



T.C.
MALTEPE ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

**MÜŞTERİ MEMNUNİYETİ BELİRLEMEK İÇİN METİN
MADENCİLİĞİ TABANLI BİR YAZILIM ARACI**

Kemal KUZUCU

Yüksek Lisans Tezi

Tez Danışmanı

Yrd. Doç. Dr. Şenol Zafer ERDOĞAN

İSTANBUL – 2015

**T.C.
MALTEPE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI**

**MÜŞTERİ MEMNUNİYETİ BELİRLEMEK İÇİN METİN
MADENCİLİĞİ TABANLI BİR YAZILIM ARACI**

YÜKSEK LİSANS TEZİ

Kemal KUZUCU

**Tez Danışmanı
Yrd. Doç. Dr. Şenol Zafer ERDOĞAN**

İSTANBUL – 2015

Bu tez çalışması, Maltepe Üniversitesi Fen Bilimleri Enstitüsü Yönetim Kurulu'nun / / tarih ve / sayılı kararıyla oluşturulan jüri tarafından ***Bilgisayar Mühendisliği Yüksek Lisans Tezi*** olarak kabul edilmiştir.

JÜRİ

Yrd. Doç. Dr. Şenol Zafer ERDOĞAN
Danışman

Yrd. Doç. Dr. Volkan TUNALI
Üye

Yrd. Doç. Dr. Fatih YÜCALAR
Üye

ÖZET

Yüksek Lisans Tezi, Müşteri Memnuniyeti Belirlenmesi için Metin Madenciliği Tabanlı Bir Yazılım Aracı, T.C. Maltepe Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı.

İşletmenin veri tabanlarında bulunan veriler tek başına anlam ifade etmezler ve işletmeler gelişen teknoloji ile birlikte bu veri yığınlarını mümkün olduğunca kullanıp, kendilerine faydalı bilgi haline getirmek isterler. Bu işlemi yapabilmek için öncelikli olarak yapısal halde olmayan verileri yapısal hale getirmeleri gerekir.

Bu tez çalışmasında bir işletmenin 2002 yılından bu yana saklanan “Müşteri Şikayet Verileri” üzerinde işlem yapılmıştır. Her şikayet kaydı Mutlu veya Mutsuz şekilde kapanmaktadır. Burada amaç şikayet kaydını açan ilgili uzmanın, daha kaydın girişini yaparken Mutlu ya da Mutsuz kapanacağını tahmin eden bir sistem geliştirmek ve şikayet kaydının Mutsuz şekilde kapanacağını tahmin eden yazılım aracının yardımıyla müşteriye nasıl davranacağı konusunda yol göstermektir.

Bu tez çalışması kapsamında bir yazılım aracı geliştirilmiştir. Bu yazılım aracına son yıllarda girilen kayıtlar ile sistem eğitilmiş ve kategorilere (Mutlu/Mutsuz) ayrılmıştır. Bundan sonra gelecek olan kayıtların kategorisinin tahmin edilmesi sağlanmıştır.

Bu tez 2015 yılında yapılmıştır ve 48 sayfadan oluşmaktadır.

Anahtar Kelimeler: Metin Madenciliği, Naive Bayes, Memnuniyet Analizi

ABSTRACT

Master Thesis, A Text Mining Based Software Tool Development for Detection of Customer Satisfaction. T.C. Maltepe University, Graduate School of Natural and Applied Sciences, Department of Computer Engineering.

Data in a firm database does not mean anything alone. While technologies used in the businesses developing rapidly, firms want to use this big data the as much as possible so that makes it an advantage for the firm. To do this, first they have to convert non-structural data into structural form for this process.

In this thesis, a firm's "Customer Complaints Data" (recorded since 2002) has processed. As a fact, each complaint record can be closed with Happy or Unhappy customer feeling. The purpose of this study; to develop a software that could help call center agents by predicting the result of customer's feelings at the end, while recording customer's request at the beginning. Also it's expected that the software leads agents to proper behaviors that makes happy customer feeling.

In this study Customer Satisfaction Software has development. In this software, the system has improved with the records in recent years and results has divided into two categories (happy/unhappy). Then the software could predict results of future records.

This thesis has been completed in 2015 and consists of 48 pages.

Keywords: Text Mining, Naive Bayes, Satisfaction Analysis

TEŐEKKÜR

Bu tez alıőmasında her tŒrlŒ bilgi ve tecrŒbelerinden faydalandıđım ve hibir zaman yardımı esirgemeyen tez danıőmanım Sayın Yrd. Do. Dr. Őenol Zafer Erdođan'a ve Yrd. Do. Dr. Volkan Tunalı'ya sonsuz teőekkŒrlerimi sunarım.

Ayrıca tez alıőmam boyunca teknik bilgi, fikir, zaman ve emeklerini esirgemeyen Otoko Otomotiv Bilgi Teknolojileri BŒlŒmŒnden; Yazılım Danıőmanı Didem Ham, Bilgi İőlem YŒnetmeni Onur Őzatalay ve Bilgi Teknolojileri MŒdŒrŒ Dr. Erdal Kemikli'ye ve tŒm Bilgi Teknolojileri bŒlŒmŒne ayrıca Neslihan GŒneő'e teőekkŒrŒ bir bor bilirim.

Son olarak tŒm hayatım boyunca hibir zaman desteklerini esirgemeyen ve her durumda yanımda olan aileme sonsuz teőekkŒr ederim.

İÇİNDEKİLER

ÖZET.....	IV
ABSTRACT.....	V
TEŞEKKÜR.....	VI
İÇİNDEKİLER.....	XI
ŞEKİLLER.....	XII
ÇİZELGELER.....	Error! Bookmark not defined.
1. GİRİŞ.....	1
1.1. Tezin Amacı.....	2
1.2. Tezin Kapsamı.....	2
2. VERİ MADENCİLİĞİ.....	3
2.1. Veri Madenciliği Uygulama Süreci.....	5
2.1.1. Problemin Tanımlanması.....	5
2.1.2. Verilerin Hazırlanması.....	6
2.1.3. Modelin Belirlenmesi.....	7
2.1.4. Modelin Değerlendirilmesi.....	7
2.1.5. Dağılım.....	8
2.2. Veri Madenciliği Teknikleri.....	8
2.2.1. Sınıflandırma.....	9
2.2.2. Kümeleme.....	10
2.2.3. Birliktelik Kuralları.....	11
2.3. Metin Madenciliği.....	12
2.3.1. Metin Madenciliği Kullanım Alanları.....	13
2.3.2. Metin Madenciliği Adımları.....	13
2.3.2.1. Metin Koleksiyonu Oluşturma.....	13
2.3.2.2. Metin Ön İşleme.....	14
2.3.2.3. Özellik Seçimi.....	14
2.3.2.4. Kategorilere Ayırma.....	15
2.3.2.5. Değerlendirme ve Yorumlama.....	15
2.3.3. Metin Sınıflandırma.....	15
2.3.3.1. KNN Algoritması.....	16
2.3.3.2. Karar Ağacı.....	17

2.3.3.3. K-Means	18
2.3.3.4. SVM	20
2.3.3.5. Naive Bayes	20
3. MÜŞTERİ MEMNUNİYETİNİ BELİRLEMEK İÇİN METİN MADENCİLİĞİ TABANLI BİR YAZILIM ARACI	24
3.1. Giriş	24
3.2. Müşteri Memnuniyeti Yazılım Aracının Veri Tabanı Tasarımı.....	24
3.3. Müşteri Memnuniyeti Yazılım Aracının Tasarımı	28
4. UYGULAMALAR VE SONUÇLARI	34
4.1. Uygulama 1	34
4.2. Uygulama 1 Sonuçları	36
4.3. Uygulama 2	38
4.4. Uygulama 2 Sonuçları	40
5. SONUÇ	41
6. KAYNAKLAR	42
7. ÖZGEÇMİŞ	48

ŞEKİLLER

Sayfa

Şekil 2.1. Veri Sınıflandırılması	10
Şekil 2.2. Kümeleme Resmi.....	11
Şekil 2.3. Metin Madenciliği Adımları	13
Şekil 2.4. Karar Ağacı.....	17
Şekil 2.5. K-Means Örneği	19
Şekil 3.1. NB_KELIMELER_DETAY tablosu	25
Şekil 3.2. NB_KELIMELER_DETAY tablosu	26
Şekil 3.3. NB_KELIMELER_MASTER tablosu	26
Şekil 3.4. NB_KELIMELER_MASTER tablosu.....	27
Şekil 3.5. Memnuniyet Analizi Programı	28
Şekil 3.6. Memnuniyet Analizi Programına Kayıt Yükleme	29
Şekil 3.7. Memnuniyet Analizi Programına Kayıtların Yüklenmesi	29
Şekil 3.8. Memnuniyet Analizi Programı ile Mutlu/Mutsuz Oranları	30
Şekil 4.1. Veri Tabanına Veri Yükleme İçin Metin Dosyasının Hazırlanması	34
Şekil 4.2. Memnuniyet Analizi Yazılımın Uygulama Sonuçları	35
Şekil 4.3. Uygulama 1 Sonuçları	36
Şekil 4.4. Memnuniyet Analizi Yazılımının Uygulama Sonuçları	37
Şekil 4.5. Uygulama 2 Sonuçları	37

ÇİZELGELER

	Sayfa
Çizelge 2.1. Veri Madenciliğinin Tarihsel Gelişim Süreci	5
Çizelge 2.2. Naive Bayes Örnek Uygulama	21
Çizelge 4.1. Karışıklık Matrisi	37
Çizelge 4.2. Uygulama 1 Sonuçları.....	38
Çizelge 4.3. Uygulama 2 Sonuçları.....	40

1. GİRİŞ

Gün geçtikçe büyüyen bilişim sektörü alanında giderek artan verilerden elle tutulur kullanışlı veri elde etmek günümüzde işletmeleri en büyük sorunlarından biri olarak göze çarpmaktadır. İşletmeler elindeki veriyi mümkün olduğunca kullanıp bundan kazanç sağlama yoluna gitmektedirler.

Büyük veri tabanları içerisindeki verilerden anlamlı ve ilginç bilgilere ulaşmak için yeni algoritmalar ve metotlar geliştirilmiştir. Veri Madenciliği ve Metin Madenciliği algoritmaları örnek olarak verilebilir.

Yapısal Veri, içinde bulunduğu ortamda tanımlanabilen ve üzerinde işlem yapılabilen, organize edilebilen verilerdir. En çok kullanılan yapısal veri kaynağı SQL, Access gibi veri tabanlarıdır. Bu veri tabanları satır ve sütun bazında tanımlanıp istenilen verilere kolay ve birkaç işlemle hemen ulaşılabilir verilerdir. Ancak işletmelerin veri tabanlarının da bulunan veriler genellikle yapısal olmayan verilerdir ve metin madenciliği içerisinde bu haliyle kullanılmazlar. Genellikle işletmelerde text, word, jpeg, pdf uzantılı dosyalarda veriler saklanmaktadır. Bunlarda yapısal olmayan verilere örnek gösterilebilir. Öncelikle bu verilerin yapısal hale getirilmesi gerekir. Veriler yapısal hale getirildikten sonra üzerinde işlemler yapılabilir.

1.1. Tezin Amacı

Bu tez çalışmasının amacı, müşteri şikayet kayıtlarının mutlu veya mutsuz şekilde kapanacağını tahminini yapan bir yazılım aracı geliştirmektir.

Bir firmanın son 10 yıllık müşteri şikayet kayıtları üzerinde çalışma yapılmıştır ve bu kayıtlar ile sistem eğitilerek kelimelerden yapısal veri elde edilmiştir. Visual Studio .NET platformu üzerinde C# programlama dili ile bir Windows Form uygulaması geliştirilmiştir. Geliştirilen bu yazılım aracı ile son 2 yılın verileri sistem üzerinde test edilerek programın doğru çalıştığı gösterilmeye çalışılmıştır.

1.2. Tezin Kapsamı

Tez çalışmasında kullanılan veri, yapısal türde olmayan veridir ve bu veriyi kullanabilmek için yapısal veri şekline çevirmek gerekir. Metin madenciliği sürecinde bu çevrim yapılmaktadır.

Tezin ilk bölümünde giriş yapılmakta ve tezin amacı açıklanmaktadır. İkinci bölümde, bu alanda temel olan Veri Madenciliği konusu anlatılmaktadır. Üçüncü bölümde, Metin Madenciliği konusu ve Metin Madenciliği adımları anlatılmaktadır. Tezin dördüncü bölümünde ise tezin konusu olan “Müşteri Memnuniyeti Belirlemek için Metin Madenciliği Tabanlı Yazılım Aracı” anlatılmakta ve bu aşamadan sonra gerçekleştirilen uygulamalar gösterilmektedir. Tezin sonuç kısmıyla tez bitirilmektedir.

2. VERİ MADENCİLİĞİ

Veri yalnız başına bir anlam ifade etmez. Ancak işlendiği durumda ya da anlam yüklenmesi durumunda kullanılabilir bir bilgi olacaktır. Bu amaç doğrultusunda bilgiye ulaşmak için veri üzerinde çeşitli işlemler veya çevrimler gerçekleştirilebilir.

Veri Madenciliği veri yığınları arasından altın bilgiye ulaşma yöntemidir. Veri madenciliği, işletmeler bu veri yığınlarını kullanmaya ve bunlardan kazanç elde etmeye başladıkça popülerliği artan bir teknoloji olmuştur.

Veri madenciliği, devasa büyüklükteki veri yığınlarının içerisinde, gizli kalmış, önceden bilinmeyen işe yarar bilgilerin çıkarılması ve bu elde edilen verilerle gelecekle ilgili tahminler yapmamızı sağlayan kurallar olarak nitelendirebiliriz.

Veri madenciliği en basit tanımıyla, büyük ölçekli veriler arasından değeri olan bir bilgiyi elde etme işidir [1].

Başka bir ifade ile veri madenciliği yapısal veri tabanlarında depolanmış verilerden geçerli, yeni, potansiyel olarak yararlı ve nihayetinde anlaşılabilir örüntülerin tanımlanması işlemidir [2].

Belirtilen tanımlarda üzerinde durulduğu gibi, geleceğe dönük tahminler veri madenciliği ile yapılır. Veri madenciliği, istatistik işlemlerine dayalı bir tekniktir. Bu iki unsurun amaçları birbirine paralel doğrultuda gitmektedir ve birbirine hizmet etmektedirler. İstatistik ile veri setleri oluşturur, sınıflara ayırır ve grafik halinde sunulup analizi yapılır. Böylece ana konu için bilgi elde edilmesi süreci yapılır. Şirketlerin içerisinde çok fazla kullanılmasının başlıca sebebi de budur. Kısaca elimizdeki veriden daha önce elde edilmemiş bilgi hazinesine ulaşma yöntemidir. Bu iki yöntemin de ortak paydaları veriden bilgiye ulaşmaktır [4].

Veri Madenciliği, gelecekte elde etmek istediğimiz davranışları geçmişteki verilerden faydalanarak tahminler üreten bir karar destek unsurudur. Frawley ve

Piatetsky veri madenciliğini şöyle ifade etmiştir. "Veride saklı halde bulunan ve geçmişte bilinmeyen ancak faydalı olacağını düşündüğümüz bilginin açığa çıkarılmasıdır" [5].

Veri Madenciliği yöntemleri kapsamlı bir araştırma ve geliştirme çalışmaları neticesinde meydana gelmiştir. Bu gelişim işletmenin sahip olduğu veri yığınlarının oluşmasıyla başlar, veri girişlerinin gelişimiyle devam eder ve son kullanıcıya sahip olduğu veriyi yönetme imkanı veren bir yapı oluşturur. Şimdilerde ise şirketler tarafından kullanılan bir teknik olarak görülmektedir. Bunun sebebi ise aşağıda belirtilen teknolojileri kapsamaktadır [6];

- Çok büyük veri yığınlarını depolayabilme imkanı,
- Çok yüksek kapasiteye sahip bilgisayar,
- Veri Madenciliği Sınıflandırma ve algoritmaları,

Veri madenciliğinin tarihsel gelişim süreci 4 adım da incelenebilir. Bu süreçler Çizelge 2.1'de gösterilmektedir.

Çizelge 2.1. Veri madenciliğinin tarihsel gelişim süreci [7].

	Gelişme Aşaması	Şirket Sorusu	Teknolojiler	Tedarikçiler	Özellikler
1. AŞAMA	Veri Toplam (1960' lı Yıllar)	"Son beş yıldaki toplam gelirin ne?"	Bilgisayarlar, diskler	IBM, CDC	Geçmişe ait, statik veri dağıtım
2. AŞAMA	Veri Girişi (1980' li Yıllar)	"Geçen Martta İngiltere'deki satış miktarı ne?"	İlgili veri tabanları (RDMBS, SQL, ODBC)	Oracle, Sysbase, Informix, IBM, Microsoft	Kayıt seviyesinde geçmişe ait, dinamik veri dağıtımı
3. AŞAMA	Veri Depolama ve Karar Destekleme (1990' lı Yıllar)	"Geçen Martta İngiltere'deki satış miktarı ne? Boston'u geçti mi ?"	Çok boyutlu veri tabanları veri depoları	Pilot, Comshare, Arbor, Cognos, Microstrategy	Farklı seviyelerde geçmişe ait, dinamik veri dağıtımı
4. AŞAMA	Veri Madenciliği (Günümüzde Gelişmekte)	"Gelecek ay Boston' daki satış miktarı ne olabilir? Neden?"	Gelişmiş algoritmalar, çoklu işlemci bilgisayarlar, büyük veri tabanları	Pilot, Lockheed, IBM, SGI	Muhtemel ve etkin veri dağıtımı

2.1. Veri Madenciliği Uygulama Süreci

Veri Madenciliği uygulanırken, problemin tanımlanması, verilerin hazırlanması, modelin belirlenmesi, modelin değerlendirilmesi ve modelin dağılımı gibi birçok süreçten geçmektedir. Bu süreçlerin detayları aşağıdaki gibidir.

2.1.1. Problemin Tanımlanması

Veri madenciliğinde, ilk olarak problem açık ve net bir şekilde ortaya konulmalı ve buna uygun çözüm metotları da tasarlanmalıdır. Problemin analizi iyi bir şekilde yapılmalı, başarılı sonuçlar alındığında faydaları ve elde edilecek başarısız sonuçlarda dezavantajları, işletmeye olan maliyeti bu adımda yer verilmelidir [3].

Bu tez çalışmasında tanımlanan problem, son yıllarda gerçekleşen olumsuz (mutsuz) şekilde kapatılan müşteri şikayet verilerinin önüne geçmek ve bunları önceden tahmin etmeye çalışmaktır.

2.1.2. Verilerin Hazırlanması

Verinin hazırlanma süreci; verinin temizlemesi, verinin dönüşümü ve veri için özellik seçimi adımları ile yapılır.

Veri temizlemenin hedefi; girilen veri içerisinde aynı formatta olmayan veya yanlış girilmiş verilerin kaldırılmasıdır. Eksik veri alanları, o alanın ortalaması alınarak da doldurulabilir ya da bu eksik veriler tamamen kaldırılır. Bu tez çalışmasında, bazı uygulamalarda “Yaş” bilgisinin girildiği alanlar da eksik girilen veya hatalı girilen kayıtlar sonuca etki etmemesi açısından ilgili alanın ortalaması ile doldurulmuş ve bu şekilde uygulamalar gerçekleştirilmiştir.

Veriler temizlendikten sonra veri dönüşümü işlemi yapılır. Elimizdeki verilerde farklı biçim veya sayısal değerlere dönüştürülerek işlem yapılabilir. Bu tez çalışmasında ki bazı uygulamalarda “Cinsiyet” bilgisi alanında erkek için “E”, kadın için “K”, tüzel kişiler için “T” olan sözel verileri hesaplamalara dahil etmek için erkek için “1”, kadın için “2”, tüzel kişiler için “0” olarak dönüştürüp uygulama gerçekleştirilmiştir.

Veri için özellik seçimi aşamasında, bir unsuru tahmin etmede belirgin değişkenlerin kullanılması hedeflenmektedir. Kategori için olasılık hesaplama işleminde birden çok nitelik belirlenebilir. Ancak bu özelliklerin hepsi üzerine işlem yapmak zor olabilir. Bu gibi durumlarda belirgin parametreleri seçerek bunların üzerinde işlem yapmak hesaplamayı kolaylaştıracaktır.

2.1.3. Modelin Belirlenmesi

Belirtilen probleme uygun modelin belirlenmesi, çok fazla sayıda modelin kurup test etmek ile olur. Bu sebeple veri hazırlama ve modelin belirlenmesi adımları, en etkin model bulununcaya kadar devam eder. Modelin belirlenmesi, denetimli ve denetimsiz öğrenme olarak ikiye ayrılır [28].

Denetimli öğrenmede sınıflar belirlenir. Bu belirlenen sınıflara göre kriterlere ayrılır. Her sınıf için çeşitli örneklemeler yapılır. Sistem bu örneklerden yola çıkarak her bir sınıfa ait özellikleri ve bu özellikleri ifade eden nitelik cümleleri ile tanımlar. Örnekler yeteri kadar tanımlandığında nitelik cümleleri yeni gelen örneklere uygulanır ve yeni örneklerin hangi sınıfa ait olduğu model yardımıyla belirlenir. Burada kurulan modelin doğruluğunu test etmek için elimizdeki örnek verilerin bir kısmı test için ayrılabilir. Model kurulduktan sonra test verilerinden alınan sonuçlara bakılır. Bu şekilde model üzerinde doğrulama testi yapılabilir. Bu tez çalışmasında elimizdeki verilerin %5'lik kısmı test verisi için ayrılıp daha sonra bu verilerle modelin doğruluğu test edildi.

Denetimsiz öğrenmede, denetimli öğrenmede yapıldığı gibi önceden bilinen sınıflandırmalar yoktur. Örnekler gözlemlenir, çıktı veya sonuca bağlı değişken yoktur. Bu örneklerin arasındaki benzerlikler bulunmaya çalışılır. Bu modelde benzer örüntü kurallarına sahip veriler gruplamaya alınır. Birliktelik tekniklerine göre işlem yapılır [29].

2.1.4. Modelin Değerlendirilmesi

Modelin değerlendirilmesi son kontrol noktasıdır. Burada belirlenen modelin işletmenin amacını tam anlamıyla gerçekleştirdiğinden veya ilk adımda belirlenen problemin tam olarak analiz edildiğinden emin olunmalıdır. Bu sebeple modelin

değerlendirme aşaması çok önemlidir. Gerekirse tüm adımlar en baştan gözden geçirmeli ya da değişiklik yapılmalıdır. Bu aşamanın sonunda veri madenciliğinde elde edilen sonuçların kullanılıp kullanılmayacağıyla ilgili karar verilir [30].

2.1.5. Dağılım

Modelin değerlendirilmesi adımından sonra dağılım adımı gelir. Burada önceki adımlarda gerçekleştirilen işlemler artık son kullanıcının kullanacağı ve anlayacağı şekilde kullanıma sunulur. İlgili uzmanlarca kontrol altında tutularak yapılması gerekir [30].

2.2. Veri Madenciliği Teknikleri

Veri madenciliğinde iki tip model vardır. Bunlar tahmin edici ve tanımlayıcı modeldir. Tahmin edici modeller de amaç; daha önceden tanımlayıp sonuçları belirlenen veri setlerini kullanarak yeni model kurmak ve bu modeli sonuçları bilinmeyen veri setlerine uygulayıp sonuçları konusunda fikir yürütmektir. Örneğin, bir müşteri bankadan kredi çekmek istediğinde, banka müşterinin önceden vermiş olduğu kredilerin ödenip ödenmediğini inceleyerek müşteriye yeni vereceği kredinin ödeme durumunu tahmin edebilir. Bu modelin başka bir kullanım alanı da veri setinde bulunan boş alanların önceki kayıtlara bakarak veya diğer kayıtların durumuna göre doldurulması ya da tahminde bulunulmasıdır [8].

Tanımlayıcı model ise; karar vermeye yardımcı olacak verilerin tanımlaması işlemidir. Veri tabanlarında bulunan veriler arasındaki önceden bilinmeyen ilişkilerin bulunması, strateji geliştirme ve karar verme gibi süreçlerde kullanılmaktadır. “Çocuk bezi alan müşterinin, mama alma olasılığı diğerlerinden üç kat fazladır” yargısına varılabilmesi tanımlayıcı modele örnek olarak verilebilir [9].

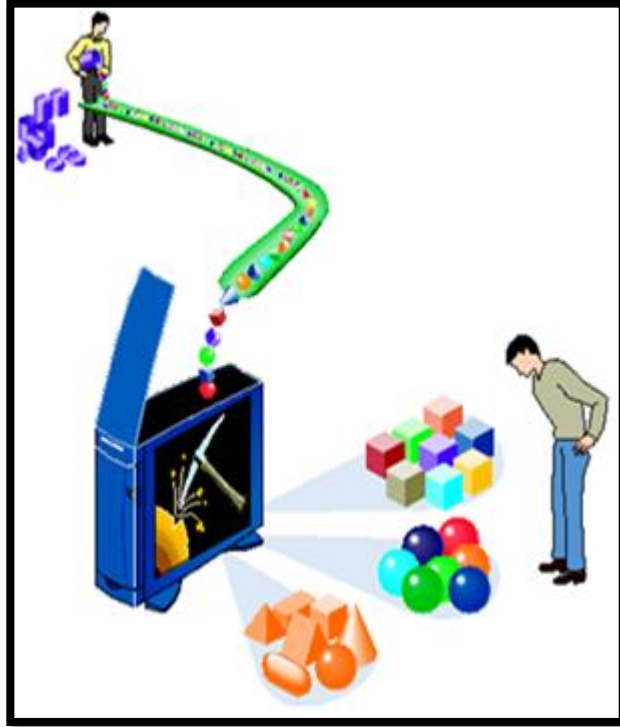
2.2.1. Sınıflandırma

Sınıflandırma, en çok kullanılan veri madenciliği tekniklerinden birisidir. Sınıflandırma ile nesnenin özellikleri incelenerek uygun olan sınıfa atanır. Burada dikkat edilmesi gereken husus atanacak olan sınıfın özelliklerinin daha önceden belirlenmiş olması gerekir. Sınıflar ise daha önceden tanımlanan eğitim verileri göz önüne alınarak ayrılır.

Örnek olarak kredi kartı başvurusu yapacak olan kişilerin düşük, orta ve yüksek riskli olarak sınıflandırılmasının yapılması verilebilir [10].

Kadınlar alışveriş yapmayı çok sever, erkekler ise futbolu çok sever [11].

Şekil 2.1’de veri sınıflandırılması gösterilmektedir

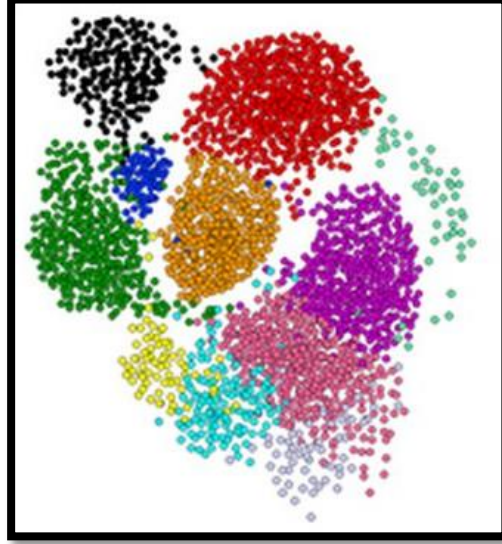


Şekil 2.1. Veri Sınıflandırması [11].

Şekil 2.1’de görüldüğü gibi Huni, Top ve Dikdörtgen şekillerden oluşan veri yığını görülmektedir. Sonrasında daha önceden tanımlanan şekillerine göre sınıflara ayrılmaktadır.

2.2.2. Kümeleme

Kümeleme analizi de benzer özelliklere sahip olan nesnelerin belirlendiği bir tekniktir. Nesneler birbirine olan yakınlık ve uzaklıklarına göre kategorilere ayrılır. Öncelikle eldeki veriler incelenir. Birbirine benzeyen nesneler bir kümeye, benzemeyenler ise başka bir kümeye ayrılır. Böylece ilk olarak kümelemeye hangi nesnelerin dahil edileceği belirlenir. Burada dikkat edilmesi gereken husus kümeleme yapılacak kriterlerin iyi belirlenmesidir. Genellikle nüfus bilimi ve astronomi alanında kullanılmaktadırlar. Şekil 2.2’de kümeleme tekniği uygulanmış bir yapı bulabilirsiniz.



Şekil 2.2 Kümeleme resmi [13].

2.2.3. Birliktelik Kuralları

Birliktelik kurallarında, verinin diğer veriler ile yakınlığı göz önüne alınarak birlikte olma durumu üzerine olasılık işlemleri yapılır. Müşteri bir ürün aldığı zaman bu ürünle birlikte başka hangi ürünleri aldığına tespitinin yapılması için bu teknikten yararlanılır [31].

Birliktelik kurallarına market sepeti uygulaması örnek olarak gösterilebilir. Bu uygulamada, müşterilerin alışveriş alışkanlıklarını değerlendirmek için, müşterilerin aldıkları ürünlerin birbiriyle olan ilişkilerini tespit eder. Bu kapsamdaki ilişkilerin değerlendirmesi neticesinde, müşterilerin alışveriş davranışları belirlenebilir ve edinilen bu bilgiler neticesinde daha etkin satış taktikleri geliştirilebilir. Bu şekilde tespit edilen müşteri bilgileri ile ürünlerin yerleşimini düzenleyerek satışlarını yükseltebilirler. Örnek vermek gerekirse, müşterilerinin meyve suyu ve çikolata alma oranı fazlaysa, market yönetimi meyve suyu ve çikolatayı aynı rafa koyarak satış miktarını yükseltebilir [14].

2.3. Metin Madenciliği

Veri madenciliği ile arasında kavramsal bütünlük vardır ve alt dalı olarak nitelendirebiliriz. Metin madenciliği elimizde bulunan veri setleri içerisinde daha önceden bilmediğimiz bilginin bilgisayar yardımıyla keşfedilme sürecidir. Veri madenciliği ile arasındaki en büyük fark ise, metin madenciliğinde kalıplar doğal dil metinlerinin içerisinde çıkarılmaktadır [26].

İnternetin hayatımıza girmesiyle ve dijital depolama aygıtlarının kapasitelerinin hızla artmasıyla depolanan veri miktarı çok büyük boyutlara ulaşmıştır. Bilgisayarlar hayatımıza girmeden önce dokümanlarla ilgili her türlü işlem elle yapılmaktaydı. Bugün internette 2 milyardan fazla web sayfası olduğu düşünülürse bu bilgilere eski yöntemlerle ulaşmak neredeyse imkansızdır [15].

Yapısal veri, SQL, Access gibi buldukları yapı içinde organize edilip tanımlanabilen verilerdir. Bu veriler satır ve sütun düzeyinde tanımlanıp daha sonradan istenildiğinde bu bilgelere erişim kolayca yapılabilmektedir.

Yapısal olmayan veriler ise, JPG, Word, Text dökümanları gibi olup tanımlanabilir bir yapısı yoktur.

Günümüzde birçok işletmenin verileri yapısal olmayan formattadır. Bu kapsamda Merrill Lynch, potansiyel olarak kullanabilecek verilerin yüzde 80'inin yapısal olmayan formatta olduğunu belirtmiştir. Metin madenciliği uygulamaları, veri madenciliği algoritmalarının verileri işleme için kullanılan bir ön aşama gibi tanımlanabilir. Metin madenciliği çok büyük kapasitedeki belgelerin analiz edilip gizli örüntülerin elde edilmesi şeklinde tanımlanabilir [16].

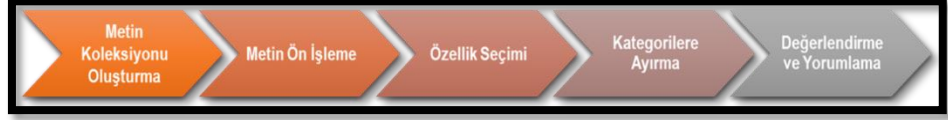
2.3.1. Metin Madenciliği Kullanım Alanları

Metin Madenciliği yeni bir teknoloji olmasına rağmen günümüzde birçok sektörde kullanılmaktadır. Bu sektörler aşağıdaki gibidir [17,18,19].

- Finans Alanı
- Pazarlama – Müşteri İlişkileri Yönetimi Alanı
- Sağlık ve Biyoloji Alanı
- Eğitim Alanı
- İnternet Yazılımı Alanı
- Diğer Ticari Uygulamalar

2.3.2. Metin Madenciliği Adımları

Metin madenciliği 5 adımdan oluşur. Öncelikle metin koleksiyon seti oluşturur, metin ön işlemeden geçirilir, özellik seçilir, kategorilere ayrılır ve değerlendirme yapılır. Şekil 2.3'te bu adımlar görülmektedir.



Şekil 2.3. Metin Madenciliği Adımları

2.3.2.1. Metin Koleksiyonu Oluşturma

Yapılacak olan çalışmada veriler içerisinde kullanılacak olanları ifade eder. Bu veriler genellikle internetten, işletmenin dosya sunucularından, kullanıcıların bilgisayarlarından temin edilir. Metin koleksiyonu ise bu toplu haldeki veriler içerisinden kullanacağımız alandır.

Örneğin, bu tez çalışmasında “Müşteri Şikayet Kayıtları” ele alınırken veri tabanlarında dağınık halde bulunan yaklaşık son 15 senenin kayıtları bulunmaktadır. Bu veriler içerisinden son 10 senenin kayıtları üzerinde çalışma yapılmıştır. Böylece metin koleksiyonu olarak bu 10 senelik veriler kabul edildi.

2.3.2.2. Metin Ön İşleme

Metin haldeki veriyi kelimelere ayırma işlemi, kelimeleri anlamsal değerini bulma işlemi, kelimenin köklerini bulma ve gereksiz kelimeleri temizleme işlemi, imla hatalarını, yazım yanlışlarını düzeltme işlemi gibi işlemler bu ön işleme adımında yapılır. Metin belgelerinin yapıtaşı olan kelime üzerindeki işlemler burada yapılır.

Bu tez çalışmasındaki şikayet kayıtlarında bulunan tüm kelimeler köklerine ayrılmıştır. Anlam bütünlüğünü bozan kelimeler temizlenmiş, yazım yanlışı olan kelimeler veri setine katkıda bulunması için düzeltilmiştir. Bu tür işlemler metin madenciliğinde “Metin Ön İşleme” adımı içerisinde yapılır.

2.3.2.3. Özellik Seçimi

Metin koleksiyonlarının içinde bulunan önemsiz ve gürültülü kelimelerin temizlenmesi işleminde “Özellik Seçimi” adımı uygulayıcılara fayda sağlamaktadır. Özellik seçimi yapılırken, metin koleksiyonu içinde bulunan ve sistemi eğitmede faydalı olacağı düşünülen önemli kelimelerin belirlenmesi ve metin koleksiyonu içinde bulunan ve sisteme katkısı olmayan sadece birkaç yerde geçen kelimelerin temizlenmesi işlemleri yapılır.

Tez çalışmasında bulunan şikayet metinleri içerisindeki tarih, mail adresi, para birimi gibi alanlar sisteme katkısı olmayacağı düşünülerek çıkartılmıştır. Aynı şekilde şikayet kayıtları içerisinde bulunan bölüm bilgisi, şube bilgisi gibi bilgiler bize faydalı olacağını düşünülerek veri tabanında tutulmuştur.

2.3.2.4. Kategorilere Ayırma

Metinlerin, veri madenciliği teknikleri ile analiz sürecidir. Önemli nokta metinler içerisindeki veriler yapılandırılmış biçimde olmalıdır. Karar ağaçları, yapay sinir ağları, genetik algoritmaları gibi teknikler ile veri analiz edilir [30].

2.3.2.5. Deęerlendirme ve Yorumlama

Veriler, analizi yapıldıktan sonra sonuçlar deęerlendirmeye alınır. Bu işlemin uzman kişilerce yapılması gerekir. Çıkan sonuçlarda istenmeyen durumlar oluşmuşsa “Kategorilere Ayırma” adımına gidilir ve tekrar analiz işlemleri yapılır. Eğer sonuçta istenmeyen hiçbir durum söz konusu deęil ise uygulamaya alınır ve son kullanıcının kullanımına hazır hale getirilir [30].

2.3.3. Metin Sınıflandırma

Metin sınıflandırmasında öncelikli olarak her bir kaydın önceden belirlenmiş bir sınıfta olması amaçlanmaktadır. Sınıf olmak için her bir kaydın belirli ortak özellikleri olması gerekir. Ortak özelliklere sahip olan bu verilerin hangi özellikleriyle bu sınıfa girdiğini tanımlayan algoritma, sınıflandırma algoritmasıdır. Sınıflandırma algoritması denetimli öğrenme kategorisine giren bir öğrenme biçimidir.

2.3.3.1. KNN Algoritması

KNN(K-Nearest Neighbor) K-En Yakın Komşu sınıflandırma algoritmasında birbirine yakın olan nesnelere aynı kategoride olarak deęerlendirilir. Algoritmada daha önceden tanımlanan eğitim verilerinin özellikleri tutulur ve yeni bir nesne geldiğinde bunun özelliklerine bakılarak sınıflandırma yapılır.

K birbirine yakın komşuları bulmaya çalışırken sorgu örneğini dikkate alır. Daha önceden tanımlanan eğitim verileri arasındaki en küçük mesafeler baz alınır. Birbirine yakın komşular bulunduktan sonra bunların kategorisi veri setinin kategorisini bulmada kullanılır [20].

Algoritmanın avantajları aşağıda verilmektedir [20]:

- Kolay ve basit bir algoritmadır.
- Eğitim dokümanlarında bulunan gürültülü veri oranını mümkün olduğunca aza indirmeye çalışır.
- Sistem ne kadar fazla eğitim dokümanı ile eğitilirse doğru sonuç oranı da o kadar artar.

Algoritmanın dezavantajları aşağıdaki gibidir [20]:

- K değişkenini önceden bulmamız gerekir.
- En yakın komşuları bulmak için kullanılacak olan özelliklerin ve mesafenin uzaklığı konusunda kesin bilgi vermez.
- En yakın komşuları bulmada kullanılan hesaplama maliyeti fazladır. Bunun nedeni de her bir sorgu örneğinin sistemde tanımlı bütün eğitim verilerine olan mesafesi hesaplanarak yapılır.

2.3.3.2. Karar Ağacı

Karar ağaçları tahmin edici modeldir. Akış şemalarına benzer şekilde tasarlanırlar, bu sebeple kolay anlaşılır, yorumlaması kolay bir yöntemdir.

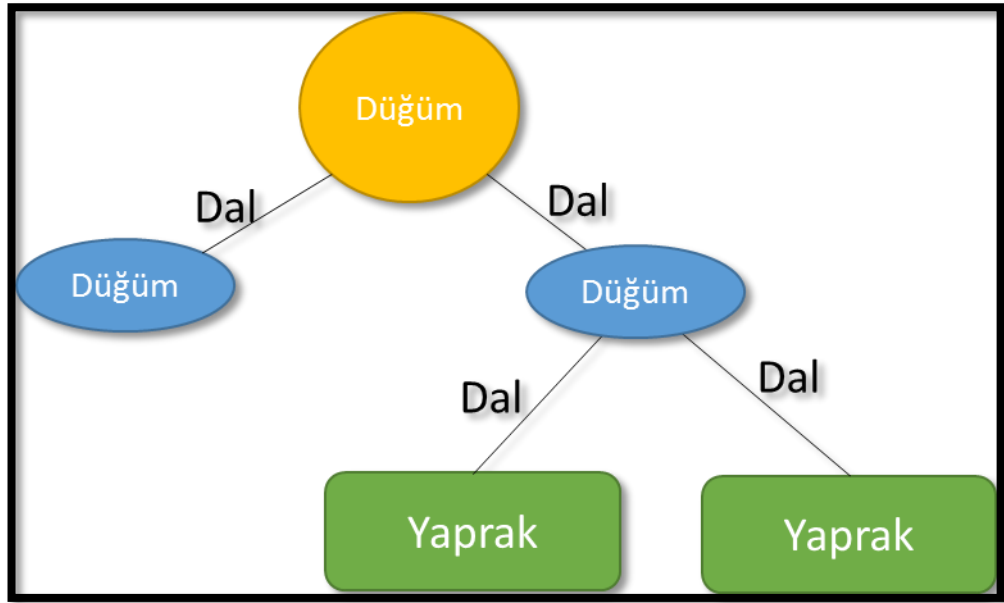
Tahmin edici değişkenlerle içinde bulunduğu bağlantıyı göz önünde bulundurup ağaç dalı şeklinde segmentlere ayıran bir yöntemdir. Bu sınıflandırmanın amacı, bağımlı değişkendeki ayrımları en yüksek seviyeye çıkartacak şekilde veriyi belirli bir sırayla farklı sınıflara bölmektir [21].

Şekil 2.4'de görüldüğü gibi dal ve yaprak ağacın elemanıdır.

Şekil 2.4’de yer alan ağaç şemasında her düğüm yeni bir testi belirtir. Test sonucunda ağacın dalları oluşur. Bu dalların oluşma sürecinde veri kaybetmemek adına verilerin hepsini içerecek miktarda değişik dal meydana getirilmelidir.

Dal; testin sonucudur. Her dal oluşumunda tanımlanacak sınıfın belirlenmesi istenir. Dal oluşumunun sonucunda sınıflandırma yapılamıyorsa tekrar karar düğümü oluşturulur.

Dalın sonucunda bir sınıflandırma ortaya çıkıyorsa yaprak oluşur. Yaprak, verileri kullanarak ulaşılmak istenen sınıflandırmanın sınıflarından birini gösterir.



Şekil 2.4. Karar Ağacı

2.3.3.3. K-Means

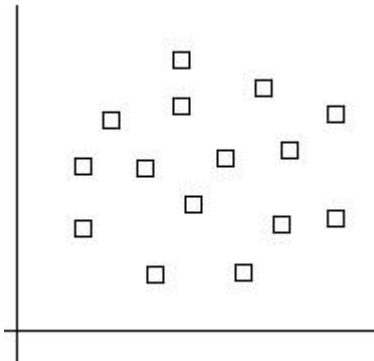
K-Means algoritması, sınıflandırmada kullanılan denetimsiz öğrenme metotlarından biridir. K-Means'in amacı özellik çıkarımı oluşturulmuş veri kümesinin birden fazla sınıfa göre sınıflandırılmasıdır. Kullanılan teknik ile tüm sınıflar için bir merkez belirlenir. Bu belirlenen merkeze uzaklığına göre yeni sınıflara atama yapılır [22].

K-Means dört adımdan oluşur :

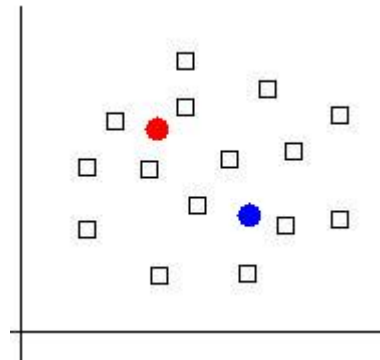
- Sınıf merkezleri belirlenir,
- Sınıf merkezinden uzak nesnelerin aralarındaki mesafelere göre sınıflandırma yapılır,
- Yeni belirlenen sınıflandırmalar ile merkezler tekrardan belirlenir,
- Kararlı yapı oluşana kadar 2. ve 3. adımların tekrarlanması sağlanır.

K-Means algoritmasını bir örnekle açıklayalım:

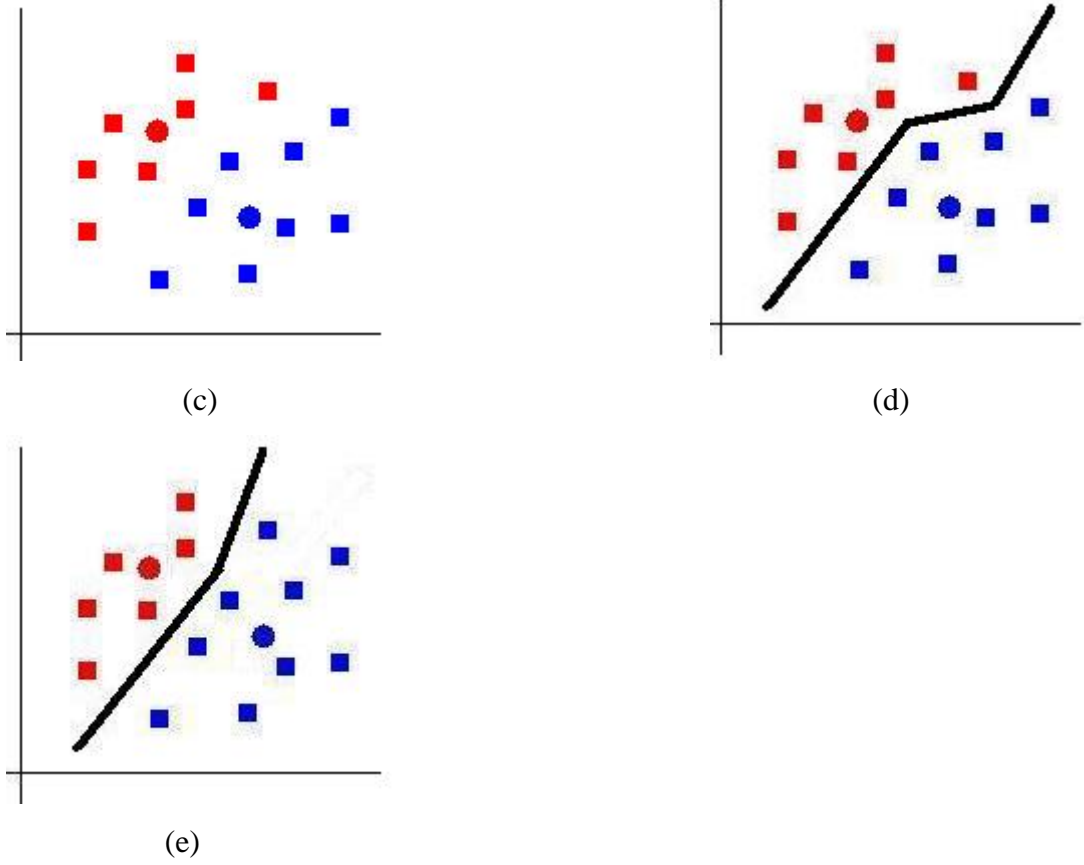
1. Aşağıda verilen örnekleri uzaya dağılmış olarak ele alalım (a)
2. Uzayda kodlanmış iki adet hedef küme tanımlanır (b)
3. Renk olarak hangisine yakınsa buna göre sınıf tanımlarının mesafelerine bütün örneklerimizi sınıflandırıyoruz (c)
4. Oluşan yeni sınırı bir hat ile ayırıyoruz (d)
5. Daha önceden sınıflandırılan örneklerin merkezi bulunur (e) yuvarlak olarak gösterilen sınıf karakteristiğini ifade eden ilk örneklerin yerine değiştirmek de diyebiliriz [22].



(a)



(b)



Şekil 2.5. K-Means Örneği [22]

2.3.3.4. SVM (Support Vector Machines)

SVM (Support Vector Machines), sınıflandırmadaki basit ve etkin tekniklerden birisidir. Bu sınıflandırmada, doğrusal alan da yer alan iki grup arasında bir ortak alan belirlenir, bu alan iki grubunda nesnelere en uzak yerdir. SVM burada bu alanın nasıl belirleneceğine karar verir.

Bu alanın belirlenmesi adına her iki gruba da eşit uzaklıkta ve paralel iki alan çizgisi çekilir ve bu alan çizgileri birbirine yaklaştırılır ve ortak bir alan çizgisi meydana getirilir [27].

2.3.3.5. Naive Bayes

Adını Thomas Bayes'den almıştır. Kolay uygulanabilir bir teknik olduğundan en çok kullanılan sınıflandırma veya kategorilere ayırma araçlarından biridir.

Sistemi eğitmek için öncelikli olarak belirli bir oranda veri girişi yapılır. Burada dikkat edilecek husus sisteme tanımlanan verilerin mutlaka bir sınıfı/kategorisi olmak zorundadır. Sistem eğitildikten sonra bu veriler üzerinde yapılan olasılık işlemleri ile bundan sonra gelecek test verileri bu orana göre işlem yapar ve test verilerinin kategorilerini tahmin etmeye çalışır. Burada sisteme tanımlanan verilerin miktarı ne kadar çoksa sunulan test verisinin tahmin etme oranı da o derece yükselmektedir. Bu tez çalışmasında görüldüğü gibi öğretilmiş veri miktarını çoğalttıkça tahmin etme oranı da o derece yükselmiştir.

Naive Bayes sınıflandırmasında Denklem 1'de belirtilen formül kullanılır.

$$P(X|Y) = \frac{P(Y|X) P(X)}{P(Y)} \quad (1)$$

$P(X|Y)$; Y etkeni gerçekleştiği anda X etkeninin ortaya çıkma ihtimalidir

$P(Y|X)$; X etkeni gerçekleştiği anda Y etkeninin ortaya çıkma ihtimalidir

$P(X)$ ve $P(Y)$; X ve Y etkenlerinin önsel ihtimalidir [23]

Bu tez çalışmasında geliştirilen yazılım aracında Multinomial(Çok Terimli) Naive Bayes yöntemi kullanılmıştır. Multinomial model olasılıklar üzerinde duran bir öğrenim tekniğidir. Veri setinde her bir dokümanın tekrarlanma sıklığıyla ilgilenir [32]. Hata oranı düşüklüğü ve kullanım kolaylığı sebebiyle bu teknik kullanılmıştır. Bu yöntemin kullanımı ile ilgili bir örnek aşağıda anlatılmaktadır.

Örneğin, bir işletmede bölümlerin bulunduğu bir veri setimiz olsun. Veri setinin içeriği aşağıdaki gibidir. Bu bölümlere A ve B gibi atanmış kategorilerimiz olsun.

Bu veri setine göre yeni gelecek olan bir bölüm dizisinin hangi kategoride olacağını tahmini yapılacaktır.

Çizelge 2.2. Naive Bayes Örnek Uygulama

Doküman No	Öğretilecek Veri Seti			Kategori
1	Bilgi İşlem	Muhasebe	Bilgi İşlem	A
2	Bilgi İşlem	Bilgi İşlem	İnsan Kaynakları	A
3	Bilgi İşlem	Finans		A
4	Pazarlama	Müşteri İlişkileri	Bilgi İşlem	B

İlk olarak, Multi Nominal sınıflandırmaya göre kategorilerin tüm bölümlere oranı bulunur.

$P(A)=3/4=0,75$ (Öğretilecek olan veri setinde A kategorisinde olan satırların tüm satırlara oranı)

$P(B)=1/4=0,25$ (Öğretilecek olan veri setinde B kategorisinde olan satırların tüm satırlara oranı)

Bu aşamadan sonra, öğretilecek olan bölümlerin atandığı kategoriye göre koşullu olasılıkları hesaplanır.

$P(X|Y)=(Y \text{ kategorisindeki satırlarda "X" ifadesinin tekrar sayısı} +1) / (Y \text{ kategorisindeki satırlarda bulunan tüm kelimelerin sayısı} + \text{Öğretilen veri sayısı})$

Koşullar:

$P(\text{Bilgi İşlem}|A)=(5 + 1)/(8+6)=0,428$

$$P(\text{Pazarlama}|A)=(0 + 1)/(8 + 6)=0.071$$

$$P(\text{Müşteri İlişkileri}|A)=(0 + 1)/(8 + 6)=0.071$$

$$P(\text{Bilgi İşlem}|B)=(1 + 1)/(3 + 6)=0.222$$

$$P(\text{Pazarlama}|B)= (1 + 1)/(3 + 6)=0.222$$

$$P(\text{Müşteri İlişkileri}|B)=(1 + 1)/(3 + 6)=0.222$$

Koşullarımızı bu şekilde belirlemiş olduk. Şimdi ise bu kategorilere göre rastgele bir bölüm bilgileri dizisinin hangi kategoride olacağını bulalım.

Test Verimiz şu şekilde olsun: Bilgi İşlem, Bilgi İşlem, Servis, Bilgi İşlem, Pazarlama, Müşteri İlişkileri

Test cümlemizde görüldüğü gibi sisteme daha önceden tanımlanmamış yani eğitilmemiş bir bölüm ismi yer almaktadır. Dolayısıyla bu bölüm ismi için sistem olasılık değerini hesaplayamayacaktır. Bu sebeple sonuca etki etmemesi için çarpmada etkisiz eleman olan “1” olarak kabul edilir ve bununla hesaplama formülüne dahil ederiz.

$$P(A|\text{Test}) = P(A) \times P(\text{Bilgi İşlem}|A) \times P(\text{Bilgi İşlem}|A) \times P(\text{Servis}|A) \times P(\text{Bilgi İşlem}|A) \times P(\text{Pazarlama}|A) \times P(\text{Müşteri İlişkileri}|A)$$

$$P(A|\text{Test})=0,75 \times 0,428 \times 0,428 \times 1 \times 0,428 \times 0,071 \times 0,071=0,000296$$

$$P(B|\text{Test}) = P(B) \times P(\text{Bilgi İşlem}|B) \times P(\text{Bilgi İşlem}|B) \times P(\text{Servis}|B) \times P(\text{Bilgi İşlem}|B) \times P(\text{Pazarlama}|B) \times P(\text{Müşteri İlişkileri}|B)$$

$$P(B|\text{Test})=0,25 \times 0,222 \times 0,222 \times 1 \times 0,222 \times 0,022 \times 0,022=0,000001$$

$P(A|\text{Test}) > P(B | \text{Test})$ olduğundan dolayı test verisinin A kategorisinde olma ihtimali daha yüksektir.

3. MÜŞTERİ MEMNUNİYETİNİ BELİRLEMEK İÇİN METİN MADENCİLİĞİ TABANLI BİR YAZILIM ARACI

Bu tez çalışmasında, metin madenciliğinde kullanılan Naive Bayes sınıflandırma tekniği kullanılmaktadır. Çalışma içerisinde bir yazılım aracı geliştirilerek bir şirketin müşteri şikayet kayıtlarından hareketle görüşmenin mutlu veya mutsuz şekilde sonlanacağı tahmin edilmeye çalışılmaktadır.

3.1. Giriş

Müşteri memnuniyeti, genellikle müşterilerin ürün ya da hizmeti satın alma öncesinde beklentisi ile satın alma sonucunda tatmin olma oranıdır ya da aldığı hizmetten memnun kalma derecesidir [24]. Ayrıca müşterinin bir ürünü aldıktan sonra beklediği faydaya, üründen beklediği performansa, sosyo-kültürel değerlerin en uygunluğuna bağlı bir fonksiyon olduğu da belirtilir [25].

Bu tez çalışması kapsamında, bir müşteri memnuniyet analizi sınıflandırma aracı tasarımı gerçekleştirilmiştir. Bu yazılım aracının sayesinde, sisteme tanımlanmış olan Mutlu/Mutsuz müşteri şikayet kayıtlarının bundan sonra oluşacak kayıtlarda kategorisini belirlemeye veya tahmin etmeye çalışılacaktır.

3.2. Müşteri Memnuniyeti Yazılım Aracının Veri Tabanı Tasarımı

Veri toplama öncesinde, metin sınıflandırması için kullanılacak yöntem seçilir. Bunun üzerinden kategorilere ayrılarak veri toplama işlemi yapılır.

Bu tez çalışmasında, istatistiksel bir yöntemle sahip olan, kullanım kolaylığı ve elde ettiği başarılı sonuçlar sebebiyle metin sınıflandırma araçlarından olan Naive Bayes sınıflandırma aracı kullanılmıştır [23]. Müşteri memnuniyetini analiz etmek için

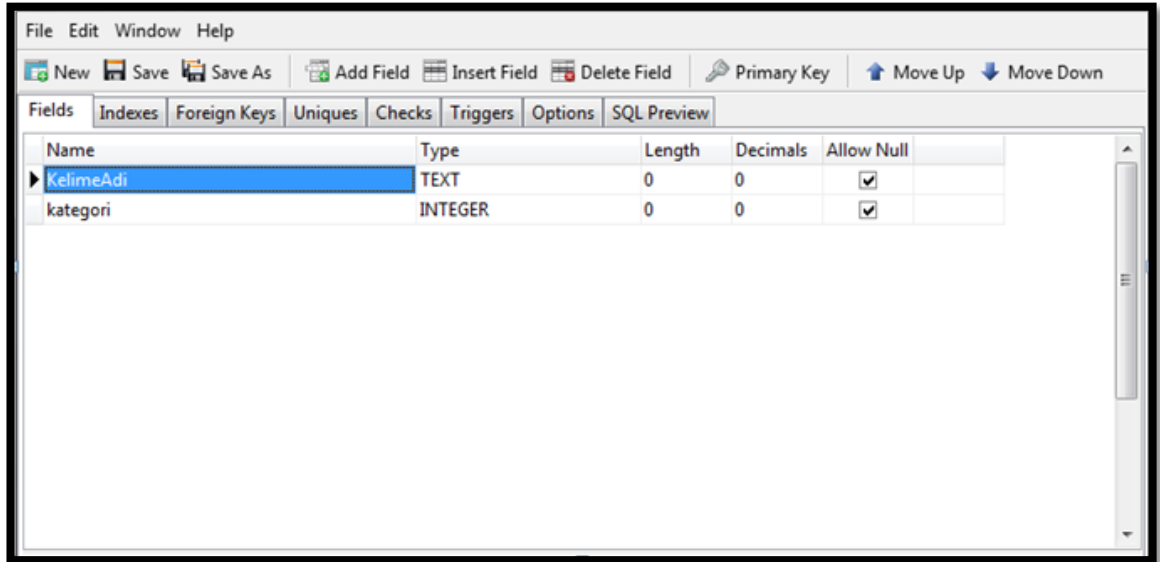
kullanılacak müşteri şikayet kayıtları ve bu şikayetlerin mutlu veya mutsuz sonuçları algoritmanın temel bileşenlerini oluşturmaktadır. Programa girilen şikayet cümleleri öncelikle kelimelere bölünür. Sonrasında kelimenin köküne inilir. Bu kök haline indirgenmiş kelimeler de mutlu veya mutsuz kategorisi ile birlikte veri tabanına yüklenir.

Bu çalışmada veri tabanı olarak SQLite ve derleyici olarak da Navicat Premium programı kullanılmaktadır.

Bu tez çalışmasında yapılan Memnuniyet Analizi programının veri tabanında iki tablo yer almaktadır;

1. NB_KELIMELER_DETAY tablosu
2. NB_KELIMELER_MASTER tablosu

Şekil 3.1’de NB_KELIMELER_DETAY Tablosu görülmektedir:



Name	Type	Length	Decimals	Allow Null
KelimeAdi	TEXT	0	0	<input checked="" type="checkbox"/>
kategori	INTEGER	0	0	<input checked="" type="checkbox"/>

Şekil 3.1. NB_KELIMELER_DETAY tablosu

Bu tabloda, kelimeler tekilleştirilmeden her bir kelimenin geçtiği cümlenin hangi kategori(mutlu/mutsuz) ile sınıflandırıldığı tutulmaktadır. Her bir kelimenin Mutlu ve Mutsuz olasılığını belirlemek için bu tablo kullanılmaktadır. Şekil 3.2’de NB_KELIMELER_DETAY tablosunda mutlu ve mutsuz kategorilerin gösterimi yer almaktadır.

The screenshot shows a database query editor with the following query: `select * from NB_KELIMELER_DETAY`. The results are displayed in a table with the following data:

KelimeAdi	kategori
aracı	mutsuz
stop	mutsuz
ol	mutsuz
bu	mutsuz
sıkıntı	mutsuz
ol	mutsuz
var	mutsuz

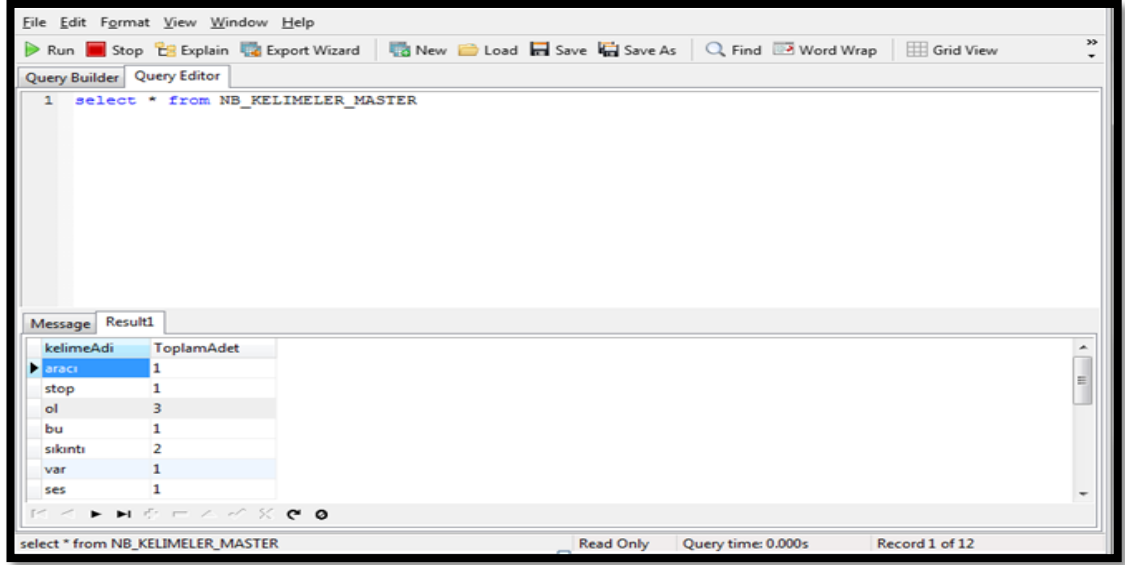
Şekil 3.2. NB_KELIMELER_DETAY tablosu

Şekil 3.3’de NB_KELIMELER_MASTER Tablosu görülmektedir:

Name	Type	Length	Decimals	Allow Null
kelimeAdi	TEXT	0	0	<input checked="" type="checkbox"/>
ToplamAdet	INTEGER	0	0	<input checked="" type="checkbox"/>

Şekil 3.3. NB_KELIMELER_MASTER tablosu

Bu tabloda, girilen şikayet kayıtları içerisindeki her bir kelimenin tekilleştirilmiş hali ve her bir kelimedenden bütünsel veri içerisinde kaç adet tutulduğu bilgisi yer almaktadır. Şekil 3.4’de NB_KELIMELER_MASTER tablosunda kelime adetleri görülmektedir.

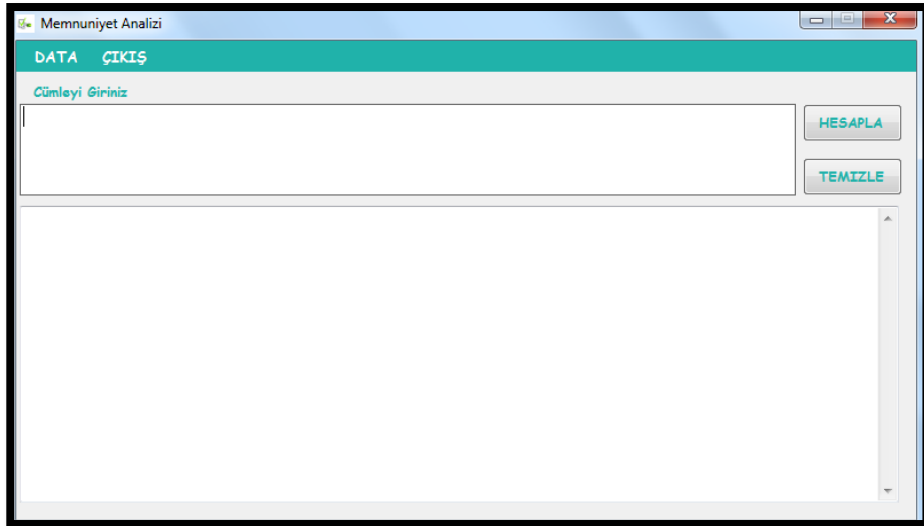


kelimeAdi	ToplamAdet
araci	1
stop	1
ol	3
bu	1
sikinti	2
var	1
ses	1

Şekil 3.4. NB_KELIMELER_MASTER tablosu

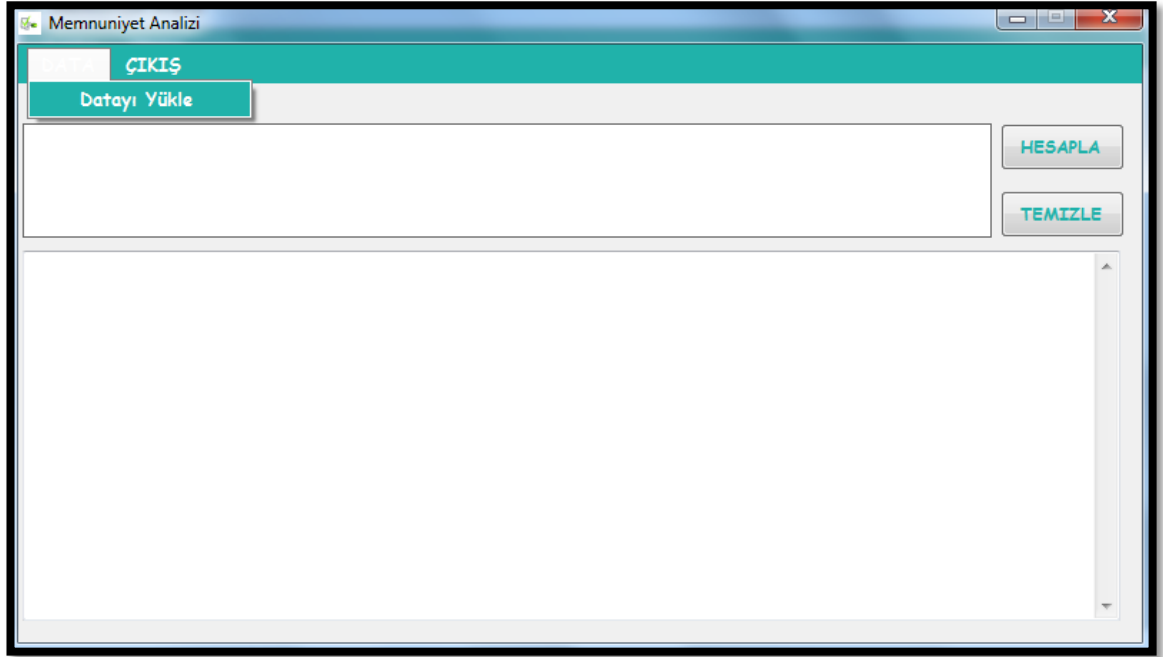
3.3. Müşteri Memnuniyeti Yazılım Aracının Tasarımı

Tez çalışması içerisinde geliştirilen yazılım aracı Visual Studio. Net platformu üzerinde C# programlama dili ile bir Windows Forms uygulaması olarak tasarlanmıştır. Şekil 3.5’de yazılım aracının ekran görüntüsü verilmektedir.



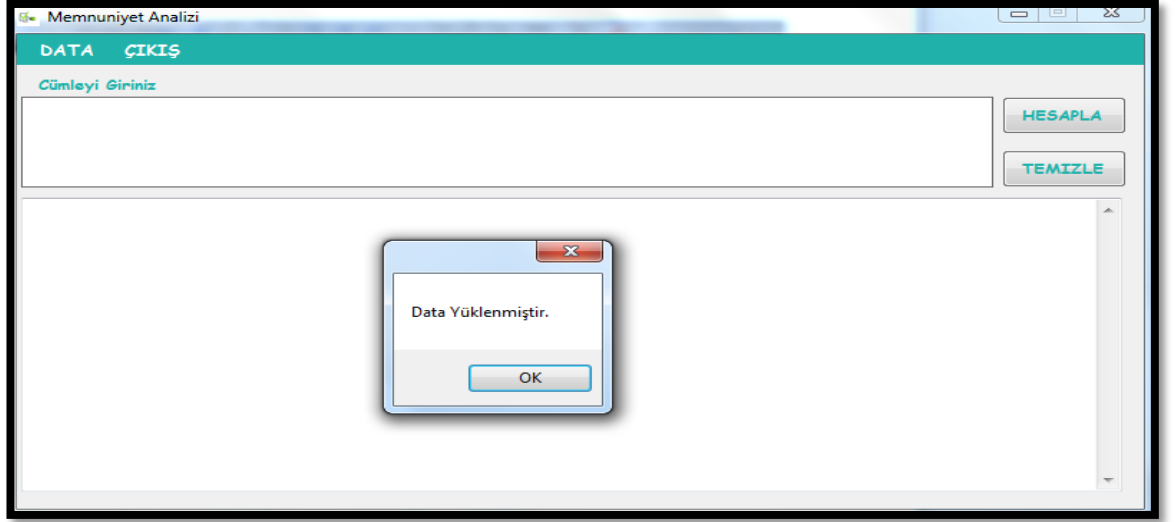
Şekil 3.5. Memnuniyet Analizi Programı

Yazılım aracı için öncelikli olarak veri tabanına yüklenecek kayıtlar hazırlanır. Excel programında bir sütun şikayet cümleleri ve diğer sütun kategori(mutlu/mutsuz) olmak üzere şikayet kayıtları hazırlanır. Hazırlanan bu Excel sayfası “tab delimited” olarak kaydedilip programa yüklenmeye hazır hale gelir. Dosyayı, programa yükleme işlemi Şekil 3.6’da gösterilen “Datayı Yükle” butonu ile yapılır.



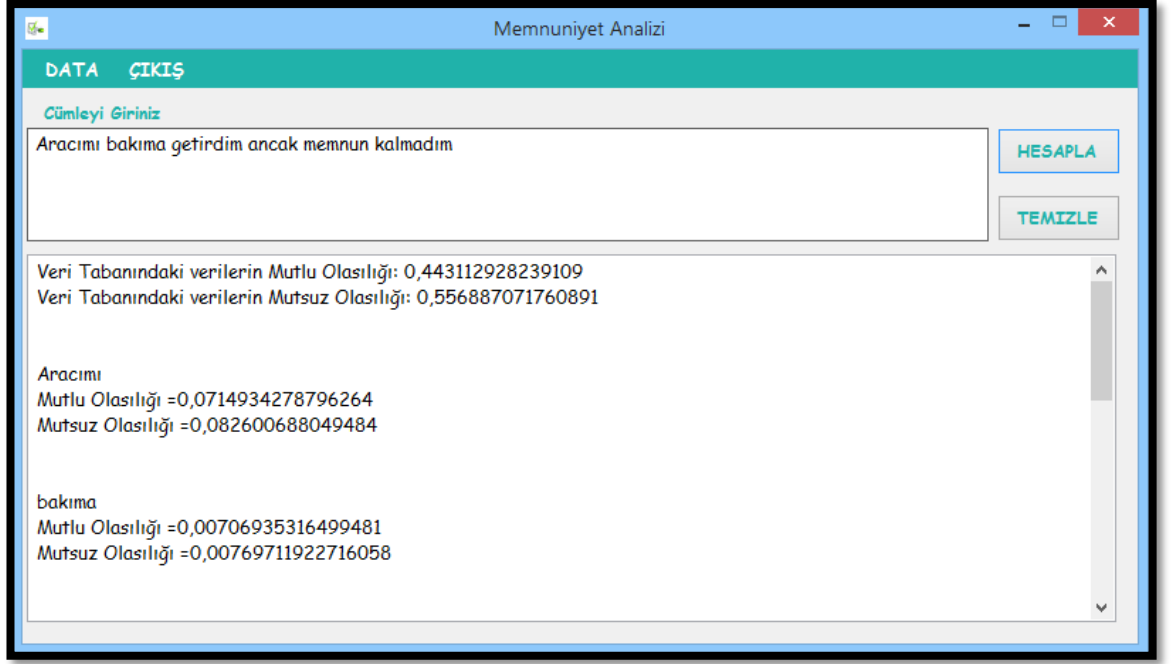
Şekil 3.6. Memnuniyet Analizi Programına Kayıt Yükleme

“Datayı Yükle” işlemi tamamlandığında Şekil 3.7’de görüldüğü gibi bir bildirim penceresi gelir ve bu bildirimle kayıtların yükleme işleminin tamamlandığını gösterir.



Şekil 3.7. Memnuniyet Analizi Programına Kayıtların Yüklenmesi

Kayıtların yüklenmesi ile program yeni gelen cümlelerin olasılıklarını hesaplayabilir hale gelmiştir. "Cümleyi giriniz" alanına yeni gelen şikayet kaydı girilir ve "Hesapla" butonuna basılır. Alt kısımda çıkan sonuçlara göre "Veri tabanındaki verilerin mutlu olasılığı ve Veri tabanındaki verilerin mutsuz olasılığı" hesaplanır, bu işlemden sonra cümle içinde girilen her kelimenin "mutlu ve mutsuz olasılıkları" hesaplanır. En son olarak da girilen cümlenin "mutlu ve mutsuz olasılıkları" hesaplanır. Şekil 3.8'de Memnuniyet Analizi Programının görünümü yer almaktadır.



Şekil 3.8. Memnuniyet Analizi Programı ile Mutlu/Mutsuz Oranları

Mutlu ve Mutsuz olasılıkların belirlenmesi; Mutlu ve Mutsuz olasılıklarını hesaplamak için nb_kelimeler_detay tablosundaki toplam mutlu kelime adedi ve toplam mutsuz kelime adedi çekilir. Bu adetler nb_kelimeler_detay tablosundaki toplam kelime adedine bölünür, mutlu ve mutsuz olasılıklar hesaplanır. Elde edilen olasılıklar programın sonuç bölümüne "veri tabanındaki verilerin mutlu olasılığı" ve "veri tabanındaki verilerin mutsuz olasılığı" olarak yazılır. Denklem 2 ve Denklem 3'de Mutlu Olma Olasılığını ve Mutsuz Olma Olasılığında hesaplanan formül yer almaktadır.

$$\text{Mutlu Olasılığı} = \frac{\text{toplam mutlu kelime adedi}}{\text{toplam kelime adedi}} \quad (2)$$

$$\text{Mutsuz Olasılığı} = \frac{\text{toplam mutsuz kelime adedi}}{\text{toplam kelime adedi}} \quad (3)$$

Kelimelerin olasılıklarının belirlenmesi; “Cümleyi giriniz” alanına yazılan şikayet kaydı boşluk karakterine göre kelimelere ayrılır. Her bir kelimenin “mutlu ve mutsuz olasılıkları” veri tabanındaki iki tabloda kullanılarak hesaplanır. Her bir kelimenin olasılıkları kelimenin kendisi ile birlikte programın sonuç bölümüne yazılır.

- Kelimenin mutlu olasılığının hesaplanması;

Kelimenin mutlu olasılığı Denklem 4’te gösterilen formülle hesaplanır.

$$\text{Kelimenin mutlu olasılığı} = \frac{\text{kelimenin mutlu kategorisindeki toplam adedi} + 1}{\text{mutlu kategorisindeki toplam kelime adedi} + \text{öğretilmiş kelime adedi}} \quad (4)$$

Denklem 4’te belirtilen “kelimenin mutlu kategorisindeki toplam adedi” NB_KELIMELER_DETAY tablosunda bu kelimenin kökünün, mutlu kategorisinde kaç adet tutulduğu bilgisini verir.

Denklem 4’te “Mutlu kategorisindeki toplam kelime adedi” bilgisi NB_KELIMELER_DETAY tablosundaki mutlu kategorisindeki kelime adedir.

Denklem 4’te “Öğretilmiş kelime adedi” olarak istenen değer ise NB_KELIMELER_MASTER tablosundaki tüm kelimelerin adedir.

- Kelimenin mutsuz olasılığının hesaplanması;

Kelimenin mutsuz olasılığı Denklem 5’de gösterilen formülle hesaplanır.

$$\text{Kelimenin mutsuz olasılığı} = \frac{\text{kelimenin mutsuz kategorisindeki toplam adedi} + 1}{\text{mutsuz kategorisindeki toplam kelime adedi} + \text{öğretilmiş kelime adedi}} \quad (5)$$

Denklem 5’de belirtilen “kelimenin mutsuz kategorisindeki toplam adedi” bilgisi NB_KELIMELER_DETAY tablosunda bu kelimenin kökünün, mutsuz kategorisinde kaç adet tutulduğunun bilgisini verir.

Denklem 5’de “Mutsuz kategorisindeki toplam kelime adedi” bilgisi NB_KELIMELER_DETAY tablosundaki mutsuz kategorisindeki kelime adedir.

Denklem 5’de “Öğretilmiş kelime adedi” olarak istenen değer ise yine NB_KELIMELER_MASTER tablosundaki tüm kelimelerin adedir.

- Girilen şikayet kaydının mutlu ve mutsuz olasılıklarının hesaplanması:

Kelimelerin mutlu ve mutsuz olasılıkları hesaplandıktan sonra girilen şikayet kaydının mutlu veya mutsuz hesaplanması adımına geçilir. Bu işlemi yapabilmek için Denklem 6 ve Denklem 7’de belirtilen formüller ile hesaplanır.

$$\text{Cümlenin mutlu olma olasılığı} = (P_1 \times P_2 \times P_3 \dots) \times (\text{mutlu kategorisindeki toplam kelime adedi}) \quad (6)$$

Denklem 6’da belirtilen P_n girilen cümlenin n. kelimesinin mutlu olma olasılığını gösterir.

Cümlenin mutsuz olma olasılığı= $(K_1 \times K_2 \times K_3 \dots) \times$ (mutsuz kategorisindeki toplam kelime adedi) (7)

Denklem 7'de belirtilen K_n girilen cümlenin n. kelimesinin mutsuz olma olasılığını gösterir.

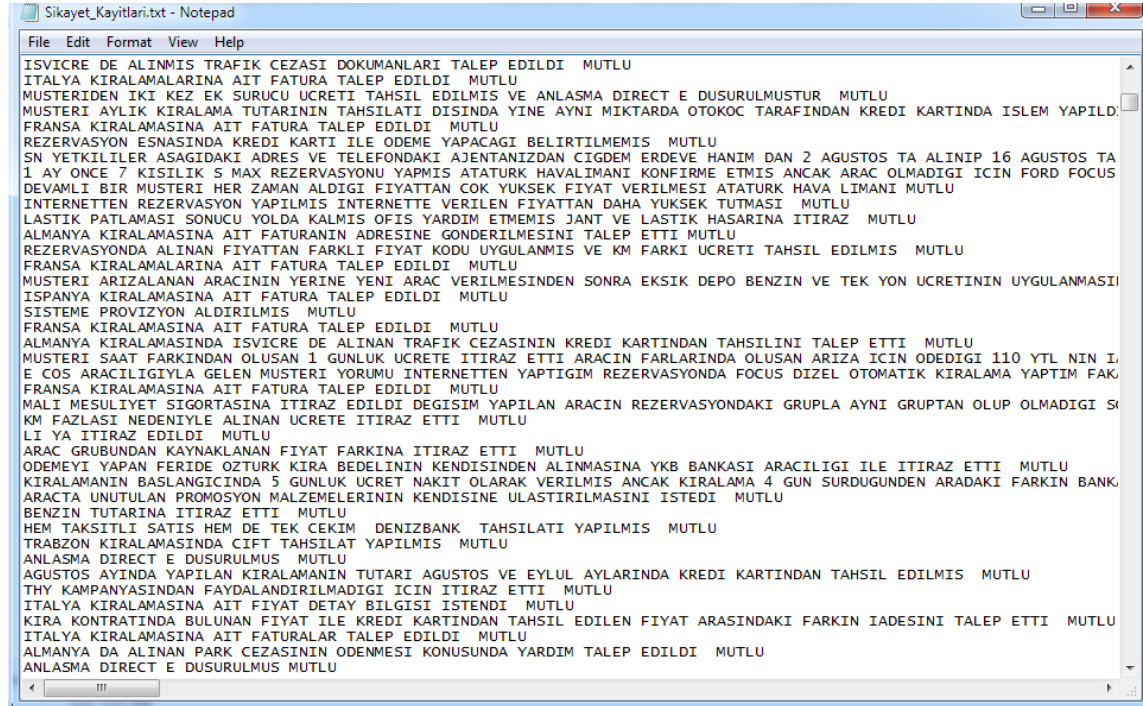
4. UYGULAMALAR VE SONUÇLARI

Tez çalışmasında, geliştirilen müşteri memnuniyeti yazılım aracı ile 2 uygulama gerçekleştirilmiştir ve sonuçları değerlendirilmiştir.

4.1. Uygulama 1

Bu uygulamada, 2002 – 2012 yılları arasında gerçekleşen toplam 5000 adet Mutlu veya Mutsuz şeklinde kapanan şikayet kaydı ile memnuniyet analizi yazılım aracı eğitilecektir. Sadece şikayet metni yazılım aracına verilecektir. Sonrasında ise 2013 ve 2014 yılları arasında gerçekleşen kayıtlar ile yazılım aracı test edilecektir ve Mutlu veya Mutsuz oranının hesaplanması istenecektir.

Uygulamada öncelikli olarak memnuniyet analizi yazılımına kayıtları yüklenmesi gerekmektedir. Bunun için Şekil 4.1’de görüldüğü gibi kayıtların yer aldığı metin dosyası hazırlanmaktadır.



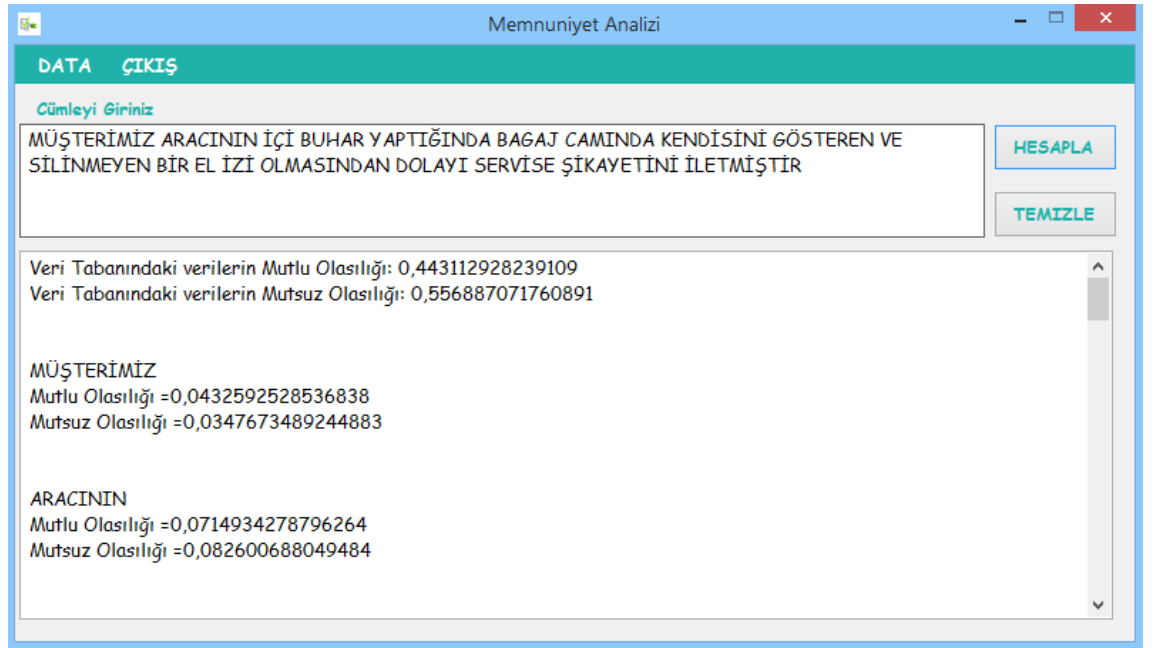
Şekil 4.1. Veri tabanına veri yüklemek için metin dosyasının hazırlanması

Bu adımdan sonra memnuniyet analizi yazılımı aracı çalıştırılır ve Datayı Yükle butonuna basılarak kayıtlar yüklenir. Kayıtlar yüklendikten sonra artık test etme aşaması başlamaktadır.

Memnuniyet analizi programında “Cümleyi Giriniz” alanına istenilen cümle girilir. Burada 2013 ve 2014 yılları arasında ki gerçek şikayet kayıtları test edilmektedir.

Girmiş olduğumuz şikayet kaydı: “MÜŞTERİMİZ ARACININ İÇİ BUHAR YAPTIĞINDA BAGAJ CAMINDA KENDİSİNİ GÖSTEREN VE SİLİNMEYEN BİR EL İZİ OLMASINDAN DOLAYI SERVİSE ŞİKAYETİNİ İLETMİŞTİR”.

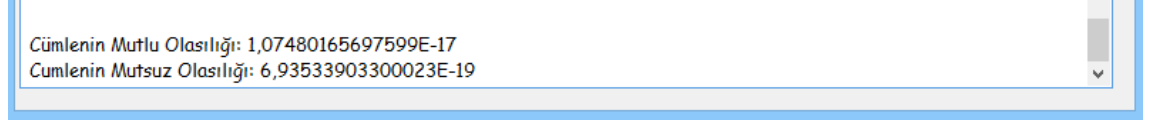
Şekil 4.2’de istenilen cümle girilmekte ve sonrasında “HESAPLA” butonuna basılmaktadır.



Şekil 4.2. Memnuniyet Analizi Yazılımının Uygulama Sonuçları

Hesaplama işleminden sonra Şekil 4.2’de görülen sonuçlarda ilk satırda yazan “Veri Tabanındaki verilerin Mutlu Olasılığı:” ve “Veri Tabanındaki verilerin Mutsuz Olasılığı:” cümlesi programın başında metin dosyasından yüklenen tüm kayıtların Mutlu/Mutsuz olasılığını vermektedir.

Programın sonunda yer alan ‘‘Cümlenin Mutlu Olasılıđı.’’ ve ‘‘Cümlenin Mutsuz Olasılıđı.’’ deđerleri ise girilen cümlenin Mutlu veya Mutsuz olma olasılıđını göstermektedir. Bu deđerler ařađıdaki Őekil 4.3’te gösterilmektedir.



Őekil 4.3. Uygulama 1 Sonuçları

Őekil 4.3.’te görüldüğü gibi girilen cümlenin ‘‘Mutlu’’ ile kapanma olasılıđı daha yüksektir. Girilen Őikayet kaydının kapanma Őekli sonrasında kontrol edildiğinde Őikayetin ‘‘Mutlu’’ Őekilde kapandıđı görülmüřtür.

4.2. Uygulama 1 Sonuçları

Uygulama 1 içerisinde tanımlama yaptığımız veri tabanını üzerinde 1000 tanesi mutlu ve 1000 tanesi mutsuz olmak üzere 2000 tane Őikayet kaydı test için sisteme girilmiřtir. Kullanılan tekniğin dođruluđunu deđerlendiren kavramlardan bir tanesi F-Measure(Ölçütü)’dür. Dođru sınıfa girilen test dokümanıyla yanlış sınıfa girilen test dokümanın nitelikleriyle ilgilidir.

F-Measure deđerini bulabilmek için Karıřıklık Matrisinden yararlanılır. Karıřıklık matrisinde satırlar test verisindeki örneklerin gerçek deđerini, sütunlar ise tahmin deđerini gösterir [33]. Çizelge 4.1’de Karıřıklık Matrisinin görünümü yer almaktadır.

Çizelge 4.1. Karışıklık Matrisi

		Tahmin Edilen Sınıf	
		Mutlu	Mutsuz
Doğru Sınıf	Mutlu	x	y
	Mutsuz	z	t

x: TP(True Pozitif)

y: FN(False Negatif)

z: FP(False Pozitif)

t: TN(True Negatif)

F-Measure değerini bulabilmek için öncesinde Kesinlik ve Duyarlılık değerlerinin bulunması gerekir.

Kesinlik; sınıfı doğru olarak tahmin edilmiş True Pozitif örnek sayısının, sınıfı 1 olarak tahmin edilmiş tüm örnek sayısına oranıdır [33].

$$\text{Kesinlik} = \frac{TP}{TP+FP} \quad (8)$$

Duyarlılık; sınıfı doğru olarak tahmin edilmiş örnek sayısının toplam pozitif örnek sayısına oranıdır[33].

$$\text{Duyarlılık} = \frac{TP}{TP+FN} \quad (9)$$

F-Measure; kesinlik ve duyarlılık sonuçlarını harmonik ortalamasıdır [33].

$$F\text{-Measure} = \frac{2 \times \text{Duyarluluk} \times \text{Kesinlik}}{\text{Duyarluluk} + \text{Kesinlik}} \quad (10)$$

Sisteme test için girmiş olduğumuz 1000 tane mutlu olasılıklı 1000 tane mutsuz olasılıklı kayıtların Karışıklık Matrisine göre dağılımı Çizelge 4.2’de olduğu gibidir.

Çizelge 4.2. Uygulama 1 Sonuçları

		Tahmin Edilen Sınıf	
		Mutlu	Mutsuz
Doğru Sınıf	Mutlu	793	207
	Mutsuz	263	737

$$\text{Duyarluluk} = \frac{TP}{TP+FN} = \frac{793}{793+207} = 0,793$$

$$\text{Kesinlik} = \frac{TP}{TP+FP} = \frac{793}{793+263} = 0,750$$

$$F\text{-Measure} = \frac{2 \times \text{Duyarluluk} \times \text{Kesinlik}}{\text{Duyarluluk} + \text{Kesinlik}} = \frac{2 \times 0,793 \times 0,750}{0,793+0,750} = 0,7709 \text{ olarak bulunur.}$$

4.3. Uygulama 2

Uygulama 1 de olduğu gibi 2002-2012 yılları arasında gerçekleşen yaklaşık 10000 adet şikayet kaydı ile deney yapılmıştır. Ancak Uygulama 1’den farklı olarak yeni alanlar eklenmiştir. Uygulama 1’de sadece Şikayet Metni üzerinde yapılan testi bu kez yeni özellikler eklenerek yapılacaktır. Eklediğimiz özellikler şunlardır; Şirket Adı, Lokasyon Adı, Şikayet Metni, Şikayet Tipi, Şube, Müşteri Sadakat Segmenti, Müşterinin İl ve ilçesi, Müşterinin Gerçek Kişi mi yoksa Tüzel Kişi mi olduğu bilgisi,

Cinsiyet, Yaş, Arabasının Marka ve Modeli, Şikayet Ettiği Bölümün Adı. Bu verilerle Uygulama 1’de olduğu gibi text dosyamızı hazırlıyoruz. Sonra yine aynı şekilde Memnuniyet Analizi Yazılımını açıp Data Yükle’ye tıklayarak verilerimizi yüklüyoruz. Daha sonra test etmek için 2013 ve 2014 yılları arasındaki Şikayet kayıtlarından rastgele bir cümle giriyoruz.

Girmiş olduğumuz Cümle: “MÜŞTERİMİZ ARACINI SORUN OLMASI NEDENİ İLE SERVİSİMİZE BIRAKTIĞINI ANCAK ARAÇ SERVİSTEN ÇIKTIKTAN SONRA ARAÇTA DAHA ÇOK PROBLEM ÇIKTIĞINI ARACINDA İSTENİLEN ARIZALARIN GİDERİLMEDİĞİNİ BUNDAN RAHATSIZ OLDUĞU İÇİN ŞİKAYETİNİ İLETTİ. KONU İLE İLGİLİ YARDIMLARINIZI RİCA EDİYORUM”.

Memnuniyet Analizi

DATA ÇIKIŞ

Cümleyi Giriniz

MÜŞTERİMİZ ARACINI SORUN OLMASI NEDENİ İLE SERVİSİMİZE BIRAKTIĞINI ANCAK ARAÇ SERVİSTEN ÇIKTIKTAN SONRA ARAÇTA DAHA ÇOK PROBLEM ÇIKTIĞINI ARACINDA İSTENİLEN ARIZALARIN GİDERİLMEDİĞİNİ BUNDAN RAHATSIZ OLDUĞU İÇİN ŞİKAYETİNİ İLETTİ. KONU İLE İLGİLİ YARDIMLARINIZI RİCA EDİYORUM

HESAPLA

TEMİZLE

Veri Tabanındaki verilerin Mutlu Olasılığı: 0,506292009905494
Veri Tabanındaki verilerin Mutsuz Olasılığı: 0,493707990094507

MÜŞTERİMİZ
Mutlu Olasılığı =0,0241876994487961
Mutsuz Olasılığı =0,0247844827586207

ARACINI
Mutlu Olasılığı =0,0479877923793473
Mutsuz Olasılığı =0,0669244720236222

Şekil 4.4. Memnuniyet Analizi Yazılımının Uygulama Sonuçları

Veri tabanında yer tüm cümlelerin Mutsuz olasılığının biraz daha fazla olduğunu görüyoruz.

Girmiş olduğumuz cümlenin Mutlu/Mutsuz olasılığı ise aşağıdaki gibidir

Cümlenin Mutlu Olasılığı: 1,01009983579117E-35 Cumlenin Mutsuz Olasılığı: 2,71783329050173E-33

Şekil 4.5. Uygulama 2 Sonuçları

Cümlenin Mutsuz olasılığı daha çok çıkmıştır ve gerçek kayıt da kapanma şeklinde baktığımızda Mutsuz şekilde çıkması Yazılım aracının doğruluğunu teyit etmiştir.

4.4. Uygulama 2 Sonuçları

Uygulama 2 içerisinde yeni alanlar eklenerek sistem eğitilmiştir. Bu şekilde veri tabanında sistem tarafından bilinen kelime sayısı da yükselmiştir. Uygulama 1 içerisinde test edilen 1000 tane mutlu ve 1000 tane mutsuz kaydın Uygulama 2 içerisinde yeni alanlar eklenen veri tabanı üzerinde test edilmiştir. Sisteme girilen test kayıtların Karışıklık Matrisi Çizelge 4.3'te görüldüğü gibidir.

Çizelge 4.3. Uygulama 2 Sonuçları

		Tahmin Edilen Sınıf	
		Mutlu	Mutsuz
Doğru Sınıf	Mutlu	896	104
	Mutsuz	188	812

Karıřıklık Matrisine göre Duyarlılık ve Kesinlik deęerlerinin hesaplaması yapılması gerekir.

$$\text{Duyarlılık} = \frac{TP}{TP+FN} = \frac{896}{896+104} = 0,896$$

$$\text{Kesinlik} = \frac{TP}{TP+FP} = \frac{896}{896+188} = 0,826$$

$$\text{F-Measure} = \frac{2 \times \text{Duyarlılık} \times \text{Kesinlik}}{\text{Duyarlılık} + \text{Kesinlik}} = \frac{2 \times 0,896 \times 0,826}{0,896 + 0,826} = 0,8595 \text{ olarak bulunur}$$

5. SONUÇ

Bu çalışmada, bir şirketin müşteri hizmetlerine gelen ve mutlu veya mutsuz şekilde kapanan şikayet kayıtları üzerinde çalışılmıştır. Alınan bu şikayet kayıtları ile büyük bir veri tabanı oluşturulmuştur. Oluşturulan veri tabanında şikayet kayıtları ve diğer öz nitelikler normalizasyon işleminden geçirilmiştir. Sistemi eğitmek için 2003 yılı ile 2012 yılları arasındaki şikayet kayıtları ele alınmıştır. Sonrasında ise testler 2013 ve 2014 verileri üzerinde yapılmıştır.

Sisteme eğitmek için ilk etap da sadece şikayet kayıtlarının bulunduğu veriler programa yüklenmiştir ve bu şekilde test edilmiştir. Daha sonra ise şikayet kayıtlarına ek olarak yeni öz nitelikler eklenerek test edilmiştir. Daha fazla kayıt girildikçe programın tahmin etme olasılığı da yükselmiştir.

Müşteri Şikayeti alanında yapılan bu çalışmada, veri madenciliği işleminde başarıya ulaşabilmek için, bol miktarda kaliteli veriye ihtiyaç olduğu saptanmıştır. Eksik veri olması durumunda, sağlıklı sonuçlar çıkartılamamıştır. Metin madenciliği bu bağlamda bir adım değil süreçtir.

6. KAYNAKLAR

- [1] Özkan, Yalçın. "Veri Madenciliği Yöntemleri". İstanbul: Papatya Yayıncılık, 2008.
- [2] Alpaydın, Ethem. "Zeki Veri Madenciliği: Ham Veriden Altın Bilgiye Ulaşma Yöntemleri". Bilişim 2000 Eğitim Semineri, İstanbul: Boğaziçi Üniversitesi, 2000.
- [3] Coşlu, Eda. "Veri Madenciliği". Akademik Bilişim 2013-XV. Akademik Bilişim Konferansı Bildirileri, Antalya: Akdeniz Üniversitesi, 2013
- [4] Küçüksille, Engin. "Veri Madenciliği Süreci Kullanılarak Portföy Performansının Değerlendirilmesi ve İmkb Hisse Senetleri Piyasasında Bir Uygulama". Yayınlanmış Doktora Tezi. Isparta: Süleyman Demirel Üniversitesi, 2009
- [5] Özgürbaş, Nermin. Koyuncugil, A.Serhan. "İmkb'de İşlem Gören Kobi'lerin Veri Madenciliği Karar Ağaçları Algoritmalarından Chaid ile Profillendirilmesi ve Küreselleştirme Sürecinde Güçlü ve Zayıf Yönlerin Belirlenmesi". Ankara, 2008
- [6] Thearling, Kurt. "An Introduction to Data Mining", 2008
- [7] Thearling, Kurt. "The Art and Science of Customer Relationship Management". Yayınlanmış Yüksek Lisans Tezi. University of Michigan, 2009
- [8] Yılmaz, Emrah. "Kütahya İlinde Sosyal Sınıfların Belirlenmesi ve Veri Madenciliği İle Tüketici Profiline Çıkarılmasına Yönelik Bir Uygulama". Yayınlanmış Yüksek Lisans Tezi. Kütahya: Dumlupınar Üniversitesi, 2006, s91
- [9] Silahtaroglu, Gökhan. "Kavram ve Algoritmalarıyla Temel Veri Madenciliği". Papatya Yayıncılık, İstanbul, 2008.

- [10] Argüden, Yılmaz. " Veriden Bilgiye, Masraftan Değere". İstanbul, Arge Danışmanlık, 2008
- [11] Arslan, Yusuf. "Veri Madenciliği". Çözümпарк Web Sitesi, 2011
- [12] Kaya, Ersin. Bulun, Mustafa. Arslan, Ahmet. "Tıpta Veri Ambarları Oluşturma ve Veri Madenciliği Uygulamaları". Konya: Selçuklu Üniversitesi, 2010
- [13] Puthier, Denis. "Data Mining And Integration of Genomic Data For Gene regulatory Network Analysis". 2008
- [14] Özekeş, Serhat. "Veri Madenciliği Modelleri ve Uygulama Alanları". Yayınlanmış Yüksek Lisans tezi. İstanbul: İstanbul Ticaret Üniversitesi, 2010, s.76.
- [15] Keziban, Seçkin. "Metin madenciliğinde kullanılan yöntemlerin karşılaştırılması: Siyasi parti liderlerinin grup genel toplantı konuşmaları ile bir uygulama". Yayınlanmış Yüksek Lisans Tezi. Sakarya: Sakarya Üniversitesi, 2011
- [16] Döven, Selçuk. "Metin madenciliği ile dokümanlar arasındaki benzerliklerin bulunması". Yayınlanmış Yüksek Lisans Tezi. İstanbul: Bahçeşehir Üniversitesi, 2013
- [17] Francis, Louise. "Introduction to Text Mining". 2008
- [18] Fan, Weiguo. Wallace, Linda. Rich, Stephanie. Zhang, Zhongju. "Tapping the Power of Text Mining". Virginia Polytechnic Institute and State University, Blacksburg, 2006
- [19] Çelikyay, E.Kübra. "Metin madenciliği yöntemiyle Türkçe'de en sık kullanılan ve birbirini takip eden harflerin analizi ve birliktelik kuralları". İstanbul: Beykent Üniversitesi, 2010

- [20] Pilavcılar, İ.Ferhat. "Metin Madenciliği ile Metin Sınıflandırma". İstanbul: Yıldız Teknik Üniversitesi, 2007, s.14.
- [21] Koyuncugil, A.Serhan. "Bulanık Veri Madenciliği ve Sermaye Piyasalarına Uygulanması". Yayınlanmış Doktora Tezi. Ankara: Ankara Üniversitesi, 2006
- [22] Şeker, Ş.Evren. "K-Ortalama Algoritması (K-Means Algorithm)". Bilgisayar Kavramları Web Sitesi, 2008
- [23] Korkem, EBRU. "Mikroarray Gen Ekspresyon Veri Setlerinde Random Forest ve Naive Bayes Sınıflama Yöntemleri Yaklaşımı". Yayınlanmış Yüksek Lisans Tezi. Ankara: Hacettepe Üniversitesi, 2013
- [24] Karpat, Işıl. "Müşteri Tatmininin Sağlanması". Pazarlama Dünyası Dergisi, İstanbul, 1998, Sayı: 71
- [25] Acuner, Taner. Keskin, H.Dilara. "Toplam Kalite Yönetiminde Pazarlamanın Yeri". Pazarlama Dünyası Dergisi, İstanbul, 2000
- [26] İlhan, Sevinç. Duru, Nevcihan. Karagöz, Şenol. Sağır, Merve. "Metin Madenciliği ile Soru Cevaplama Sistemi". Kocaeli: Kocaeli Üniversitesi, 2011
- [27] Şeker, Ş.Evren. "İş Zekası ve Veri Madenciliği". İstanbul, Cinus Yayınevi, 2013
- [28] Hand, D. Mannila, H. Smyth, F. "Principles of Data Mining". England: Cambridge Ma and London. The MIT Press: A. Bradford Book, 2011
- [29] Oğuzlar, Ayşe. "Temel Metin Madenciliği". Bursa: Dora Yayıncılık Ltd. Şti. 2011
- [30] Dolgun, M. Özgür. "Büyük alışveriş merkezleri için veri madenciliği uygulamaları". Yayınlanmış Yüksek Lisans Tezi. Ankara: Hacettepe Üniversitesi, 2006

[31] Akçay, M. Aksel. "Dünya Değerler Anket Verilerinin Veri Madenciliği Yöntemiyle İncelenmesi". Yayınlanmış Yüksek Lisans Tezi. İstanbul: Aydın Üniversitesi, 2013

[32] Kalıyeva, Samal. "Bilimsel Makalelerin Metin İşleme Yöntemleri İle Sınıflandırılması". Yayınlanmış Yüksek Lisans Tezi. Ankara: Gazi Üniversitesi, 2013

[33] Coşkun, Cengiz. Baykal, Abdullah. "Veri Madenciliğinde Algoritmaların Bir Örnek Üzerinde Karşılaştırılması". Diyarbakır: Dicle Üniversitesi, 2013

7. ÖZGEÇMİŞ

Kemal Kuzucu, 1987 yılında İstanbul'da doğdu. İlk ve orta öğrenimini Pendik'te tamamladı. 2001 yılında Tuzla Anadolu Teknik Lisesinde Bilgisayar(Donanım) bölümüne başladı ve 2005 yılında mezun oldu. Ön Lisans eğitimini 2005-2007 yılları arasında İstanbul Üniversitesi Bilgisayar Teknolojisi ve Programlama bölümünde tamamladı. 2007-2010 yılları arasında Anadolu Üniversitesi İşletme bölümünde Lisans eğitimini tamamladı. Şubat 2012 tarihi itibariyle Maltepe Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği bölümünde Yüksel Lisans eğitime başlamıştır. 2007 tarihi itibariyle Bilgi teknolojileri bölümünde Sistem tarafında çalışmakta olan Kemal Kuzucu, son olarak bir otomotiv şirketinde Sistem Yöneticisi pozisyonunda çalışmaktadır. Kemal Kuzucu, iyi derecede İngilizce bilmektedir.