynaklarda gösterildiğini ve çalışmanın içinde kullanıldıkları her yerde atıf yapıldığını belirtir ve bunu onurumla doğrularım.

İMZA

Aday: Çağdaş KEKLİK

TEŞEKKÜR

Çalışmam süresince fikir ve düşünceleriyle beni yönlendiren öncelikle Bilgisayar Mühendisliği Bölüm Başkanımız Sayın Yrd. Doç. Dr. Ediz ŞAYKOL 'a, değerli danışmanım Sayın Yrd. Doç. Dr. Cengiz ÖRENCİK 'e ve Sayın Prof. Dr. Yücel SAYGIN 'a sonsuz teşekkürlerimi sunarım.

Ayrıca, çalışmam boyunca hoşgörü ve desteğini esirgemeyen değerli aileme teşekkürlerimi sunarım.

EKSİK DEĞERLERİ EN OLASI DEĞER İLE DOLDURMANIN SINIFLANDIRMA ALGORİTMALARI ÜZERİNDEN KARŞILAŞTIRILMASI

Tezi Hazırlayan : Çağdaş KEKLİK

ÖZET

Günümüz bilgi çağında gözümüze çarpan veri madenciliği en temel makine öğrenmesi yöntemlerinden biri olarak dikkat çekmektedir. Gün geçtikçe bilgisayarların devamlı ucuzlama durumu ve güç performansının dur durak bilmeden artışı, bilgisayarlarda çok fazla miktarlarda verinin saklanabilmesine olanak vermektedir. Veri madenciliği, bu büyük hacim ve çeşitlilikteki veriden anlamlı bilgi edinebilmenin hemen hemen tek çözüm yolu şeklinde bakılmaktadır. Bu sebepten ötürü çok miktarda verileri işleyebilen metotları kullanabilmek, hayati olabilecek bir öneme sahiptir. Veri madenciliğinin asıl amacı birçok veri içerisinde saklı durumda mevcut olan örüntü ve eğilimleri bulup çıkartma işlemidir.

Çok büyük veri ambarlarının içinde tutulan veriler tek olarak kullanıldıklarında değersiz olarak görülebilseler de, bu veriler toparlanıp bir hedefe odaklı olarak kullanıldığı zaman anlamlı hale dönüşmektedirler. Asıl amaç veriyi uygun bilgiye çevirme işidir ve bu veri madenciliği ile gerçekleştirilmektedir. Veri madenciliğinde esas olan şey kısaca verilerin işlenmesi metodudur. Dünya üzerinde durmaksızın artış gösteren ve inanılmaz boyutlara ulaşan veriyi en yüksek performansı sağlayacak şekilde kullanmanın yolu veri madenciliğinden geçmektedir. Bu olay diğer alanlarda görüldüğü gibi tıp alanında da çok büyük ilgi odağı haline gelmiştir.

Veri madenciliği yaparken karşılaşan en temel problemlerden biri üzerinde çalışılan verinin düzenlenmesidir. Verinin bazı satırları eksik değerler içerebilir. Bu değerlerin eksik olması o verinin işleme sokulmasını ve diğer değerler ile karşılaştırılmasını imkansız kılar. Bu tezde bu eksik değerlerin olası en uygun değerler ile doldurularak işleme sokulmasının sonuca etkileri analiz edilmiştir. Eksik değer içeren satırları toptan yok saymak, belli bir sınır değerden çok eksik veri içeren satırları yok sayıp kalan değerleri olası tahmini değerler ile doldurmak ve her türlü eksik veriyi olası en uygun değer ile doldurarak analize dahil etmek senaryoları ayrı ayrı test edilerek başarımları test edilmiş ve birbirlerine olan üstünlükleri değerlendirilmiştir.

Bu analizlerimizde kanser verisi örnek test kümesi olarak seçilmiştir. Veri madenciliğinin tanımı ile başlayarak sonrasında veri madenciliği tekniklerinin ve algoritmalarının kullanılıp kanser hastalığının bu kapsamda irdelenmesi ve erken teşhisin çıkarılabilmesi ve ayrıca bu algoritmaların performanslarının weka adlı program kullanılarak elde edilen çıktılar doğrultusunda karşılaştırılması hedeflenmiş.

Üzerinde çalışılacak olan Wisconsin veri setinde kanser verileri irdelenecektir. Karar ağacı algoritmalarından olan J48, Bayes ile sınıflandırma yapılan algoritmalarından biri olan Naive – Bayes, regresyon esasında olan algoritmalardan biri olan lojistik şekilde olan regresyon ve örnek tabanlı şekilde sınıflandırma algoritmalarından biri olan KStar biçiminde olan algoritmaları dikkate alınarak oluşan modeller ortaya getirilmiş ayrıca oluşturulan modellerin başarım dereceleri birbirleri arasında karşılaştırılmıştır.

Anahtar Kelimeler: Veri Madenciliği, Eksik Değer Tahmini, Sağlık Verisi, Sınıflandırma, Kanser, Göğüs Kanseri

COMPARISON OF FILLING MISSING VALUES WITH THE BEST FIT OVER CLASSIFICATION ALGORITHMS

Presented By: Çağdaş KEKLİK

ABSTRACT

In the era of information age, data mining is notable as one of the most fundamental machine learning methods. The continuous increase in the computation power and storage capacities of computers leads an increased development in data analytics and data mining resulting several research and methods on the field. The main aim of data mining is to extract valuable knowledge from large amounts of diverse data that can be used in decision making. Data mining can be used in different areas such as predicting future events, describing interesting patterns or clustering similar data elements which gives knowledge that can be used in the decision making process.

While individual data elements have little or no value, when large amounts of data collected together it becomes quite valuable. Valuable information and goal-oriented knowledge can be extracted from this large data through data mining methods. The continuous rise of data production in the world requires efficient data mining tools to control on the huge amounts of data. Therefore, data mining has become one of the most essential parts in medical researches as also occurred in several other fields.

One of the fundamental problems in data mining is to prepare and preprocess the data for the mining operation. In this concept, missing values is an important issue. The collected data may contain some missing fields. As the data contains null values, it is impossible to make any comparison with those values. A possible solution is to fill those missing values with the best fitting value. In this theses, we compare three scenarios where, in the first one we omit all the lines that contains any missing value, in the second one we omit the lines that have missing values larger than a threshold and fill the rest with best fitting values, and in the third case we fill

all the missing values with the best fit. We then compare the success rates of those scenarios using different algorithms and different success metrics.

During those analyses we use a cancer database as test set. Starting from the definition of data mining, we explain some well-known data mining algorithms. Next, we apply those techniques on a publicly available health record data to predict cancer related diseases and provide analysis and comparison of the performances of different methods utilizing a software program named Weka.

In this thesis, the breast cancer related data of the Wisconsin data set is used as the publicly free health record data. For the algorithms, we select J48 algorithm as a decision tree based approach, the Naive - Bayesian method as a Bayesian classification approach, logistic regression method and the K-star algorithm as a sample based classification method. The performance of each test scenario is compared according to accuracy and efficiency metrics.

Keywords: Data Mining, Filling Missing Values, Health Data, Medical Informatics, Classification, Cancer, Breast Cancer.

İÇİNDEKİLER

1 GİRİŞ

Veri madenciliği bilgisayarların gelişmeye ve hızlanmaya başladığı senelerden geçerek şuan ki yeni ve güncel teknolojilerden bir tanesi olma istikametinde çok büyük bir yönelim göstermiştir, gelen verilerin devamlı ve büyük hızda çoğalması meydana gelen veri analizi ihtiyacına paralel olarak çok süratli şekilde gelişmeye başlamıştır. Bu olaylar bu zaman zarfında pek çok araştırma ve geliştirmeyi ardından getirmiştir.

Veri madenciliğindeki teknolojide esas amaç büyük miktardaki verileri işleyebilme tekniklerini kullanarak, içeride kalmış olan değerli bilgileri bulmak, ileriye yönelik çıkarımlarda bulunmak, karar ve destek konularında ise katkılar sağlamak gibi birden çok işlevi vardır. Birden fazla kurumsal uygulamalarda veriler üzerinde çalışılarak karar verebilmeye yönelik adımlar atılmakta ve bu veriler ile ilgili anlamlı bilgiyi üretmeye çalışmaktadır.

Son zamanlarda verilerden çıkartılan çok fazla öz bilgi ile alakadar olunmaktadır. Bilgi teknolojilerinin bu olaya paralel gelişmesi ve günlük hayatımızın her katmanında işimize yarayabilecek duruma gelmesiyle beraber, tüm alanlarda aşırı derecede çok miktarda veri birikmeye başlamıştır. Bu şekilde, birden çok yerin verileri veri tabanlarında tutulmuştur bu yerlerin bazıları şöyle ifade edilebilir; banka, üniversite, okul, tur şirketleri, hastane, devlet kurumları vb.

Eski şekilde yapılan sorgular ve bu sorguları raporlama araçlarının veri yığınları karşısında problem yaşıyor olarak kalması Veri Tabanlarında gerekli olan Öz Bilgi Keşfi adı altında, devamlı ve yeni arayışlara bakmaya insanları yöneltmişlerdir. Veri tabanlarının içerisinde öz bilgi keşfi aşamalarında, öz bilginin modeli olarak geçen keşif kısmının meydana geldiği aşama olarak değerlendirilebileceği gibi bağımsız şekilde oluşan bir süreç olarak da dikkate alınıp incelenmektedir.

Veri madenciliğinde esas amaç, geçmişi analiz ederek gelecekte olabilecek olan olayları tahminine yönelik karar verme şekillerini ve modellerini oluşturabilmektir. İlk olarak müşteri ilişkilerinde veri madenciliği hayat bulmuştur. Veri madenciliği firmalar ile ilgili organize edilmiş hedeflerin elde edilmesinde ve daha birçok alanlarda kullanım alanına sahiptir. Bununla birlikte, bankacılık sektöründe finansal olarak göstergelere ilişkin gizli olan ilişkilerin ortaya çıkartılmasında, pazarlama alanında müşterilerin satın alma davranışlarının belirlenmesinde ve sigortacılık bölümünde risk kapsamında olan vatandaşların örüntülerinin belirlenmesinde veri madenciliği sıklıkla kullanılmıştır. Veri madenciliği sayesinde büyük küçük demeden firmalara fiyatlar bölümünde, üretim planlaması bölümünde personelin performansı hakkında bilgi sağlanmış olur.

Şuanda güncel olarak veri madenciliği firmalar tarafından öncelikle müşteri hakkında bilgi almak amaçlı olarak kullanılmaktadır. Veri madenciliği herkesin de bildiği üzere birden çok sahada kullanılıp çıktılar elde edilebilmektedir. Veri madenciliği günümüzde çok yaygın şekilde kullanıldığından farklı yeni alanlar da bu işlere başlayabilirsiniz.

Veri madenciliği sıkça sağlık verilerinin içerisinde kullanılan yaygın bir yöntemdir. Bu çalışmada metotları işlevlerine göre Kümeleme, Regresyon, Sınıflama ve Birliktelik Kuralları başlıkları çatısı altında dikkate alınmakta ve uygulama bölgeleri ifade edilmektedir. Bu kapsamda hızlı şekilde araç ve metotların ilerlemesiyle iş sektörlerinden ortaya çıkan, olayla alakalı oluşan çeşitli isteklerden dolayı, algoritmaların ve program materyallerinin geliştirilmesi hakkında, hem iş sektörlerinde ayrıca akademik camialarda olayla alakalı çok fazla dikkat meydana gelmiş verilerin devamlı şekilde büyümesi ve algoritmaların karmaşık olmasından ötürü daha iyi sonuçlar almanın yolları araştırılmıştır. Yapılan araştırmadan elde edilen birçok metottan hangisinin diğerlerine göre daha iyi olduğu gibi sorular meydana çıkmıştır.

Sağlık verileri olarak düşünecek olursak veri madenciliği bu alan için hiç kuşkusuz biçilmiş kaftandır. Bu çalışmada veri madenciliği ile ilgili olarak modelleri ve bu modellere göre sınıflandırma, regresyon işlemleri, kümeleme yapılması ve birlikte olma kuralları kategorileri altında değerlendirilmekte ve uygulama alanları belirtilmektedir.

Veri madenciliği günümüzde adını çok duyurmuş olsa dahi literatüre bakacak olursak bu konuya 1980'ler den itibaren bakmamız gerekir. 3 temel başlıkta Veri Madenciliği kavramını ifade edebiliriz.

İkinci olarak söyleyebileceğimiz Veri Madenciliğinde yapay zeka (AI) dır. Amacı sezgisel bir şekilde olan heuristic yaklaşımı esas alarak, insan gibi düşünebilme ilkesiyle, istatistikten çok değişik uygulamalarla, istatistiksel olan sorunlara çözüm getirmeye çalışır. Yapay zeka 1980'ler de pratik uygulamalarda yer edinememiştir.

Üçüncü başlığı makine öğrenmesidir. Makine öğrenme, yapay zekanın istatistiksel yöntemlerle yoğurup evrimleşerek geliştiği çok ileri safhası denilebilir. Makine öğrenme kısaca istatistiksel ve yapay olacak olan zeka algoritmaları kullanılarak mevcut olan verinin en iyi şekilde faydalanılmasına, bunlardan çıktılar elde etmesine ve bu sonuçlardan bir karara varılmasına imkan sağlar.

Temelinde veri madenciliği, öğrenme yöntemlerinin iş ve bilimsel bilgilerin verileri kullanılarak mantıklı ve anlamlı bilginin ortaya çıkartılması işidir. Veri madenciliği, yapay zeka, istatistik ve makine öğrenme yollarının gelişmesiyle meydana gelen, veriden bir şeyler çıkarma ve bunları öğrenme taktiğiyle gizli olan mesajları ve örüntüleri elde ederek geleceğe yönelik tahminler çok fazla olan veri öbeklerinden elde edilmesi mümkün olamayan veriyi elde etmede gün ilerledikçe kabul görmektedir.

1.1.Veri Madenciliği ’nin Kullanıldığı Sektörler Hakkında Kısa Bilgiler

1.1.1.Sağlık ve Biyoloji Sektörü

Bu alanda güncel virüs çeşitlerinin tespit edilmesi ve bunların sınıflandırılması, inceleme sonunda hastalığın çeşit ve özelliğinin belirlenerek teşhislerin kolay hale getirilmesi, tıbbi veriler, yanında alınan ilaçların yan etkilerinin araştırılması, test raporlarının tahmini, ürün geliştirme ve tedavi aşamalarının tespit edilmesi gibi işlemleri kapsamaktadır.

Tıp alanı esasında en çok verilerin saklandığı alanların başında gelmektedir. Sebebi ise ilerleyen teknolojinin yardımlarıyla gen haritaları ile hastalıkları sınıflandırma ve ne çeşit bir gene sahip olduğu bireyin ya da bireylerin bunları tespit etmedir. Hangi hastalıklara yakalanma olasılığı bu kapsamda mümkün olduğuna dair detaylı çalışmaları fazla miktarda gerek gazetelerden gerekse televizyon kanallarından ve internetten bilgi ediniyoruz.

Bu durum gibi virüslerde yapısal olarak incelenerek onlarda kategorilere ve sınıflara ayrılıyor. İlaçların üretimi, kullanımı ve yan etkileri aşamalarında veri madenciliği ile gerçekleştirilebilmektedir.

Sağlık sektöründe bilgide meydana gelen değişiklikler sağlıkta hastaya bakım hizmetinde değişikliğe yol açmıştır. Sağlık bakım hizmetinin hastalara uygulanmasında bilgisayar kullanımı artmıştır. Bu kapsamda bilginin paylaşımı ve ekip yaklaşımını, veri ve bilgi temelli uygulama gibi kavramları hızla yaygın hale getirmeye başlamıştır. Hastaların bilgisayarlar ile takibi ve destekleme, bu bakım hizmetlerinin değerlerinin değerlendirilmesi gibi yardımların yanı sıra bu gibi hizmetlerin sunulmasında kullanılmasının yanı sıra, karar verme yönetim, planlama

ve tıbbi araştırmalar gibi yönetsel ve akademik fonksiyonların yerine getirilmesinde daha fazla kullanılmaya başlanılmıştır.

1.1.2. Telekomünikasyon Sektörü

Hatlar ile ilgili olarak yoğunluğu hakkında tahminlerde bulunma. İnternette siteleri ziyaret edenler ile ilgili olarak profil analizi. Eksik olan bölümleri iyileştirme ve kalite ile alakalı analizleri kapsar.

İletişim çağında yaşadığımız ve önemini her gün arttıran telekomünikasyon sektörünün inanılmaz boyutlara ulaştığı herkes tarafından görülmektedir. Buradan yola çıkarak kişilerin kullanım sıklıklarıyla alakalı bilgi, kişilerin burayı kullanım amaçları ve kişilerin hatlar ile alakalı yoğunlukların tahminleri yapılarak firmalara öneriler öngörüler altyapı güncellemeleri hakkında verilebilir, firmalar müşteriye ilişkin özel kampanyalar oluşturabilirler.

1.1.3.Finans (Borsa, Bankacılık Sektörü) Sektörü

Değişik finans ile alakalı olarak mevcut olan göstergeler içerisinde korelasyon tespiti, bu olayla birlikte kredi kartı ile ilgili olarak dolandırıcılıkların belirlenmesi, krediler ile ilgili olarak bankalara gelen başvuruların incelenmesi, kredi kartı harcamasına göre kişilerin harcama alanında bir çıktı belirlenmesi, sigortalar ile ilgili olarak dolandırıcılık vakalarının belirlenmesinde, poliçe isteğinde bulunacak müşterilerin tahmininde genellikle bu alandan faydalanılmaktadır. (Cengiz Coşkun1, 2008)

1.1.4. Pazarlama ve Sigortacılık Sektörü

Gerek internet üzerinden satışlarda gerek normal satışlarda müşterilerin satışlarının takip edilmesi buna göre kampanyaların belirlenmesi, mevcut potansiyel müşteriyi kaybetmeden yeni müşterilere ulaşma ve bu örüntülerden yola çıkarak satışlar hakkında verilerden çıktı alma hakkında bundan faydalanılır. Alanları en yaygın şekilde veri madenciliği uygulama alanıdır.

Sigortacılık sektörü ile ilgili olarak poliçelerin talep edecek olan müşterilerin belirlenmesi, dolandırıcılıkların tahmin edilmesi ve bunlara karşı önlem alınması, fiyat tablosunun tespitinde ülkenin coğrafi şartlarının ön görülmesi.

1.1.5. Astronomi, Biyoloji, Tıp, Genetik, Kimya Sektörleri ve Alanları

Yeni galaksilerin tespiti, mevcut gezegenler ile ilgili verilerin toplanıp ileriye yönelik incelenen gezegenler hakkında öngörülerde bulunulması, yıldızların pozisyonlarına bakılarak gruplara bölünmeleri ve bunların takip edilmesi sürecinde veri madenciliği büyük rol oynar.

Bitkiler ile ilgili olarak gen haritası çıkartılması ve hastalıkların tespiti, bazı bitkilerin ıslah edilmesi sürecinin tespiti. Genler ile ilgili olarak baktığımızda ise hangi bireylerimizin suç işlemeye yönelik eğilimlerinin genlerden yola çıkılarak irdelenmesi buradan yola çıkarak olayların önlenmesi ya da kullanıcıların yazı karakterlerini inceleyerek birden çok olasılığın hesap edilmesinden çıkın veri madenciliğinin en fazla kullanıldığı alanlardan biridir kriminoloji alanı.

Kimya ile ilgili olarak yeni moleküllerin keşfi ve sınıflandırılması sırasında kullanılması, yeni gelecek ilaçlar ile ilgili türlerin keşfi gibi alanlarda kullanılmaktadır.

1.1.6.Yüzey Çözümlemesi, Coğrafi Bilgi ve Robot Görüş Sistemleri, Görüntü Tanıma

Bölgeler hakkında coğrafi olarak sınıflandırma yapılması, kentleri inceleyerek yerleşim yerlerinin tespit edilmesi, kentlerin içerisinde gerçekleşen suç oranlarının tespiti aşaması, yoksul olma oranı, zengin olma oranı, kişilerin kökenlerinin tespit edilmesi aşaması, otobüs duraklarının yerleştirilmesi düzeni, bankaların ATM para makinelerinin yerleştirilme düzenlerine kadar birçok konuda veri madenciliği kullanılmaktadır.

Algılayıcılar yardımı ile görüntülerden yola çıkılarak engel tanıma, yolu tanıma, yüzleri tanıma kişilerin kimliklerini belirleme, parmak izini tanıma ve kişilerin kimlikleri ile ilgili bilgi edinmede kullanılmaktadır.

1.1.7.Meteroloji, Atmosfer, Sosyal ve Davranış Bilimleri

Bölgelere bağlı olarak değişen iklimsel farklılaşmalar, yağışlar hakkında tahmin haritaları yapma, hava tahminleri yaparak bunlar ile ilgili doğal afetlerin olabilecek yerler ile ilgili uyarılarda bulunup yetkililerin önlem almalarını sağlama ayrıca çeşitli okyanus hareketlerinin takibi ve belirlenmesinde kullanılmaktadır. Ayrıca, kamuoyu içinde inceleme yapma, seçimler hakkında öngörüler tespit etme gibi birçok alanda da kullanılmaktadır.

1.1.8.Metin ve İnternet Madenciliği

Büyük ve ilk bakışta bir anlam veremediğimiz metinler içerisinden anlamlı ilişkiler elde etmekte kullanılmaktadır. Sadece düz mevcut olan metinler dışında resimden daha değişik başka ilerleyen (streaming) ve sayısal biçimde olan veriler hakkında verilerde web verileri içerisinde mevcuttur. İnterneti belli başlı kategorilere ayırıp bunlardan istenilen veriye ulaşmak web madenciliğinde asıl hedeftir.

2 KURAMSAL ALTYAPI

Veri Madenciliği alanında bilgiyi elde etmede çok farklı yollar kullanılmaktadır. Bu yollara ait birçok algoritma mevcuttur. Bu algoritmalardan hangisi hangisinden iyidir bunlar üzerinde birden fazla araştırma yapılmıştır, yapılan çalışmalar ışığında çok farklı sonuçlara ulaşılmıştır. Bu durumun sebebi ise, işlem başarısının, veri kaynağına, veri üzerinde gerçekleştirilmiş ön işleme, algoritma seçimine bağlı olmasıdır. Farklı veriler ve farklı parametreler belirlenerek yapılan işlemlerden farklı sonuçların çıkması çok doğaldır. Lakin, benim yapmış olduğum çalışmada "benzer veri kümelerinde hali hazırda mevcut olan yöntemlerin daha başarıya ulaştığı" şeklindeki çıkarıma (D. Michie, 1994) uygun olacak biçimde, başka çalışmalarla (Cengiz COŞKUN, 2004) (Abdelghani Bellaachia, 2003) benzer sonuçlar vermiştir. Göğüs kanseri olaylarının değişik yıllarını kapsayan Wisconsin veri kümesi bilgilerinin işleme alındığı çalışmada, karar ağacı algoritması olarak ifade edilen C4.5 algoritmasının diğerlerinden çok güzel çıktılar elde edildiği çıktısına ulaşıldığı belirtilmiştir (Cengiz COŞKUN, 2004) Weka tanımlaması olan J48 karar ağacı algoritması, benzer şekilde diğerlerine göre daha çok başarılı olduğu tespit edilmiştir.

Diğer tez çalışmaları sonucunun ise kullanıcının yatkın olduğu modele bağlı olarak değiştiği, bu yüzden güvenilir şekilde bir objektif sonuca ulaşılamadığı görülmüştür. Bunların haricinde bazı bir takım çalışmalarda karmaşık algoritmaların diğer klasik algoritmalara karşı çok daha iyi performansta olduğu şeklinde ortaya atılan iddiaların ise gerçekte birer hayal ürününden ibaret olduğu açıklanmaktadır (Hand, 2006).

Deney ile yapılan bu çalışmalara gelen eleştirilerin doğruluk oranı gayet yüksektir. Buradan yola çıkılarak gerçekleştirilecek olan bir karşılaştırma işlemine bakılarak bir algoritmanın bir diğer algoritmaya üstünlüğünden bahsetmek kesinlikle doğru bir yaklaşım olmayacaktır. Ancak model hakkında ulaşılan başarı ile alakalı olarak yapılan karşılaştırmaların, bir veri madenciliği alanındaki çalışmalarda önemli katkıları görüleceği çok net ve açıktır. Bir kullanıcının bir problem üzerinde gerçekleştireceği modeli oluşturma işleminde çok farklı algoritmaları toplayarak ve bunları karşılaştırarak en iyi ve en başarılıyı tespit etmesinin ve biçimini o algoritma ile oluşturmasının sonuçları içerisinde pozitif bir göstergesi olacağı açıktır. Lakin, dikkatin verilmesi gerekli olan asıl nokta öğrenme kümesinin belirlenmesidir. Nedeni ise farklı öğrenme kümeleriyle gerçekleştirilen farklı karşılaşmalar çok farklı sonuçları bize sunabilir (Hand, 2006). Bunun yanında yeni geliştirilen algoritma ya da algoritmaların bilimsel açıdan geçerliliğinin belirlenmesinde deneysel verilerin çıktıları bu tespitte önemli bir yer tutmaktadır.

2003 yılında gerçekleştirilen bir çalışma neticesinde 1991 yılında alınan Meme Kanseri Wisconsin bilgileri içerisinde mevcut olan göğüs kanseri olan hasta bilgileri kullanılarak işleme alınan çalışmada eğiticisi olan, eğiticisi olmayan nöral algoritmalar göğüs kanseri tespitini belirleme hedefiyle karşılaştırılmıştır. RBF ve SOM eğitme setlerinde RBF eğitme setindeki en güzel sınıflandırıcı olarak tespit edilmiş .Lakin, SOM ise en iyi sınıflandırma sonuçlarını vermektedir. Dışarıdan genel olarak bakıldığında WBCD verisi sınıflandırılması aşamasında en güzel nöron ağı modeli olarak RBF ve SOM tespit edilmiştir. Bununla birlikte sonuçlar eğiticisi

olan ve eğiticisi olmayan algoritmaların göğüs kanseri tespitinde çok iyi sonuçlara ulaştığını ortaya koymuştur (Tüba Kıyan, 2003).

2007 yılında yazılan bir tezde yapılan araştırmada bir grup sağlık verisinin veri madenciliği ile incelenmesi gerçekleştirilmiştir. Tezin asıl amacı en yaygın veri madenciliği algoritması kimliğini araştırıp tespit etme. Bunu yaparlarken 8 tane algoritmadan faydalanılmıştır. Bunların isimleri kısaca ; J48, Karar Ağacı (Decision Tree), Karar Destek Makinesi (Support Vector Machine), Radyal Temel Fonksiyonu (Radial Basis Function), Çok Katmanlı Perceptron (Multilayer Perceptron),K-En Yakın Komşu (K-Nearest Neighbors), Naive Bayes,Bayes Ağı ve Lojistik Regresyon. (Kurum, 2007). Birden çok farklı hastalık verileriyle çalışılarak yapılan çalışma neticesinden en iyi olan algoritmanın tespit edilmesi hedeflenmiştir.

2012 yılında yazılan bir makalede Göğüs Kanseri Teşhisinin üç tane farklı veri seti kullanılarak çoklu sınıflar eşliğinde tespit edilmesi amaçlanmaktadır. Bu makalede kullanılan veri setleri Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC) ve son olarak Wisconsin Prognosis Breast Cancer (WPBC) şeklindedir. (Gouda I. Salama, 2012).

2013 yılında yazılan bir tezde ise Zaman Serilerine Dayalı Salgın Tespit Algoritmalarını Karşılaştırma Amaçlı Bir Yazılım Aracının Tasarım ve Gerçekleştirimi adlı bir tez yazılmıştır. Burada amaç salgın ile alakalı olarak tespit mekanizmalarını bulmaya yönelik çalışmalar üzerinde yoğunlaşmaktadır. (Şahin, 2013)

2.1.Veri Madenciliği İle İlgili Genel Bilgiler

Veri madenciliği devasa büyüklükteki veri kaynaklarında saklı, çok fazla öneme sahip ve faydalı olan bilgilerin bilgisayar kullanılarak ortaya çıkartılmasıdır. Veriler kapsamındaki benzerliklerin, örüntülerin ya da birbirleri aralarındaki ilişkilerin belirlenmesi hedefiyle uygulanan işlemlerin bütününü meydana getirir. Veri madenciliğinin ekonomik alanda araştırmaları, müşteri profilinin elde edilmesi, sepet analizi; risk analizi bankacılıkta olan, sahtekarlık tespiti; bilişimde web verilerinin analiz çalışmaları, ağ güvenliği, belgelerin sınıflandırılması gibi uygulamaları vardır. Bütün bu olayların dışında meteoroloji sektöründe, tıp sektöründe, temel bilimler sektöründe, ilaç bilimi sektöründe ve farklı sektör alanlarında da çalışmaları vardır. Veri madenciliği günümüzde yeni bir sektör olsa da, esasında çok önceleri ekonomistler, istatistikçiler, hava raporu tahmin edenleri, eldeki mevcut olan verileri kullanarak ileriye yönelik çıkarımlar yapmaya çalışıyorlardı. Son yıllarda veri miktarlarının aşırı derecelerde büyümesi, çok farklı verilerin daha farklı algoritma ihtiyacı, bu gelişen durumun kendi kendini yettirebilme ihtiyacı gereksinimine neden olmuştur.

İlerleyen teknoloji ile veriler her gün artmakta, öncelerde kilobaytlarla ifade edebildiğimiz bilgisayarlarımızdaki verileri artık megabaytlar, gigabaytlar ile açıklanmaya çalışılmaktadır. Eskiden dikkate alınmayacak küçüklükteki veriler şimdi gelişen teknoloji ile depolanmakta ve analiz çalışmaları vb. çalışmalarda kullanılmaktadır. Günlük olarak kullandığımız bankacılık işlemlerinde, online sistemler, internet ortamının genişlemesiyle, bilgiye ulaşma olayı ve bilgiyi bir yerden bir yere aktarma olayında bir faktör olmuştur. Olayla alakalı gerçekleştirilen araştırmalara göre veri miktarı her 20 ayda bir çifte katlanarak çoğalmaktadır. Bu nedenden ötürü bu denli şekilde devasa olan verilerin içerisinden beklentileri karşılayacak olan istenilen anlamlı veriyi çekip almak için algoritmalar geliştirilmiştir. Farklı tarzda olan veriler için farklı tarzda algoritmalar geliştirilmesine sebep olmuştur. Büyük şirketlerin, okulların, üniversitelerin, hastanelerin, bankaların, kamu kurumlarının elleri altında bulunan veri bankaları çok devasa verilerden meydana gelmektedir. Bu veriler işlenerek bir plan çizmeye, geleceğe yönelik sistem tahminlerinde bulunmaya geleceği tahmin edebilmeye ve

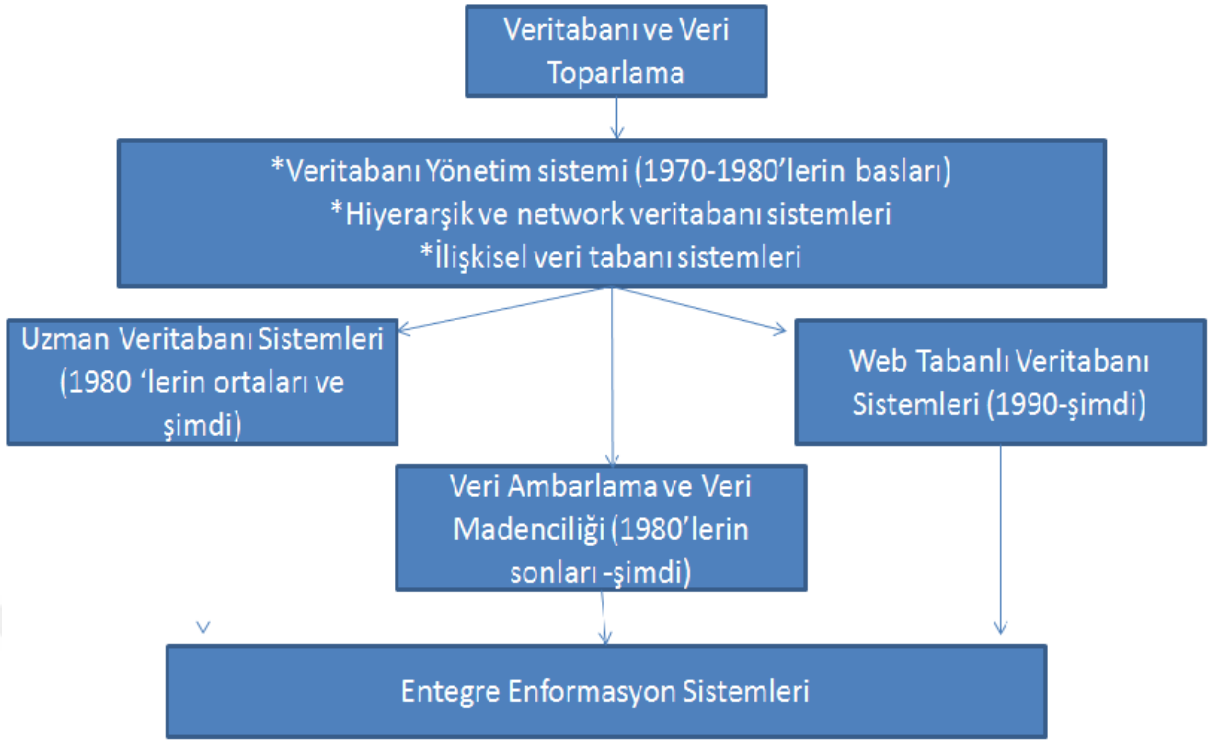
ciddi kararlar alacak mekanizmalarda önemli rol oynarlar. Veri madenciliği aslında çok devasa büyüklükteki karmaşık ya da düzenli biçimde olan verilerin analiz edilmesi için yapılması gerekli olan aşamaların tamamını içerir.

2.1.1.Veri ve Veri Tabanı Teknolojisi

Verileri yönetme sistemleri aracılığıyla büyük ve karışık olarak gördüğümüz verilere erişmek çok kolaydır. Verilerin toplandığı dosyalar, veri miktarları genişledikçe aradığımız veriye erişmede zorluklar yaşamaktayız. Bu olayla birlikte kişi ya da kişilerin aynı olan bilgileri birbirini etkilemeden kullanmasına da erişim sağlamaktadır.

Ortak nokta olarak veri madenciliği ile ilgili büyük miktarda verilerden bahsetmesidir. Büyük miktarda verinin temeli ise veri tabanlarına dayanmaktadır. Veri ile ilgili disiplin mekanizmasının gelişmesinde veri tabanı teknolojisinde gerçekleşen gelişmelerin önemi oldukça büyüktür. Veri tabanı esasında ilişkili verilerin tekrar olasılığına yer vermeden birden çok şekilde kullanımına imkan sağlayacak tarzda depolanması şeklinde ifade edilebilir. Buradan çıkarılacak olan şey, veri tabanı bir veri kümesi olmakla birlikte, kullanıcıların ihtiyaçlarına göre sınıflandırma yapılması gerekmektedir. Ayrıca raporlanma ve analiz edilmesi olayları gerçekleşmesi gerekmektedir. Bu durum veri analizi yardımıyla elde edilmektedir. (Babadag, 2006)

Üst kısımda ifade edilen veri tabanı gelişimi ve bu gelişmeye dahil olan veri madenciliği rolü Şekil 1'de gösterilmiştir. (Jiawei Han, Data Mining: Concepts and Techniques, 2001). Şekil 1'de veri tabanı gelişim sürecinin bir ürünü olduğu ifade edilmektedir.



Şekil 1 Veri tabanı Teknolojisinin Gelişimi ve Veri Madenciliği

2.1.2.Veri Tabanları İçerisinde Bilgi Keşfi Aşamaları

Veri tabanlarında bilgi keşfi kısaca VTBK (KDD-Knowledge Discover in Databases) işleminin ana bileşenlerinden bir tanesidir. VBTK sadece veri madenciliğinden oluşmamaktadır.

a. Veride Ön İşlenme Aşaması

Birim olarak düşünürsek, nitelikler bilgileri, sayısal verileri, nominal verileri ya da katarları şeklinde ifade alabilirler. Bu aşamada birincil olarak veriler içerisinde bulunan gürültüler, tutarsızlık bölümleri ve düzensizlik arz eden kısımlar giderilir. Verilerin analizimiz aşamasına gelecek yapıya dönüştürülmesi işlemine veri ön işlemesi denir.

Bu aşama esasında bir veri madenciliği çalışmasının çok kapsamlı bir bölümünü içine dahil eder ve analizin çok doğru çıktılar elde etmesini ve çok etkili şekilde uygulanmasında da büyük öneme sahip olur.

Bu aşama veri madenciliğinde ilk ve en uzun aşaması olarak görülür. Verileri temizleme, verileri birleştirme, verilerin dönüşümü ve verilerin azaltılması işlemlerini içine almaktadır.

b. Veride Temizleme Aşaması

Kullanıcı hatalarından oluşan hatalar, program hatalarından oluşacak hatalar, bazı otomatik hale getirilebilecek işlemleri kullanıcıya bırakmayın, verinin girişinin önemsenmemesi tarzında sebepler ile veri kümelerinde tam olmayan ya da gürültülü veriler meydana gelebilirler. Veri üzerinde olan bir takım nitelikler yanlış değer taşıma, eksik, geçersiz veriler de olabilir.

Bölme işleminde mevcut olan eldeki veri sıralanarak eşik olabilecek bölümlere parçalanması ve her kategorinin kendisine ait çıktılar ya da ileri derecede verilmiş sonuçlarla açıklanmasıdır. Bu şekilde bilgilerdeki problemli bölümlerin sayılarının en aza indirilmesi hedeflenir.

c. Veride Birleştirme Aşaması

Farklı kaynaklardaki verileri aynı olan başlıkların elde edilen farklı değerleri, ölçü olan birimleri ya da bu kısma bağlı olan derecelendirmeler kullanılmış olabilir. Farklı kaynaklarda aynı olacak şekilde farklı niteliklere sahipmiş gibi dikkate alınmış olabilir ya da birleştirme neticesinde elde edilen sonuç neticesinde şekil itibarıyla veriler içerisinde gereksiz olanlar ortaya çıkabilir. Bu türde verilerin tespit edilmesi, lüzum olmayan verilerin bir kenara konulup ayrılması gerekir.

d. Veride Seçim ve Dönüştürülmesi Aşaması

Bazı veriler uyumlu olarak algoritmada kullanılırlarken bazı veriler kullanılamayabilirler. Bu tip tarzda veriler algoritmaya uygun olmayabilir ya da verilerde bulundukları nitelikler belirleyici olmayabilir. Bu tarzda verilerin dönüşümü yapılarak nitelikleri algoritma ile uyumlu olabilecek hale getirilir ve nitelikleri daha fazla belirleyici olacak biçimde dönüştürülebilir.

Verinin seçim ve dönüştürülmesi aşamasında aşağıdaki yöntemlerden yararlanılır;

* Veri madenciliği çalışması hakkında bilgi tespit edilmesi.

* Madencilikte üzerinde çalışılacak olan veri türünün tespit edilmesi aşaması.

* Verilerin aralarında hiyerarşik düzen ve genellemelerin tespit edilmesi.

* Madencilik sonunda elde edilecek bilgi elde etmek için yenilik ve ilginç olma metodlarının belirlenmesi.

* Veri madenciliği tespit edilecek veri için olası sunumun görselleştirme araç ya da araçlarının tespit edilmesi.

e. Veride Verilerin Azaltılması Aşaması

İncelenecek olan verinin devasa derecede olması, yapılması düşünülen algoritmanın daha çok zaman harcayarak tamamlanmasına ve esasında sonucu kesinlikle etkilemeyecek düzeyde gereksiz işlemlere sebebiyet verir. Veri ön işleme bölümünde uygulanması gereken olayda gerek görülmeyen bilgilerin yok edilmesi, birleştirilmesi ya da diğer metotlarla daha anlamlı şekilde daha verimli durumda bir hale.

Sorun alanında ki veriyi işleme alarak yapabileceği şeklinde istatistiksel metotlarla, karar ağaçlarıyla ya da bilgi elde edinimi sonuçlarıyla belirlenebilir. Verilerin sıkıştırılmasında ki esas amaç verilerin küçük hale getirilerek daha fazla verinin saklanması ve veri erişimimin en yüksek düzeyde hızlandırmayı amaçlar.

f. Örüntü Değerlendirme Aşaması (Pattern Evaluation Stage)

Bu aşamada tespit edilen ilginç durum ölçüm metotları kullanılarak veri madenciliği ile ilgili bulunana verilerin ne kadar farklı durumu içinde barındırdığı ya da yararlı duruma sahip olduğu belirlenir.

g. Bilgi Sunumu Aşaması (Knowledge Presenation Stage)

Birden farklı şekilde tespit edilmiş ve rapor haline getirme araçları kullanılarak bulunmuş olan veriler ilgili kullanıcılara sunulurlar. Veri tabanlarında bilgi keşfi ne doğrusal bir işlem nede tek aşamalı işlemdir. Bu süreç düzgün çıktı elde edilene kadar denenmesi işleme alınması gereklidir. Veri madenciliği aşamasına odaklanılmakta lakin diğer tüm aşamalar Veri Tabanlarında Bilginin Keşfi işleminin birlikteliği yönünden çok az olacak şekilde veri madenciliği olacak bölüm kadar önemlidir. (Fayyad & Piatetsky-Shapiro)

3 VERİ MADENCİLİĞİNDE TEKNİKLER

Veri madenciliği iki temel kategoride altında incelenmektedir.

*Öngörüsel(Predictive)

*Tanımlayıcı(Descriptive)

Öngörüsel veri madenciliği tekniği verileri ileriye odaklı çıkarımlar yapma, sonuç elde etme esasına dayalı olacak olan işlemlerde kullanılırlar. Tanımlayıcı veri madenciliği tekniği ise, veri tabanındaki verinin karakterini tespit, mevcut durumu anlamaya çıkarmaya yönelik metotları öne getirir. Bunların içerisinde en çok kabul edileni J.Han'ın savunduğu kategorilerdir. (M, 2001) Ancak aşağıdaki başlıklar şeklinde ifade edilmiştirler.

- Tanımlama ve Ayrımlama Bölümü
- Birliktelik Analizi Bölümü
- Sınıflandırma ve Öngörü Bölümü
- Kümeleme Analizi Bölümü
- Sıra dışılık (istisna) Analizi Bölümü
- Evrimsel Analiz Bölümü

3.1.Tanımlama ve Ayrımlama Bölümü

Veriler aslında göstermiş oldukları ortak olan detaylara bakılarak aralarında kategorilere bölünme imkanı vardır. Bir banka firmasının alışveriş yapan müşterilerini sınıflandırması Profesyonel Müşteri ve Genel Müşteri şeklinde ifade edebilir. Bu şekilde olan ifadeler veri kümesinin elemanlarının birlikte olan benzerliklerini ya da farklılıklarını gösterecek şekilde yapılabilmektedir.

3.1.1.Tanımlama Yapma Bölümü

Kısaca açıklamak gerekirse veri kümesi içerisindekilerin genel olan özelliklerini kısaca açıklamak hedefli kullanılır. Örnek vermek gerekirse mağazada olan %60 ile %80 arasında olan kişilerin alışveriş yapmadan çıktığı yapılan çalışmalar sonucunda tespit edilmiş. Buradan yola çıkılarak bir alışveriş merkezinde satışı yükselen mallar açıklaması bir Tanımlama olarak söylenebilir.

3.1.2.Ayrımlama Yapma Bölümü

Veri kümelerinin birbirleri arasında mevcut olan farkları ortaya koyma işlemidir. Konuyla alakalı olarak kısa bir örnek vermek gerekirse Türkiye'de trafik kaydı yeni olarak yaptırılan %58,8 sonucunda dizel olan araçların yüzdesi ve %41,2 sonucunda benzinli araçların yüzdesi olma durumların karşılaştırılması biçiminde ayrımlama yapabilme olayın veri madenciliğidir. Her iki şekilde olan veri madenciliği yöntemleri aralarında fazla derecede yakın metotlar kullanırlar. Her iki yolla elde edilmiş olan çıktılar pasta şeklinde olan grafiği, sütun şeklinde olan grafiği, eğriler ve birden fazla boyutlu küpler şekil ve grafikleriyle ifade edilirler.

3.2. Birliktelik Analizi Yapma Bölümü

Verilerin kümesinde kendi başına gerçekleşen ya da devamlı olacak biçimde gerçekleşen, bir arada ya da aynı zaman içinde satın alınma, yapılmış olanlar şeklindekileri tespit etme tabanına dayanır. Bu yol bankacılık kullanılan işlemlerdeki analizde ya da pazar sepeti incelemesi şeklinde herkesçe bilinen biçimde gerektiğinde kullanılmaktadır. Pazar sepeti analizini ifade etmek gerekirse kısaca, bir alışveriş esnasında veya birbirini takip eden alış veriş olaylarında müşterinin ne çeşit mal veya hizmetleri aldıklarına yönelik bulgular tespit edilmesiyle müşteriye fazla ürün satışının yapılması hedeflerindendir. (Akpınar, 2000).

Pazarda olan sepet analizine örnek vermek gerekirse müşteriler bira şişesi ya da bira kutusu satın aldığında yüzde yetmiş beş olasılıkla cips gibi farklı ürünlerde satın alabilirler biçiminde bir çıktı meydana gelebilir. Bu işlemin neticesinde bira ile cips aynı bölgedeki raflara yerleştirilebilir ya da cips alan kişi ya da kişilere birada indirim yapılacak şekilde kampanyalar oluşturularak satışlar yükseltilebilir.

Bu şekilde müşteri profili de bir nevi çıkarılmış olmaktadır. Örnek vermek gerekirse bankaya ait olan bir kredi kartı kayıtları dikkate alınıp bakıldığında yaşları yirmi ile yirmi dokuz arasında farklılık gösteren müşterilerinden, gelirleri 3000 -

4000 Türk Lirası sınırlarında olan müşterilerinin masaüstü bir pc satın almaları ile ilgili harcamaları görülmüştür. Bu kural, birlikte olma analizi metoduyla şu şekilde açıklanabilir;

Yaş (X , ”yirmi .. yirmi dokuz”) ^ Gelir(X, “üç bin…. dört bin”) alır (X,”pc”)

3.3.Sınıflandırma ve Çıkarım Bölümü

Sınıflara ayırma aşaması insanların düşüncesine müsait olan veri madenciliği yöntemidir. İnsanlar etrafındaki objeleri ayrıca meydana gelen olayları daha iyi ifade edebilmek ayrıca başkalarına açıklayabilmek için nerdeyse her şeyi sınıflandırma yönelimindedir. Örneğin, insanları tavırlarına ve hareketlerine bakılarak, hayvan çeşitlerini türlerine bakılarak, evleri görünüş tiplerine bakılarak kategorize edilip sınıflandırılmaktadır.

Veri madenciliğinde sınıflandırma işlemi, elde olan verileri önceden tespit edilen bir özelliğe bakılarak kategorilere bölmek ve yeni ilave edilecek verilerin hangi kategoriye dahil olacağını tespit edebilme olayıdır. Başka ifadeyle, yeni görülen bir girdinin hangi kategori ve kategoriye ekleneceğini tespit etme olayıdır.

Sınıflandırma olayına örnek vermek gerekirse bankaların kredi başvurularını düşük, orta ve yüksek riskli şekline uygun kategoriye ayrılması, bir okulda sıfır olacak şekilde başlayan öğrencilerin hangi sınıfta başlayıp eğitimini sürdüreceğinin tespit edilmesi örnek olarak verilebilir.

İleriye yönelik tahmin işlemi esasında kategorilere ayırma olayında fazla aynı şekilde benzer. Lakin ileriyi tahmin etme işleminde sınıflandırma, gelecekle alakalı yapılan tahminlerin belirli bir davranışa ya da belirli bir sonuca bakılarak işleme alınır. İleriye yönelik tahmin olaylarında yapılan kategori sisteminin doğru şekilde yapılıp yapılmadığını test etmenin tek çaresi “bekleyin ve görün” kuralıdır. (Jiawei Han, Data Mining: Concepts and Techniques, 2001)

Sınıflama kategorik değerlere çıkarımda bulunurken, regresyon süreklilik sağlayan değerlerin tahmin edilmesinde kullanılır (Jiawei Han, Data Mining:

Concepts and Techniques, 2001). Örneğin, bir sınıflama modeli banka kredi uygulamalarının güvenli veya riskli olmalarını kategorize etmek amacıyla kurulurken, regresyon modeli geliri ve mesleği verilen potansiyel müşterilerin bilgisayar ürünleri alırken yapacakları harcamaları tahmin etmek için kurulabilir. Sınıflama ve regresyon modellerinde kullanılan başlıca teknikler şunlardır (Jiawei Han, Data Mining: Concepts and Techniques, 2001): Karar ağaçları, k-en yakın komşu, bellek temelli nedenleme, naive bayes mevcut olmakla birlikte bu tez kapsamında biz Karar Ağaçları (Decision Trees) ve Naive Bayesian sınıflandırma yöntemleri içerisinden yararlandık.

3.3.1. Karar Ağacı Bölümü

Karar ağaçları, veri madenciliğinde kuruluşlarının ucuz olması, yorumlanmalarının kolay olması, veri tabanı sistemleri ile kolayca entegre edilebilmeleri ve güvenilirliklerinin iyi olması nedenleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahip tekniktir.

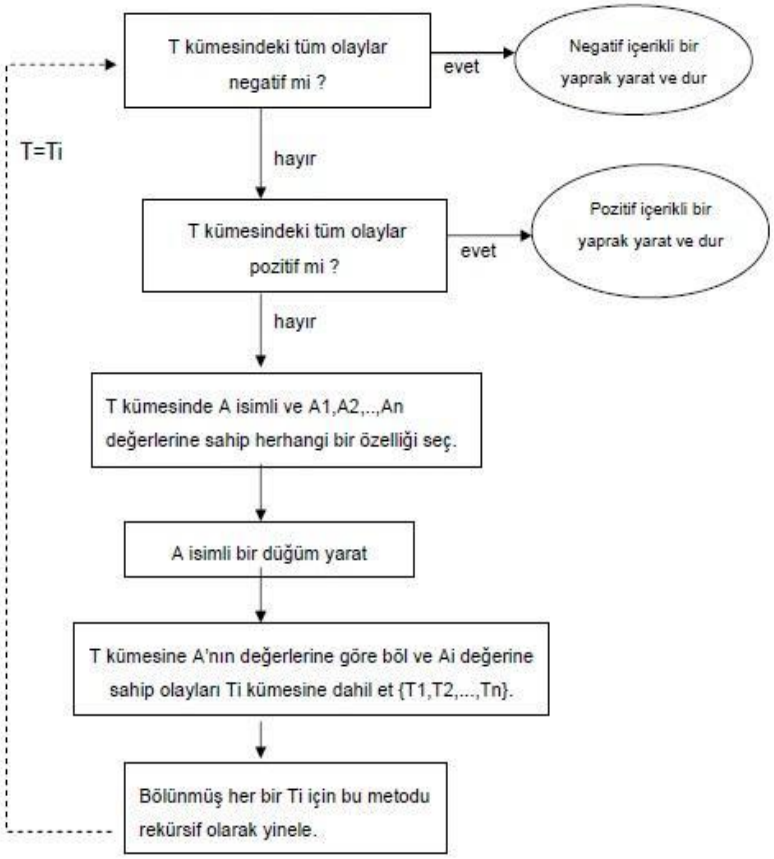
Karar ağacı, adından da anlaşılacağı gibi bir ağaç görünümünde, tahmin edici bir tekniktir. Ağaç yapısı ile kolay anlaşılabilen kurallar yaratabilen, bilgi teknolojileri işlemleri ile kolaylıkla uyum sağlayabilen en popüler sınıflama tekniklerinden biridir. Karar ağacı karar düğümleri, dallar ve yapraklardan oluşur.

3.3.2.Karar Ağacı Meydana Getirme Bölümü

Ağacın meydana getirilmesine yönelik olarak farklı ağaç oluşturma yöntemleri mevcuttur. Ağacı meydana getirmede en önemli nokta belli özelliklere yönelik toplanmış, güvenilir ve yeterli miktarda olay örneklerinin mevcut olmasıdır. Bu iki gereklilik ağaç oluşturmanın temelini meydana getirir. Ağaç meydana getirmedeki en önemli kısım böl ve ele geçir sekmesidir.

3.3.2.1.Böl ve Ele Geçir Bölümü

Görülen yöntem Hunt'ın uygulayıp kullandığı bir metottur. Bu metotta örnek uzay T ve sınıflar "+" ve "–" olsunlar. Gelinen aşamada karar ağacı oluşturma Şekil 2'de görüldüğü şekilde meydana gelecektir.



Şekil 2 Hunt'ın ağacı meydana getirme metodu

Bu görülen algoritma esasında en temel ağacı yaratma ile ilgili algoritmadan oluşmuştur. Algoritmanın ileri götürülmesine yönelik olarak hali hazırda çalışmalar sürmektedir.

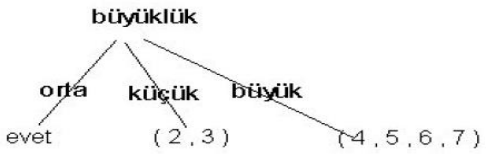
Örneklendirmek gerekirse, maddenin bizler tarafından uyumlu mu değil mi incelemek istediğimizi varsayalım. Bu maddenin şekil, renk son olarak büyüklük gibi özelliklere sahip olsun ve yedi tane örnek olay meydana gelsin. Bu türde örnekler

evet ya da hayır olarak ikili sınıflandırılmış şekilde olacak biçimde Şekil 3'de gösterilmektedir.

Şekil 3 Bir Olay Kümesi Örneği

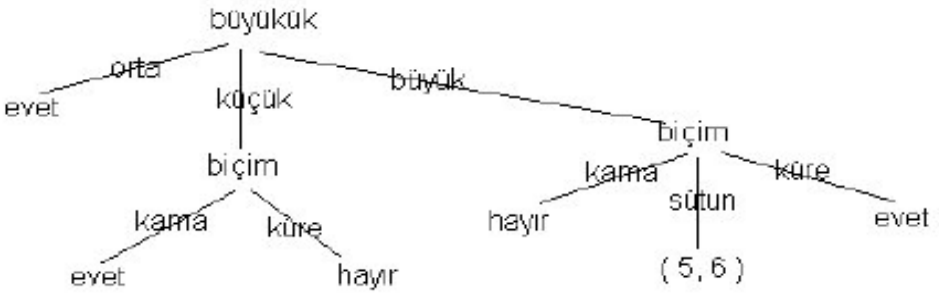
	Büyüklük	Renk	Biçim	Sonuç
1	Orta	Mavi	Tuğla	Evet
2	Küçük	Kırmızı	Kama	Hayır
3	Küçük	Kırmızı	Küre	Evet
4	Geniş	Kırmızı	Kama	Hayır
5	Geniş	Yeşil	Sütun	Evet
6	Geniş	Kırmızı	Sütun	Hayır
7	Geniş	Yeşil	küre	Evet

7'den 1'e kadar sıralı şekilde örnekler gelişi güzel şekilde seçilen özelliği ile alt kısımdaki kümelere bölünmüş olsun. Şekil 4'de gösterildiği şekilde büyüklüğün muhtemel üç tane değeri ortaya çıkar ve üç tane dal meydana gelir.



Şekil 4 Büyüklük niteliğine bakılarak sınıflara ayrılma aşaması

Bu aşamada büyüklük = küçük dalına ve büyüklük = büyük dalına yönelik olacak şekilde aynı işlem meydana getirilsin. Bölme işlemi gene rasgele şekilde biçim özelliğine göre yapıldığı takdirde Şekil 5 'de meydana gelen ağaç önümüze çıkar.



Şekil 5 Büyüklük niteliğine göre ayrılmış ağacın biçim niteliğine göre tekrar ayrılması sonucu meydana gelen durum.

3.3.2.2.ID3 Algoritması Bölümü

Hunt'ın ortaya çıkardığı çalışmadaki büyük eksiklik özelliklerin gelişi güzel seçilmesi olayıdır. Halbuki bu tercih sırasında belirsizliği en aza indirgeyen özelliği dikkate alırsak oluşacak olan karar ağacı o ölçüde sade ve anlaşılır biçimde olacaktır.

Örnek verecek olursak X düğümü bölümünde 5 pozitif ve 3 negatif olay meydana geliyor. Bu aşamadan sonraki noktada yapılacak bir sınıflandırmanın işleminin pozitif olma olasılığı 5/8 negatif olma olasılığı ise 3/8'dir. İşte bu ortaya çıkan olasılık şeklinde sınıflandırmayı meydana getirme yeteneğinin açıklaması şudur: Doğru şekilde sınıflandırma yapılmış bir örneğin ifade ettiği mesajın içeriği şimdi hesaplanabilir hale getirilmiştir.

Öyle ki tablodaki sonuçlar mesaj şeklinde olsun ve bu eldeki mesajlar iki değeri içersinler. Bu değerlere ilave olarak p bilgisinin pozitif şekilde olma olasılığını, q bilgisinin negatif biçimde olma olasılığını ifade eder. Bu mevcut iki çıktının genel olarak toplamı zaten 1 (p + q) olmak şarttır. Doğru şekilde kategoriye sokma ile elde edilen herhangi mesajın bilgilerinin içeriği aşağıda verilen formül yardımıyla bulunur.

$$I\,(p,n) = -\,p\ \log_2\ p\ -\ q\ \log_2\ q$$

Bu formül esasında bilgi içerik formülünün daha özel hale getirilmiş şeklidir. İki adet olasılık hali hazırda mevcuttur. Bunlara değinilecek olursa: "+" ve "-" $\{A_1,A_2,..,A_n\}$ değerlerini barındıran A özellikli ağacın bölümlere ayrılması için kullanıldığında, T kümesi $\{T_1,T_2,….,T_n\}$ şeklinde ifade edilerek ayrılacaktır. Yapılan ayrımlar da T kümesin içerisindeki A özelliğinin A_i olduğu yerlere T_i söylenirse. "+" biçimde olan olayların rakamını p_i ifade etsin dersek, "-"biçimde olan olayların rakamını n_i ifade etsin dersek. Sonuçta yapılan olayda T_i alt kısımdaki ağacı için gerçekleşmesi beklenen bilgi ihtiyacı ise $I(p_i,n_i)$ şeklinde olur. T ağacı için çıkması istenen bilgi ihtiyacı bütün T_i şeklinde olan ağaçlarının içinde beklenen bilgi ihtiyaçlarının ağırlıklı olacak olan ortalamalarının toplamı sonuç olarak ifade edilir ve aşağıdaki şekilde açıklanır.

$$E(A) = \sum_{i=1}^{n} \frac{p_i + n_i}{p+n} I(p_i + n_i)$$

Bu kapsamda A özelliğine bakılarak bunun üzerinden elde edilen bilgi Bilgi kazancı(A) = I(p,n) - E(A) biçiminde açıklanabilir.

Bilgi ihtiyacı ve bilgi elde edilmesi ID3 algoritmaları için 2 tane olan ve çok önem arz eden özelliktir. Tespit edici olacak şekilde sınıflandırma için bilgiye gereksinime gerek olan aslında doğru sınıflandırmayı elde edecek mesajın bilgi detayından başka bir olay değildir. Buna kapsama yönelik şekilde gerçekleştirmesi amaçlanan karar ağaçlarının ana amacı doğru soruları tespit edip sorması. En son bölümde öyle bir pozisyona ulaşılmalı ki, bu durumun karar için bilgi gereksinimi 0 olsun. İşte bu bölümde ID3 algoritmasının ortaya çıkardığı iş, ağacı güzel düzgün ve doğru meydana getirmektir. Meydana getirilen karar ağacının her aşamasında geriye bilgi ihtiyacı minimize yapılır. (Tüba Kıyan, 2003)

3.3.2.3.C4.5 Karar Ağacı Eğitme Algoritması Bölümü

ID3 algoritmasında birtakım noksanlıklar, problemler mevcuttur. Bu durumlar alt kısımda ifade edilmektedir. Bu problemler tekrar Quinlan'ın meydana getirdiği C4.5 algoritmasıyla çözüldü. C4.5 Algoritması ID3 algoritmasının tüm detaylarını kendi içinde tutarak meydana gelmiş bir algoritmadır. Üst kısımda anlatılan bütün detayın içerisine yeni olacak olan kavramlar dahil edilmiştir. Bölünme ve Dağılma Bilgisi (Split-Info), özelliklerin bulunamayan öğeler ile mücadele edilmesi, sayısal özellik verilerinin hesaplara dahil edilmesi bu başlıklardan yüksek öneme sahip olanlarıdır.

ID3 algoritmasının verileri çok sayıda gruba bölme eğilimini kırmak için dağılma bilgisi (Split-Info) bu algoritması eklenerek C4.5 algoritması oluşturulmuştur. Dağılma bilgisi temelde bilgi kazanımı bilgisini normalize ederek daha kaliteli bir karar ağacı oluşturmayı sağlar.

3.3.3. Sayısal Özellikler

Veri kümesi içerisinde 2 çeşit veri mevcuttur; Nominal(kategorik şekilde) ve sayısal olacak biçimde. Nominal biçimde olan daha erken kısımlarda işleme sokulan veri tipleri şeklinde ifade edilebilir.

Başta sayısal özelliklerle uğraşmak ve onların bilgi edinmesini hesaplamak çok sıkıntılı olabilir. Ancak bu iş esasında abartılacak kadar zor bir şey değildir. Halledilmesi lazım olan iş yalnız bu özelliğin sayısal değerleri içerisinde düzgün ve uygun eşik değerini saptamaktır. Sınır eşik sayısı tespit edildikten sonra ikili bir ayrılma neticesinde veri kümesi bölünebilir. Bu eşik değeri tespit edildikten sonra ikili bir bölünme ile veri kümesi ayrılabilir; Bu eşik değerinden yüksekte olan veriler ile bu eşik değerinden alçakta olan veriler. Bu anlamda algoritma çok sade bir şekilde ifade edilebilir. Başta tüm sayısal olan veriler küçükten büyüğe sıraya konulur. Bu sıra $\{V_1,\ V_2,\ \ldots.\ V_m\}$şeklinde ifade edilsin. Bu durumda tespit edilen eşik sınırı vi ve V_i + 1 aralarında olursa $\{V_1, V_2\ ,\ldots,\ V_i\}$ile $\{V_i$ + 1, V_i + 2, …. , $V_m\}$ şeklinde

olacak biçimde farklı bölümler meydana gelir. Bu aşamada açıkça görülmektedir ki m-1 adet eşik sınırı tespit edilebilir. Elde edilen tercih etme işlemi bölümü aşaması için gerekli olan olası tüm eşit bulunan verileri

$$\frac{V_i + V_i + 1}{2}$$

Formülüne eklenerek hesaplanır. Bu kompozisyon ile sanki mevcut özellik büyük – küçük değerleri mevcut olan nominal bir özelliktir. Bu anlayışla nominal değerler üzerine uygulanan bilgi oranı formülü bütün eşik değerlerinde geçerli olması için uygulanır ve bilgi edinimi en çok olan eşik değeri söz konusu özelliğin sınırı olarak kabul edilir. Eğer en güzel sınır değeri e ise ve bahsi geçen özelliğin sayısal değerleri $\{V_1, V_2, \ldots, V_n\}$kümesi şeklinde açıklanıyor ise, bu kümedeki V_i<e şartını yerine getiren değişkenler ufak kategori grubunda ve V_i>e şartını yerine getiren elemanlar büyük kategori grubuna sokulurlar. Bu şekilde mevcut olan bir ayrım yapma olayı birçok veri içerisinde işlem yapıldığı zaman güzel ve pozitif sonuçlar gözlemlenmiştir. Lakin eksik kısmı yani sadece ikili bir ayırma işlemi meydana getirmesidir. Halbuki bu tip bir ayrım yapma üçlü ya da çok daha fazla miktarda olursa veri yığını içerisine saklanmış olan kuralları tespit etme olasılığı daha çok yükselir. (Tüba Kıyan, 2003)

3.3.4.Lojistik Regresyon(Logistic Regression)

Kullanımındaki hedef, istatistiksel şekilde kullanılan başka bir model elde etme teknikleriyle aynı şekildedir. En az değişkenleri görüp bunları lehine kullanarak en güzel biçimde uyuma elde edebilecek biçimde bağımlı ya da bağımsız biçimdeki değişkenler içerisindeki bağlantıyı düzgün biçimde olacak şekilde ifade edebilen bir yapı tasarlamaktır. Lojistik regresyon modelleri, günümüzdeki zamanlarda biyoloji alanında, tıp alanında, ekonomi alanında, tarım alanında ve veterinerlik tasıma alanlarında çok yaygın biçimde işleme alınmaktadır (Bircan, 2004).

3.4.Kümeleme Analizi

3.4.1. Kümeleme Analizi Tanımı

Veri kümesinde olan bilgileri belirli yakınlık olma uzaklık olma kriterleri dikkate alarak bölümlendirme işlemidir. Bunların her birine ifade olarak "küme" ismi takılır. Kümeleme kullanılarak yapılan analizi kısaca isimlendirmek gerekirse "kümeleme" ismi verilir. Kümeleme şeklinde yapılan işlemde küme içinde bulunan elemanların benzerliği fazla, kümelerin arasındaki benzerlik oranı ise az olmak zorundadır.

Kümeleme işlemi esasında sınıflandırma ve ileriye yönelik tahmin işleminin tersine, veri kümesini daha önceden sınıflara bölmez, bu işlemin yerine veriler ayırımlarına göre incelenerek saf şekilde kategorilere ayırmalar meydana getirilir. Kümeleme olayının sınıflandırma olayından ayırt edici farkı şudur durumdan önce tespit edilmiş sınıflar ya da sınıf tanımları (etiketleri) mevcut olmamasıdır. Bu nedenden dolayı kümeleme işlemi gözetimsiz (unspervised) şekilde yapılan veri madenciliği yöntemi olarak adlandırılır.

Kümeleme olayı, gözetimsiz şekilde sınıflama (unsupervised calassification) yöntemidir (Kamber, Data Mining:Concepts and Techniques, 2001). Gözetimli şekilde sınıflandırma operasyonunda veriler önceden sınıflandırılmış örüntüler halindedir.

Burada asıl hedef, yeni gelmesi düşünülen ve henüz hangi sınıfta mevcut olduğu bilinmeyen verilerin mevcut olan sınıflardan en uygun seçilene yerleştirilmesidir. Gözetimsiz şekilde yapılan sınıflamada ise hedef, başlangıçta verilen ve henüz sınıflandırılmamış bir küme veriyi anlamlı alt kümeler biçiminde

oluşturacak biçimde toplanmaktadır. Kümeleme işlemi tamamen gelen verinin özelliklerine göre yapılır.

Kümeleme işlemi neticesinde ele geçen verilerin kümeler ile kullanılan yönteminin giriş yapma değerlerine bağımlı olsa da, giriş değerlerinden bağımsız kümeleme işlemleri ilerletmeler devam etmektedir (Berkhin, 2002).

Kümeleme analizi yalnızca veri madenciliğinde değil, örüntü ifade edebilme açıklama yöntemleri, görüntü değerlendirme yöntemleri, coğrafi olarak bilgi edinme sistemleri yöntemleri şeklinde birden fazla bölümde fazla şekilde olarak işleme alınmaktadır.

Literatürde içerisinde birden fazla kümeleme algoritması mevcuttur. Kullanılacak olan kümeleme algoritması tercihi, veri şekline ve hedefe bağlı şekilde kullanılmaktadır.

Genel şekilde kümeleme işlemleri alt kısımdaki gibi sınıflandırılabilir.

1- Bölümleme yolları
2- Hiyerarşik yollar
3- Yoğunluk tabanlı yollar

1-Bölümleme yöntemleri

Bölümleme metotlarında amaç, n adet nesneden meydana gelen veri tabanını, giriş parametresi olacak şekilde belirlenen k adet kısma (k<=n) bölümleme temel alınarak yapılır. Veri tabanında mevcut olan her bir eleman bir değişiklik fonksiyonuna bakılır ve buna bakılarak k tane bölümden bir tanesine sokulur. Bu sokulan bölümlerden her bir tanesi birer küme olacak şekilde isimlendirilir.

Bölümlere ayırma metotları k sayısı doğru biçim ve şekilde tahmini yapılabilirse benzer şekilde olan dışbükey kümeleri tespit etmede oldukça başarılı veriler elde etmektedir.

Ayrımlama yapılarak oluşturulan metotların bilinen sıkıntısı k giriş parametresine bağlı şekilde olmaları ve düzgün şekilli mevcut olmayan kümeleri tespit edememeleridir. Bölümleme metotları olarak k - means, k - medoids ve CLARA - CLARANS şeklinde tanınan algoritmaları kullanırlar (Kamber, Data Mining:Concepts and Techniques, 2001).

2-Hiyerarşik yöntemler

Hiyerarşik olan modeller bir ağaç yapısını meydana getirip oluşturarak kümeleme işlemini ortaya çıkarmaktadır. Elde edilen kümeleme ağacının tamamı ağaç yapılarında oluştuğu gibi bir kök (root) düğümü ve çocuk düğümleri vardır.

Bu işlemde, başlangıç olarak küme sayısı miktarı ifade edilmemektedir. Algoritma, x: veri seti s: uzaklıklar matrisi şeklinde ifade edilmek üzere; (x,s) girdi şeklinde olacak biçimde tanımlanmaktadır. Sonuçta çıktı şeklinde olması planlanan ve elde edilen kümeler arasında hiyerarşi mevcuttur. Hiyerarşik biçimde oluşturulan kümeleme metotlarının çoğunda süreç optimizasyon tabanlı yapılmış biçimde olmaz. Bu metotlardaki esas hedef, birleşme yapılıncaya kadar bölmenin ileri aşamaya gitmesi ve tekrarlamalar kullanılarak bazı tarzda yaklaşımlar tespit etmektedir.

Hiyerarşik şekilde olan kümeleme teknikleri esasında ağaca benzer (dendogram) bir grafik şeklini oluştururlar. Hiyerarşik biçimde kümeleme tekniklerini, hiyerarşik birleşme operasyonunun yönü aşağıdan üst kısımdaki (bottomup), ayrıca yukarıdan aşağıya (top - down) işlemi uygulanmasına paralel

olacak şekilde bir araya getirici (aggolomerative) ve bölümlere ayırıcı (divise) hiyerarşik biçimde kategorilere ayırma teknikleri şeklinde ikiye bölmek mümkündür.

Alt kısımdan üst kısma olacak şekilde, toplaşım kümeleme algoritmaları ve yukarıdan aşağıya tarzda kümeleme algo*r*itmaları şeklinde 2 grupta bir araya getirilebilir. Toplaşım ve kümeleme algoritmaları, ilk başta veri tabanındaki her bir bölümü bir küme şeklinde olacak biçimde görür. Bu mevcut kümeleri bir araya getirip birbirinden farklı kümeler meydana getirilir. Bölünür şekilde olan kümeleme algoritmaları ise başlangıçta veri tabanındaki tüm noktaların bölümlerini tek bir kümeymiş tarzında gibi görür. Veri tabanına bakıp araştırdıkça, birbirinden farklı bölümleri kümeden çıkartarak önceden ifade edilmiş olan, k olacak kadar büyüklüğündeki kümeye dağıtılır (Silahtaroğlu, 2008).

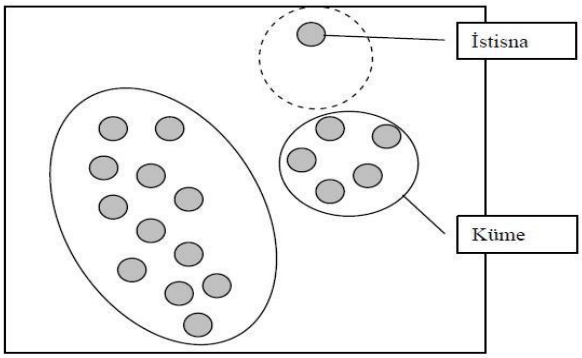
Hiyerarşik şekilleri sıkça kullanan ünlü algoritmalar şu şekildedir. SLINK algoritması, CURE algoritması, CHAMELEON algoritması, BIRCH algoritması.

3.Yoğunluk Tabanlı Yöntemler

Dağılmış verileri barındıran veri tabanlarının sadece uzaklığı merkez olarak merkez alan bölümlü algoritmalar ile kategorilere ayrılması çok fazla zordur. Nedeni ise hiçbir şekilde bir kümeye dahil edilmeyen sınırdaki noktaları içerisinde bulunduran bu dağılmış veri tabanlarının bölümlere ayrılmış şekilde algoritmalar ile kümelenme işlemi sonucunda doğru kümeler meydana gelmeyecektir. Bu olayda birlikte bir yoğunluk meydana getiren verilerin aynı kümeye toplanmasını destekleyen yoğunluğu merkez olarak alan algoritmalar kullanılmaktadır. Bu şekildeki algoritmalara basit örnek vermek gerekirse DBSCAN, OPTICS ve DENCLUE algoritmaları örneklendirilebilir.

Bu metotta gösterilen bütün noktanın etrafındaki komşularıyla birlikte olan yakınlıkları incelenir ve işlenir hesaba katılır. Yakınlık mesafesini hesaplamada genel olarak öklit mesafesi formülü işleme alınsa da veri tipine bakılarak yakınlık

hesaplayabilme şekli değişiklik gözlemlenebilir. Bu metodun ana prensibi “ yeterli derecede yakınlığı olmayan bölümleri ” belirlemektedir. Bu olay Şekil 6’da görsel şekilde görülmektedir.



Şekil 6 İstisna ve küme oluşumları

3.4.2. Kümeleme Analizinin Özellikleri

Yüksek başarımlı bir kümeleme analizi metodu şekli şu özellikleri elde etmiş olmalıdır (Kamber, Data Mining:Concepts and Techniques, 2001):

*Ölçeklenebilir şekilde gerçekleştirilmelidir. Birkaç yüz kayıttan meydana gelerek oluşan veri kümesine de tonlarca kayıt içeren kümeye de uygulanabilmelidir.

* Farklı şekilde olan veri çeşitleri ile kullanılabilmelidir. Hem sayısal şekilde hem de kategorik şekildeki verileri bünyesinde bulunduran veri tabanlarında kullanılabilmelidir.

*Düzgün şekilde çeşitli olmayan kümeleri de bulabilmelidir.

*En az sayıda giriş değişkeni sayısı gerekir. Bir yöntem ile alakalı olarak ne kadar az giriş değişkeni gerektiriyor ise o derecede olan ölçüde kullanıcının kararlarından bağımsızdır.

*Gürültülü veriler ile değerlendirilmelidir.

*Veri kümesindeki kayıtların tek tek sıralamasından bağımsız şekilde olmalıdır. Kümenin hangi elemanından başlangıç yapılırsa yapılsın sonuç hiçbir şekilde değişmemelidir.

*Çok farklı boyutlarda veri tabanlarına uygulanabilmelidir.

*Veri kümesinin kapsamında olan sınırları dikkate alabilmelidir.

*Kolay şekilde açıklanabilir çıktılar üretebilmeli ve işlevsel şekilde olmalıdır.

Bu sıralanan maddeler ideal şekilde oluşturulmuş bir kümeleme algoritmasının özellikleridir. El altındaki algoritmaların hiç biri bu niteliklerin hepsine maalesef sahip değildir. Kümeleme analizi devamlı kendini geliştirmekte olan bir konudur ve ilerideki senelerde hedefe yakın metotların ilerletileceği düşünülmektedir.

3.5. Sıra Dışılık Analizi

Bir veri kümesinde olanların ana tavırlarından veya veri modelinden değişiklik ortaya çıkartan nesneler olağan dışı (outlier) denir. Birden fazla veri madenciliği metodu istisnaları gürültü veya aşırı olaylar şeklinde gözlemlemektedir, bu durumdan dikkate almaz. Lakin bir takım benzer olaylarda istisna kısımlar diğerlerine bakılarak olduğundan çok bilgi barındırır. Örnek bilgisayardaki anti virüs programları, sızmaları tespit edilmesi (intrusion detection), kanser teşhisi, kanser teşhisi başlangıcını belirlemede ya da hiç karşılaşılmamış istisna olan durumlar analiz edilip incelenirler.(Kamber, Data Mining: Concepts and Techniques, 2001).

4 VERİ MADECİLİĞİ PROGRAMLARI

Veri Madenciliği işleminde esas olarak amaçlanan, veriden bilgi elde etme amaçlı kullanılan tekniklerinin tamamını kapsar. İstatistiksel analiz yollarının ve yapay zekâ algoritmalarının birlikte olacak biçimde işleme alınarak veri içerisindeki saklı olan bilgilerin olacak biçimde bilgiye döndürülmesi aşamasıdır. Veri

Madenciliği ile ilgili uygulamaları gerçekleştirmek için bilgisayar programı kullanmanız gerekmektedir. Bu olaydan yola çıkılarak birden fazla yazılım tasarlanmış ve geliştirilmiştir. Bu tezde Açık Kaynak Kodlu olacak biçimde Veri Madenciliği yazılımlarından meydana getirilen WEKA'ya değinilmiştir.

4.1.Weka

WEKA (Waikato Environment for Knowledge Analyses), Waikato Üniversitesi tarafından tasarlanıp yazılan 1996‟da ilk resmi sürümü yayınlanmış şekilde bulunan bir makine öğrenme ve veri madenciliği yazılımıdır. Akademik araştırmalar, eğitim ve endüstriyel uygulama bölümlerinde kullanım yerine sahip olan WEKA, veri analizi ve veriler içerisinden tahmin yürütücü şekilde modelleme için yapılmış algoritma ve araçların görebileceğimiz şekilde bir araya gelmelerini içerir. WEKA bir proje şeklinde olacak biçimde başlayıp günümüzde dünya üzerinde birden çok insan tarafından dikkate alınıp kullanılan bir Veri Madenciliği uygulaması geliştirme yazılımıdır. WEKA programı esasında bakılacak olursa java içerisinde yazılarak programlanmış açık kaynak kodlu bir programdır. WEKA açıldıktan sonra çalıştırılıyor sonra Tablo 16'da gözlemlendiği şekliyle, uygulama (application) menüsünde görülüp çalışma yapılabilecek olan modlar liste halinde görülmektedir. Bu olanlar komut modunda çalışmasını gerçekleşmeyi yapan Simple CLI, projeyi bizlere aşamalar şeklinde görsel ortamda etkinleştirmeyi gerçekleştiren Explorer ve projeyi itekleyerek bırak yöntemiyle hayata geçirmeyi sağlayan Knowledge Flow seçenekleridir. Explorer seçeneği tercih edildikten sonra bu konu üzerinde çalışılacak verilerin tercih edilmesi, bu veriler üzerinde temizleme, ayıklama ve dönüştürme işlemlerinin gerçekleştirilebilmesini yapan ekran ile birbirleri arasında karşılaşılmaktadır.

Arff, Csv, C4.5 şekil formatında mevcut olan dosyalar WEKA'da kullanılabilir yani import edilebilir. Herhangi bir metin şeklinde olan verileri WEKA ile işlemek olanaksızdır. Ayrıca JDBC aracılığı ile veri tabanına bağlantı kurulup bu kısımda da işlemler gerçekleştirilebilir. WEKA'nın içinde mevcut olan Veri işleme, Veri Sınıflandırma, Veri Kümeleme, Veri ilişkilendirme özelliklerini kullanabilmekteyiz ve özellikler mevcuttur. Bu aşamadan sonra yapılacak olan proje

kapsamında amacına hitap edecek biçimde açılan sayfadaki uygun kısımdaki (Sınıflandırma, Kümeleme, İlişkilendirme) uygun algoritma veya algoritmalar tercih edilerek veriler üzerine denenmekte ve en doğru en düzgün sonucu veren algoritma tercih edilebilmektedir.

Şekil 7 Weka'nın Ön Menüsü

WEKA nasıl meydana gelmiştir diye soranlara paket programı yardımıyla elde edilmiştir şeklinde cevap verilebilir. WEKA paket programında aslında veri kümesi için sırasıyla olmak koşuluyla Naive Bayes, Kstar, RBFNetwork, J48, JRIP, Ridor algoritmaları tercih edilerek program çalıştırılmış ve elde edilen sonuçlar neticesinde hazırlanmıştır. Bununla birlikte HyperPipes, VFI gibi birçok algoritma test edilmiştir.

4.1.1.Veri Ön İşleme Aşaması

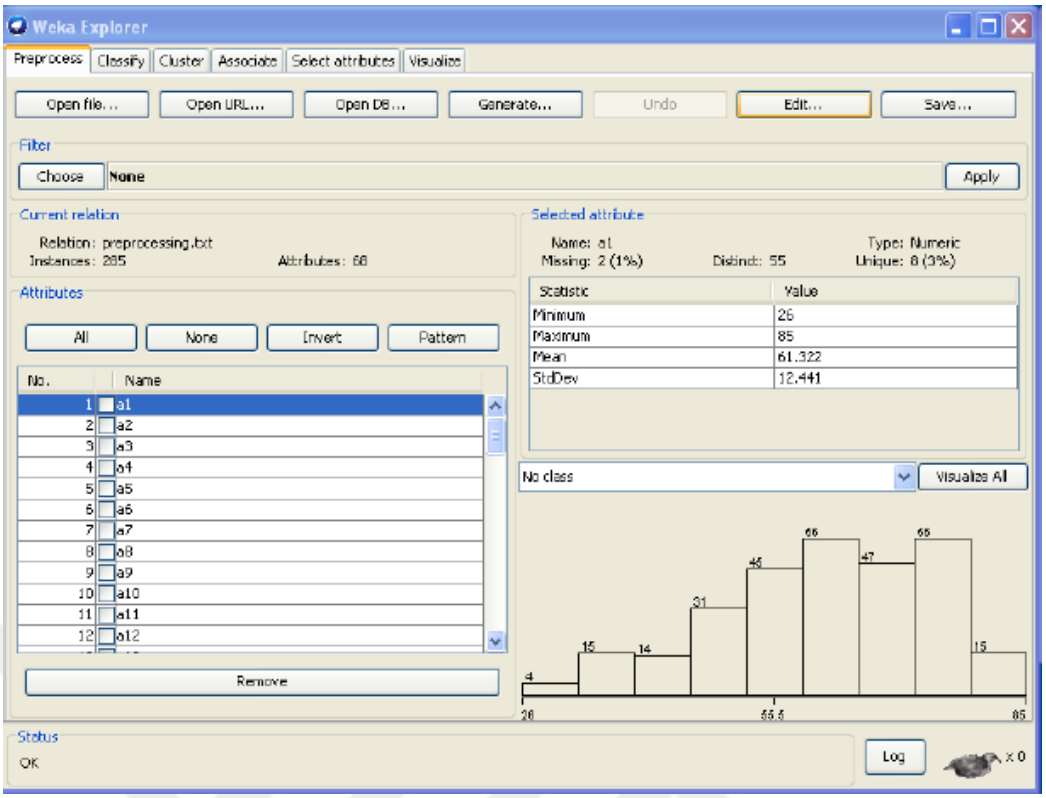
Algoritmaların birbirleri arasında karşılaştırılarak içlerinde hangi algoritmanın daha başarılı olduğunu tespit etmeye yönelik işlemlere yapılan eleştirilerden biri tanesi uygulama sırasında gerçekleştirilen veri önişleme aşamasıdır. Bu aşamada şunlar yapılmaktadır veri temizleme, veri birleştirme, veri dönüşümü, veri azaltma yöntemleri kullanılarak, veri analiz aşamasına hazır biçime getirilir. Bu bütün işlemler oluşacak muhtemel modelin başarı yüzdesini doğrudan etkileyebilir. Elde edilen işlemlerin neticesi uygulamacının görüşüne bağlıdır. Veri kümesi

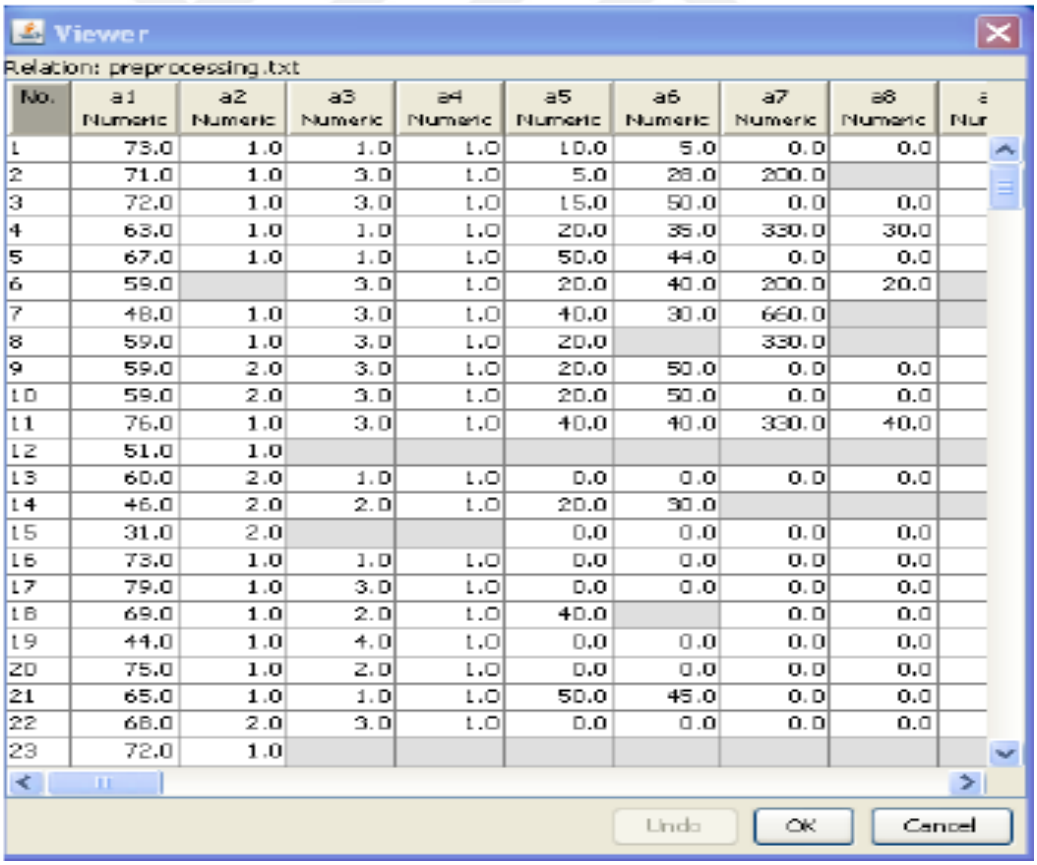
içerisinde uygulanan birkaç değişiklik farklı algoritmalar üzerinde değişik sonuçlar ortaya çıkartabilir. Elde edilecek çalışmanın çok güzel çıktılar elde etmesi için uygulamacının uygulama yapılan bölüm hakkında iyi derecede bilgi sahibi olmasını ya da bu alanla ilgili uzmanlarla bir araya gelerek birlikte çalışılmasını gerektirir.

4.1.2. Kayıp Veriler ve Bu Verilerin Yaratacağı Problemleri Çözmek

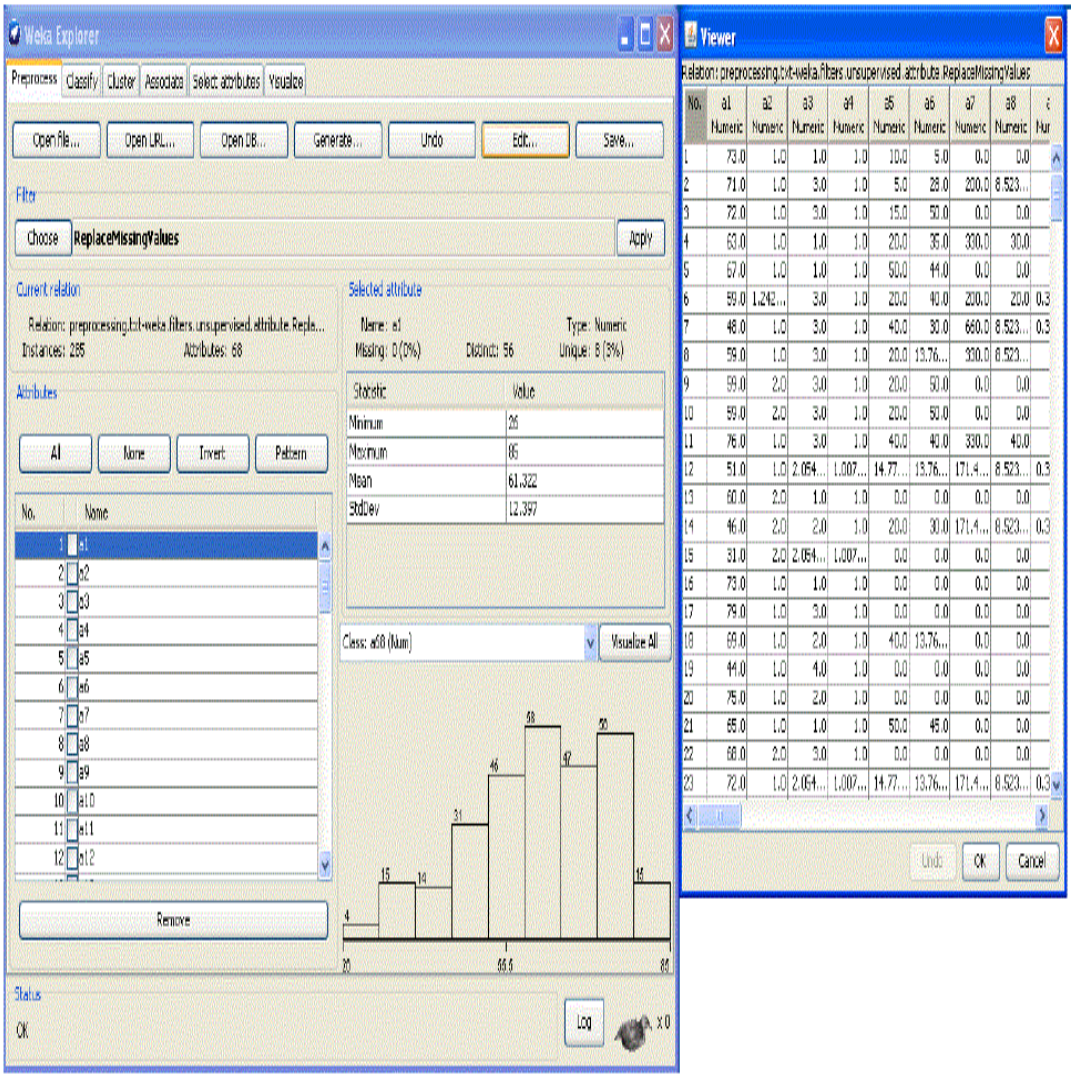
Bu olay için kullanılan metotlara örnek verilmesi gerekirse Replace Missing Values metodu kullanılmıştır Şekil 8 ve Şekil 9. Bu şekilde kullanılan yolla veri tabanında olan mevcut olan kayıp veri veya değerler, ait oldukları değerlerin diğer verilerinin aritmetik olarak ortalaması ya da modülerini alma yoluyla değişikliğe uğramaktadır.

Kayıp veriler şöyle ifade edilebilir; içerisinde eksik ve hasara uğramış ya da kaybolmuş olan veri değerleri içerisinde barındıranlardan oluşmaktadır. Veriler arasında karşılaştırma yapabilmede sıkıntı yaşarız. Sonuçlarımız net ve kesin olmaz. Bu durumu ortadan kaldırmada ilk olarak bütün eksik olan verileri veri setinden çıkartmak bir çözüm olabilir. İkinci olan olay ise verilerin kullanılabilir bölümüne kadar olanlarını sisteme dahil edip en uygun ve yüksek sonucu elde etmeye çalışmak. En son olarak ise veri kaybına uğramış bütün veri setlerinin hepsini kayıp bölümlerini Replace Missing Values modülü kullanılarak iyileştirilmiş ve işleme girdiğinde sonuç elde edebilecek hale getirilmişlerdir.





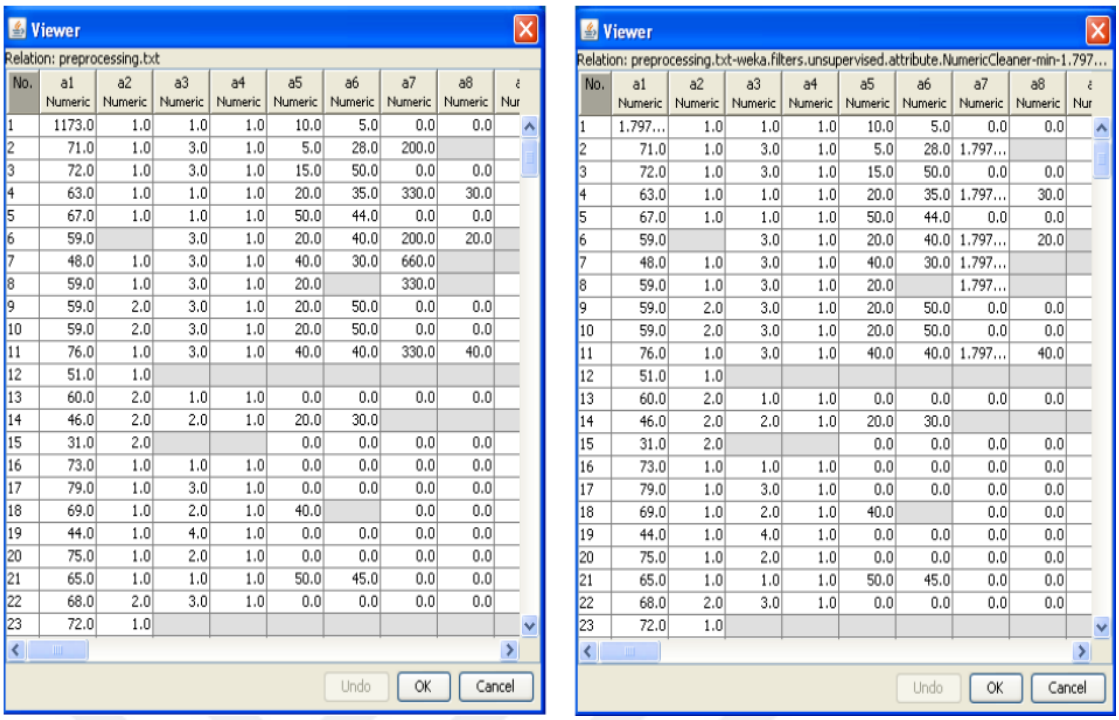
Şekil 8 Elimizin Altında Olan Veritabanındaki Kayıp Veriler



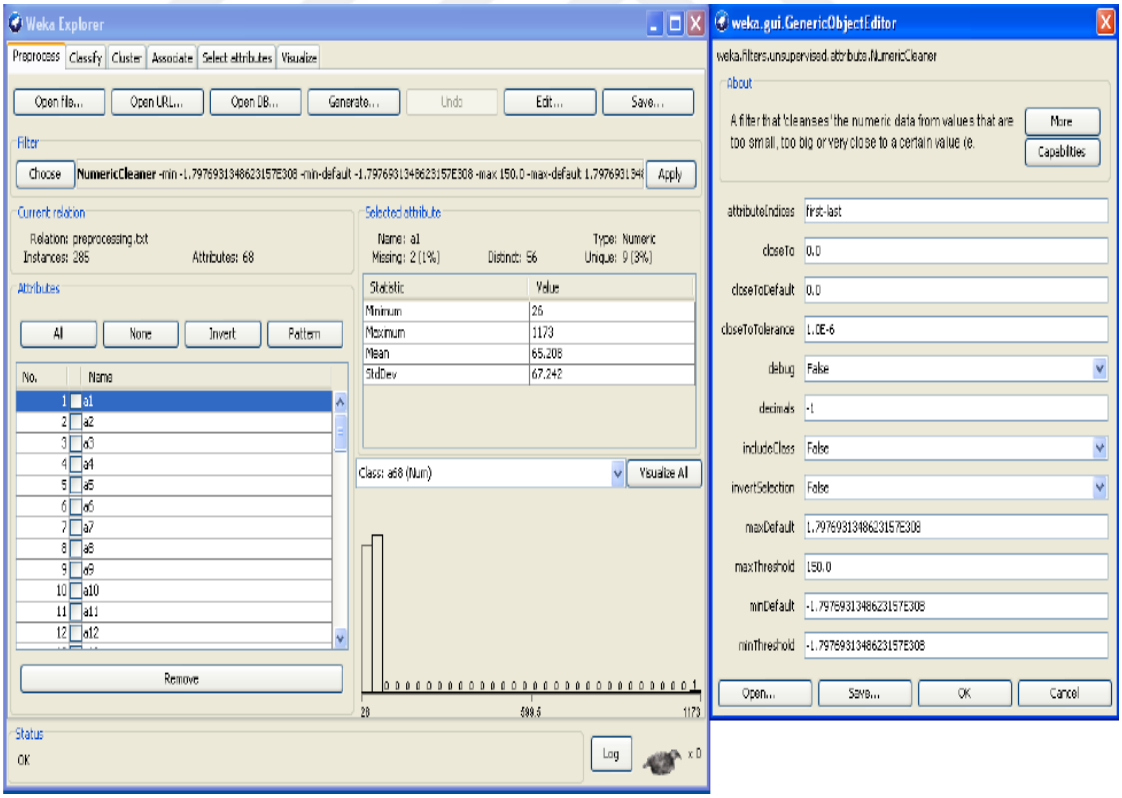
Şekil 9 ReplaceMissingValues Adlı Modülün Yapılmış Olan Kullanım Şekli

4.1.3. Yanlış ya da Uç Veriler Hakkında Yapılan İşlemler

Bu tipte veriler için ise Numeric Cleaner adlı bölümü dikkate alınmıştır. (Şekil 10 ve Şekil 11) Bu şekilde verideki diğer değerlere kıyasla çok büyük çok küçük olan değerlerin silinerek bu değerlerin yerine önceden tespit edilmiş değişik değerin geçirilmesini barındırır.



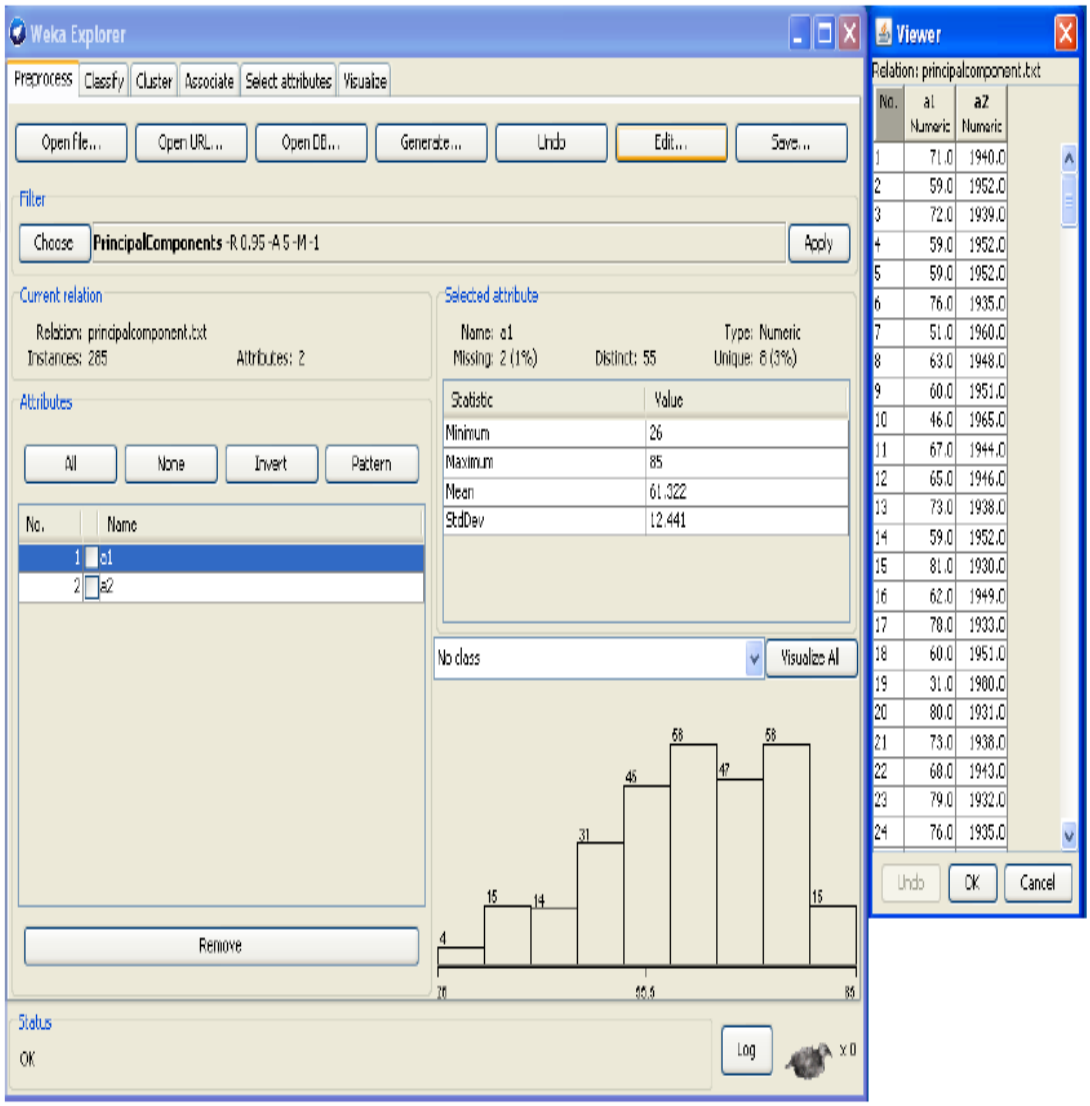
Şekil 10 Aşırı Uç Biçimde Mevcut Olan Verilerin Numeric Cleaner Kullanılmamış Biçimde Önce ve Sonrasındaki Son Durumlarını Göstermektedir



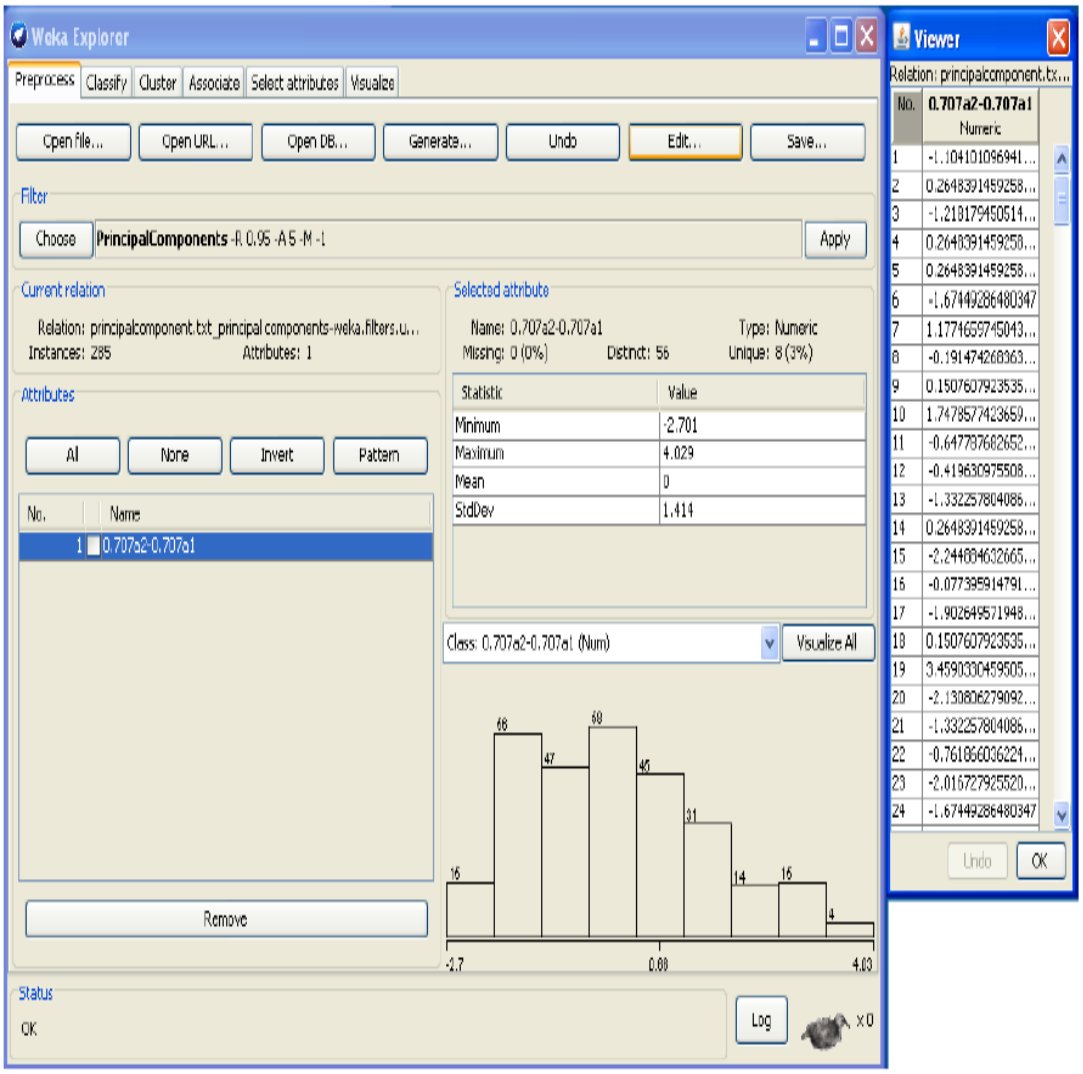
Şekil 11 Numeric Cleaner Modülünün Weka'da Kullanılarak Elde Edilenler

4.1.4. Gereksiz Olan Veriler

Aynı veritabanı dahilinde olan yaş kriteri ayrıca doğum tarihi kriteri bilgilerinin elde edilmesi durumu aşamasında oluşan lüzumsuz olan verilerin bilgisayarın çalışma performansı, zamanını ve sonuçların kalite oranlarına müdahale etmemesi için Principal Components isimli modülü kullanılarak verilerdeki boyut küçültülmüştür(Şekil 12 ve Şekil 13).



Şekil 12 Principal Components Modülü Kullanıma Geçmeden Önceki Hali

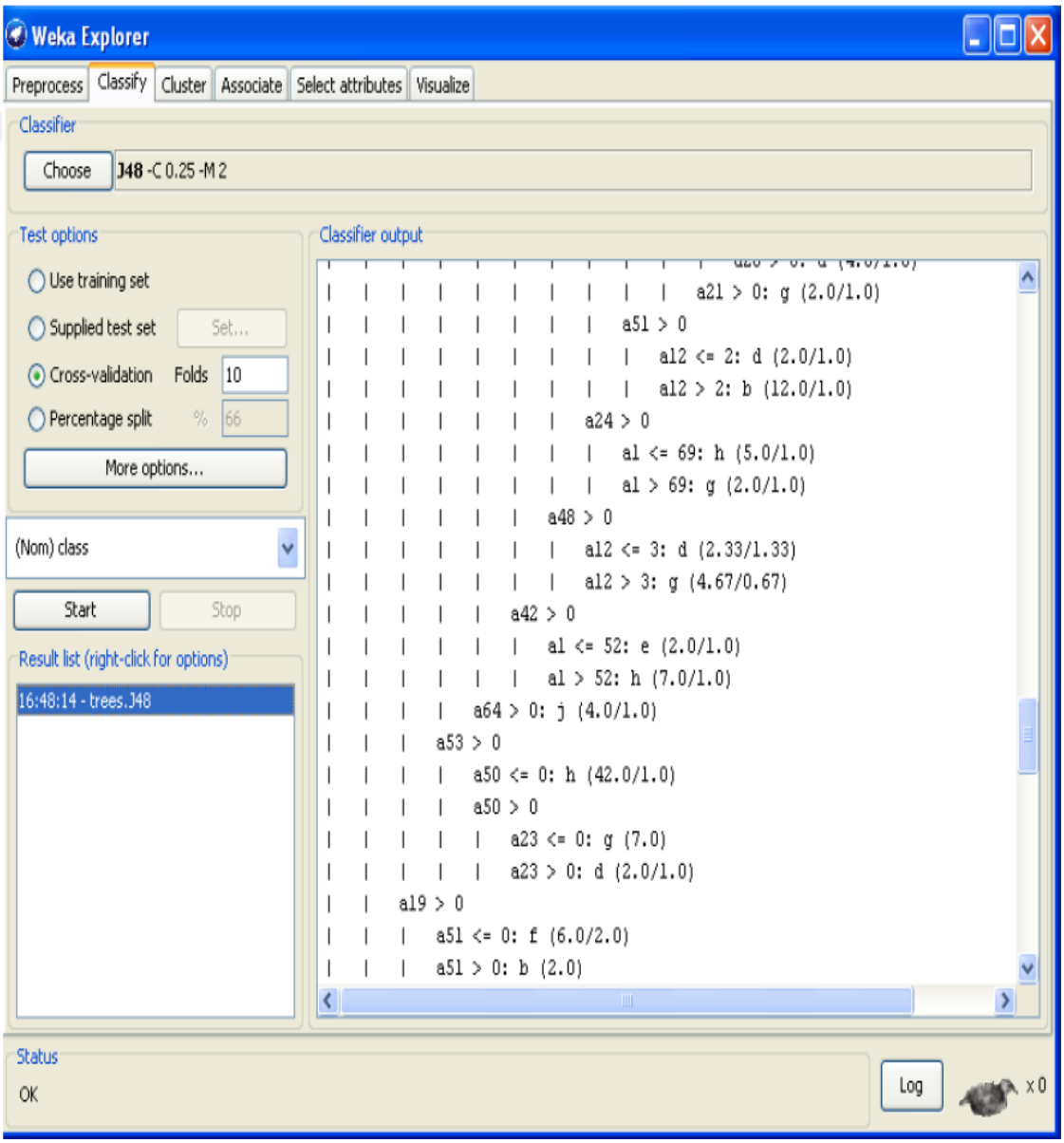


Şekil 13 Principal Components Modülü Kullanılarak Boyutun İndirgenmesi (Pınar Tapkan, 2011)

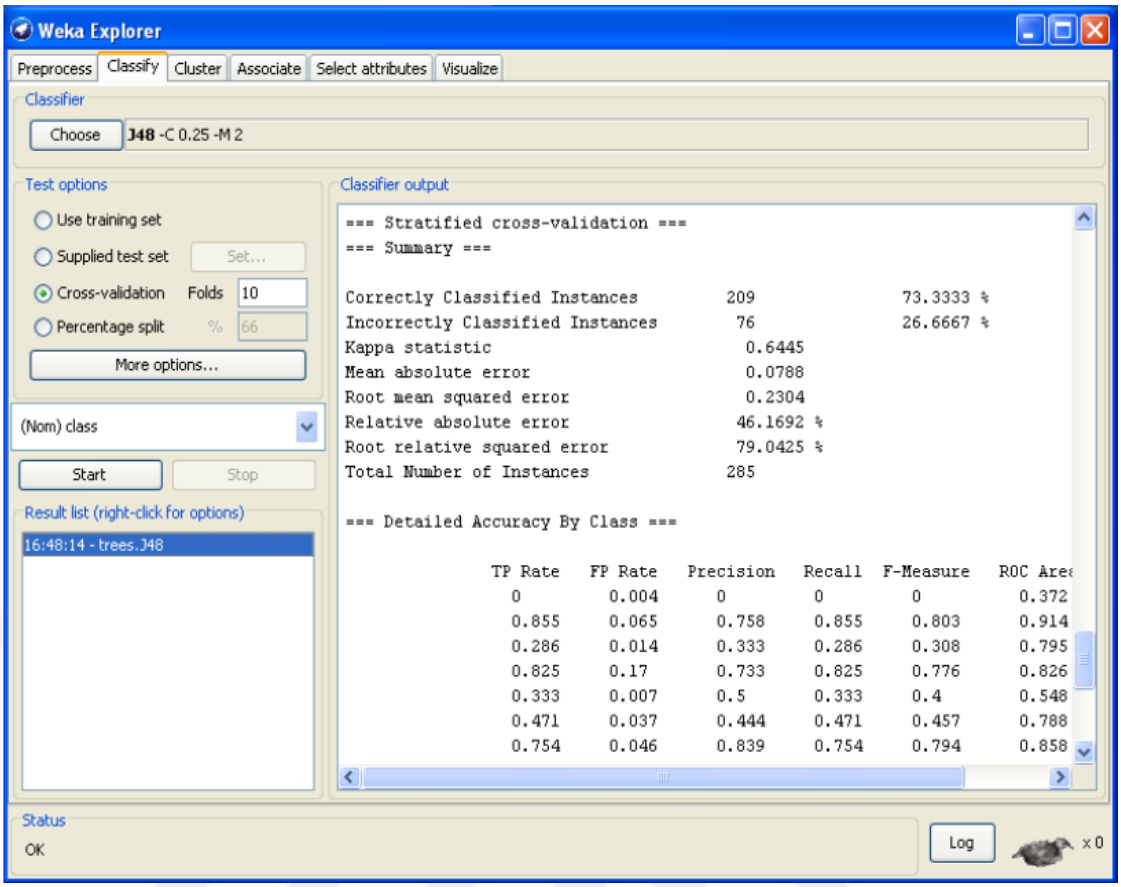
4.1.5.Sınıflandırma

Sınıflandırma, esasında bir örüntü tanıma bölümünde sisteminin en alt denilebilecek kısmı olan son aşamasında yer almaktadır ve bu işlemde Wisconsin göğüs kanseri örnek verilerine teşhis (kötü türde / iyi türde tümör) tespit edip yerleştirmek maksadıyla kullanılmıştır. Doğrusal olan analizi kullanılarak sınıflandırıcı meydana getirirken, gruplar içerisinde farklılığın yüksek, grup içi farklılığın aritmetik ortalamasının ise düşük olması lazımdır. Bu sebepten dolayı bir en iyileştirme gerçekleştirilmektedir. Sınıflandırıcıların, aktif şekilde kullanılmadan

önce eğitime tabii tutulmaları gerekmektedir. Verilerin sınıflara ayrılması durumu dikkate alındığında elde olması gereken bölümlerden biri de maliyetlerdir. Bu kapsamda WEKA'da maliyet esasına bakılarak sınıflandırmayı yapıp gerçekleştiren Cost Sensitive Classifier modülü kullanılır. Bu olayın esas amacı gelecek olan yanlış sınıflandırma masrafını minimize yapacak biçimde en güzel sınıflandırmayı tahmin etmeyi gerçekleştirmektedir. Sınıflandırma şeklinde olan algoritmaları arasında C4.5 karar ağacına dayalı olan J48 modülü, programda işleme sokulmak için seçilmiştir (Şekil 14, Şekil 15).



Şekil 14 J48 Modülünün Kullanımı İle Elde Edilen Kurallar



Şekil 15 J48 Modülü ile Elde Edilen Sınıflandırma Doğrulukları (Pınar Tapkan, 2011)

4.1.5.1. Öznitelik Tercih Bölümü

Sınıflandırmada kullanılması beklenen öznitelik vektörünü meydana getirmek için, toplamda dokuz öznitelikten en güzel tespit edilenler ($p < 0.001$), t - testle kontrole tabii tutulmuştur. Bu kapsamda p-seviyesi, gözlemlenen sonucun geçerliliğindeki mevcut olan hata olasılık verisindeki değeri bizlere göstermektedir. Bütün öznitelikler test' de güzel biçimde başarılı sonucu vermiştir. Ardışık biçimde ileriye yönelik tercih metotlarında ilk öncelik en iyi öznitelik seçilmiş, sonrasında, her seferinde bir öznitelik dahil edilerek kriter fonksiyonunu en iyi hale getirilen öznitelik biçimde bileşimi tespit edilmiştir. Sonuç olarak tespit edilen vektör 5 bileşeni içinde bulundurur ve bu öznitelikler, en güzelden başlayarak, hücrenin ana bölümünde olan çekirdeğinin çevre sitoplazmaya oranı, bölgedeki hücre ebatlarının

eşitliği, Hücrenin bütün hale gelmesindeki kalınlığı, normal şekilde olacak çekirdekçik ve kromatin dağılımı biçimleri olacak şekilde sıraya konmaktadır.

4.1.5.2.Sınıflandırma Algoritmalarının Aralarında Karşılaştırılması Konusunda Dikkat Edilmesi Gereken Konular

Verilerin önişleme aşamalarındaki işlemlerinde, parametrenin ya da parametrelerin tercih edilmesi ve test yapılacak olan kümesinin tercihi veri madenciliği uygulamasında ortaya çıkacak olan modelin başarısını doğrudan etkiler. Bu sebepten ötürü ile yapılan karşılaştırma çıktıları çok fazla şekilde ölçüde uygulamacıya bağlıdır.

4.2. Verilerin Ön İşlenmesi

Algoritmaların birbirleri arasında karşılaştırılarak hangi türde algoritmanın birbiri arasında çok daha iyi olduğunu tespit etmeye yönelik yapılan işlemlerde eleştirilerden biri uygulama esnasında yapılan verilerin önişleme aşamasıdır. Bu basamakta verilerin temizlenmesi, veri birleştirme, veri dönüştürülmesi, veri indirgeme metotları kullanılarak, veri analize tamamen hazır biçime getirilir. Bu işlemler gerçekleşecek olan modelin başarısını doğrudan ya da dolaylı etkileyebilir. Oluşturulan programlar ve işlemler programı kullanan kişinin durumuna bakış açısına göre değişmektedir. Veri kümesi içinde gerçekleştirilen bazı türde farklı müdahaleler değişik türde algoritmalarda değişik türde sonuçları ortaya çıkartabilir. Gerçekleştirilecek çalışmanın güzel sonuçlar vermesi uygulamacının uygulamayı gerçekleştiren alan hakkında detaylı veriye sahip olmasını ya da bu sektördeki alan uzmanlarıyla bir araya gelip çalışmasını gerektirir.

4.3.Parametre Tercihi Aşaması

Veri madenciliğinde kullanılan değişik tarzda olan algoritmaların değişik parametreleri mevcut olabilir. Örnek vermek gerekirse yapay olarak oluşturulan sinir ağlarında saklı halde olan nöron sayısı, karar ağaçlarındaki mevcut olan budama operasyonunun parametreleri, algoritmaların kullanacağı parametrik sonuçları tespit ederler. Bu tanımlı olan parametreler algoritmadan algoritmaya farklılık gösterebilir, ya da kullanılan veri madenciliği yazılımlarında farklılıklar olabilir. Bunların hepsinin seçimi tespit edilerek oluşacak olan modelin başarı yüzdesini doğrudan etkileyecektir.

4.4.Test Kümesinin Belirlenmesi

Model meydana getirilirken kullanılan öğrenme ve test kümelerinin tespit edilmesinde de modelin başarılı olması üzerinde büyük etkisi mevcuttur. Eldeki verinin öğrenme kümesi şeklinde ve test kümesi şeklinde olacak biçimde bölümlere ayrılmasında farklı yollar izlenebilir. Kullanılan veri madenciliği yazılımında bu gerçekleşen işlem için çok değişik seçenekler olabilir. Öğrenme kümesi ve test kümesi değişik olan dosyalardan programa aktarılabileceği gibi, programın herhangi bir veri dosyasını tespit edilen bir rakamda test kümesi olacak şekilde kullanması ya da n-fold metodu ile programın veri kümesini n tane sayıdaki parçalara ayırarak sırasıyla her bölümü test kümesi şeklinde olacak biçimde kullanması gerçekleştirilebilir.

4.5.Modelin Başarıya Ulaşma Parametreleri

Model başarısını belirleyip değerlendirirken kullanılan ana kavramlar hata yüzdesi düzeyi, kesinlik durumu, duyarlılık ve F-ölçütüdür. Modelin başarısının tespiti, doğru sınıfa gönderilen örnek miktarı ve yanlış sınıfa gönderilen örnek miktarı nicelikleriyle ilişkilidir. Test sonunda ulaşılan çıktıların başarım detayları karışıklık matrisi ile açıklanabilir. Karışıklık matrisinde satır ile ilgili bölümleri test

bölümünde olan örneklere dayanılacak olan gerçek sayıları, kolonlar şeklinde olanlar ise modelin sadece tahmin edilmesini açıklar. (Cengiz Coşkun1, 2008)

Şekil 16 İki Sınıflı Biçimde Oluşturulmuş Bir Veri Kümesindeki Modelin Karışıklık Matrisi Şeması

		Öngörülen Sınıf	
		Sınıf=1	Sınıf=0
Doğru Sınıf	Sınıf=1	a (TP)	b (FN)
	Sınıf=0	c(FP)	d (TN)

a: TP (True Pozitif)
c: FP (False Pozitif)
b: FN (False Negatif)
d: TN (True Negatif)

4.5.1.Doğruluk Bulma ve Hata Oranı Tespiti

Başarının ölçülmesi ile ilgili olarak kullanılan en duyulmuş ve kolay yöntem, modele ait doğruluk oranıdır tespitidir. Doğru şekilde sınıflandırılmış olan örnek sayısının tespiti (TP + TN), toplam olacak biçimde örnek sayısını tespit için (TP + TN + FP + FN) oranı alınmıştır. Hata oranı tespitine gelecek olursak bu değerin 1'e tamlayanıdır. Başka şekilde açıklamayla yanlış sınıflandırılmış örnek değerinin (FP + FN), toplam şekilde örnek sayısının tespiti (TP + TN + FP + FN) oranıdır. (Cengiz Coşkun1, 2008)

$$Doğruluk = \frac{TP+TN}{TP+FP+FN+TN} \qquad Hata\ Oranı = \frac{FP+FN}{TP+FP+FN+TN}$$

4.5.2.Kesinlik Tayini

Kesinlik olan, sınıfı 1 olacak şekilde tahmin edilmiş olan örnek rakamının, gerçekte sınıfı 1 olan örnek rakamına oranlanmışıdır. Yani sınıf 1'dir dediklerimizin yüzde kaçı gerçekten sınıf 1'e aitmiş. (Cengiz Coşkun1, 2008)

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

4.5.3.Duyarlılık Değerinin Belirlenmesi

Doğru biçimde sınıflandırılmış olan pozitif şekildeki örnek sayısının toplam pozitif olan örnek sayısına oranı ile tespit edilmektedir. Bir diğer değişle sınıf 1 olarak tahmin ettiğimiz ve gerçektende sınıf 1'e ait örnekler, toplam gerçekte sınıf 1'e ait örneklerin yüzde kaçını oluşturuyor (Cengiz Coşkun1, 2008).

$$Duyarlılık = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

4.5.4.F-Ölçütü Belirlenmesi

Kesinlik ve duyarlılık parametrelerini birbirleri ile ters orantılıdır ve tek başlarına değerlendirmek için yeterli olamayabilir. Duyarlılıktan feragat edip kesinliği yükselmek ve ya kesinlikten feragat edip duyarlılığı yükseltmek mümkündür. Dolayısıyla, her iki parametreyi beraber değerlendirmek daha doğru sonuçlar ortaya çıkartır.

$$F - \mathrm{Ölçütü} = \frac{2 \text{ x Duyarlılık x Kesinlik}}{\mathrm{Duyarlılık} + \mathrm{Kesinlik}}$$

5 UYGULAMA MEME KANSERİ VERİLERİNİN SINIFLANDIRILMASI

5.1.Meme Kanseri

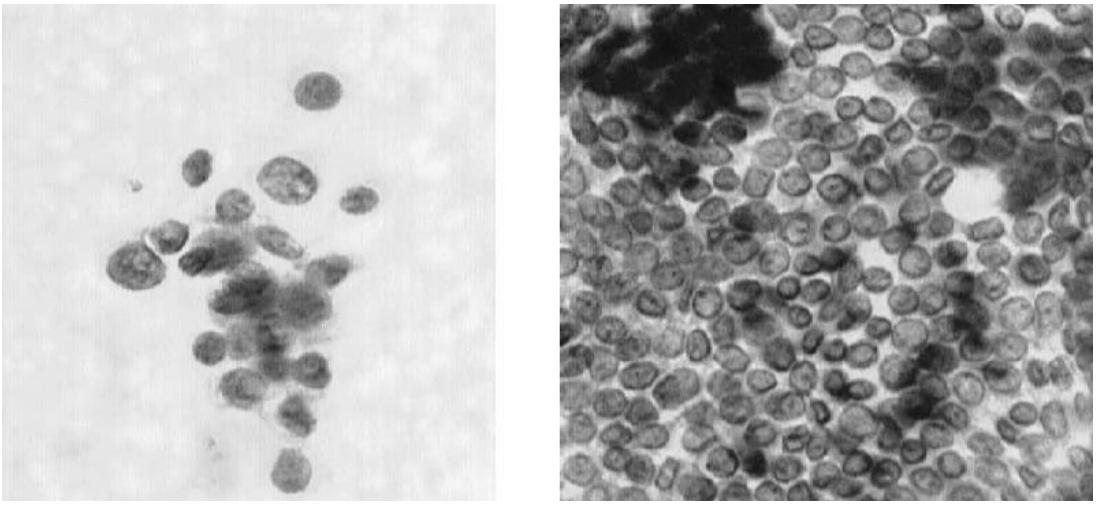
İnsan vücudunda mevcut olan sağlıklı şekilde yer alan hücrelerin, kas ve sinir hücreleri hariç olmakla birlikte, parçalara ayrılma yeteneği mevcuttur. Hücreler, yenileme hücreleri ve yaralanan hücre dokularının tamir edilmesi sebebiyle bölünebilme yeteneklerini kullanırlar. Her hücrenin yaşama süresince belli bir bölünebilme miktarı vardır. Sağlıklı biçimde yaşamını sürdüren bir hücre gerektiği pozisyonda ve gerektiği miktar kadar bölüneceğini bilir. Buna duruma karşın kanserli

olan hücreler, bu bilinci kaybedip maalesef kontrolsüz şekilde olacak biçimde bölünmeye başlayarak çoğalırlar. Bütün bu olan olaylardan ve bilgilerin ışığında kanser ile ilgili olarak aşağıdaki tanımlama işlemlerini ifade edebiliriz,

Kanser ile ilgili düzgün açıklamaları alt kısımdaki şekilde ifade edebiliriz. Kanser, bir dokuda veya organın içerisindeki hücrelerinde sağlıklı olmayan şekilde olacak biçimde değişme ortaya çıkarak bu hücrelerin denetimi olmayacak biçimde şekilde üremeye başlamasıyla birlikte ortaya çıkan rahatsızlıkların hepsi için uygulanan genel olacak biçimde kavram olarak ifade edilir.

- Genellikle kontrolümüzden çıkan hücrelerin çabuk ve hızlı bölünerek çoğalması durumudur. Vücudumuzun her bir organı içerisinde birden fazla şekilde farklı hücre yapılarından meydana gelmektedir. Normal olacak şekilde bu hücreler vücut için çok lazım olacak şekilde olduğu aşikar süreç gözetilerek büyürler bölümlere ayrılırlar. Bu olan olay büyük bir sistematik şeklinde ilerlemeye devam eder ve vücudumuzdaki sağlığın devam etmesine imkan sağlar. Eğer, yeni olan hücrelere gerek kalmadığı durumda hücreler çoğalmaya başlarsa, lazım olandan çok dokular meydana gelmeye başlar. Bu fazla olan dokulara tümör adı verilen bir hastalığın meydana gelmesine neden olurlar. Böylelikle ortaya çıkan fazla doku iyi türde olacak biçimde veya kötü türde olacak biçimde meydana gelebilir.

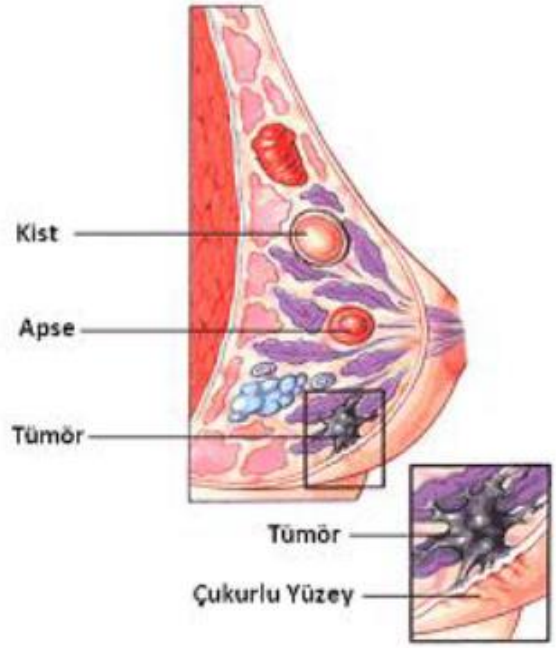
- Kötü türde mevcut olan tümörlere malignant tümörler şeklinde tanımlanır; yani iyi huylu biçimde olan benign tümörlerin tam tersine başka biçimde ve şekilde dokulara girme ve yayılma özelliği gösterirler. İyi huylu şekilde olan tümörler: Kanser kesinlikle değildirler. Bunların normal biçimde olacak şekilde yok edilme seçeneği vardır ve hiçbir zamanda meydana gelmezler. Kötü huylu içeren tümörler hakkında: Kanser hücreler anlamına gelmektedir. Kanserli hücrelerin denetimsiz biçimde büyüyerek bölünmeye başlarlar. Bunlar vücutta bulunan yakınlarındaki sağlıklı olan dokuların içerisine hücum ederek bunları bozabilirler. Kanser hücreleri bununla birlikte ayrı olacak biçimde ilk tümörden ayrılarak kopup kan dolaşımına veya lenf bölümüne hücum edip sızabilirler bu bölüme de girebilirler. Göğüs kanseri de kan yolundan çıkarak yayılarak vücudun başka bölümlerinde ya da parçalarında yeni tümörler meydana getirirler. Kanserli olan dokunun doku ya da dokuların yayılmasına metastaz oluşturması denmektedir. (Temiz, 2007).



Şekil 17 Meme Kanseri İle İlgili Hücre Örneği

Maalesef bazı durumlarda kanserin önceki aşamalarında olmasına karşın kanser hücreleri kan damarlarında olan yola hücum ederek kana karışabilir. Lakin kana karışmış olan kanser bölümlerinin birçoğu kanın içerisinde tahribata yol açar, bundan ötürü kan yolu ile yayılım olayı erken safhalarda genellikle meydana gelmez. Kanser, zaman içerisinde ilerledikçe kana bulaşan kanser hücrelerinin miktarı da yükselir, bu artışa paralel olacak şekilde kan yolu vasıtasıyla metastaz işleme alma olasılığı da yükselir. Kanser hücreleri kan yolu vasıtasıyla bütün vücuda dağılabilir şekilde yerleşip metastaz olabilmeleri için belli olacak bir seviyede oksijene gereksinim hissederler. (Temiz, 2007)

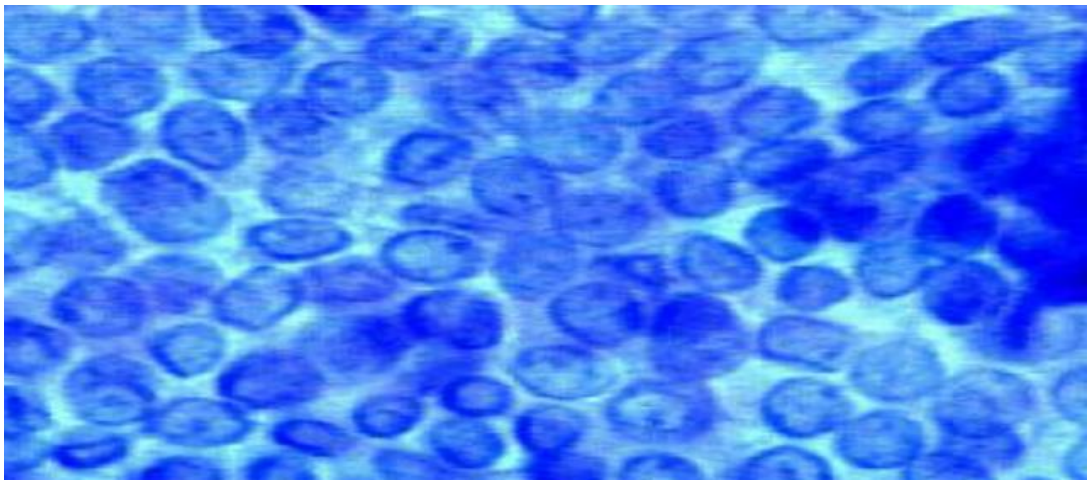
Şekil 17'de meme kanseri ile ilgili olan hücreleri görülebilmektedir Kadınlar arasında meme fizyonomisi, süt bezleri ve bulunan bölümde meydana gelen sütü meme ucu bölümüne aktaran bölümlerden meydana gelmektedir. Bu süt bezlerinde mevcut olan ve kanalları döşeyen hücrelerin; erkeklerde ise, nadir şekilde olacak biçimde görülse de, meme hücrelerinin kontrolden çıkacak şekilde çoğalmasına meme kanseri adı verilir. 2004 yılı içerisinde Dünya Sağlık Örgütü kayıtlarında geçen bilgiye göre aynı yıl içerisinde olan ölümlerin %13'ü kansere hastalığına bağlıdır. 7.4 milyon insan kanserden dolayı hayatını kaybetmiştir ve meme kanseri sebebinden ötürü 519 bin insan hayatını yitirmiştir. (Danacı; Mustafa Danacı, 2010)



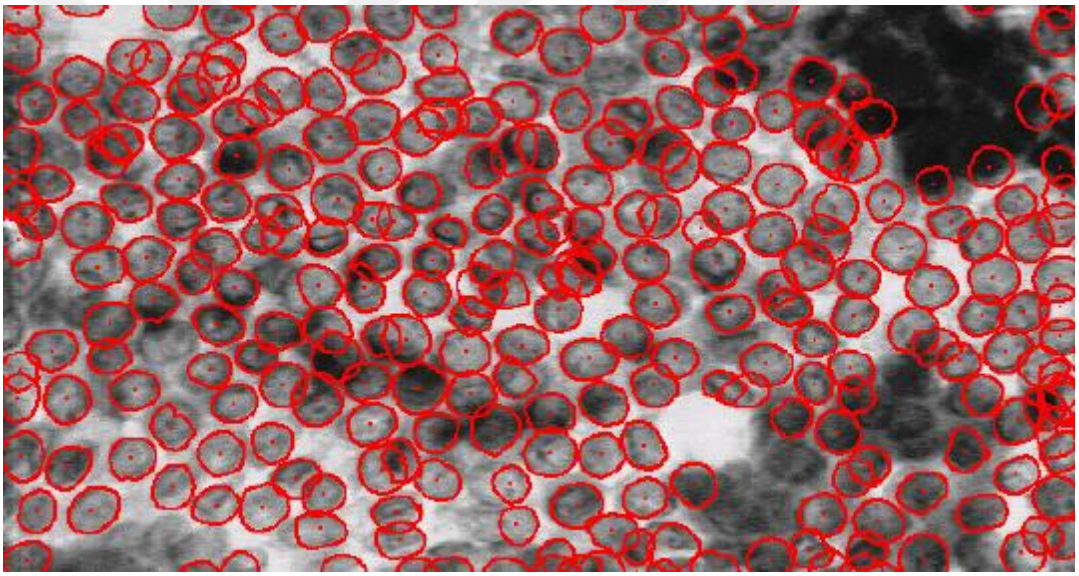
Şekil 18'de Memede Kist, Apse ve Tümör Hücrelerine Örnek

5.2.Örnek Uygulama

1990 senesinin başında Amerika Madison Eyaleti Wisconsin Üniversitesi'nde temeli atılmış olan bir yazılım projesinin ismidir. Yazılım, meme kanserindeki kritik iki önemli evresi olan tanı (diagnosis) ve tahmin (prognosis) safhalarını gerçekleştirmeyi amaçlayacak şekilde bir proje şeklinde ortaya çıkmıştır. Şekil 19'da kanser riski potansiyeli olan bir hastanın göğüs bölgesinden alınan doku örneğinin mikroskop çatısı altında 63 defa yakınlaştırılıp büyütülmesi ile elde edilen görüntü gözlemlenmektedir. Bu görüntüden yola çıkılarak kenarlardan çıkarma ve merkezlerden saptama algoritmaları vasıtasıyla Xcyt programına hücrelerin detaylarından tutun dokunun genel şablonuna kadar detaylar çıkarttırılmaktadır. Xcyt programı kullanılarak Şekil 20'de görüldüğü biçimdeki şekilde önce verilen hücre topluluğu gri skalaya döndürülmüştür. (Xcyt : A System for Remote Cytological Diagnosis and Prognosis of Breast Cancer)



Şekil 19 Meme Dokusu Altından Tespit Edilip Alınan 63 Kez Büyütülmüş Olan Hücre Topluluğu



Şekil 20 Merkez Tespit Etme ve Çerçeve Çıkartma

Hücrelerin büyük bölümü program tarafından tespit edildikten sonra hücrelere ait olan genel veriler ele geçmiştir.

Radius: Tüm hücrelerin yarıçaplarının ortalaması, standart sapması ve en berbat sonucu olarak ifade edilir.

Texture: İç yüz bölümlerinin gri skaladaki değişim çıktılarının ortalaması, standart sapması ve en berbat değeri olarak ifade edilir.

Perimeter: Hücrelerin çevre uzunluklarının ortalaması baz alınır, standart sapması ve en kötü değeri olacak şekilde ifade edilir.

Area: Hücrelerin yüzey alanları ortalamasına bakılarak, standart sapması ve en kötü değeri olarak ifade edilir.

Smoothness: Komşu hücrelerin yarıçap uzunluklarının ortalamasına bakılarak, standart sapması değeri ve en kötü değeri ifade edilir.

Compactness: Çevre²/Alan = Yoğunluk ortalaması değeri, standart sapması değeri ve en kötü değeri şeklinde ifade edilir.

Concavity: Hücre çevresindeki girinti ve çıkıntıların büyüklükleri ortalaması, standart sapması ve en kötü değerini ifade eder.

Concave Points: Hücre çevresindeki girintili ve çıkıntılı noktaların sayısının ortalaması, standart sapması ve en kötü değerini ifade eder.

Symmetry: Hücrelerin elips biçimdeki değişikliğin ortalaması, standart sapması ve en kötü değerini açıklar.

Fractal Dimension: İç içe geçmiş bir biçimde düzensiz olan hücrelerin bütün normal hücrelere oranının ortalaması, standart sapması ve en kötü değeri şeklinde ifade edilir.

5.3.Kullanılan Meme Kanseri –Wisconsin Veri Kümesi Özeti

Veri madenciliğin de dijital biçimde olan devasa veri yığınları içinden istenileni elde etmeyi ve elde edilen bilgiyi doğru biçimde dikkate almayı hedefler. Bu hedefle işleme alınan çalışmalarda yüz yüze gelinen en kapsamlı problem veri tabanının hatalı bir şekilde bilgiler dahil edilmesidir ya da birden fazla sayıda nitelik özelliklerinin tam olmayacak biçimde girilmesidir. Göğüs Kanseri - Wisconsin veri kümesi veri kaynağı gerçekten düzgün şekilde toplanmış ve dünyada tamamen kabul edilen ve iyi derecede bilinen; istatistiksel açıdan yapılan çalışmalarda sık bir şekilde kullanılan veri kaynağı olarak ifade edebiliriz. Veri kaynağı içerisindeki nitelikler belirli bir şekil dahilindedir ve nitelikler ile ilgili bilgiler hakkında ifadeleri veri kaynağı ile bir araya getirilerek kullanıma sunulmaktadır. Böyle nedenlerden ötürü

Wisconsin veri kümesinin bu şekilde olacak olan programda kullanılması uygulamanın çıktılarını çok daha sağlam olduğunu ortaya koyacaktır. Veri kaynağınız ne denli düzgün ve sağlıklı bir şeyden meydana gelirse gelsin, veri madenciliği yazılımında ya da programında işleme sokabilmek için veriler üzerinde hali hazırda önişlem işini yapmak gerekmektedir.

Meme-kanseri-Wisconsin alt bölümünde mevcut olanda altı yüz doksan dokuz hasta ile ilgili örnekleri dahil olmaktadır. Bu ayrıntılar 9 değişik çeşitteki skalada dış görünüm ve kromozom farklılıklarını ölçmektedir. Tüm veriler bir ile on içerisinde değişen çıktıları elinde bulundurur.

Dokuz adet tamsayı şeklinde olan özellik ve öznitelikleri bilgileri alt kısımdaki şekilde ifade edilmiştir:

Alanlar

1. Clump Thickness , -> Clump Kalınlığı 1 - 10 1 - 10
2. Uniformity of Cell Size, ->Hücre Boyutu Düzenliliği 1 - 10 1 - 10
3. Uniformity of Cell Shape, ->Hücre ġekli Uniformity 1 - 10 1 - 10
4. Marginal Adhesion, -> Marjinal Yapışma 1 - 10 1 - 10
5. Single Epithelial Cell, ->Tek Epitel Hücre Boyutu 1 - 10 1 - 10
6. Bare Nuclei , ->Çıplak çekirdeklerin 1 - 10 1 - 10
7. Bland Chromatin, ->Bland Kromatin 1 - 10 1 - 10
8. Normal Nucleoli , ->Normal nükleol 1 - 10 1 - 10
9. Mitoses , ->Mitoz 1 - 10 1 - 10
10. Classes, ->Sınıflar

Bu durum esasında bir meme içinde olan bir kitlenin kanser olmayan (benign) veya kanser (malignant) olduğunu tespit etmek ve değerlendirmek için ince iğne aspirasyonları bir patoloji sonuçlarında olan terimleri kapsar.

Wisconsin veri kümesi değişik kanser bölümlerini içeren ve bilimsel araştırmalarda yüksek derecede öneme sahip olan, güvenilir, iyi belgelenmiş, çok nadir olan bir veri kümesidir. Bu mevcut olan veri kaynağı olduğunca düzenli ve iyi belgelenmiş olsa da üzerinde yaptığım çalışma ile ilgili olarak bir önişlemden geçirilmesi gerekmiştir. Bu üzerinde durulan çalışmada yıllık olacak biçimde güncellenen Wisconsin veri kümesinin bin dokuz yüz doksan bir yılında mevcut versiyonu işleme alınmıştır. Örnek vermek gerekirse kanser hücreleri boyut açısından ve şekil itibarıyla değişebilen eğilimindedir.

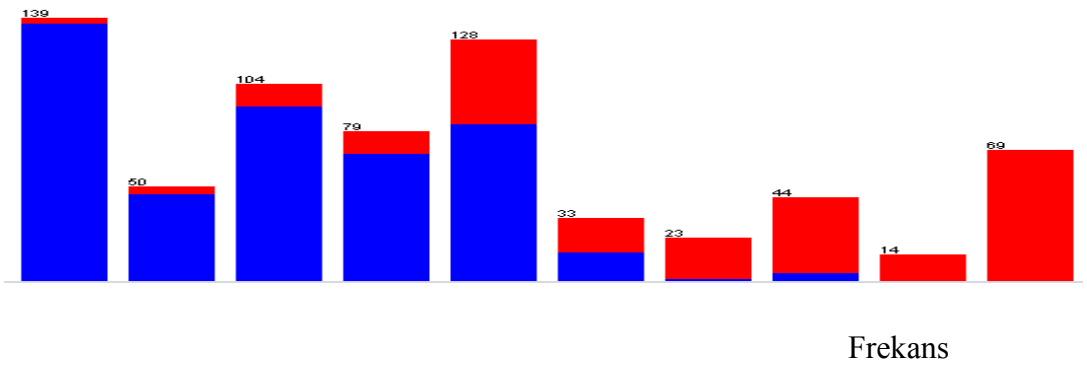
Yapılan çalışma Cross Validation 10 Fold değerleri seçilerek işleme alınmış olmakla birlikte bütün verinin %90'ı öğrenme %10'u da test kümesinden meydana gelmektedir.

Yapılan çalışmada üç analiz yapılıyor. Bu kapsamda yapılan çalışmalardan ilkinde 16 tane olan ve içinde kayıp veriler barındıran settekiler tamamen çıkartılarak işleme alınıyor. Sonraki yapılan çalışmada kısmen fazla kayıp olmayan veri setleri tespit edilip bunlar 683 veriye eklenerek 690 veri ile işleme girmiş ve sonuçlar elde edilmiştir. En son çalışmada Missing Values Replace modülü kullanılarak 16 tane eksik verileri iyileştirerek tam veri seti ile işlem yapılmıştır. Elde edilen çıktılar neticesinde tüm sonuçlar birbirleri arasında karşılaştırılmıştır.

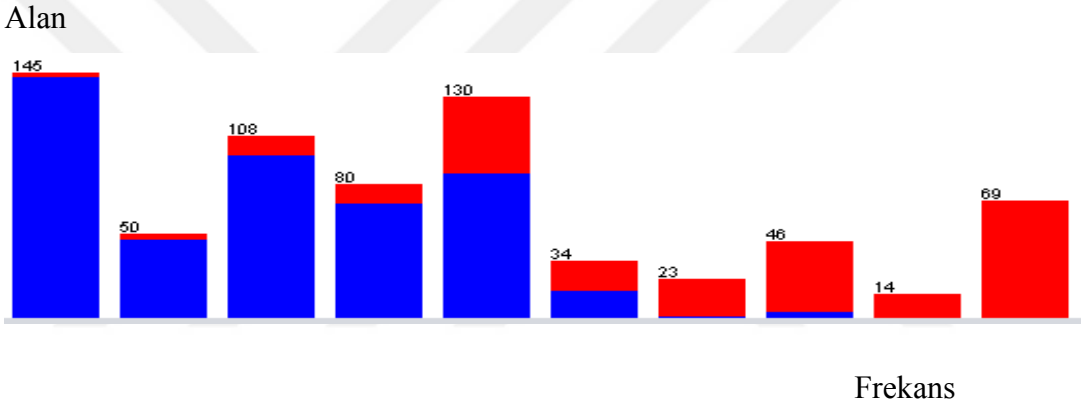
5.4.Clump Kalınlığı

Clump kalınlığı: iyi türde olacak şekilde olan hücreler kanserli biçimde olan hücreleri genel olarak fazla katmanlardan oluşarak gruplandırılmıştır, tek olacak şekilde olan katmanların gruplandırılması şeklindedir. (Wang, 2003)

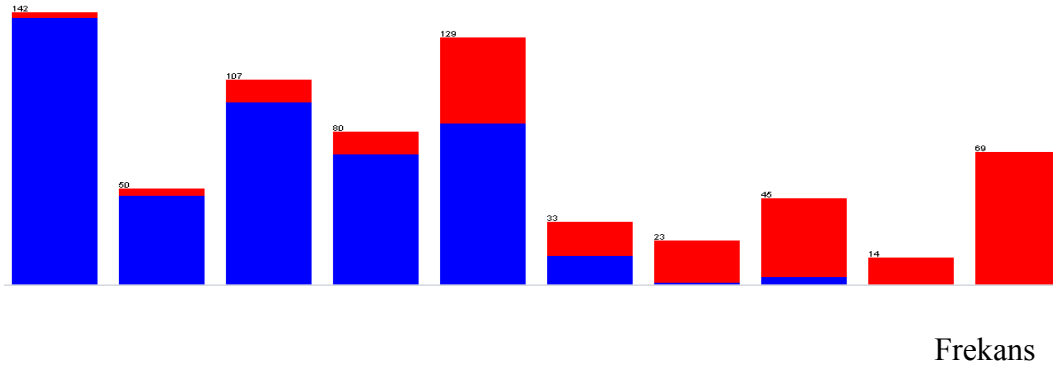
16 Tane hatalı verinin çıkartılmasıyla elde edilen grafik



16 Tane verinin boşlukları doldurulup 699 tam veri setiyle elde edilen grafik



9 Tane Eksik Verinin Çıkarılmasıyla Elde Edilen Grafik



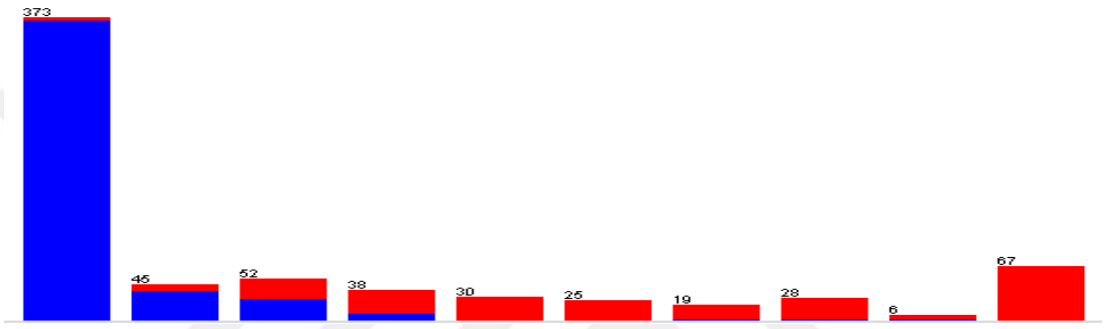
Tablo 26 Clump Kalınlığını Gösteren Grafikler

5.5.Hücre Boyutu Düzenliliği

Kanser biçiminde mevcut olan hücrelerinin ebat ve boyutu değişme şekli eğilimindedir. Bu elde edilen veriler ışığında hücrelerin kanserli biçimde olup olmadığını tespit etmede çok önemli pozisyon teşkil eder. (Wang, 2003)

16 Tane hatalı verinin çıkartılmasıyla elde edilen grafik

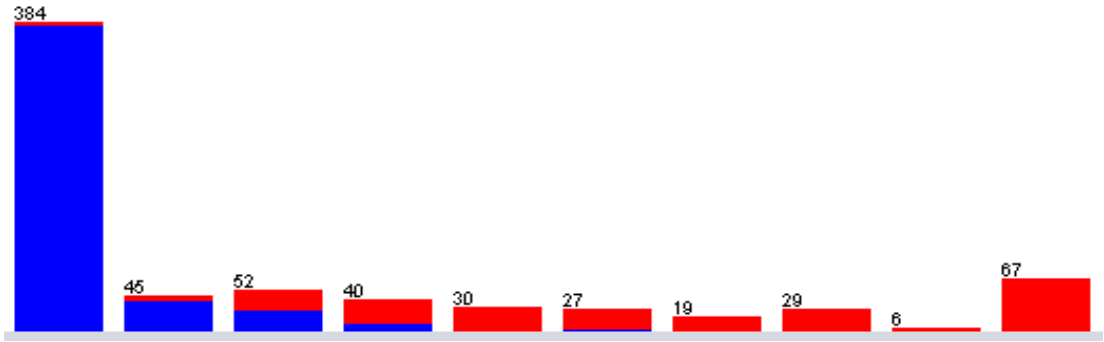
Alan



Frekans

16 Tane verinin boşlukları doldurulup 699 tam veri setiyle elde edilen grafik

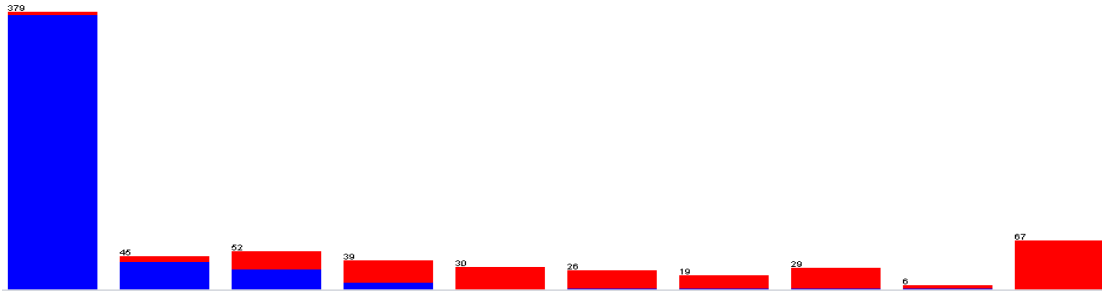
Alan



Frekans

9 Tane Eksik Verinin Çıkarılmasıyla Elde Edilen Grafik

Alan



Frekans

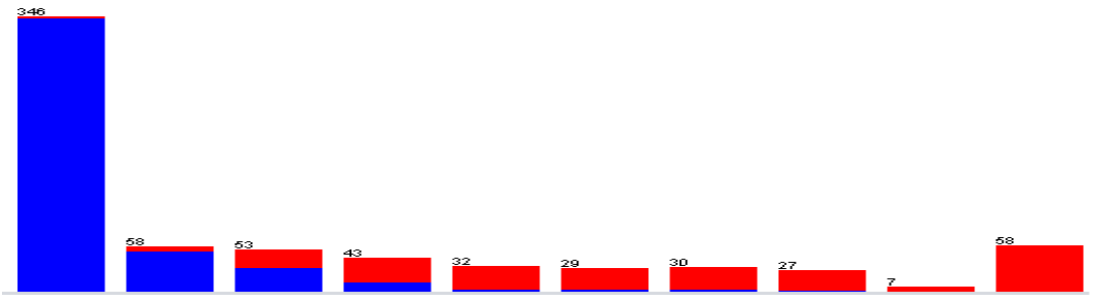
Şekil 21 Hücre Boyutu Düzenliliği Grafiği

5.6.Hücre Şekil Düzenliliği

Kanser şeklinde olan hücrelerin şekli (sınırların net biçimde olup olmaması yüzey kısmının engebeli biçimde bulunup bulunmaması) de değişme yönelimindedir. Bu parametrelere bakılarak bölgelerin kanserli biçimde olup olmadığını tespit etmede çok büyük önemli yer alır.

16 Tane hatalı verinin çıkartılmasıyla elde edilen grafik

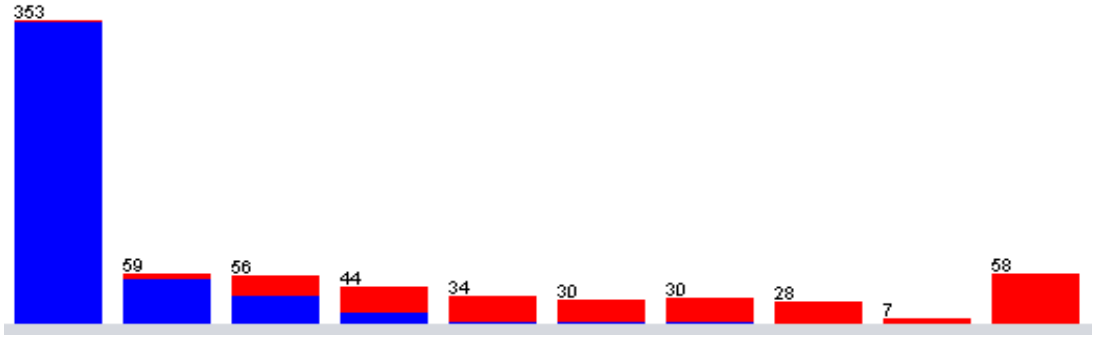
Alan



Frekans

16 Tane verinin boşlukları doldurulup 699 tam veri setiyle elde edilen grafik

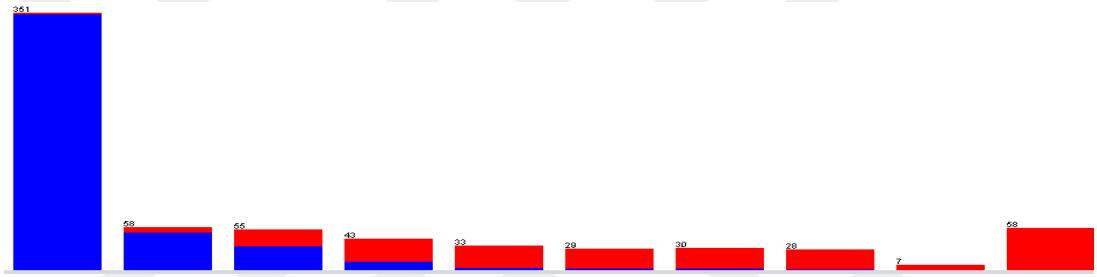
Alan



Frekans

9 Tane Eksik Verinin Çıkarılmasıyla Elde Edilen Grafik

Alan



Frekans

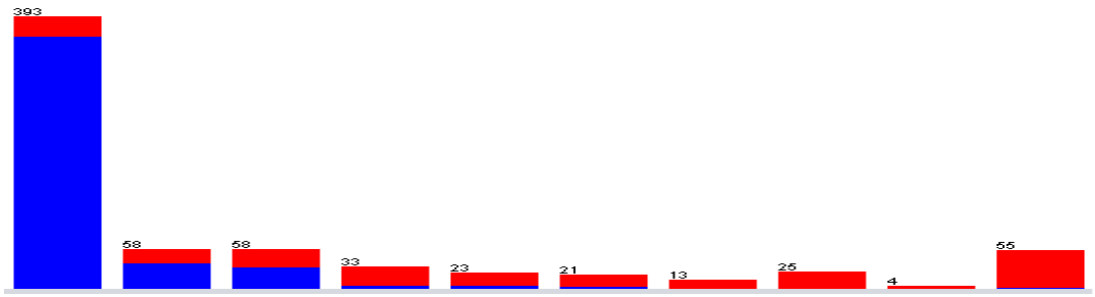
Şekil 22 Hücre Şekil Düzenliliği Grafiği

5.7.Marjinal Yapışma

Normal olan hücreler birbirine yapışma gibi bir yöne yönelirler. Kanserli biçimde mevcut olan hücreleri bu yeteneği loos'un yönelimindedir. Bu yüzden ötürü yapışma gibi bir kaybı malignite işaretidir. (Wang, 2003)

16 Tane hatalı verinin çıkartılmasıyla elde edilen grafik

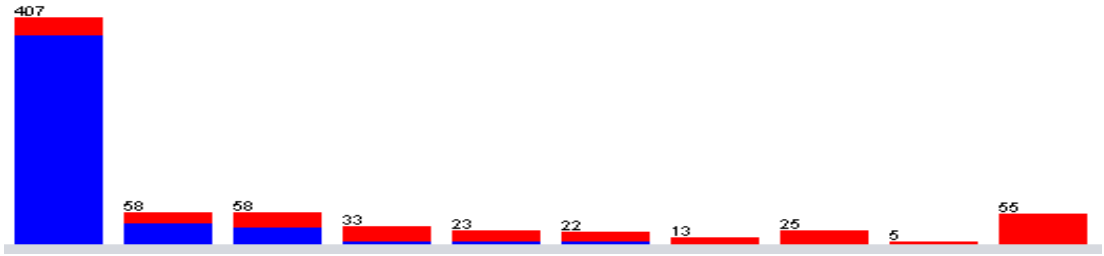
Alan



Frekans

16 Tane verinin boşlukları doldurulup 699 tam veri setiyle elde edilen grafik

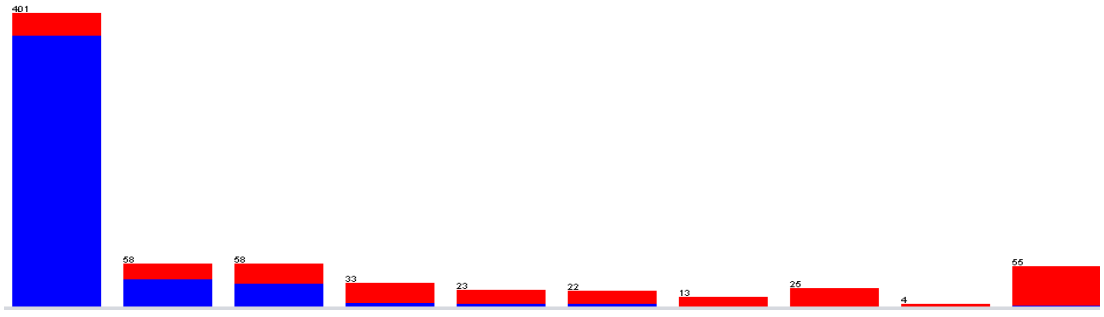
Alan



Frekans

9 Tane Eksik Verinin Çıkarılmasıyla Elde Edilen Grafik

Alan



Frekans

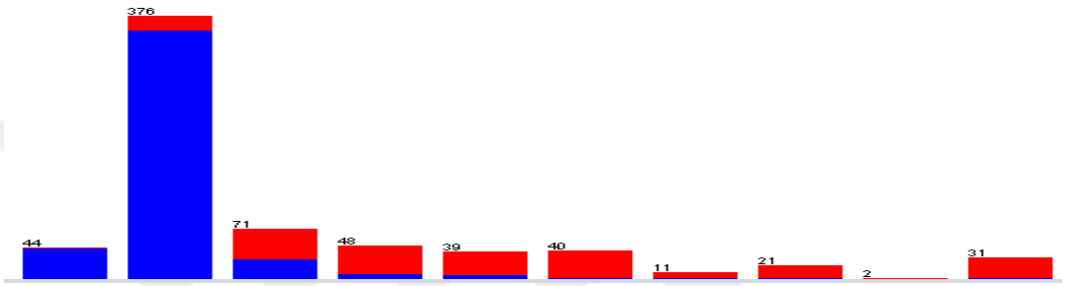
Şekil 23 Marjinal Yapışma Grafiği

5.8.Tek Epitel Hücre Boyutu

Önemli sayıda büyütüldüğü epitel olan hücreleri kötü olacak biçimde hücre olabilir.

16 Tane hatalı verinin çıkartılmasıyla elde edilen grafik

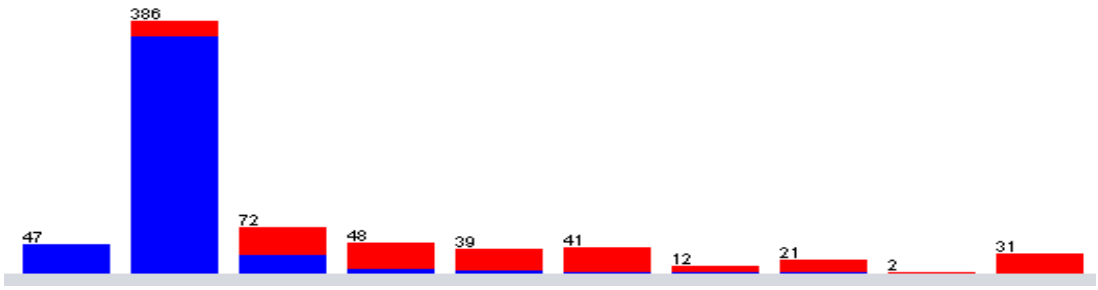
Alan



Frekans

16 Tane verinin boşlukları doldurulup 699 tam veri setiyle elde edilen grafik

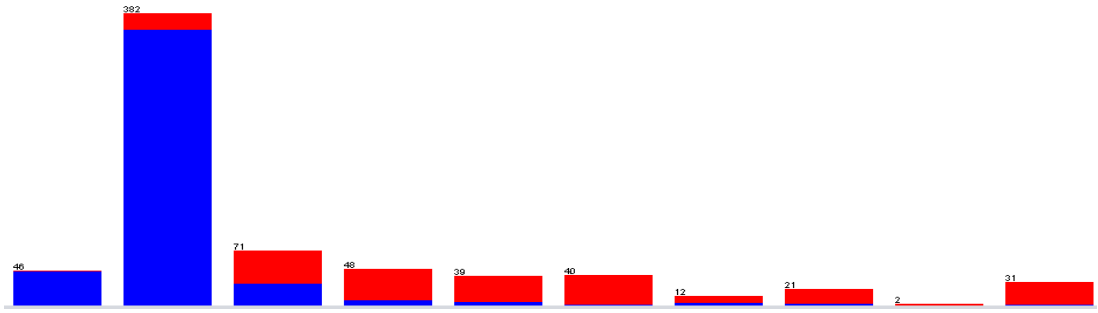
Alan



Frekans

9 Tane Eksik Verinin Çıkarılmasıyla Elde Edilen Grafik

Alan



Frekans

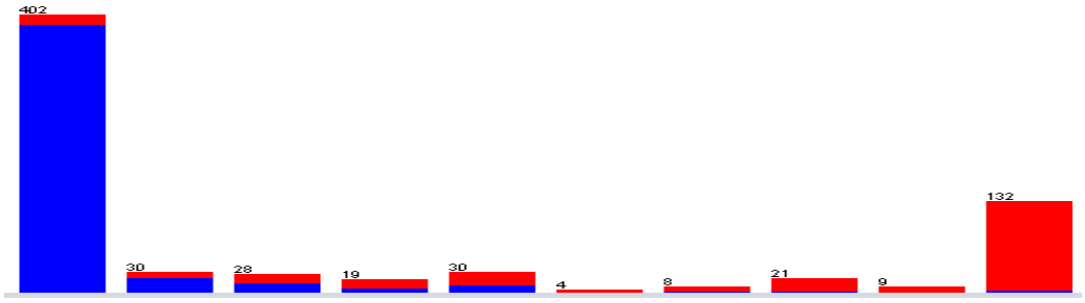
Şekil 24 Tek Epitel Hücre Boyutu Grafiği

5.9.Çıplak Çekirdekler

Bu olayda sitoplazması (hücrenin kalanı) ile çevrelenmiş bir durum söz konusu değildir. Esasında çekirdeklerini ifade etmek için kullanılan bir ifadedir. Bunlar genel olarak iyi türde biçimde tümörler görülebilir.

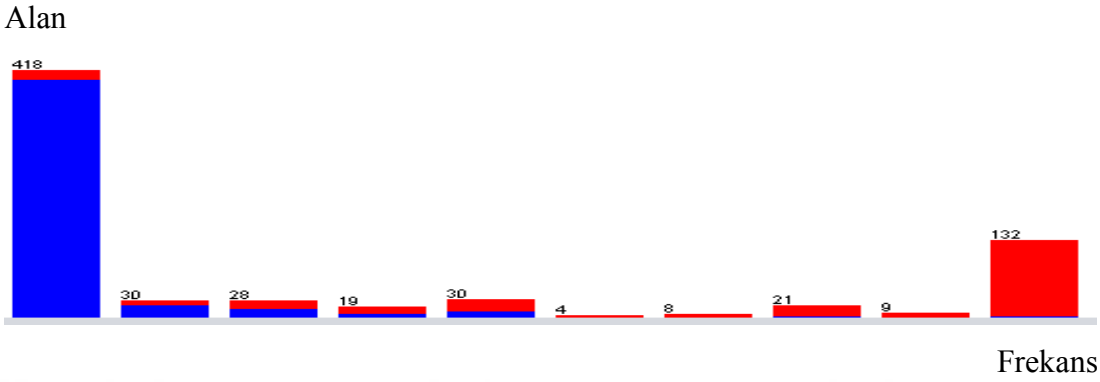
16 Tane hatalı verinin çıkartılmasıyla elde edilen grafik

Alan

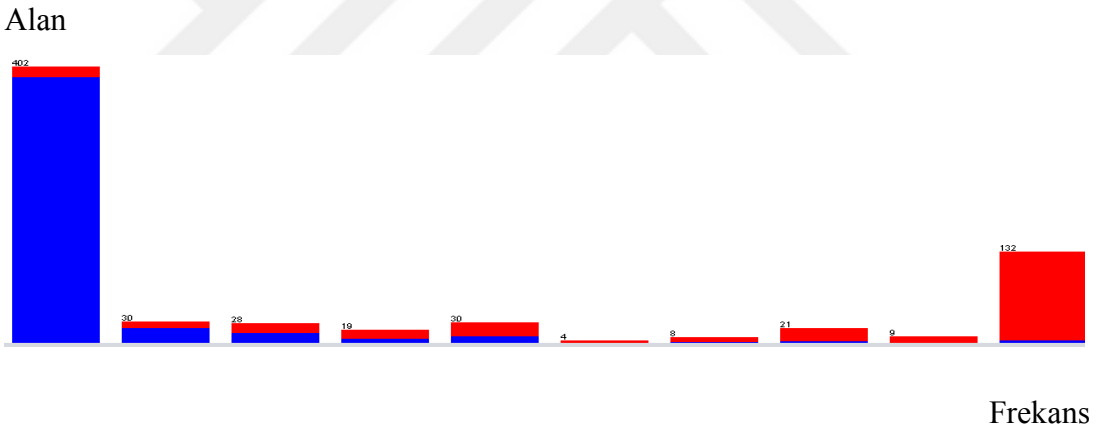


Frekans

16 Tane verinin boşlukları doldurulup 699 tam veri setiyle elde edilen grafik



9 Tane Eksik Verinin Çıkarılmasıyla Elde Edilen Grafik

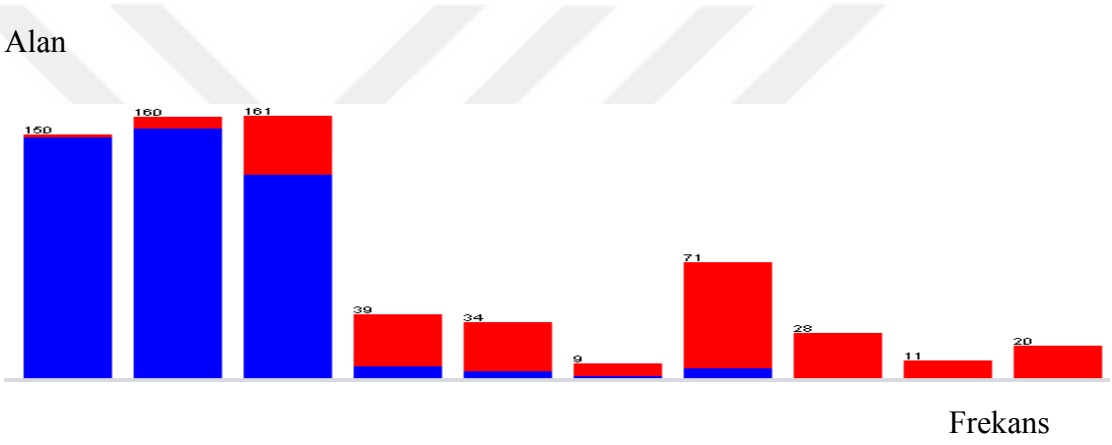


Şekil 25 Çıplak Çekirdekler Grafiği

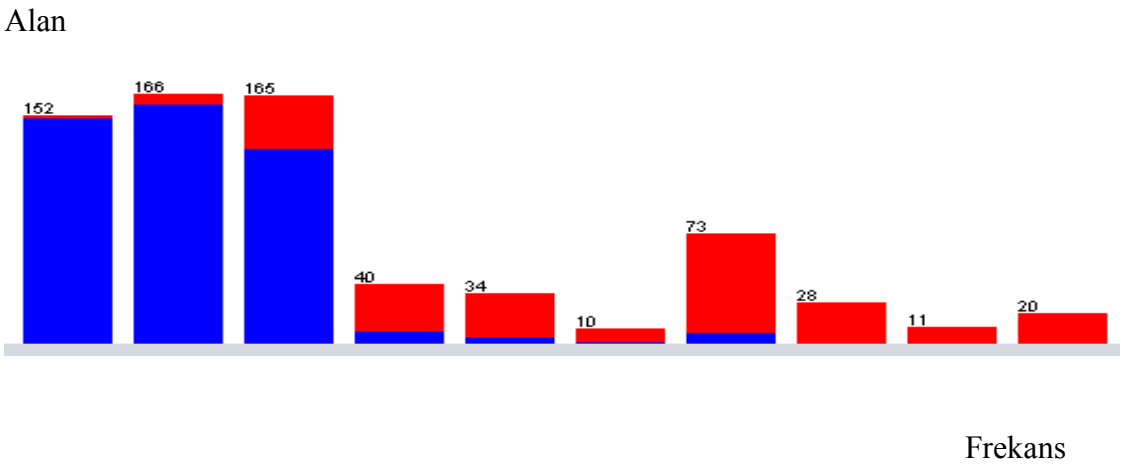
5.10.Bland Kromati

Huylu biçimde olan ve davranan hücrelerde gözlemlenen çekirdek yeknesak bir "doku" olarak açıklanabilir. Kanser olan bölümlerinde kromatinin ebatlarının çok daha büyük hale gelme eğiliminde olabilirler.

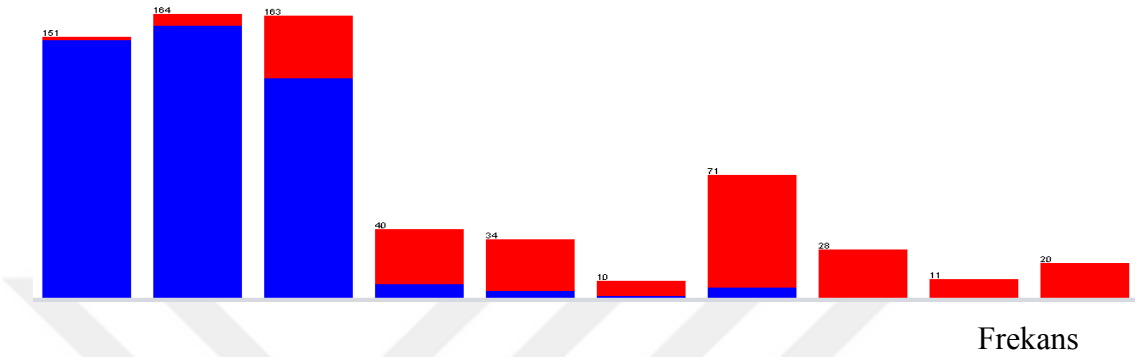
16 Tane hatalı verinin çıkartılmasıyla elde edilen grafik



16 Tane verinin boşlukları doldurulup 699 tam veri setiyle elde edilen grafik



9 Tane Eksik Verinin Çıkarılmasıyla Elde Edilen Grafik

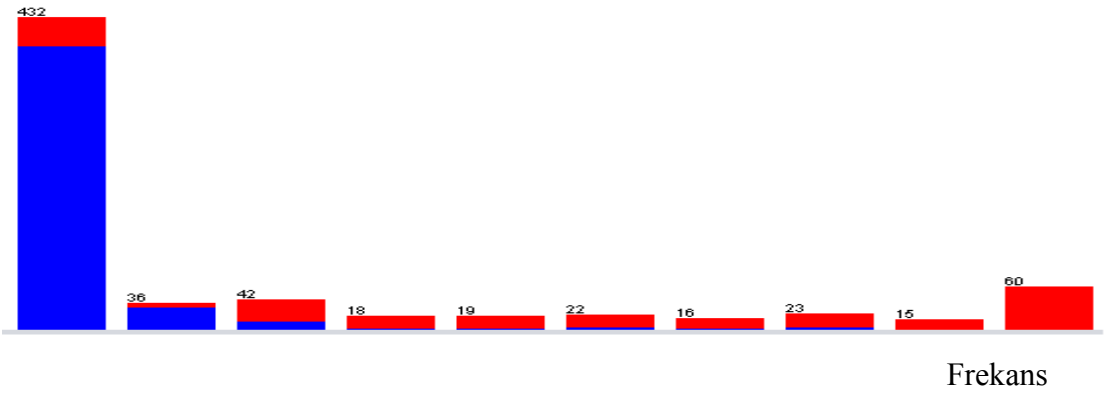


Şekil 26 Bland Kromati Grafiği

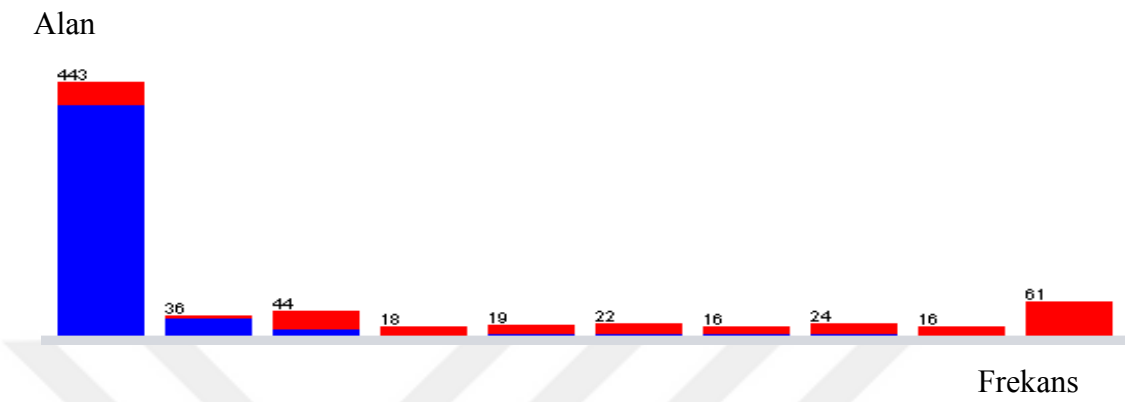
5.11.Normal Nükleol

Nükleol çekirdeği içerisinde tespit edilen minik bölümlerdir. Görünmüyorsa, normal olan hücrelerde çekirdekçik genel olarak fazla minik olur. Kanser hücrelerinde nükleol oranı daha fazla belirgin biçime gelir . (Wang, 2003)

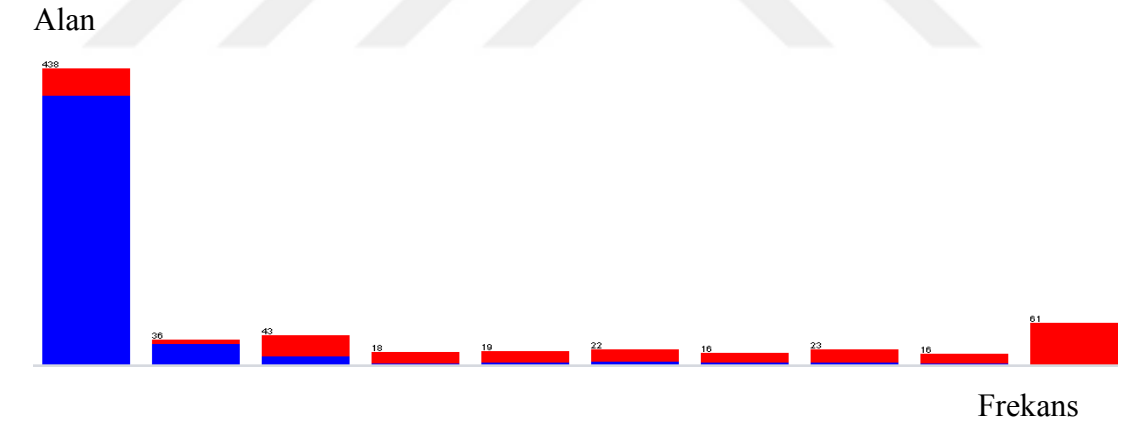
16 Tane hatalı verinin çıkartılmasıyla elde edilen grafik



16 Tane verinin boşlukları doldurulup 699 tam veri setiyle elde edilen grafik



9 Tane Eksik Verinin Çıkarılmasıyla Elde Edilen Grafik

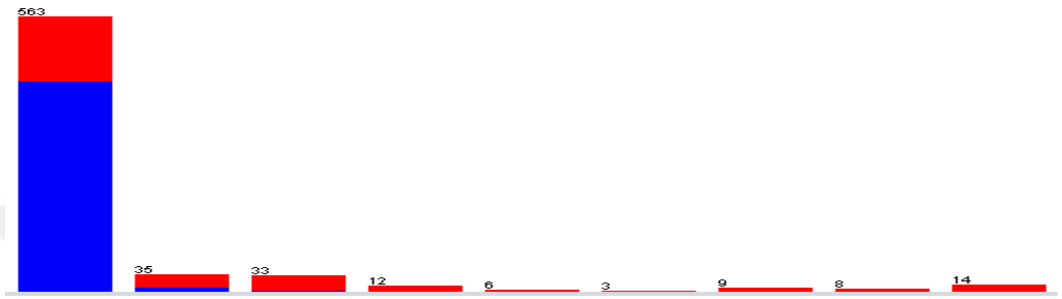


Şekil 28 Normal Nükleol Grafiği

5.12.Mitoz

16 Tane hatalı verinin çıkartılmasıyla elde edilen grafik

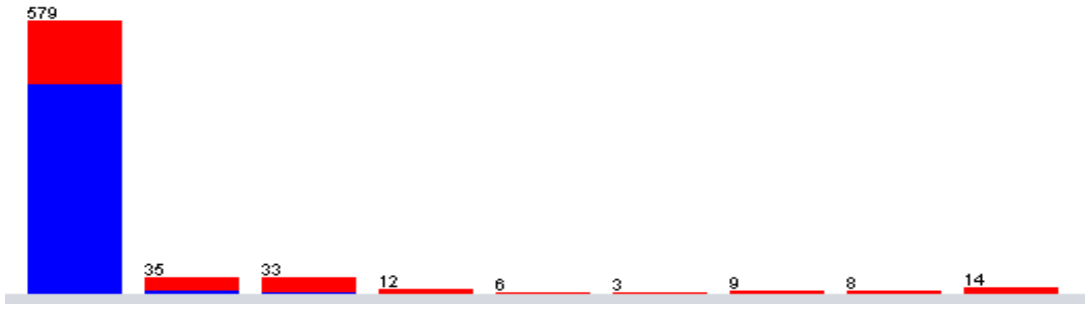
Alan



Frekans

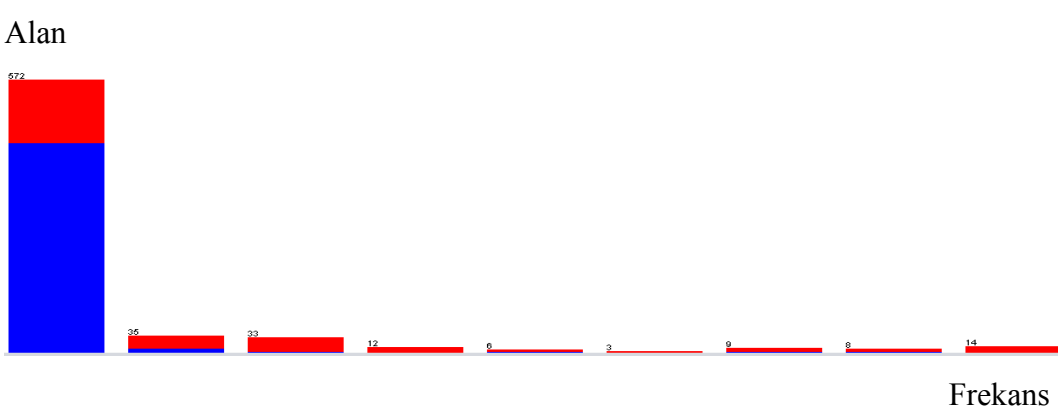
16 Tane verinin boşlukları doldurulup 699 tam veri setiyle elde edilen grafik

Alan



Frekans

9 Tane Eksik Verinin Çıkarılmasıyla Elde Edilen Grafik



Şekil 29 Mitoz Grafiği

6 WEKA KULLANILARAK MEME KANSERİ HÜCRELERİNİN TAHMİNİ

Verilerin önişleme bölümünde aşamasında gerçekleştirdiğim şey nitelik tercihi, verilerin tamamlanması tarzında olan analiz çıktılarını doğrudan etkileyici işlemlerde fark etmeden model çıkarımını etkileyici biçimde işlem ya da işlemler gerçekleştirilmiş olabilir. Farklı biçimlerde yapılan ön işlemelerde meydana getirilen verilerin analiz çıktılarının değişik biçimde elde edilmesi kaçınılmazdır. Yapılan işlemlerde meme kanserinden rahatsız olanların kayıtları detaylı şekilde tetkik edilmiş, hastaların yaşayıp yaşamadıkları, hayatta olmayanlar için ne kadar zaman yaşayabildikleri ve ölüm nedenleri dikkate alınarak tutulan sıradan bir hastanın bu mevcut olan hastalıktan kurtulup kurtulamayacağı sınıflandırılarak ileriye yönelik olacak tahminlerde tespit edebilme hedefi ile farklı algoritmalarla meydana getirilen modellerin elde ettiği başarıların dereceleri aralarında karşılaştırılmıştır. Programda

kullanılan bir karar ağacı algoritması olan ve temeli itibarıyla ID3 ve C4.5 algoritmalarını temel alan J48, istatistiksel bir algoritmadan meydana gelen Bayes sınıflandırma algoritmalarından biri olan Naive-Bayes, regresyon şeklinde olan algoritmalardan lojistik regresyon ve örnek biçimde olan sınıflandırma algoritmalarından birisi Kstar algoritmaları ele alınarak modeller meydana getirilmiş ve oluşturulan biçimdeki modellerin başarım dereceleri birbirleri arasında karşılaştırılmıştır.

A-6.1.Karar Ağacı Modelinin Başarım Ölçütleri

Weka yazılımının tanıyabileceği Arff uzantılı formata döndürüldükten sonra Weka ana menüde de bunlar önişlemden geçirilmiştir. Önişlemlerden yapılanlardan sonra sonuç olarak elde edilmiş Arff formatında olan altı yüz seksen üç kayıt bulunduran göğüs kanseri hastalıkları Wisconsin verileri içerisinde incelenerek sadece bir tane algoritma tercih edilerek başarıları seviyeleri tespit edilerek birbirleriyle karşılaştırılmıştır. Karşılaştırılacak olan bu algoritmalar tercih edilip seçilirken dikkate alınan bu algoritmaların popülerliği ve bu kapsamda literatürde aynı şekilde ve konuda gerçekleştirilen işler dikkatli şekilde ele alınmıştır. Weka yazılımı kullanılarak model meydana getirilirken kullanılabilecek birden fazla karar ağaçları algoritmaları vardır. Şekil 30'da meydana getirilen modelin test çıktılarına ait istatistiksel veriler ve karışıklık matrisi tespit edilmektedir. Şekil 31'de ise gerçekleştirilecek olan karışıklık matrisini kullanarak yapılan hesaplama sonucu yapılan karşılaştırma verileri verilmiştir.

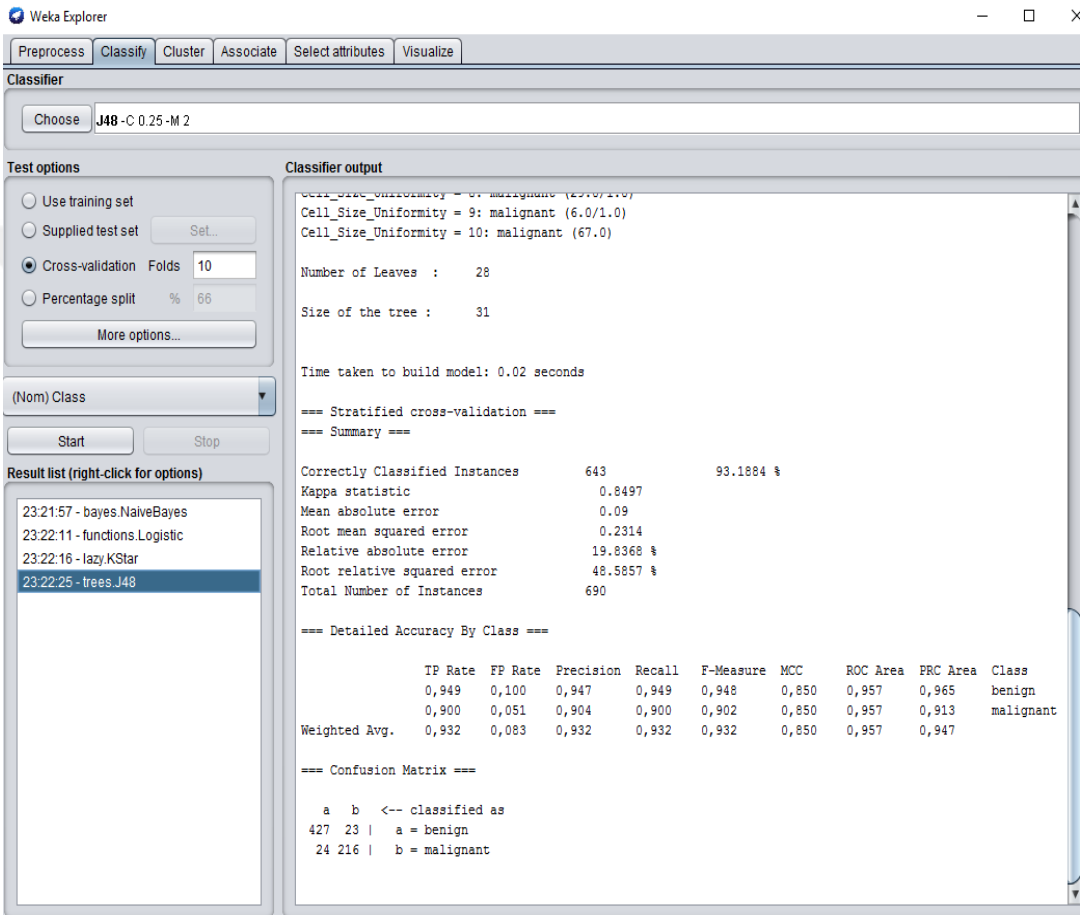
```
J48 pruned tree
------------------

uniformity <= 2
|   bareNuclei <= 3: 2 (394.0/2.0)
|   bareNuclei > 3
|   |   clump <= 3: 2 (11.0)
|   |   clump > 3
|   |   |   blandChromatin <= 2
|   |   |   |   marginalAdhesion <= 3: 4 (2.0)
|   |   |   |   marginalAdhesion > 3: 2 (2.0)
|   |   |   blandChromatin > 2: 4 (8.0)
uniformity > 2
|   uniformityofcellshape <= 2
|   |   clump <= 5: 2 (19.0/1.0)
|   |   clump > 5: 4 (4.0)
|   uniformityofcellshape > 2
|   |   uniformity <= 4
|   |   |   bareNuclei <= 2
|   |   |   |   marginalAdhesion <= 3: 2 (11.0/1.0)
|   |   |   |   marginalAdhesion > 3: 4 (3.0)
|   |   |   bareNuclei > 2: 4 (54.0/7.0)
|   |   uniformity > 4: 4 (174.0/3.0)

=== Summary ===

Correctly Classified Instances         653               95.7478 %
Incorrectly Classified Instances        29                4.2522 %
Kappa statistic                          0.9074
Mean absolute error                      0.0581
Root mean squared error                  0.2006
Relative absolute error                 12.7551 %
Root relative squared error             42.0437 %
Total Number of Instances              682
```

Şekil 30 J48 Karışıklık Matrisi Tablosu Grafiği

		Ön Görülen Sınıf	
		a=2	b=4
Doğru Sınıf	a=2	424	19
	b=4	10	229

Şekil 31 J48 Algoritmasına Ait Olan Modelin Karşılaştırma Yapacak Olan Ölçütleri Tablosu

Doğruluk	Kesinlik	Duyarlılık	F-Ölçütü
% 95.74	%97.7	%95.7	%96.7



Şekil 32 Dataset İçinde 9 Adet Kayıp Olan Veri Silindikten Sonra J48 Algoritmasının Verdiği Sonuç

Yapılan çalışma neticesinde elde edilen karışıklık matrisi ve doğruluk, kesinlik, duyarlılık, F-ölçütleri aşağıdaki gibi değişime uğramıştır.

Şekil 33 Dataset İçinde 9 Adet Kayıp Olan Veri Silindikten Sonraki Yeni J48 Karışıklık Matrisi Tablosu Grafiği

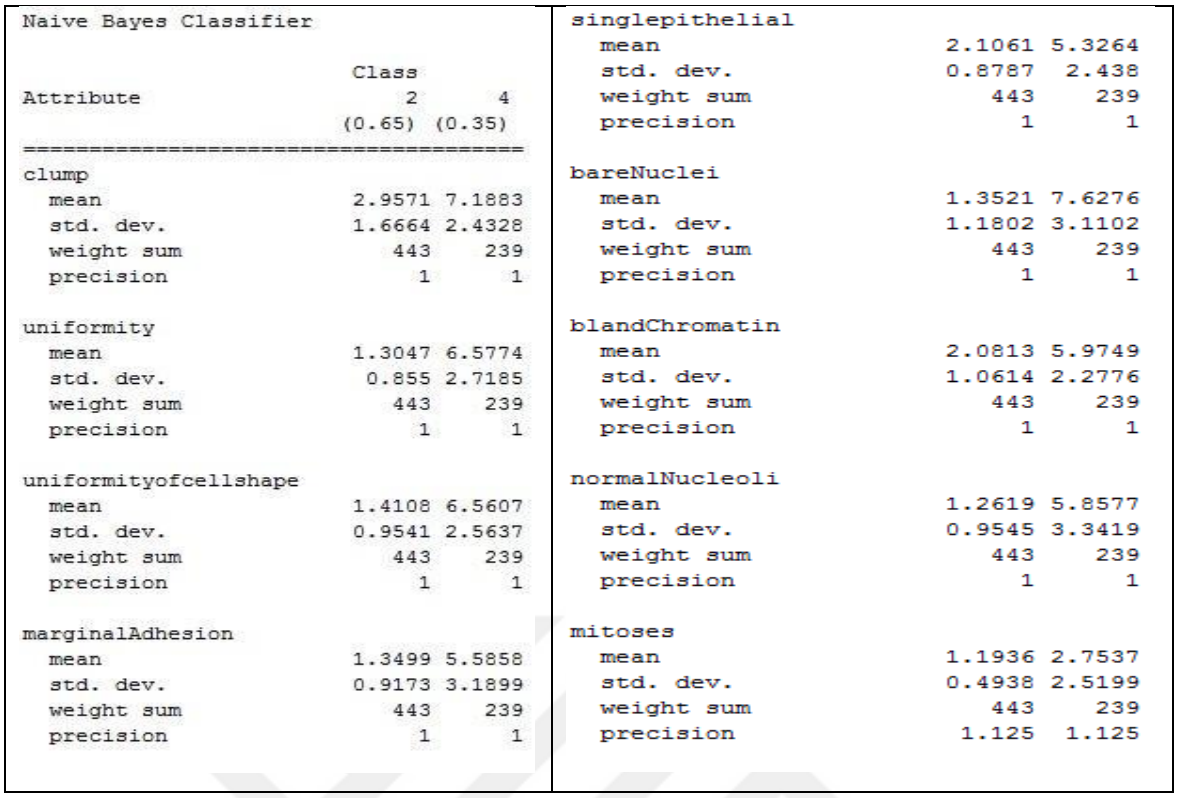
		Ön Görülen Sınıf	
		a=2	b=4
Doğru Sınıf	a=2	427	23
	b=4	24	216

Şekil 34 Dataset İçinde 9 Adet Kayıp Olan Veri Silindikten Sonraki Yeni J48 Algoritmasına Ait Olan Modelin Karşılaştırma Yapacak Olan Ölçütleri Tablosu

Doğruluk	Duyarlılık	Kesinlik	F-Ölçütü
%93.18	%93.2	%93.2	%93.2

A-6.2.Bayes(İstatistiksel) Sınıflandırma Modelinin Başarım Ölçütleri

Weka ’da mevcut halde bulunan algoritmaların içerisinden NaiveBayes algoritması tercih edilerek veri kümesi içerisinde çalıştırılmıştır. Şekil 35’te meydana getirilen modelin test çıktılarına ait olan istatistiklerin karışıklık matrisi çıktıları gözlemlenmektedir.



```
Naive Bayes Classifier

                          Class
Attribute                     2      4
                           (0.65) (0.35)
=======================================
clump
  mean                   2.9571 7.1883
  std. dev.              1.6664 2.4328
  weight sum                443    239
  precision                   1      1

uniformity
  mean                   1.3047 6.5774
  std. dev.               0.855 2.7185
  weight sum                443    239
  precision                   1      1

uniformityofcellshape
  mean                   1.4108 6.5607
  std. dev.              0.9541 2.5637
  weight sum                443    239
  precision                   1      1

marginalAdhesion
  mean                   1.3499 5.5858
  std. dev.              0.9173 3.1899
  weight sum                443    239
  precision                   1      1

singlepithelial
  mean                   2.1061 5.3264
  std. dev.              0.8787  2.438
  weight sum                443    239
  precision                   1      1

bareNuclei
  mean                   1.3521 7.6276
  std. dev.              1.1802 3.1102
  weight sum                443    239
  precision                   1      1

blandChromatin
  mean                   2.0813 5.9749
  std. dev.              1.0614 2.2776
  weight sum                443    239
  precision                   1      1

normalNucleoli
  mean                   1.2619 5.8577
  std. dev.              0.9545 3.3419
  weight sum                443    239
  precision                   1      1

mitoses
  mean                   1.1936 2.7537
  std. dev.              0.4938 2.5199
  weight sum                443    239
  precision                1.125  1.125
```

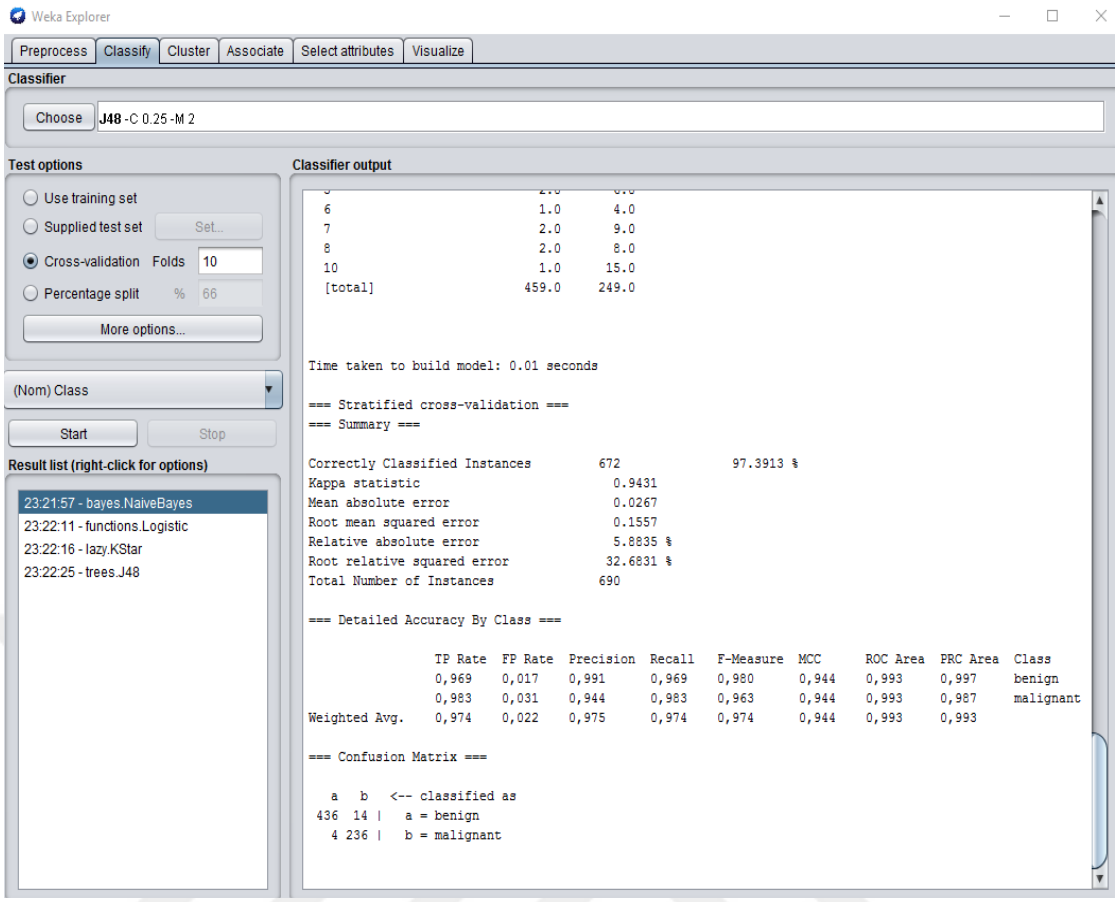
Şekil 35 Naive Bayes Karışıklık Matrisi Grafiği

		Ön Görülen Sınıf	
		a=2	b=4
Doğru Sınıf	a=2	424	19
	b=4	6	233

Şekil 36 Bayes (İstatistiksel) Sınıflandırma Modelinin Algoritmasına Ait Modelin

Şekil 37 Karşılaştırma Ölçütleri Tablosu

Doğruluk	Kesinlik	Duyarlılık	F-Ölçütü
% 96.33	%98.6	% 95.7	% 97.1



Şekil 38 Dataset İçinde 9 Adet Kayıp Olan Veri Silindikten Sonra Naive Bayes Algoritmasının Verdiği Sonuç

Yapılan çalışma neticesinde elde edilen karışıklık matrisi ve doğruluk, kesinlik, duyarlılık, F-ölçütleri aşağıdaki gibi değişime uğramıştır.

Şekil 39 Dataset İçinde 9 Adet Kayıp Olan Veri Silindikten Sonraki Yeni Naive Bayes Karışıklık Matrisi Tablosu Grafiği

		Ön Görülen Sınıf	
		a=2	b=4
Doğru Sınıf	a=2	436	14
	b=4	4	236

Şekil40 Dataset İçinde 9 Adet Kayıp Olan Veri Silindikten Sonraki Yeni Naive Bayes Algoritmasına Ait Olan Modelin Karşılaştırma Yapacak Olan Ölçütleri Tablosu

Doğruluk	Duyarlılık	Kesinlik	F-Ölçütü
%97.39	%97.5	%97.4	%97.4

A-6.3.Regresyon Modelinin Başarım Ölçütleri

Karşılaştırma amaçlı bir şekilde olacak olan regresyon tabanlı tekniklerden lojistik regresyon algoritması tercih edilerek veri kaynağına tatbik edilmiştir. Şekil 42'de meydana getirilen yapının test çıktılarına sahip olunan bilgileri ve karışıklık matrisi açıkça görülebilmektedir. Şekil 41 ' da ise karışıklık matrisini işleme alınarak karşılaştırma çıktı sonuçları verilmiştir.

```
Logistic Regression with ridge parameter of 1.0E-8
Coefficients...
                           Class
Variable                       2
=================================
clump                    -0.5353
uniformity                0.0065
uniformityofcellshape    -0.3235
marginalAdhesion         -0.3302
singlepithelial          -0.0965
bareNuclei               -0.3826
blandChromatin           -0.4462
normalNucleoli           -0.2127
mitoses                  -0.5335
Intercept                10.0958


Odds Ratios...
                           Class
Variable                       2
=================================
clump                     0.5855
uniformity                1.0065
uniformityofcellshape     0.7236
marginalAdhesion          0.7188
singlepithelial            0.908
bareNuclei                0.6821
blandChromatin            0.6401
normalNucleoli            0.8084
mitoses                   0.5866
```

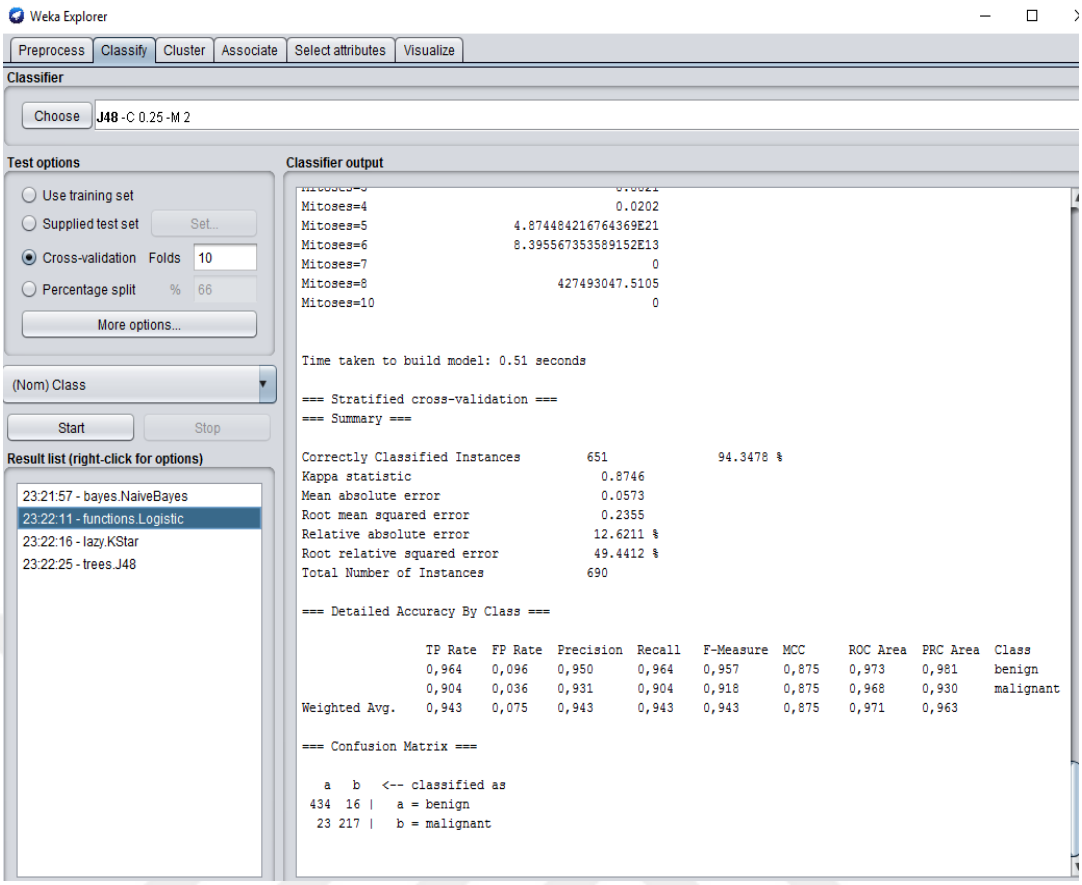
```
Time taken to build model: 0.11 seconds

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

Correctly Classified Instances         661               96.9208 %
Incorrectly Classified Instances        21                3.0792 %
Kappa statistic                          0.9323
Mean absolute error                      0.0444
Root mean squared error                  0.1594
Relative absolute error                  9.7599 %
Root relative squared error             33.4087 %
Total Number of Instances              682
```

Şekil 42 Lojistik Regresyon Algoritması Karışıklık Algoritması Grafiği

		Ön Görülen Sınıf	
		a=2	b=4
Doğru Sınıf	a=2	433	10
	b=4	11	228

Şekil 43 Lojistik Regresyon Algoritmasına Ait Modelin Karşılaştırma Ölçütleri Grafiği

Doğruluk	Kesinlik	Duyarlılık	F-Ölçütü
%96.92	%97.5	%97.7	%97.6



Şekil 44 Dataset İçinde 9 Adet Kayıp Olan Veri Silindikten Sonra Lojistik Regresyon Algoritmasının Verdiği Sonuç

Yapılan çalışma neticesinde elde edilen karışıklık matrisi ve doğruluk, kesinlik, duyarlılık, F-ölçütleri aşağıdaki gibi değişime uğramıştır.

Şekil 45 Dataset İçinde 9 Adet Kayıp Olan Veri Silindikten Sonraki Yeni Lojistik Regresyon Karışıklık Matrisi Tablosu Grafiği

		Ön Görülen Sınıf	
		a=2	b=4
Doğru Sınıf	a=2	434	16
	b=4	23	217

Şekil 46 Dataset İçinde 9 Adet Kayıp Olan Veri Silindikten Sonraki Yeni Lojistik Regresyon Algoritmasına Sahip Olan Modelin Karşılaştırma Yapacak Olan Ölçütleri Tablosu

Doğruluk	Duyarlılık	Kesinlik	F-Ölçütü
%94.34	%94.3	%94.3	%94.3

A-6.4.Örnek Tabanlı Sınıflandırma Modelinin Başarım Ölçütleri

Verilebilecek örnek şekilde olan tabanlı biçimdeki tekniklerden Weka‟ da mevcut olan KStar algoritması işleme alınarak yapı meydana getirilmiştir. Şekil 47'de meydana getirilen sistemin test çıktılarına ait istatistiksel verileri ve karışıklık matrisi göz önüne serilmiştir. Şekil 48'de ise karışıklık matrisini kullanıp yapılan hesaplamada gerçekleştirilen karşılaştırma ile yapılan ölçütleri ifade edilmiştir.

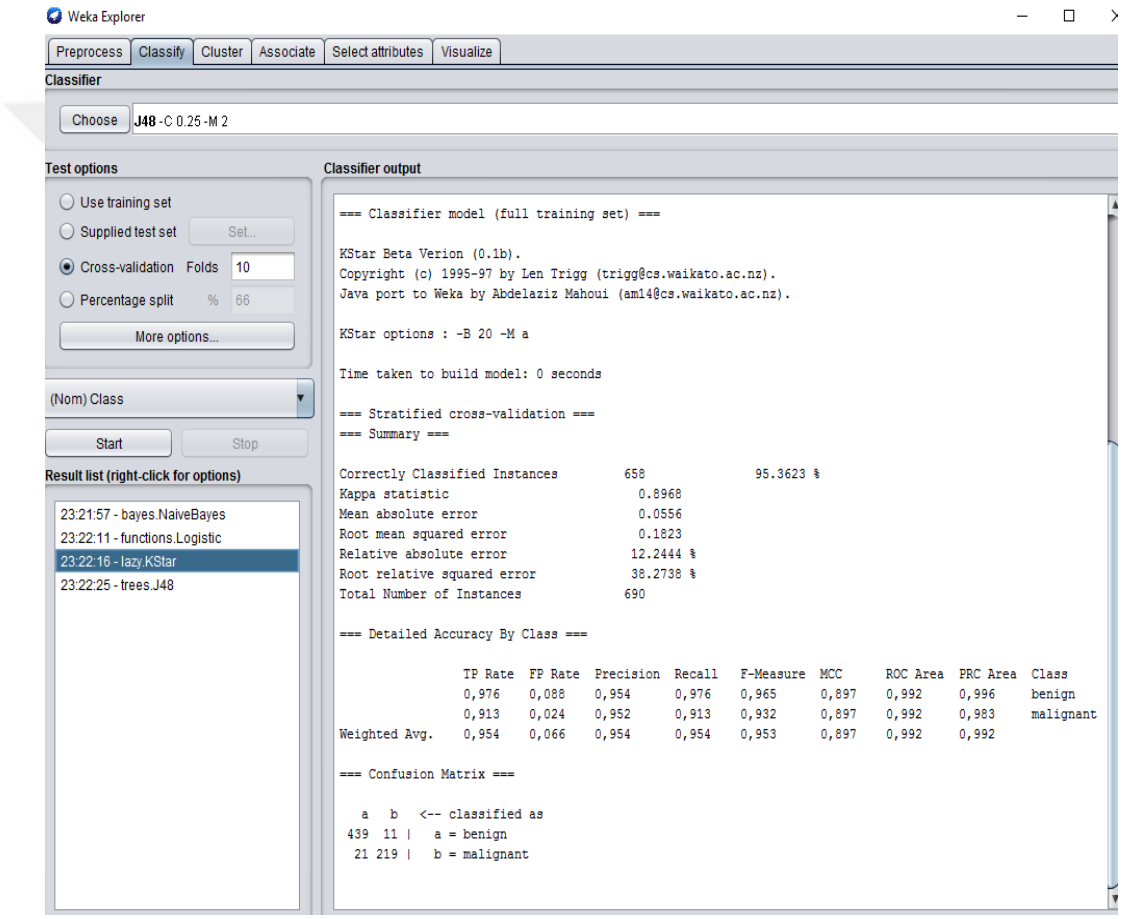
```
Correctly Classified Instances         653               95.7478 %
Incorrectly Classified Instances        29                4.2522 %
Kappa statistic                          0.9056
Mean absolute error                      0.0514
Root mean squared error                  0.1831
Relative absolute error                 11.284  %
Root relative squared error             38.3684 %
Total Number of Instances              682
```

Şekil 47 KStar Algoritması Karışıklık Matrisi Tablosu

		Ön Görülen Sınıf	
		a=2	b=4
Doğru Sınıf	a=2	434	9
	b=4	20	219

Şekil 48 KStar Algoritmasına Ait Olan Modelin Karşılaştırma Ölçütleri Gösteren Tablosu

Doğruluk	Kesinlik	Duyarlılık	F-Ölçütü
%95.74	% 95.6	% 98.00	% 96.8



Şekil 49 Dataset İçinde 9 Adet Kayıp Olan Veri Silindikten Sonra KStar Algoritmasının Verdiği Sonuç

Yapılan çalışma neticesinde elde edilen karışıklık matrisi ve doğruluk, kesinlik, duyarlılık, F-ölçütleri aşağıdaki gibi değişime uğramıştır.

Şekil 50 Dataset İçinde 9 Adet Kayıp Olan Veri Silindikten Sonraki Yeni KStar Karışıklık Matrisi Tablosu Grafiği

		Ön Görülen Sınıf	
		a=2	b=4
Doğru Sınıf	a=2	439	11
	b=4	21	229

Şekil 51 Dataset İçinde 9 Adet Kayıp Olan Veri Silindikten Sonraki Yeni KStar Algoritmasına Ait Olan Modelin Karşılaştırma Yapacak Olan Ölçütleri Tablosu

Doğruluk	Duyarlılık	Kesinlik	F-Ölçütü
%95.36	%95.4	%95.4	%95.3

A-6.5.Oluşturulan Modellerin Karşılaştırılması

Ön aşamadaki işlemlerinden geçen J48, NaiveBayes, Lojistik Regresyon ve KStar algoritmaları ile analiz gerçekleştirilerek bütün algoritmalar için yapılmış bulunan modele dahil olan test çıktı ve bulgu verileri önde olan kısımda izah edilerek gösterilmiştir. Karşılaştırma gerçekleştirebilmek için bütün modele ait karşılaştırma ölçüt sonuçları Şekil 52' de genel olan bir tabloda tekrardan ifade edilmiştir.

Şekil 52 Elde Edilen Modellerin Karşılaştırılması

16 Eksik Olan Verinin Tamamen Çıkartılması Sonucunda Bulunan Veriler

Ölçütler/Algoritmalar	J48	Naive Bayes	Lojistik Regresyon	Kstar
Doğruluk	95.74%	96.33%	96.92%	95.74%
Kesinlik	97.70%	98.60%	97.50%	95.60%
Duyarlılık	95.70%	95.70%	97.70%	98.00%
F-Ölçütü	96.70%	97.10%	97.60%	96.80%

Dataset İçinde 9 Adet Kayıp Olan Veri Silindikten Sonraki Yeni Karşılaştırma Tablosu

Ölçütler/Algoritmalar	J48	Naive Bayes	Lojistik Regresyon	Kstar
Doğruluk	93.18%	97.39%	94.34%	95.36%
Kesinlik	93.20%	97.50%	94.30%	95.40%
Duyarlılık	93.20%	97.40%	94.30%	95.40%
F-Ölçütü	93.20%	97.40%	94.30%	95.30%

Dataet içinde 699 tam veri setiyle çalışma yapıldığında elde edilen sonuçlar

Ölçütler/Algoritmalar	J48	Naive Bayes	Lojistik Regresyon	Kstar
Doğruluk	94.4%	97.3%	94.1%	95.4%
Kesinlik	94.4%	97.4%	94.1%	95.4%
Duyarlılık	94.4%	97.3%	94.1%	95.4%
F-Ölçütü	94.4%	97.3%	94.1%	95.4%

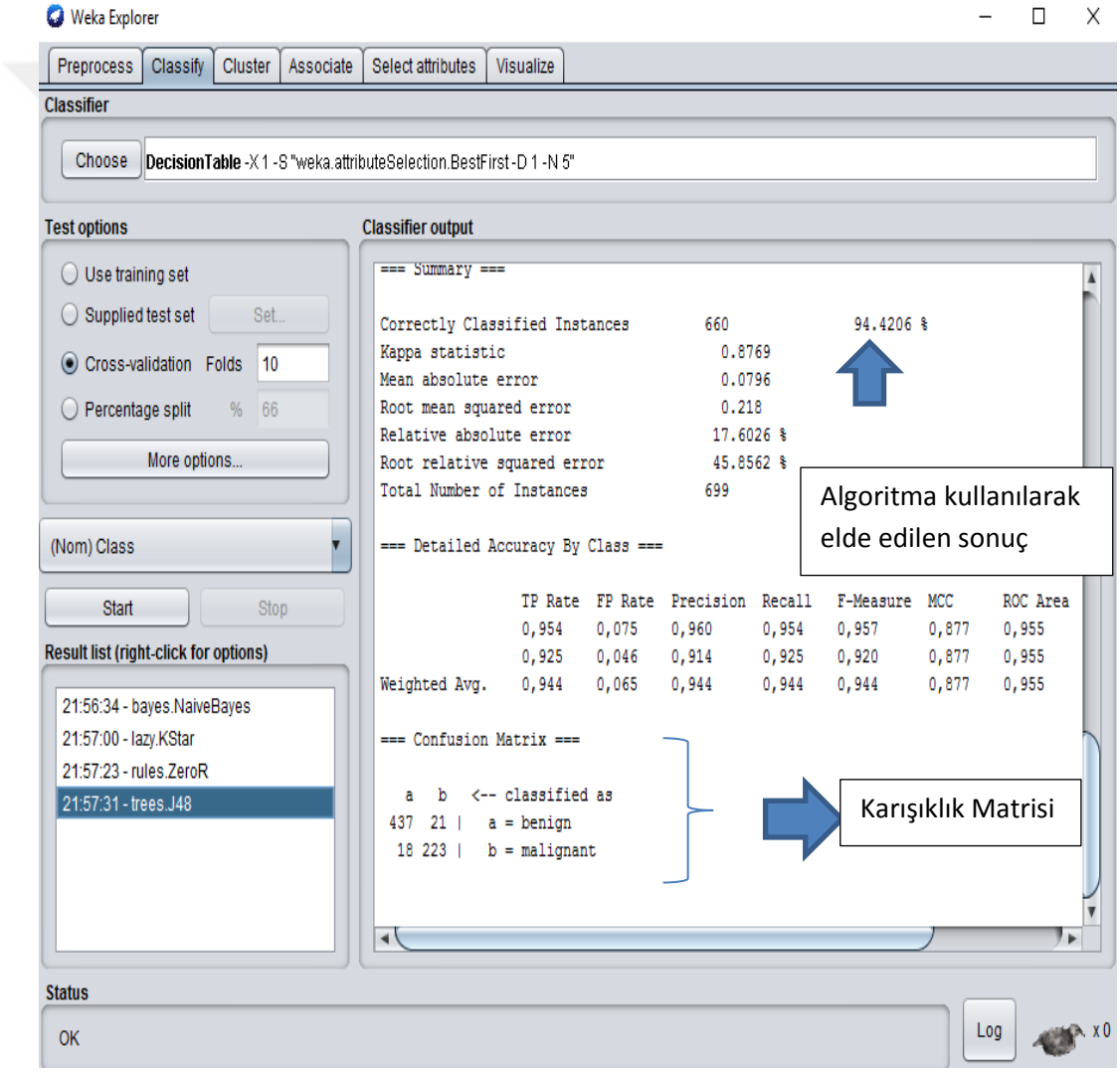
Gerçekleştirmiş olduğum çalışma neticesinde, algoritmaların üzerinde çalıştığı parametreler mevcut olan veriler olarak tespit edilerek seçilmiştir. Burada olan amacım, algoritmalar içinde pozitif biçimde yapılan bölme denebilecek olaylara meyil vermemek; hedefimin, versiyonları çok güzel biçimde yaratabilmesi hedeflenmeyen bir işlemde, yapılan araştırmanın çok değişik şekillerde bir istikamete yönlenmesini engellemektedir.

Daha önce ifade edilen bölümde gerçekleştirdiğimiz karşılaştırma olayını, verilerinin içerisinde farklı olacak olan algoritmalara bakılarak güzel tahmin çıktıları meydana getirdiği biçiminde ifade edip kısaca söyleyebiliriz. Lakin, Şekil 52'deki rakamlara bakılarak değer verileri arasında çok fazla değişiklikler bulunmadığını, en düşük olarak gözlemlendiğinde Lojistik Regresyonla çok yakın olan takipçisi Naive Bayes aralarında doğruluk ve F-ölçütü görüşünden yüzde sıfır nokta beşlik değişiklik bulunduğunu görebiliriz.

Veri madenciliği algoritmalarının birbirleri arasında karşılaştırma biçimi ile gerçekleştirilen deneysel işlemler bilim dünyasında çok sert eleştirisel yorumlara maruz kalmaktadır. Doğası neticesiyle veri madenciliği modelinin yüksek başarılı şekilde gerçekleşmesinin veriye çok bağlı olduğunu, veri üzerinde gerçekleştirilen önişleme olaylarının ve hali hazırda elimizin altında kullanılmakta olan algoritma çıktılarının meydana gelen sonuç neticesinde değişik olan farklı yan etkileri bulunacağını, kullanmaya göre yani kullanıcıya göre bağlı şekilde olacak biçimde benzer türde değişik çıktılar ele geçirilebileceğini ifade etmiştir.

B-6.1.Karar Ağacı Modelinin Başarım Ölçütleri

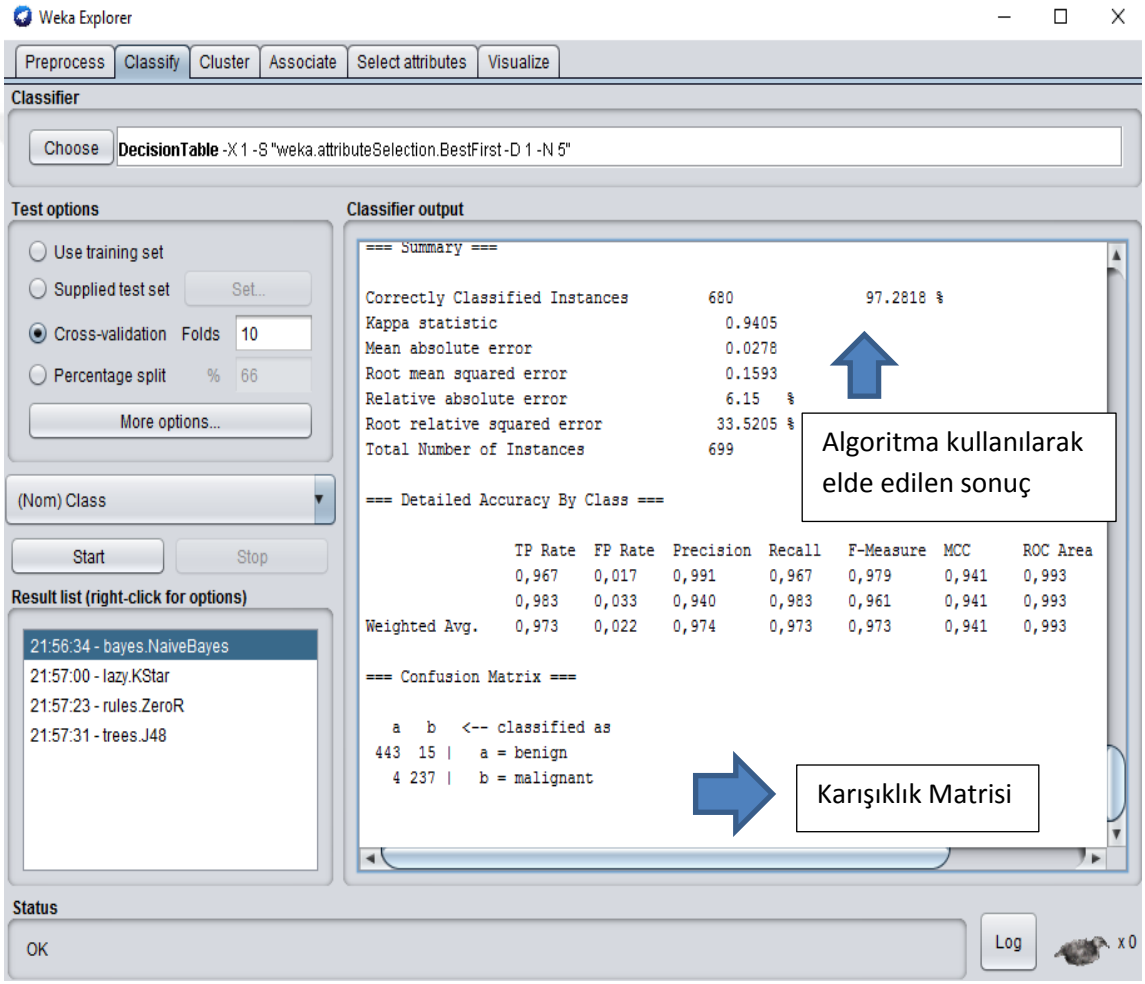
J48 algoritması incelemeden geçmiş durumda olan Wisconsin veri kaynağı içerisinde işlem yapılmıştır. Şekil 53'de aynı zamanda gerçekleştirilecek olan karışıklık matrisini kullanarak yapılan hesaplama sonucu yapılan karşılaştırma verileri gösterilmektedir.



Şekil 53 Karışıklık Matrisi ve Yapılan Çalışma Sonuçları

B-6.2.Bayes Sınıflandırma Modelinin Başarım Ölçütleri

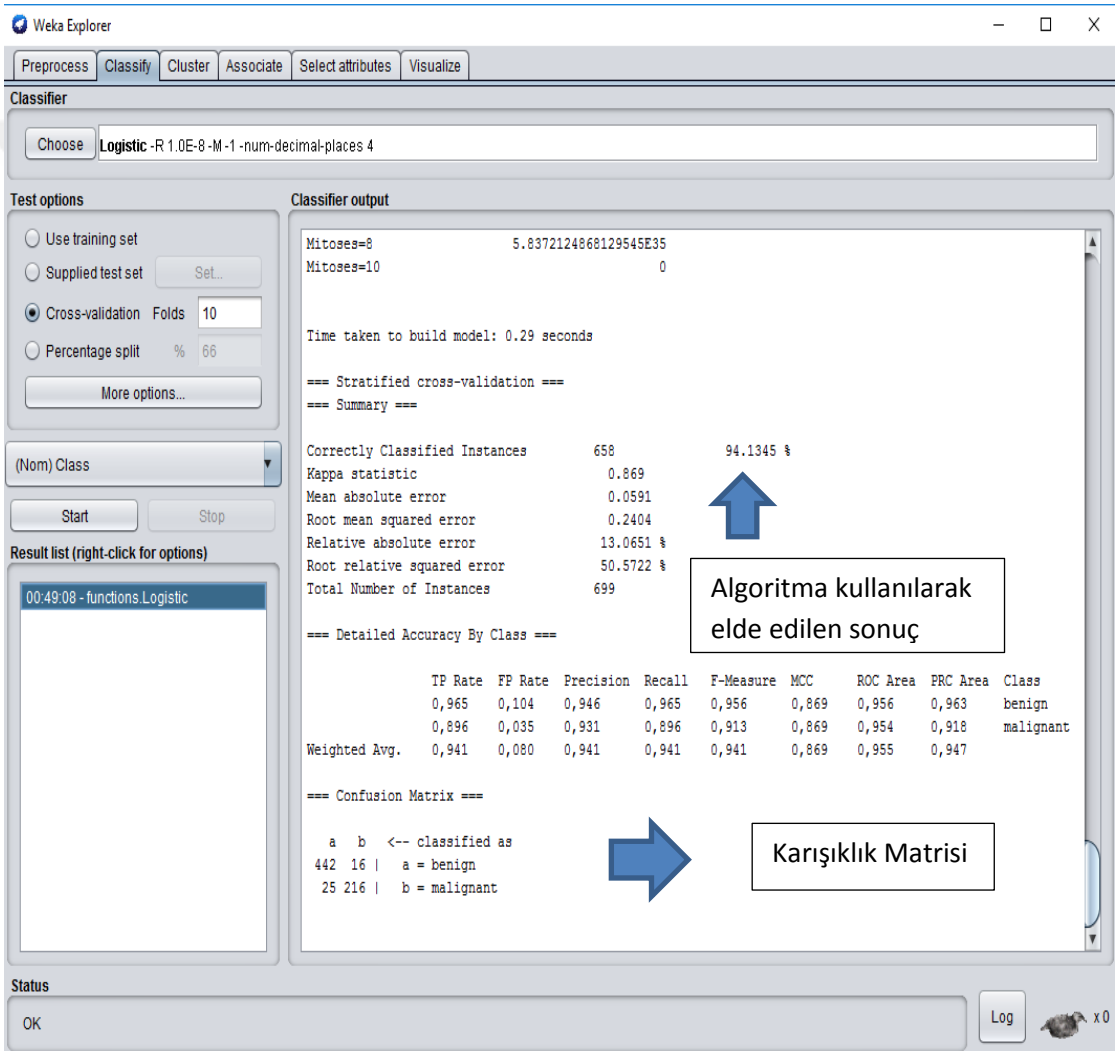
Bayes sınıflandırma yapılabilmesi için Weka‟ da mevcut halde bulunan Bayes Net, Naive Bayes, Naive Bayes Simple, Naive Bayes Updateable algoritmalarından Naive Bayes algoritması tercih edilerek veri kümesi içerisinde çalıştırılmıştır. Şekil 54'te meydana getirilen modelin test çıktılarına ait olan istatistiklerin karışıklık matrisi çıktıları gözlemlenmektedir.



Şekil 54 Karışıklık Matrisi ve Yapılan Çalışma Sonuçları

B-6.3.Regresyon Modelinin Başarım Ölçütleri

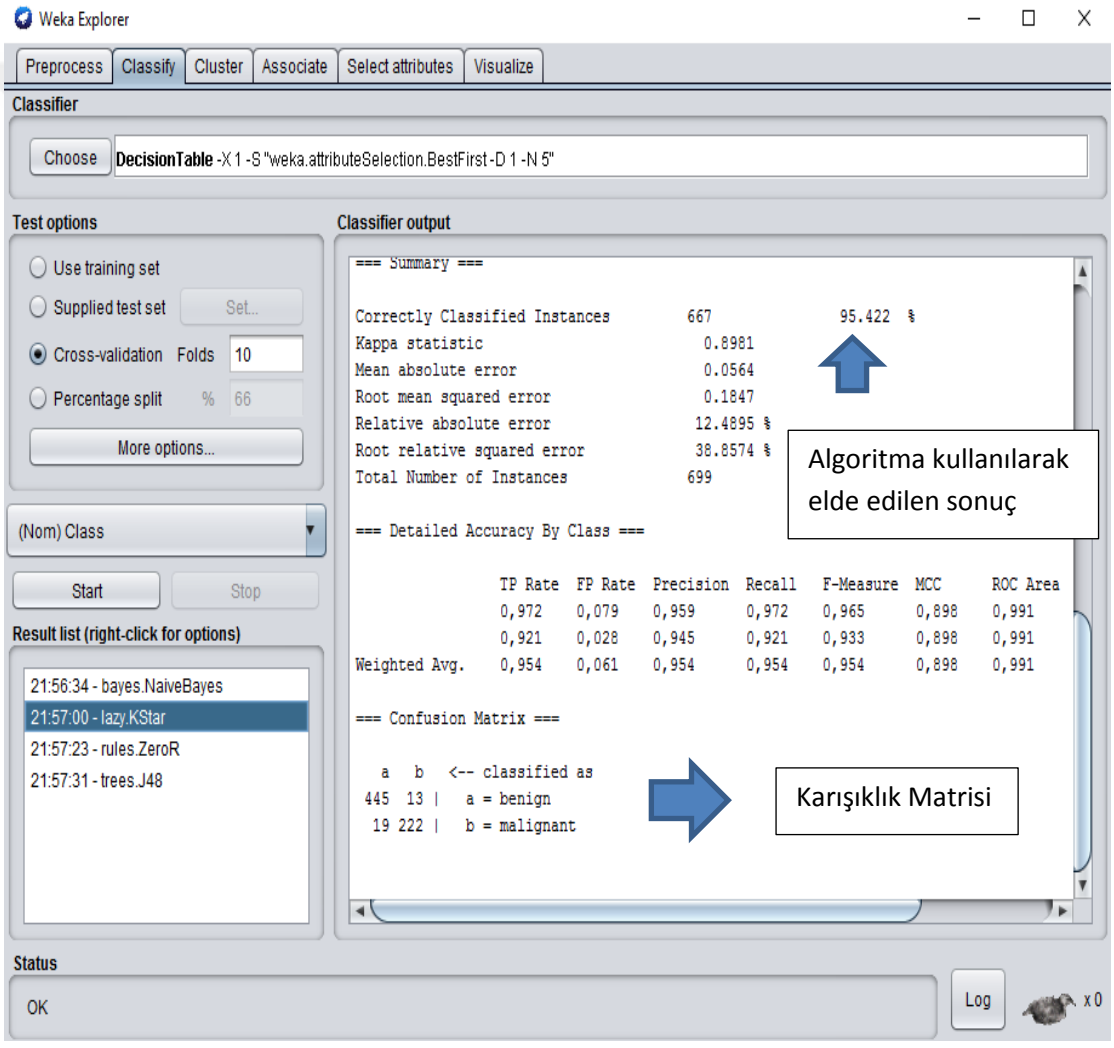
Karşılaştırma amaçlı bir şekilde olacak olan regresyon tabanlı tekniklerden lojistik regresyon algoritması tercih edilerek veri kaynağına tatbik edilmiştir. Şekil 55'de meydana getirilen modelin test çıktılarına ait bulunan istatistikleri ve karışıklık matrisi görülmektedir.



Şekil 55 Lojistik Regresyon Algoritması Karışıklık Matrisi ve Algoritmanın Başarı Yüzdesi

B-6.4.Örnek Tabanlı Sınıflandırma Modelinin Başarım Ölçütleri

Olabilecek olan örnek olarak söyleyebileceğimiz şekilde olan biçimdeki tekniklerden Weka' da mevcut olan KStar algoritması işleme alınarak yapılan model meydana getirilmiştir. Şekil 56' da ise karışıklık matrisini kullanıp yapılan hesaplamada gerçekleştirilen karşılaştırma ile yapılan ölçütleri ifade edilmiştir.



Şekil 56 KStar Algoritması Karışıklık Matrisi Sonuçları ve Algoritmanın Başarı Yüzdesi

B-6.5.Oluşturulan Modellerin Karşılaştırılması

Ön aşamayı başarıyla geçen J48, NaiveBayes, Lojistik Regresyon ve KStar algoritmalarıyla incelemeler gerçekleştirilerek her algoritma için yapılmış olan metoda dahil olan test sonuç verileri buraya gelmeden önceki başlıklarda izah edilerek gösterilmiştir. Karşılaştırma gerçekleştirebilmek için bütün modele ait karşılaştırma ölçüt sonuçları Şekil 57' de genel olan bir tabloda tekrardan ifade edilmiştir.

Şekil 57 Algoritmaların Son Durumda Karşılaştırılması

Ölçütler/Algoritmalar	J48	Naive Bayes	Lojistik Regresyon	Kstar
Doğruluk	94.4%	97.3%	94.1%	95.4%
Kesinlik	94.4%	97.4%	94.1%	95.4%
Duyarlılık	94.4%	97.3%	94.1%	95.4%
F-Ölçütü	94.4%	97.3%	94.1%	95.4%

7 SONUÇ VE ÖNERİLER

Açık Kaynak Kodlu şekilde olan Veri Madenciliği WEKA ile ilgili olarak kısa bilgiler verilmiş ve açıklanmış olup ilgili olan aşamalar ile alakalı kullanılabilecek yollar ve yöntemler ifade edilmiş ve 1991yılı içerisinde Meme Kanseri Wisconsin (Orijinal Olan) Veri Seti veri kaynağı içerisindeki göğüs kanseri biçiminde sağlıklı olmamış olan kişilerin kayıtlarının içerisinden tespit edilip seçilen yöntemlerin uygulaması WEKA yazılımı kullanılarak yapılmıştır ve algoritmaların kanseri tespit edebilme oranları ortaya çıkmış ve farklar irdelenmiştir. Sınıflandırma algoritmalarının birbirleri arasında karşılaştırma metotlarını inceleyen bu tezde veri madenciliği ile ilgili ve birbirleri arasında karşılaştırma verileri içerisinde çalışma yapılmıştır. Genel açıdan bakıldığında hangi algoritmanın çok daha güzel model meydana getirdiği biçiminde olan bir çalışmada farklı biçimde bulunan veri kaynakları içerisinde, daha fazla sayıda algoritma bir araya getirip kullanarak

birbirleri arasında karşılaştırma olayının gerçekleştirilmesi gerekecektir. Bu gerçekleştirilmiş olan çalışmada, modellerin bir araya getirilmesi için ücretsiz olarak piyasada kullanılmakta olan Weka veri madenciliği yazılım aracı kullanılmıştır. Mevcut olan diğer veri madenciliği yazılımları üzerinde bu kullanılan algoritmalar koyulup çalıştırılarak daha değişik veri madenciliği yazılımları benzer biçimde çıktılar elde edip etmediği kontrol edilebilir. Elimizin altında olarak kullanılan göğüs kanseri verilerini barındıran veritabanı University of Wisconsin Hospitals, Madison Dr. William H. Wolberg'den elde edilmiştir. Veritabanı toplamda altı yüz doksan dokuz örnekten her birinin içerisinde bulunan dokuz özellik ve bunları ayrı olarak iyi ya da kötü türde biçimde meydana gelmesi biçiminde sınıf bilgisi bulundurur. Yapılan inceleme neticesinde 16 tane eksik veriyi komple çıkartma yoluna gidilerek 683 tane veriden sonuçlar elde edilmiştir. 16 tane verinin içindeki kayıp kısımları iyileştirerek 699 tam set ile çıktılar elde edilmiştir. Son olarak kullanılabilecek miktar verinin kullanılarak diğer kısmının çıkartılması ile 690 veri ile yapılan çalışma sonuçları elde edilmiştir. Bu sonuçlar birbirleri arasında karşılaştırılmışlardır. Hangi çalışmanın bu kapsamda iyi sonuç verdiği araştırılmıştır. Veri madenciliğinde gerçekleştirilen sınıflandırma bölümlerinden olan karar ağaçları, Naive Bayes, lojistik regresyon ve örnek tabanlı olacak olan kategorilere ayırma yollarından tercih edilen 4 çeşit algoritmanın, 1991 senesinde Göğüs Kanseri Wisconsin (Orijinal) Veri Seti veri kaynağı içerisindeki göğüs kanseri hastası olacak biçimde bulunan kayıtların üzerinde gerçekleştirilen karşılaştırmalar neticesinde karar ağacı algoritması olarak bilinen lojistik regresyon algoritmasının farklı olanlara bakılarak hemen hemen diğerlerinden güzel bir yapı meydana getirdiği gözlemlenmiştir.

Şekil 52'den kolayca açılamalar takip edilebilir. Eskiden yapılan çalışmalara bakıldığında çıkartılan sonuçlarda birbirleri arasında karşılaştırma hedefli olarak regresyon tabanlı yollardan lojistik regresyon algoritması yüzde doksan altı nokta doksan iki ile en düzgün ve doğru biçimde olan sonuç elde edilmiştir, bu çıktılar içerisinde 433(TP) tane olan iyi huylu biçimde ifade edebileceğimiz 228 (TN)tanesi kötü huylu biçimde ifade edebileceğimiz sınıfa dahildir. (Poyraz, Tıpta Veri Madenciliği Uygulamaları: Meme Kanseri Veri Seti Analizi , 2012)

Ölçütler/Algoritmalar	J48	Naive Bayes	Lojistik Regresyon	Kstar
Doğruluk	95.74%	96.33%	96.92%	95.74%
Kesinlik	97.70%	98.60%	97.50%	95.60%
Duyarlılık	95.70%	95.70%	97.70%	98.00%
F-Ölçütü	96.70%	97.10%	97.60%	96.80%

Şekil 57 Eskiden Yapılan Çalışmada Bulunan Çıktılar (Poyraz, Tıpta Veri Madenciliği Uygulamaları: Meme Kanseri Veri Seti Analizi , 2012)

Lojistik regresyon algoritmasında peşinde olacak şekilde görünen takipçisi Naive Bayes algoritması yüzde doksan altı nokta otuz üç ile ikinci sırada en güzel çıktıyı elde etmektedir, yapılan işte J48 ile KStar algoritmaları birlikte bakıldığında doğruluk çıktısı olarak %95.74 biçiminde aynı olacak olan sonuçları elde etmektedir. Kesinlik ölçütü tarafından bakıldığında Naive Bayes en iyi çıktıyı vermiş olup, işleme giren diğer algoritmalar bu parametreler dikkate alınarak, J48, Lojistik Regresyon ve KStar biçiminde sıralanmaktadır. Lakin kesinlik tespiti sadece ona bakılarak yorumlanma yoluna gidilirse değerlendirme bizleri yanlış sonuçlara varmamızı sağlayabilir. Bu nedenden dolayı bu ölçütü duyarlılık ölçütüyle birlikte hep beraber dikkat etmek gereklidir. Tablodan bakılarak anlaşılacağı üzere algoritmalar, duyarlılık ölçütü gözetilecek biçime göre KStar, Lojistik Regresyon,J48, Naive Bayes şeklinde olacak biçimde sıralaya konulabilir hatta J48, Naive Bayes duyarlılık ölçütü sonuçları aynı rakamları almışlardır. Bu durumdan da anlaşılacağı üzere, kesinlik ölçütü değerleri ve duyarlılık ölçütü değerleri aralarında ters olacak şekilde sıralama meydana getirmiştir. Bu durumdan ötürü elde geçen verilerin otomatik biçimde analizinin yapılması ve sınıflara ayrılması için hem hastalar hem de sağlık sektörü bölümleri göz önüne alınarak bakıldığında çok büyük önem barındırmaktadır. (Poyraz, Tıpta Veri Madenciliği Uygulamaları: Meme Kanseri Veri Seti Analizi , 2012) .

Gerçekleştirmiş olduğum çalışmada ise aynı durum tekrar gözden geçirilmiştir. Çıktılarda farklılaşmalar sıralamalarda yer değişimleri gözlemlenmiştir. Bunlar kısaca büyükten küçüğe sıraya şu şekilde konulabilir :

Doğruluk->NaiveBayes(%97.3)>KStar(%95.4)>J48(%94.4) >Lojistik Reg.(%94.1)
Kesinlik->NaiveBayes(%97.4)> KStar(%95.4) >J48(%94.4)>Lojistik Reg.(%94.1)
Duyarlılık->NaiveBayes(%97.3)>KStar(%95.4)>J48(%94.4)>Lojistik Reg.(%94.1)
F-Ölçütü->NaiveBayes(%97.3)>KStar(%95.4) >J48 (%94.4)> Lojistik Reg.(%94.1)

Yapılan çalışmalara bakıldığında çıkartılan sonuçlarda birbirleri arasında karşılaştırma hedefli Doğruluk temel alınarak bakıldığında Naive Bayes algoritması %97.3 ile en düzgün ve doğru biçimde olan sonucu elde etmiştir. En yakın biçimde görünen takipçisi KStar algoritması %95.4 ile ikinci en güzel sonucu elde etmektedir, yapılan işte J48 ile %94.4 Lojistik Regresyon algoritması ile %94.1 biçiminde olacak olan sonuçları elde etmektedir.

Bu durumdan ötürü kazanılan verilerin otomatik biçimde analizinin yapılması ve sınıflara ayrılması için hem hasta olan kişiler hem de sağlık sektörü alanları göz önüne alınarak bakıldığında çok büyük önem barındırmaktadır. İleriye yönelik olarak daha büyük şekilde olacak olan veri tabanları ile yapılması muhtemel işlemler bilgisayar destekli tanı sistemlerinin başarı yüzdelerini üst seviyeye yükseltecektir. Algoritmaların veri kaynağı içerisinde işleme alınması aşamasında algoritma çıktıları olarak bütün algoritmaların o parametre için mevcut olan çıktısı işleme alınmıştır. Bütün algoritmalar ve konuyla alakalı bütün veri kaynakları ile ilgili bakılacak olursa başarı yüzdesini yükseltmeye yoluna gidilerek parametre değerleri bulunarak bu bulunan parametrelerle algoritma çıktılarını birbirleri arasında karşılaştırmak çok daha farklı sonuçlara bizleri taşıyabilir. Lakin, bu şekilde yapılacak olan bir karşılaştırmada maalesef yanlılık meydana gelebilecektir. Bu yapılan işlemde, algoritmaların meydana getirdiği modellerin başarıya ulaşma çıktıları birbirleri arasında değerlendirilmiştir. Benzer olacak bir biçimde, algoritmaların göstermiş olduğu hızı ve ayrıca kullandıkları hafıza miktarı kullanımı ile algoritmaların performansları aralarında karşılaştırma olayı da gerçekleştirilebilir.

*Bu yapılan çalışmada, algoritmaların çıkarttığı ve ortaya getirdiği modellerin başarılarının çıktıları birbirleri arasında karşılaştırılmıştır. Bu olaya benzer olacak biçimde, algoritmaların hızları ve hafızalarının tüketimi ile algoritmaların gerçekleştirdikleri performanslar arasındaki karşılaştırmalar da ortaya çıkabilmektedir.

*Bu işlemde değişik olabilecek olan kategorilerdeki veri grupları içerisinden üzerinde yapılma imkanı vardır.

*Daha geniş olacak biçimde ve sayıda algoritmalar işlemlere alınıp kullanılarak değişik algoritmalar birbirleri arasında karşılaştırılabilir.

*Bu yapılan işlemde Weka veri madenciliği aracı yazılımı kullanılmıştır. Bu işleme ek olarak değişik Veri Madenciliği yazılım araçları kullanılarak çalışma daha büyültülüp kompleks hale getirilebilir.

*Bütün bu olaylarda gerçekleştirilen işlemler içerisinde Algoritmanın yazım aşamasında bizlere gösterdiği başarısını en yüksek düzeye çıkartarak parametreler ele geçirilerek aralarında kıyaslama bu tarzda şekil gerçekleştirilebilir.

*Algoritmaların hedeflerini başarmalarının yanı sıra, hızı açısından ve hafızanın kullanımı açısından diğer parametreler üzerinde bir kıyaslama hakkında farklı olacak bir iş sektörü olarak ilerlenebilir.

KAYNAKLAR

1-Akademik Bilişim 2008, Çanakkale Onsekiz Mart Üniversitesi, Çanakkale, 30 Ocak – 01 şubat 2008 Hastane Bilgi Sistemlerinde Veri Madenciliği, Pınar YILDIRIM, Mahmut ULUDAĞ, Abdülkadir GÖRÜR

2-Data Mining Concepts andTechniques, Han, J.-Kamber, M., Morgan Kaufmann Publishers, 1st Ed., San Francisco, USA, 2000

3-Determination Of Breast Cancer Using ANN Armağan Ebru Temiz1,D.Ü.ZiyaGökalp Eğitim Fakültesi Dergisi 7,95-107 (2006), Veri Madenciliği Uygulama Alanları, Application Fields of Data Mining, Abdullah BAYKAL1

4-Elektrik -Elektronik - Bilgisayar Mühendisliği 10. Ulusal Kongresi453, Eğiticili Ve Eğiticisiz Nöral Algoritmalar Kullanarak Göğüs Kanseri Teşhisi, Tüba KIYAN, Tülay YILDIRIM,2003.

5-EndüstriMühendisliği Yazılımları ve Uygulamaları Kongresi | 30 Eylül-01/02 Ekim 2011 Weka ile Veri Madenciliği Süreci ve Örnek Uygulama Pınar TAPKAN Lale ÖZBAKIR Adil BAYKASOĞLU

6-Farboudi,S,Tıp Bilişiminde İstatiksel Veri Madenciliği Yüksek Lisans Tezi ,Hacettepe Üni,Fen.Bil.2009) .

7-Fayyad, U.M.; Piatesky-Shapiro, G.; Smyth, P.;Uthurusamy, R., "Advances indata mining and Knowledge Discovery", AAAI Pres,USA (1994).

8-Hand, D. J. ; Classifier Technology and the Illusion of Progress; Statistical Science, Vol. 21;Institute of Mathematical Statistics, 2006; 1-15.

9-Han, J., ve Kamber, M., "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2001

10-Jiawei Han ve Micheline Kamber, Data Mining : Concept and Techniques, USA : Morgan Kaufmann Publishers, 2001, s. 39 - 40.

11-Korhan Kadir Babadağ, "Veri Madenciliği Yaklaşımı ve Veri Kalitesinin Artması için Kullanılması", TÜĠK 15. İstatistik Araştırma Sempozyumu Bildiriler Kitabı, Yayın No. 3062, Ankara, 2006, s.242.

12-Uludağ Üniversitesi iktisadi ve idari Bilimler Fakültesi Dergisi Cilt XXIX, Sayı1, 2010, s.65-90 Veri Madenciliği ve istatistik Selim TÜZÜNTÜRK

13-Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması Cengiz COŞKUN, Yrd. Doç. Dr. Abdullah BAYKAL, Makale, 13. Akademik Bilişim Konferansı, Şubat 2011.

14-Veri Madenciliği Uygulama Alanları (Application Fields of Data Mining) Abdullah BAYKAL. 2006, D.Ü. Ziya Gökalp Eğitim Fakültesi Dergisi Sayı 7, 95-107, 2006.

15-Veri Madenciliği Yöntemleri Kullanılarak Meme Kanseri Hücrelerinin Tahmin ve Teşhisi Mustafa DANACI, Mete ÇELİK, A. Erhan AKKAYA, 2010, Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu(ASYU 2010).

16-Tıpta Veri Madenciliği Uygulamaları : Meme Kanseri Veri Seti Analizi Oğuz POYRAZ 2012, Edirne Trakya Üniversitesi, Yüksek Lisans Tezi.

ÖZGEÇMİŞ

7 Eylül 1987 tarihli, İstanbul İli Şişli ilçesi doğumluyum. İlkokulu 19 Mayıs İlköğretim okulunda tamamladım. Ortaokulu Ali Yalkın İlköğretim okulunda tamamladım. Liseyi Etiler Anadolu Lisesinde tamamladım. İstanbul Kültür Üniversitesi İngilizce Bilgisayar Mühendisliği'ne 2007 yılında kayıt oldum. Bu bölümden 2012 yılında mezun olduktan sonra çeşitli kamu ve özel sektörde alanımla ilgili sektörlerde çalıştım. Beykent Üniversitesi Bilgisayar Mühendisliği Anabilim Dalında Bilgisayar Mühendisliği Bölümü Yüksek Lisans Programına kayıt oldum.

Özel ilgi alanlarım, elektronik gelişmeleri takip etme, yazılım ve donanımsal yenilikleri takip etmedir.

Aday: Çağdaş KEKLİK