Veri Madenciliği Yaklaşımı ile Sosyal Ağ Analizi

Bu tez Bilgi Güvenliği Mühendisliği'nde
Tezli Yüksek Lisans Programının bir koşulu olarak

Emine Büşra AKDEMİR
tarafından

Fen Bilimleri Enstitüsü'ne
sunulmuştur.

İSTANBUL
ŞEHİR
ÜNİVERSİTESİ

Bu tezi okuduk, kapsam ve nitelik açısından Bilgi Güvenliği Mühendisliği alanında Yüksek Lisans derecesi için tümüyle uygun olduğu görüşüne vardık.

ONAYLAYANLAR:

Prof. Dr. Ensar Gül .....
(Tez Danışmanı)

Dr. Ahmet Fatih Mustaçoğlu .....
(Tez Eş-danışmanı)

Doç. Dr. Murat Can Ganiz .....

Dr. Öğr. Üyesi Mehmet Yasin Ulukuş .....

Bu tez İstanbul Şehir Üniversitesi, Fen Bilimleri Enstitüsü tarafından belirlenen tüm koşullara uygundur.

ONAY TARİHİ: 04.2019

MÜHÜR/İMZA:

Yazarlık Beyanı

Ben, Emine Büşra AKDEMİR, başlığı, 'Veri Madenciliği Yaklaşımı ile Sosyal Ağ Analizi' olan tezin ve içinde sunulan bilgilerin şahsıma ait olduğunu beyan ederim. Ayrıca:

- Bu çalışmanın bütünü veya esası bu üniversitede Yüksek Lisans derecesi elde etmek üzere çalıştığım süre içinde gerçekleştirilmiştir.
- Daha önce bu tezin herhangi bir kısmı başka bir derece veya yeterlik almak üzere bu üniversiteye veya başka bir kuruma sunulduysa bu açık biçimde ifade edilmiştir.
- Başkalarının yayımlanmış çalışmalarına başvurduğum durumlarda bu çalışmalara açık biçimde atıfta bulundum.
- Başkalarının çalışmalarından alıntıladığımda kaynağı her zaman belirttim. Tezin bu alıntılar dışında kalan kısmı tümüyle benim kendi çalışmamdır.
- Esaslı yardım aldığım bütün kaynaklara teşekkür ettim.
- Tezde başkalarıyla birlikte gerçekleştirilen çalışmalar varsa onların katkısını ve kendi yaptıklarımı tam olarak açıkladım.

İmza: 

Tarih: 01. 06. 2019

Veri Madenciliği Yaklaşımı ile Sosyal Ağ Analizi

Emine Büşra AKDEMİR

Öz

Günümüzde internet kullanımının yaygınlaşmasıyla geliştirilen uygulamalar hem iletişim hem de eğlence amaçlı olarak ortaya çıkmıştır. Sosyal ağlar olarak adlandırılan bu uygulamalar kişiler, toplumlar hakkında büyük miktarda veriye internet üzerinden kolay şekilde erişim imkanı sunmaktadır. Sosyal ağlar üzerinde yapılan veri madenciliği çalışmaları ise son yıllarda bu alandaki gelişmeler ile artış göstermiştir. Pek çok araştırmanın konusu olarak geniş kitleler hakkında yararlı bilgiler elde edilmeye çalışılmıştır. Bu tez çalışmasında sosyal ağlarda yapılan veri madenciliği çalışmaları ve problemleri araştırılmıştır. Twitter uygulaması üzerinden verilere erişilerek Türkçe Tweetlerin duygu analizi yapılmıştır. Duygu sınıflandırma işlemi için Naive Bayes, Destek Vektör Makineleri ve K en yakın komşu algoritmaları kullanılmıştır. Twitter kullanıcılarının belirlenen sektördeki kurumsal şirketler ile ilgili tweetleri duygu polaritesi açısından incelenerek sosyal ağlar üzerinde kurumsal itibarı en yüksek kuruluş tespit edilmeye çalışılmıştır.

Anahtar Sözcükler: Sosyal Ağlar, Veri Madenciliği, Duygu Analizi, Sınıflandırma Algoritmaları

Aileme,

Teşekkür

Tez çalışmam süresince her türlü yardım ve fedakarlığı sağlayan danışmanlarım Sayın Prof. Dr. Ensar Gül ve Sayın Dr. Ahmet Fatih Mustaçoğlu'na,

Tüm eğitim hayatım boyunca maddi, manevi desteklerini hissettiğim, daima arkamda duran biricik aileme ve tezimin hazırlanması sırasında beni cesaretlendiren, ümit veren ve her zaman yanımda olduğunu hissettiren sevgili eşime,

Ve son olarak bünyesinde çalışmaktan gurur duyduğum; hem çalışma hem de akademik hayatıma yön veren kurumum TÜBİTAK BİLGEM ve değerli yöneticilerime teşekkürü borç bilirim.

İçindekiler

Şekil Listesi

Tablo Listesi

Kısaltmalar

API	**A**pplication **P**rogramming **I**nterface
CRISP-DM	**CR**oss **I**ndustry **S**tandard **P**rocess Model for **D**ata **M**ining
ENIAC	**E**lectrical **N**umerical **I**ntegrator **A**nd **C**alculator
KDD	The **K**nowledge **D**iscovery in **D**atabases
KNN	**K** **N**earest **N**eighbor
MB	**M**arket **B**asket
NPV	**N**et **P**resent **V**alue
LTV	**L**ife**T**ime **V**alue
PE	**P**rocessing **E**lement
RFM	**R**ecency, **F**requency, **M**onetary
SPSS	**S**tatistical **P**ackage for the **S**ocial **S**ciences
SVM	**S**upport **V**ector **M**achine

Bölüm 1

Giriş

İnsanlık var olduğundan günümüze kadar geçen süre içinde iletişim hep bir ihtiyaç olmuştur. Son yıllarda internetin yaygınlaşmasıyla bu ihtiyaç uygulamalar aracılığıyla giderilmeye başlanmıştır. Geliştirilen uygulamalar sadece iletişim değil aynı zamanda insanlara eğlence imkanı da sunmaktadır.

Genel itibariyle "sosyal ağlar" olarak adlandırılan bu uygulamalar kişiler, toplumlar hakkında büyük miktarda veriye internet üzerinden kolay şekilde erişim imkanı sunmaktadır. Sosyal ağ kavramının, bu denli hızlı gelişen teknolojiler ile birlikte değerlendirildiğinde günlük hayatta ne kadar önemli bir yere sahip olduğunu anlamak kaçınılmazdır. Bu sebeple geliştirilen uygulamalar yanında daha pek çok uygulamaya açık olduğu aşikardır.

Bu tez çalışmasında sosyal ağlar üzerinde yapılan veri madenciliği çalışmaları incelenmiş ve Türkiye'de yapılan çalışmalar araştırılmıştır. Sosyal ağların insanlar üzerindeki etkisinin artmasıyla kullanıcılarının sosyal ağlarda bıraktıkları izlerin değerlendirilmesi işlemi ön plana çıkmıştır. Sosyal ağ uygulamalarının başında gelen Twitter, kullanıcıların yorumlarını içerdiğinden bu alanda yoğunlukla araştırmalara konu olmuştur. Herhangi bir konu kısıtlaması olmaması bu alandaki veri miktarını arttırmaktadır. Yorumların değerlendirilmesi ile belli olayların insanlar üzerindeki etkisinin araştırılmasına ışık tutabileceği gibi kurumsal kullanıcıların da yorumları değerlendirerek faaliyetlerine yön vermesi konusunda olanak sağlamaktadır.

Tez içerik olarak 1. Bölüm'de yapılan Giriş Bölümünün ardından, 2. Bölümde Veri Madenciliğinin tanımı, gelişim süreci ve genel bilgiler ile devam etmektedir. 3. Bölümde

Veri Madenciliği fazları anlatılırken, 4. Bölümde Sınıflandırma, Kümeleme, Birliktelik Kuralları gibi veri madenciliği işlevlerinden bahsedilmiştir. 5. Bölümde veri madenciliği algoritmaları ve 6. Bölümde uygulama alanları açıklandıktan sonra 7. Bölümde Sosyal Ağlar kısmına geçilmiştir. Genel bilgilerin verilmesinin ardından 8. Bölümde Sosyal Ağlar üzerinde yapılan veri madenciliği yani web madenciliği uygulamaları anlatılmıştır ve literatürde yapılan çalışmalar İlgili Çalışmalar başlığı altında verilmiştir. 9. Bölümde yani Uygulama kısmında Savunma Sanayi alanında faaliyet gösteren kuruluşların sosyal ağlardaki kurum itibarını ölçmek adına kuruluş isimlerini içeren tweet mesajlarının duygusal polariteleri incelenmiştir. 10. Bölümde elde edilen sonuçlar paylaşılmıştır.

Bölüm 2

Veri Madenciliği

2.1 Veri Madenciliği Nedir?

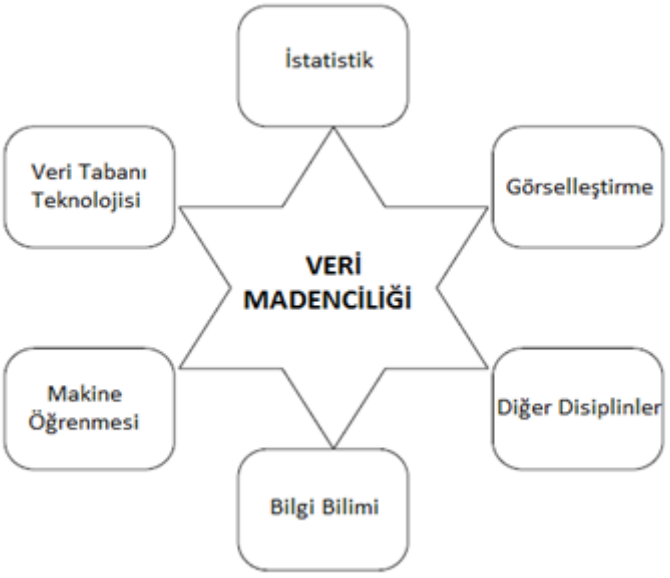
Veri Madenciliği; büyük veriden anlamlı ve yararlı bilgilerin elde edilmesi işlemidir. Başka bir ifade ile seçilen veri setinin analize uygun şekilde hazırlanması ile anlamlı çıkabilecek bilgiye ulaşma sürecidir.

Günümüzde veri miktarının artması, veriyi toplama ve saklama kapasitesindeki hızlı büyüme, insanlığı yeni arayışlara yönlendirmiştir. Bir bilgisayarın işleyebileceği veri miktarından çok daha fazlası üretilmektedir. Verilerin hızlı bir şekilde artması sonucu etkin bir veri tabanı analizi için yeni tekniklere ihtiyaç doğmuştur. Geleneksel sorgu veya raporlama araçları çok miktardaki veriler karşısında yetersiz kaldığından veri madenciliği kavramı ortaya atılmıştır [1].

Literatürde veri madenciliği ile benzer anlamları ifade eden başka terimler de bulunmaktadır [2].

- Veri Tabanlarında Bilgi Keşfi
- Bilgi Çıkarma
- Veri/Desen Analizi
- Veri Arkeolojisi
- Veri Tarama

Veri madenciliği ve veri tabanlarında bilgi keşfi kavramları birçok kaynakta birbirinin yerine kullanılmaktadır. Veri madenciliği, veri tabanlarında bilgi keşfi sürecinde bir adım olarak yer almasına rağmen birçok çalışmada tüm süreci anlatmak için kullanılmaktadır. Veri tabanlarında bilgi keşfi (KDD) daha az kullanıma sahip olduğu halde veri madenciliğinden daha açık ve daha bilgilendiricidir.



Şekil 2.1: Veri Madenciliği: Çoklu Disiplinlerin Birleşimi

Şekil 2.1'de [3] görüldüğü gibi Veri Madenciliği pek çok disiplinin birleşiminden ortaya çıktığından çok geniş kullanım alanına sahiptir.

2.2 Veri Madenciliği Gelişim Süreci

Veri madenciliğinin tarihi ilk sayısal bilgisayar olan ENIAC'a kadar dayanmaktadır. Bilgisayarların verimli bir şekilde kullanımı veri depolanması ile başlamaktadır. Bilgisayarlar ilk olarak karmaşık hesaplamaları yapmak amacıyla geliştirilmiştir ancak zamanla gelişen ihtiyaçlar doğrultusunda veri depolama işlemleri için de kullanılmaya başlanmıştır. Bunun sonucunda veri tabanları ortaya çıkmıştır. Veri tabanlarının genişlemesi sonrası donanımsal olarak bu verilerin tutulacakları depo ihtiyacı ortaya çıkmış ve bu depoların da zamanla genişlemesi gerekmiştir. Bunun sonucunda ise veri ambarı kavramı ortaya çıkmıştır. Verilerin uzun süre saklanma ihtiyacı nedeniyle fiziksel sürücülerden yararlanılmıştır. Bu süreçle birlikte veri modelleme kavramı ortaya çıkmıştır [4].

İlk basit veri modelleri olarak geliştirilen modeller Hiyerarşik ve Şebeke modelleri olmuştur [4].Hiyerarşik veri modelleri, ağaç yapısına benzeyen, temelinde bir kök bulunduran ve bu kök aracılığıyla üstünde her daim bir, altında ise n sayıda düğümün bulunduğu veri modelleri olarak tanımlanmıştır.

Şebeke veri modelleri ise kayıt tipi ve bağlantıların olduğu, kayıt tiplerinin varlık, bağlantılarınsa ilişki tiplerini belirlediği bir veri modeli olarak tanımlanmıştır. Şebeke veri modelinde herhangi bir varlık bir diğeri ile ilişki içerisine girebiliyordu. Ancak çoklu ilişki kurmak mümkün değildi. Hiyerarşik veri modellerinde ise bu durum daha kısıtlıydı. Bu nedenle kullanıcı ihtiyaçlarını tam olarak karşılanamamıştır. Bu ihtiyaçlar doğrultusunda Geliştirilmiş Veri Modelleri ortaya çıkmıştır. Bunlar Varlık-İlişki, İlişkisel ve Nesne-Yönelimli veri modelleri olarak bilinmektedir [4].

Veri madenciliği, kavramsal olarak 1960'lı yıllarda, bilgisayarların veri analiz problemlerini çözmek amaçlı kullanılmaya başlamasıyla ortaya çıkmıştır. O dönemlerde kullanılan bilgisayarlar ile yeterince uzun taramalar yapıldığında hedeflenen verilere ulaşmanın mümkün olduğu anlaşılmıştır. O dönemlerde bu işleme veri madenciliği yerine veri taraması, veri yakalanması gibi isimler verilmiştir [4].

1990'lı yıllara gelindiğinde veri madenciliği kavramı, bilgisayar mühendisleri tarafından ortaya atılmıştır. Bu kavramın ortaya atılmasındaki amaç, geleneksel istatistiksel yöntemler yerine, veri analizinin algoritmik bilgisayar modülleri tarafından değerlendirmesini vurgulamaktır. Bu aşamadan sonra bilim insanları veri madenciliğine pek çok yaklaşım getirmişlerdir. Bu yaklaşımların kökeninde istatistik, makine öğrenmesi, veri tabanları, otomasyon, pazarlama, araştırma gibi disiplin ve kavramlar yer almıştır.

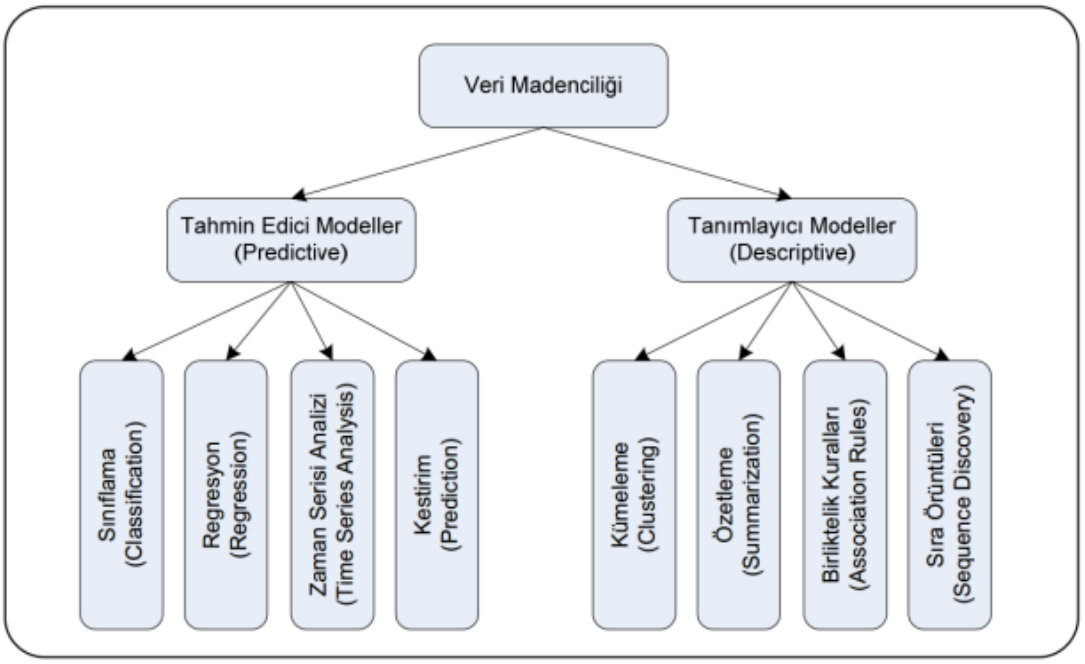
İstatistik verilerin değerlendirilmesi ve analizini sağlayan yöntemler topluluğudur. Bilgisayarların veri analizi için kullanımıyla istatistiksel çalışmalar hız kazanmıştır. Bilgisayarların gelişmesi ile birlikte daha önce yapılması mümkün olmayan istatistiksel araştırmalar yapılabilmiştir. 1990'lardan sonra istatistik, veri madenciliği ile ortak bir platformda düşünülmeye başlanmıştır. Bilginin, çok miktardaki veri yığınları içinden çıkarılması ve analizinin yapılarak kullanıma hazırlanması sürecinde veri madenciliği ve istatistik ortak olarak kullanılmıştır. Aynı zamanda veri madenciliği, veri tabanları ve makine öğrenimi disipliniyle birlikte gelişmiştir [4].

Genel anlamda değerlendirildiğinde veri madenciliğinin üç farklı disiplinden beslendiği ortaya çıkmaktadır;

- İstatistik: Veriler arasındaki sayısal ilişkilerin ortaya konması,
- Yapay Zeka: Yazılım veya makineler aracılığı ile insan benzeri teknoloji üretimi,
- Makine Öğrenmesi: Verilerden elde edilen bilgiler ile tahminler çıkartmaya yarayan algoritmalar.

2.3 Veri Madenciliği Modelleri

Tanımlayıcı ve tahmin edici model olmak üzere 2 tür model söz konusudur [5].



Şekil 2.2: Veri Madenciliği Modelleri ve Gorevleri

2.3.1 Tanımlayıcı Model

Tanımlayıcı modeller analiz yapan kişiye daha önceden bir bilgi ve hipoteze sahip olmadan, veri kümesinin içinde ne tür ilişkiler olduğunu anlama imkanı sunar. Analizi yapan kişinin çok büyük veri tabanlarındaki bilgileri incelemek, örüntüleri keşfetmek için doğru soruları sorup hipotezler geliştirmesi pratikte pek mümkün olmadığından, ilginç

örüntüleri keşfetme önceliği veri madenciliği programına bırakılır. Keşfedilen bilginin kalitesi ve zenginliği, uygulamanın kullanışlılığını ve gücünü gösterir [1].

Tanımlayıcı Model'de kullanılan yöntemler aşağıda verilmiştir, bu yöntemler ile ilgili açıklamalar Bölüm 4'te anlatılacaktır.

- Kümeleme
- Özetleme
- Birliktelik Kuralları
- Sıra Örüntüleri

2.3.2 Tahmin Edici Model

Tahmin edici modellerde, eldeki verilerden hareket edilerek bir model geliştirilmesi ve sonrasında kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuçların tahmin edilmesi yönünde çalışmalar yapılmaktadır. Örneğin bir sınıftaki öğrencilerin bir dersle ilgili almış oldukları vize ve ödev notları gibi veriler bir veri tabanında toplanır. Bu verilere uygun olarak bir model kurulur. Kurulan model sonucunda öğrencilerin o dersin final sınavından alacağı notun tahmininde kullanılır [1].

Tahmin Edici Model'de kullanılan yöntemler aşağıda verilmiştir, bu yöntemler ile ilgili açıklamalar Bölüm 4'te anlatılacaktır.

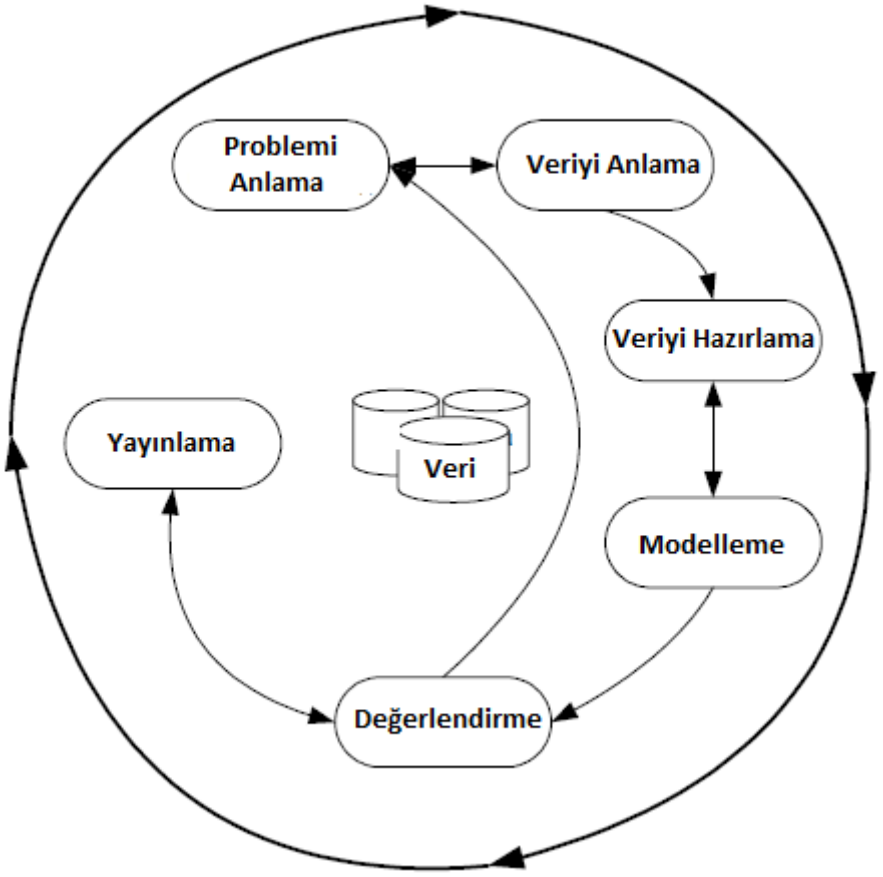
- Sınıflandırma
- Regresyon
- Zaman Serisi Analizi
- Kestirim

Bölüm 3

Veri Madenciliği Aşamaları

Veri madenciliği birçok farklı adımdan oluşur. Hemen hemen her veri madenciliği çalışmasında verilerin kaynaklardan elde edilmesi ve entegrasyonu, verilerin temizlenmesi, modelin oluşturulması, modelin değerlemesi ve sonuçların sunuma hazırlanması adımları karşımıza çıkar. Ancak veri madenciliğinin çoklu disipliner yapısı ve farklı uygulama alanlarındaki görev ve prosedürlerin çeşitliliği, endüstri ve yazılım standardı metodolojisinin oluşturulması yolunda sorunlara yol açmaktadır. Standart bir uygulama metodolojisi, veri madenciliği teknolojinin uygulanmasını daha az maliyetli, daha güvenilir, daha kolay yönetilebilir ve daha hızlı hale getirecek ve geliştiricilerin veri madenciliği çözümünü genel metodolojiye uyacak şekilde entegre etmelerini sağlayacaktır.

Veri madenciliği için CRISP-DM (Cross Industry Standard Process Model for Data Mining) referans modeli, bir veri madenciliği projesinin yaşam döngüsüne genel bir bakış sağlar. Bir projenin aşamalarını, ilgili görevlerini ve çıktılarını içerir. CRISP-DM veri madenciliği süreçlerini, kullanılan yazılımdan ve endüstriden bağımsız standartlaştırmayı amaçlar. CRISP-DM asıl üyeleri Daimler-Benz, SPSS ve NCR olan bir konsorsiyum tarafından geliştirilmiştir.



Şekil 3.1: CRISP-DM:Referans Modelinin Fazları

Veri madenciliği projesinin yaşam döngüsü, Sekil 3.1 'de gösterilen altı aşamaya ayrıştırılmıştır:

- Problemi/İşi/Sorunsalı Kavrama
- Veri/Veri Setlerini Anlama
- Veriyi Hazırlama/Ön İşleme
- Modelleme
- Değerleme
- Yayınlama/Canlıya Alma

Aşamalar arası geçişler çok katı olmamakla birlikte, oklar sadece aşamaları arasındaki en önemli ve sıkı bağları göstermektedir.

3.1 Problemi Anlama

Problemi Anlama veri madenciliği sürecinde başlangıç aşamasıdır. Proje hedeflerini ve gereksinimlerini iş perspektifinden anlamak ve daha sonra bu bilgiyi bir veri madenciliği problemi tanımına dönüştürmek ve hedeflere ulaşmak için tasarlanan bir ön proje planına odaklanır.

Herhangi bir veri madenciliği projesindeki ilk ve en önemli adım, açık ve ulaşılabilir bir hedef oluşturmak ve bu hedefe ulaşmak için bir süreç geliştirmektir. Hedefi tanımlarken, ilk olarak neyin ölçüleceği yada tahmin edileceğine karar verilmesi gerekir. Veri Madenciliği modelleri genellikle tahmin edici ve tanımlayıcı olmak üzere iki kategoriye ayrılır. Tahmin modelleri, gelecekteki aktiviteyi temsil eden bir değeri hesaplar. Tanımlayıcı bir model, göründüğü gibidir: Konuları tanımlayıcı kategorilere ayırmak için kullanılan kurallar oluşturur. Pazarlama, risk ve müşteri ilişkileri yönetiminde bugün kullanılan yaygın analitik hedeflerin bazıları aşağıda sıralanmıştır.

3.1.1 Profil Analizi

Günümüz dünyasında müşteriler ve beklentileri hakkında derinlemesine bilgi sahibi olmak, rekabetçi kalmak için gereklidir. Profil analizi, müşterileri veya potansiyel müşterileri tanımak için mükemmel bir yoldur. Bir ilgi popülasyonu ortak özellikleri ölçmeyi içerir. Ortalama yaş, cinsiyet, medeni durum ve ortalama yaşam süresi gibi demografik özellikler tipik olarak bir profil analizine dahil edilir. Diğer özellikler, müşteri ilişkileri yaşı veya ortalama risk seviyesi gibi daha spesifik özellikler de olabilir. Profiller, ilgili popülasyonun segmentleri içinde kullanıldığında en faydalıdır [2].

3.1.2 Segmentasyon

Hedefleme modelleri, pazarlama ve/veya riske dayalı eylemlerin verimliliğini artırmak için tasarlanmıştır. Ancak, hedefleme modelleri geliştirilmeden önce mevcut müşteri tabanınızı iyi anlamak önemlidir. Profil analizi, müşterileriniz hakkında bilgi edinmek için etkili bir tekniktir.

Segmentasyon analizinin yaygın kullanımı, müşterileri karlılığa ve pazar potansiyeline göre segmentlere ayırmaktır. Örneğin, bir perakende işletmesi, müşteri tabanını, tüm

perakende mağazalarındaki toplam satın alma davranışları ile ilgili olarak satın alma davranışlarını tanımlayan segmentlere ayırır. Bu yolla bir perakendeci hangi müşterilerin en fazla potansiyele sahip olduğunu değerlendirebilir.

Bir kredi veya kredi kartı portföyünde yapılan bir profil analizi, iki boyutlu bir risk ve denge matrisine bölünebilir. Bu, olası pazarlama ve/veya risk eylemleri için müşteri veri tabanının farklı bölümlerini değerlendirmek için görsel bir araç sağlayacaktır. Örneğin, bir segmentin yüksek bakiyeleri ve yüksek riski varsa Yıllık Yüzde Oranını (APR) artırmak isteyebilirsiniz. Düşük riskli segmentler için, düşük riskli müşterilerin dengelerini alma veya çekme umuduyla APR'yi düşürmek isteyebilirsiniz.

3.1.3 Yanıt Modeli

Bir yanıt modeli genellikle bir şirketin geliştirmeyi amaçladığı ilk tür hedefleme modelidir. Geçmişte bir hedefleme yapılmadıysa, bir yanıt modeli, yanıtları artırarak ve/veya posta harcamalarını azaltarak bir pazarlama kampanyasının verimliliğine büyük bir destek sağlayabilir. Amaç, bir ürün veya hizmet için bir teklife kimin yanıt vereceğini tahmin etmektir. Benzer bir popülasyonun geçmiş davranışlarına veya bazı mantıksal ikamelerine dayanabilir.

Teklif kanalına bağlı olarak bir yanıt çeşitli yollarla alınabilir. Bir posta teklifi, yanıtlayıcıyı posta, telefon veya internet ile cevap vermesi için yönlendirebilir. Sonuçları derlerken, yanıt kanalını izlemek ve kopyaları yönetmek önemlidir. Bir yanıtlayıcının bir yanıt göndermesi ve birkaç gün sonra telefon ya da internet ile cevap vermesi olağandışı bir durum değildir. Bir şirketin aynı kişiden birden fazla posta yanıtı alabileceği durumlar bile vardır. Bu durum, bir adayın, birkaç hafta arayla aynı ürün veya hizmetler için özellikle çok sayıda takip teklifi alması durumunda yaygındır. Model geliştirmede çoklu cevapların ele alınması için bazı kuralların oluşturulması önemlidir.

3.1.4 Risk

Onay veya risk modelleri, bir ürün veya hizmet sunarken kayıp potansiyeli olan belirli sektörlere özgüdür. En çok bilinen risk türleri bankacılık ve sigortacılık sektörlerinde meydana gelmektedir.

Bankalar kredi verdiklerinde finansal bir risk üstlenirler. Genel olarak, bu risk modelleri, bir potansiyel müşterinin ödünç alınan miktarı geri ödeyeceği veya geri ödeyemeyeceği olasılığını tahmin etmeye çalışır. İpotekler veya araba kredileri gibi birçok türde kredi temin edilir. Bu durumda banka, güvenlik için ev ya da otomobilin sahipliğini elinde bulundurur. Risk, ev veya araba kredisi eksi satış değeri ile sınırlıdır. Teminatsız krediler, bankanın güvenlik sağlamadığı kredilerdir. En yaygın teminatsız kredi türü kredi kartıdır. Her tür kredi için öngörü modelleri kullanılsa da, kredi kartları için yaygın olarak kullanılmaktadır. Bazı bankalar kendi risk modellerini geliştirmeyi tercih ederler. Diğer bankalar, risk skoru geliştirme konusunda uzmanlaşmış birçok şirketten standart veya özel risk puanları satın alırlar.

Sigorta sektörü için risk, talepte bulunan müşterinin riskidir. Temel sigorta kavramı riskleri havuzlamaktır. Sigorta şirketleri risk yönetiminde onlarca yıllık deneyime sahiptir. Hayat, otomobil, sağlık, kaza, kaza ve sorumluluk, fiyatlandırma ve rezervleri yönetmek için risk modelleri kullanan her tür sigortadır. Sigorta sektöründeki ağır devlet düzenlemesi nedeniyle, risk yönetimi, sigorta şirketlerinin karlılığını korumada kritik bir görevdir.

Diğer birçok endüstri, gelecekteki ödeme vaadi ile bir ürün veya hizmet sunarak risk almaktadır. Bu kategori, telekomünikasyon şirketlerini, enerji sağlayıcıları, perakendecileri ve daha birçoğunu içerir. Risk türü bankacılık endüstrisininkine benzerdir, çünkü bir müşterinin mal veya hizmet için ödeme yapma olasılığını yansıtır.

Dolandırıcılık riski birçok şirkete, özellikle de bankalara ve sigorta şirketlerine yönelik bir başka endişe kaynağıdır. Bir kredi kartı kaybolur veya çalınırsa, bankalar genellikle sorumluluk üstlenir ve borçlanan tutarların bir kısmını zarar olarak alırlar. Dolandırıcılık tespit modelleri, bankaların müşterilerinin tipik harcama davranışlarını öğrenerek kayıpları azaltmalarına yardımcı olmaktadır. Bir müşterinin harcama alışkanlıkları büyük ölçüde değişirse, durum değerlendirilinceye kadar onay süreci durdurulur veya izlenir.

3.1.5 Aktivasyon

Aktivasyon modelleri, bir potansiyel müşterinin tam teşekküllü bir müşteri olup olmayacağını öngören modellerdir. Bu modeller en çok finansal hizmet sektöründe uygulanabilir.

Örneğin, bir kredi kartı adayının aktif bir müşteri olması için, adayın yanıt vermesi, onaylanması ve hesabı kullanması gerekir. Müşteri hesabı hiç kullanmazsa, aslında bankayı yanıt vermeyenden daha pahalıya mal ediyor. Çoğu kredi kartı bankası, yeni müşterileri harekete geçirmeye motive etmek için düşük oranlı alımlar veya bakiye transferleri gibi teşvikler sunmaktadır. Bir sigorta olasılığı aynı şekilde görülebilir. Bir aday yanıt verebilir ve onaylanabilir, ancak ilk primi ödemezse, politika asla aktif değildir.

Aktivasyon modeli oluşturmanın iki yolu vardır. Birinci yöntem, yanıtı öngören bir model ve yanıt verilen aktivasyonu tahmin eden ikinci bir model oluşturmaktır. İlk teklifin son aktivasyon olasılığı bu iki modelin ürünüdür. İkinci yöntem, tek adımlı modellemeyi kullanmaktır. Bu yöntem, farklı fazları ayırmadan aktivasyon olasılığını tahmin eder.

3.1.6 Çapraz Satış

Çapraz satış modelleri, mevcut bir müşterinin aynı şirketten farklı bir ürün veya hizmet satın alma olasılığını veya değerini tahmin etmek için kullanılır. Satış sonrası modeller, bir müşterinin aynı ürünleri veya hizmetleri daha fazla satın alma olasılığını veya değerini tahmin eder.

Daha önce de belirtildiği gibi, mevcut müşterilere satış yapmak, yeni müşteri kazanımlarına göre çok daha kolaydır. Teklif dizilerinin test edilmesi, bir sonraki teklifi ne zaman ve nasıl yapacağınızı belirlemeye yardımcı olabilir. Bu, şirketlerin aşırı müşteriyi önlemek ve muhtemel müşterilerini yabancılaştırmaktan kaçınmak için teklifleri dikkatlice yönetmelerini sağlar.

3.1.7 Yıpranma

Yıpranma, birçok endüstride büyüyen bir sorundur. Genellikle şirketlerin daha iyi bir anlaşmadan yararlanmak için başka şirketlere geçiş yapan müşterilerini karakterize eder. Yıllar boyunca, kredi kartı bankaları düşük faiz oranlarını kullanarak rakiplerinden müşterileri çalmıştır. Telekomünikasyon şirketleri, müşterilerini rakiplerinden uzak tutmak için stratejik pazarlama taktiklerini kullanmaya devam etmiştir. Ve diğer pek çok endüstri mevcut müşterilerini tutmaya ve rakiplerinden yeni müşteri çalmaya çalışırken önemli miktarda çaba harcamaktadır.

Son birkaç yılda, yeni kredi kartı müşterileri için pazar önemli ölçüde azaldı. Bu şu anda bankalarının müşteri tabanını diğer sağlayıcılardan müşteri çekerek artırmaya çalıştığı ve zorlandığı anlamına geliyor.

3.1.8 Net Bugünkü Değer

Net Bugünkü Değer (NPV) modeli, bir ürünün önceden belirlenmiş bir süre boyunca genel karlılığını tahmin etmeye çalışır. Değer genellikle belirlenen bir yıl boyunca hesaplanır ve bugünün dolarına indirgenir. Net bugünkü değeri hesaplamak için bazı standart yöntemler olsa da, ürünler ve endüstriler arasında birçok varyasyon vardır.

3.1.9 Ömür Boyu Değer

Ömür boyu değer modeli, bir müşterinin (kişi veya iş) önceden belirlenmiş bir uzunluğun genel karlılığını tahmin etmeye çalışır. Ömür boyu değer de net bugünkü değere benzer şekilde, belirli bir yıl boyunca hesaplanır ve bugünkü dolara indirgenir. Kullanım ömrünü hesaplamak için kullanılan yöntemler, ürün ve sektörler arasında da farklılık gösterir.

Piyasalar küçüldükçe ve rekabet arttıkça, şirketler mevcut müşteri tabanlarından kar etme fırsatlarını aramaktadır. Sonuç olarak, birçok şirket mevcut müşterilerini arttırmak için ürün ve/veya hizmet tekliflerini genişletiyor. Bu yaklaşım, bir ürünün net bugünkü değerinin ötesine geçen ve bir müşterinin yaşam boyu değerini veya bir müşterinin ömür boyu değer (LTV) modelini tanımlayan bir modele olan ihtiyacı ortaya çıkarır.

3.2 Veriyi Anlama

Veriyi anlama aşaması, ilk veri toplama ile başlar ve veriyi tanımak, veri kalitesi problemlerini tanımlamak, verilere ilk bakışları keşfetmek veya gizli bilgi için hipotez oluşturmak üzere ilginç alt kümeleri tespit etmek amacıyla faaliyetlere devam eder. İş Anlama ve Veriyi Anlama arasında yakın bir bağlantı vardır. Veri madenciliği probleminin ve proje planının formüle edilmesi, mevcut verilerin en azından bir şekilde anlaşılmasını gerektirir.

3.3 Modelleme için Veri Seçme

Modelleme verileri, kaynakların sayısından üretilebilir. Bu kaynaklar iki kategoriden oluşur: iç kaynaklar veya dış kaynaklar. İç kaynaklar, müşteri kayıtları, web sitesi, posta veya aile kampanyalarından gelen posta kayıtları veya şirket verilerini barındırmak için özel olarak tasarlanmış veri tabanları ve/veya veri ambarları gibi şirket faaliyetleriyle oluşturulan alanlardır. Veriler aşağıda verilen dahili kaynaklardan seçilebilir;

- Müşteri Veritabanları: Müşteri veri tabanı tipik olarak müşteri başına bir kayıt ile tasarlanmıştır.
- İşlem veritabanı: İşlem veritabanı müşteri faaliyetlerinin kayıtlarını içerir.
- Sipariş/Teklif/Satın Alma Geçmişi Veritabanı: Teklif geçmişi veritabanı, potansiyel müşterilere, müşterilere veya her ikisine yapılan tekliflerle ilgili ayrıntıları içerir.
- Veri Ambarı: Özel bir amaç için tasarlanabilir ve geçmiş verileri içerebilir.

Dış kaynaklar ağırlıklı olarak liste satıcılarından oluşur, liste satıcıları müşteri listelerini satan şirketler ve derleyicilerdir.

3.4 Modelleme Metodolojisini Seçme

Veri Madenciliği sürecinde net bir hedefi tanımladıktan sonra bir modelleme algoritması kullanılmalıdır. Tezin "Veri Madenciliği Algoritmaları" bölümünde, ortak veri madenciliği algoritmaları ve özellikleri özetlenmiştir. Doğrusal regresyon veya lojistik regresyon gibi istatistiksel yöntemler kullanılabilir. Nöral ağlar, genetik algoritmalar, sınıflama ağaçları ve regresyon ağaçları gibi istatistiksel olmayan veya karma yöntemler de yaygın olarak kullanılan yöntemlerdir.

Bu aşamada, çeşitli modelleme teknikleri seçilir ve uygulanır ve parametreleri en uygun değerlere ayarlanır. Tipik olarak, aynı veri madenciliği sorun tipi için çeşitli teknikler vardır. Bazı teknikler belirli veri formatlarını gerektirir. Veri Hazırlama ve Modelleme arasında yakın bir bağlantı vardır. Çoğu zaman biri modelleme yaparken veri problemlerini fark eder veya yeni veriler oluşturmak için fikir alır.

3.5 Veri Hazırlama

Veri hazırlama aşaması, başlangıçtaki ham verilerden nihai veri kümesini (modelleme aracına beslenecek verileri) oluşturmak için tüm etkinlikleri kapsar. Veri hazırlama görevlerinin, herhangi bir reçeteli siparişte değil, birden çok kez gerçekleştirilmesi olasıdır. Görevler arasında tablo, kayıt ve özellik seçimi, veri temizleme, yeni özelliklerin oluşturulması ve modelleme araçları için verilerin dönüşümü yer alır.

3.5.1 Örnekleme

Bilgisayar teknolojisindeki gelişmeler örneklemenin önemini azaltmıştır. Numune alma olmadan, birçok analiz yapılabilir. Ancak daha profesyonel yazılım araçları ve bilgisayar donanımı gerektirir. Bu işlemin işlem ve zaman maliyetlerini artırır. Örnekleme işlemi hızlandırdığı ve genellikle aynı sonuçları ürettiği için örneklemeden kaçınmaya gerek duyulmamaktadır.

3.5.2 Veri Kalitesinin Sürdürülmesi

Veri Kalitesini korumak oldukça zordur ve veriler nadiren yüzde yüz "temiz" olabilir. Kurumsal dünyada kararların verildiği verilerin kalitesi genellikle şüphelidir. Veri madenciliğinin başarısı ise temsil ettiği verilere bağlıdır.

Aşağıda veri tabanında karşılaşılan sorunlar listelenmiştir; Listelenen sorunlardan veri temizleme işlemi Veri Temizliği olarak adlandırılır.

Yinelenen Veriler: Bu tür bir hata, yinelenen kayıtları veya imkansız doğrulukları ifade eder. Örneğin, bir müşterinin aynı üründen 100 adet alması veriden şüphe duyulmasına neden olur.

Yanlış veya Tutarsız Veriler: Verilerdeki geçersizlik veya tutarsız anlamına gelir. Aşağıda yanlış ve tutarsız verilere örnekler verilmiştir:

Müşteri listesinde anlamsız bir isim ve soyismin yer alması:

İsim: Noname

Soyadı: Noname

Bu veriler yanlıştır ve fark edilmesi zordur.

Tutarsız girişleri olan bir adres:

Ülke: ABD

Şehir: Ankara

Veri Tipleri: Birçok veri tabanı büyük / küçük harfe duyarlıdır ve büyük harfler bile sorunlara neden olur. Örneğin "Annkara, Ankara, ANKARa, anlara" gibi yazımlar farklı yazım hataları gösterir.

Eski Veri: Dinamik olarak değişen verileri ifade eder. Başka bir deyişle, ilk veri girişinden bu yana değişmiş olabilecek veriler. Adres, yaş eski verilerin tipik örnekleridir. Veri sapmalarında diğer önemli bir faktör de, dünyanın durumunun değişmesidir. Örneğin, müşteri davranışları ve eğilimleri bir süre içinde değişir.

Tanımlama Koşullarında Varyans: Veriler farklı kaynaklardan birleştirildiğinde ortaya çıkar. Veri alanlarının tanımlarında farklılıklar olabilir. Örneğin, aynı ürünleri üreten iki farklı bitkiden toplanan verileri toplandığında saha işlem süresi farklı prosedürler ve tekniklerle hesaplanabilir, böylece karşılaştırılamaz hale gelir.

3.5.3 Aykırı Değer Analizi

Aykırı Değer Analizi genellikle sürekli değişkenler ile gerçekleştirilir. Bir aykırı değer, değişkenin değerinin, bu değişkenin diğer değerlerinin çoğunun yanı sıra, ortalamadan uzak olan tek veya düşük frekanslı bir oluşumudur. Bir değerin aykırı olup olmadığını veya veri hatası olup olmadığını belirlemek bir bilimin yanı sıra bir yetenektir. Veriler hakkında gerçek bir bilgiye sahip olmak çok zordur.

Verideki aykırı değerler yanlış yazılan değerleri veya bankacılık ve finansman hileli faaliyetlerini gösterebilir. Aykırı değerler, çeşitli istatistiksel analizlerle otomatik olarak algılanabilir.

3.5.4 Kayıp Değer

Bilgi toplandığında ve birleştirildiğinde, hemen hemen her veri kümesinde eksik değerler bulunur. Birçok yazılım paketi, eksik değerleri olan kayıtları göz ardı eder ancak bir

değerin eksik olması gerçeğin tahmin edilmesinde engel teşkil eder. Bilgiyi elde etmek önemli olduğundan eksik değerler şu şekilde değiştirilebilir:

Tek Değer İkamesi: Tek değer ikamesi, eksik değerleri değiştirmek için en basit yöntemdir. Üç ortak seçenek vardır: ortalama, medyan ve mod.

Sınıf Ortalaması ile Değiştirme: Sınıf ortalama ikamesi, değişkenlerin diğer değişkenlerinin alt gruplarındaki ortalama değerleri kullanır.

Regresyon İkamesi: Sınıf ortalamasına benzer şekilde, regresyon ikamesi, diğer değişkenlerin alt gruplarıyla ortalamayı kullanır.

3.6 Değişkenlerin Seçimi ve Dönüştürülmesi

Madencilik için veri hazırlamaya ek olarak, bazı ek dönüşümler gerekli olabilir. Davranışı tahmin etmek için veri madenciliği kullanmak, verilerden türetilmesi gereken yeni değişkenler gerektirebilir. Mevcut müşteriler üzerindeki işlem verileri için, RFM değişkenleri iyi öngörücüler olabilir. RFM, yenilik, sıklık ve parasal anlamına gelir. Yenileme genellikle son işlemden sonra geçen zamanın bir ölçüsüdür. Sıklık belirlenen sürede işlem sayısı olacaktır ve Parasal, belirli bir süre içinde toplam işlem ve işlem başına ortalama değer olacaktır. Bu değişkenler, verilerin madencilik sürecinde daha anlamlı hale getirilmesi ve madencilik yazılımının yararlı ilişkiler keşfedebileceği ek parametreler sağlamak için gereklidir.

3.7 Modelin Uygulanması ve Değerlendirilmesi

Modeller, maliyet fayda analizine ve yatırım getirisine dayanan bir iş perspektifinden değerlendirilmelidir. Modelin sonuçları bazı ilginç desenler gösterebilir, ancak bunlar üzerinde hareket etmek, kullanımlarını haklı çıkaracak ek gelir veya maliyet tasarrufu sağlamamaktadır.

Bir modeli değerlendirmenin en basit yollarından biri, sonuçları gerçek dünyada test etmektir. Modelin bir tahminini test etmek için popülasyondan bir örnek seçilir ve gerçek sonuçların tahmin edilen sonuçları ne kadar iyi takip ettiği incelenir. Model, pazarın belirli bir segmentinin belirli bir promosyona cevap verme olasılığını öngörebilir.

Promosyonu sınırlı bir örnek üzerinde uygulayarak ve sonuçları tahminlere göre test ederek, modelin etkinliği ölçülebilir.

3.8 Modelin Kullanılması ve İzlenmesi

Modeller çalışır durumda olduğunda, müşteri davranışlarını ve müşteri beklentilerini anlamak için kullanılabilirler. Pazarlama amaçlı olarak kampanya yönetim yazılımı gibi yazılımlar, üretim sistemlerine dahil edilebilirler. Kampanya yönetimi yazılımı, istenen sonuçları elde etme olasılığı en yüksek olan belirli promosyonlarla müşteri kazanmak için kullanılan pazarlama kampanyalarını otomatik hale getirir. Belirli segmentleri bu şekilde hedeflemek, modelin tahminlerine dayanarak reklam kampanyasına yanıt oranını arttırmalıdır. Bu pazarlama, verimliliğini ve etkinliğini en üst düzeye çıkarır. Veri madenciliği çalışmalarından geliştirilen profiller, kuruluşa müşteri ömür boyu katma değerinde artışa yol açacak olan çapraz satış veya yukarı satış promosyonlarına yanıt verme olasılığı en yüksek olan müşterileri tanımlamalıdır.

Veri madenciliği çalışmalarından geliştirilen müşteri profilleri, kayıt bilgilerini esas alarak siteye gelen ziyaretçileri sınıflandırmak için bir Web ortamında da kullanılabilir. Daha sonra site, sınıflandırmaya dayanarak kendilerine sunulan içeriği kişiselleştirebilir. Bu, ziyaretçiyi bir müşteriye dönüştürme olasılığını artıracaktır. Web sitesinin içeriğinin kişiselleştirilmesi, sitenin rekabetten farklılaşmasına yardımcı olur ve daha yüksek seviyede müşteri hizmeti sunar.

Amaç, Web sitesi ziyaretçilerini müşterilere ve müşterilere uzun vadeli müşterilere dönüştürecek pazarlama çabalarını yönlendirmek için öngörü modellerini kullanmaktır.

3.8.1 Değerleme

Veri madenciliği çalışmalarındaki bu aşamada, veri analiz perspektifinden yüksek kalitede görünen bir veya daha fazla model oluşturulur. Modelin yayınlama aşamasına geçmeden önce, modeli daha ayrıntılı bir şekilde değerlendirmek ve modelin oluşturulması için atılan adımları gözden geçirerek, iş hedeflerini doğru bir şekilde gerçekleştirdiğinden emin

olmak önemlidir. Önemli bir amaç, yeterince dikkate alınmamış önemli bir iş konusunun olup olmadığını belirlemektir. Bu aşamanın sonunda, veri madenciliği sonuçlarının kullanımına dair bir karara varılması gerekir.

3.8.2 Yayınlama

Modelin oluşturulması genellikle veri madenciliği çalışmalarının sonu değildir. Genellikle kazanılan bilginin, müşterinin kullanabileceği şekilde organize edilmesi ve sunulması gerekecektir. Gereksinimlere bağlı olarak, yayınlama aşaması bir rapor oluşturmak kadar basit veya tekrarlanabilir bir veri madenciliği sürecinin uygulanması kadar karmaşık olabilir. Çoğu durumda, yayınlama adımlarını gerçekleştirecek olan veri analisti değil, kullanıcı olacaktır. Her durumda, hangi eylemlerin ne olacağını anlamak önemlidir. Oluşturulan modellerden gerçek anlamda faydalanmak için bunun belirlenmesi gerekir.

Bölüm 4

Veri Madenciliği İşlevleri

4.1 Karakterizasyon ve Ayırt Etme

Veri Karakterizasyonu, bir hedef sınıfın genel özelliklerinin veya özelliklerinin bir özetidir. Kullanıcı tarafından belirtilen sınıfa karşılık gelen veriler tipik olarak bir veritabanı sorgusu tarafından toplanır. Örneğin, geçtiğimiz yıl satışları yüzde on artmış olan yazılım ürünlerinin özelliklerini incelemek için, bu tür ürünlerle ilgili veriler bir SQL sorgusu yürütülerek toplanabilir.

OLAP işlemleri de dahil olmak üzere etkili veri özetlemesi ve karakterizasyonu için çeşitli yöntemler vardır.

Veri Ayrımı, hedef sınıf veri nesnelerinin genel özelliklerinin, bir veya bir dizi karşıt sınıftan nesnelerin genel özellikleriyle karşılaştırılmasıdır. Hedef ve karşıt sınıflar, kullanıcı tarafından ve veritabanı sorguları aracılığıyla alınan ilgili veri nesneleri tarafından belirlenebilir. Örneğin, kullanıcı, geçen yıl satışları % 10 artarken, aynı dönemde satışları en az % 30 azalmış olan yazılım ürünlerinin genel özelliklerini karşılaştırmak isteyebilir. Veri ayırt etme için kullanılan yöntemler, veri karakterizasyonu için kullanılanlara benzerdir.

Çıktılarının sunumu biçimsel anlamda karakteristik ve ayırt etme için benzerdir, ancak ayırt etmenin açıklamaları, hedef ve karşıt sınıfları birbirinden ayırmaya yardımcı olan karşılaştırmalı önlemleri de içermelidir. Kural formunda ifade edilen ayırt etme açıklamaları, ayırt eden kurallar olarak adlandırılır. Kullanıcı, çıktıyı karakteristik ve ayırt etme tanımlamaları bakımından yönlendirebilmelidir.

4.2 Birliktelik Kuralı

Birliktelik Kuralı, belirli bir veri kümesinde sık sık birlikte ortaya çıkan özellik-değer koşullarını gösteren ilişkilendirme kurallarının keşfidir. Birliktelik analizi, Pazar sepet analizi veya işlem verileri analizi için yaygın olarak kullanılmaktadır.

Birliktelik Kuralları şu formdadır;

X => Y ilişki kuralı: "X'deki koşulları karşılayan veritabanı kümelerinin de Y'deki koşulları karşılaması muhtemeldir" şeklinde yorumlanır.

Müşteri analizi modellerinde yaygın kullanımından dolayı birliktelik analizine genellikle pazar sepeti analizi (MB) denir. Pazar Sepeti (MB) Analizi, kataloglarda veya satılan farklı ürünler arasındaki ilişkilerden toplanabilen iş açısından yararlı bilgileri ifade eder. MB analizinin çıktısı, ürün ilişkilerini ve müşteri satın alma davranışını kullanan bilgi ve önerilerdir. [3].

4.3 Sınıflandırma

Sınıflandırma, veri sınıflarını veya kavramlarını tanımlayan ve ayıran bir model (veya işlev) bulma işlemidir. Model, bir dizi eğitim verisinin (yani, sınıf etiketlerinin bilindiği veri nesneleri) analizine dayanarak türetilmiştir. Model, sınıf etiketinin bilinmediği nesnelerin sınıf etiketini tahmin etmek için kullanılır [3].

Veri madenciliğine yönelik sınıflandırma veya denetimli öğrenme yaklaşımı iş dünyasında çok yaygındır. İnsan aklı, doğal olarak nesneleri farklı gruplara ayırır. Örneğin, insanlar bebeklerin, çocukların, ergenlerin, yetişkinlerin ve yaşlıların sınıflandırılmasına girebilir. Sınıflandırma bir haritalama sağlar. Belirtilen gruplara adresler. Örneğin, iki yaş veya daha küçük olan özellik, bebek kategorisine haritalanabilir. Veriler sınıflandırıldıktan sonra, bu belirli grupların özellikleri özetlenebilir [6].

Türetilmiş modeller, sınıflandırma kuralları, karar ağaçları, matematiksel formüller veya sinir ağları gibi çeşitli biçimlerde temsil edilebilir. Bir karar ağacı, her bir düğümün bir öznitelik değeri üzerinde bir testi, her bir dalın testin bir sonucunu temsil ettiği ve ağaç yapraklarının sınıfları veya sınıf dağılımlarını temsil ettiği, akış şemasına benzer bir ağaç yapısıdır. Karar ağaçları, sınıflandırma kurallarına kolaylıkla dönüştürülebilir.

Sınıflandırma için kullanıldığında, bir sinir ağı, tipik olarak birimler arasında ağırlıklı bağlantılara sahip nöron benzeri işlem birimlerinin bir koleksiyonudur. Sınıflandırma modelleri oluşturmak için Naive Bayes sınıflandırma, destek vektör makineleri ve k-en yakın komşu sınıflandırma gibi birçok başka yöntem vardır [3].

4.4 Tahmin

Tahmin süreci basittir. Bir dizi girdi ile belirli bir sonuç üzerinde bir tahmin yapılır. Doğrulama süreci tahmin kullanmakla birlikte, bilinen sonuçları bir doğruluk düzeyini hesaplamak için yapılan tahminlerle karşılaştırmak gerçekten önemlidir. Doğru tahminle, tahmin edilecek sonuç bilinmeyecektir [3].

Sınıflandırma ile oluşturulan modeller, ayrı kategorileri öngörür. Örneğin, birisinin kredi riskini hesaplayan bir model onları "yüksek", "orta" veya "düşük" olarak tahmin edebilir. Tahmin, sürekli değerlere sahip olan sonuçlarla çalışır (örneğin, 1 ve bir milyon arasında gerçek rakamlar). Tahmin bağlamında, istatistikçiler, ayrık değer sonuçlarının sınıflandırılması ve sürekli değer sonuçlarının "regresyon" olarak ele alınmasını gerektirmektedir.

4.5 Kümeleme Analizi

Kümeleme, benzer eğilimleri ve kalıpları paylaşan veri satırlarını gruplandırma yöntemidir. Kümeleme veya segmentasyon, bir veri kümesini ayırt edici gruplara bölme işlemidir. Denetimli öğrenme için model bağımsız değişkenleri alır, bağımlı değişken ile gerçek bağımlı değer arasında bir tahmin üretir ve bir hata düzeltme yapılır; Bu nedenle, çalışma "denetlenir." Kümelenmede böyle bir süreç yoktur, çünkü onu karşılaştırmak için bir sonuç yoktur; Bu nedenle, çalışma "denetimsiz" olarak adlandırılır [6].

Kümeleme çalışmalarının bağımlı değişkenleri yoktur. Sınıflandırma çalışmalarındaki gibi özel bir özellik profili oluşturulmaz. Bu çalışmalara denetimsiz öğrenim ve/veya segmentasyon da denir.

Sınıf etiketli veri nesnelerini analiz eden sınıflandırma ve tahminlemenin aksine, kümeleme, bilinen bir sınıf etiketine başvurmadan veri nesnelerini analiz eder. Genel olarak, sınıf

etiketleri, başlangıçta bilinmedikleri için eğitim verilerinde mevcut değildir. Kümeleme, bu tür etiketler üretmek için kullanılabilir. Nesneler, sınıf içi benzerliği en üst düzeye çıkarmak ve sınıflar arası benzerliği en aza indirgeme ilkesine dayalı olarak kümelenmiş veya gruplandırılmıştır. Yani, nesnelerin kümeleri, bir kümelenme içindeki nesnelerin, birbirine göre yüksek benzerliğe sahip olmaları, ancak diğer kümelerdeki nesnelerden çok farklı olmaları için oluşturulmuştur. Oluşturulan her küme, kuralların türetilebileceği bir nesne sınıfı olarak görülebilir. Kümelenme aynı zamanda taksonominin oluşumunu da kolaylaştırabilir, yani gözlemlerin benzer olayları birlikte gruplandıran bir sınıf hiyerarşisi halinde oluşumunu sağlayabilir.

4.6 Aykırı Veri Analizi

Bir veritabanı, verilerin genel davranışına veya modeline uymayan veri nesneleri içerebilir. Bu veri nesneleri aykırıdır. Çoğu veri madenciliği metodu, aykırı değerleri gürültü veya istisna olarak kaldırır. Bununla birlikte, dolandırıcılık tespiti gibi bazı uygulamalarda, nadir olaylar daha düzenli olarak meydana gelenlerden daha ilginç olabilir. Aykırı veri analizi, aykırı madencilik olarak adlandırılır.

Aykırı değerler, veriler için bir dağılım veya olasılık modeli üstlenen istatistiksel testler kullanılarak veya başka bir kümeden önemli bir mesafe olan nesnelerin aykırı olduğu düşünülen mesafe ölçümleri kullanılarak tespit edilebilir. İstatistiksel veya mesafe ölçümleri kullanmak yerine sapma temelli yöntemler, bir gruptaki nesnelerin ana özelliklerinde farklılıkları inceleyerek aykırı değerleri tanımlar.

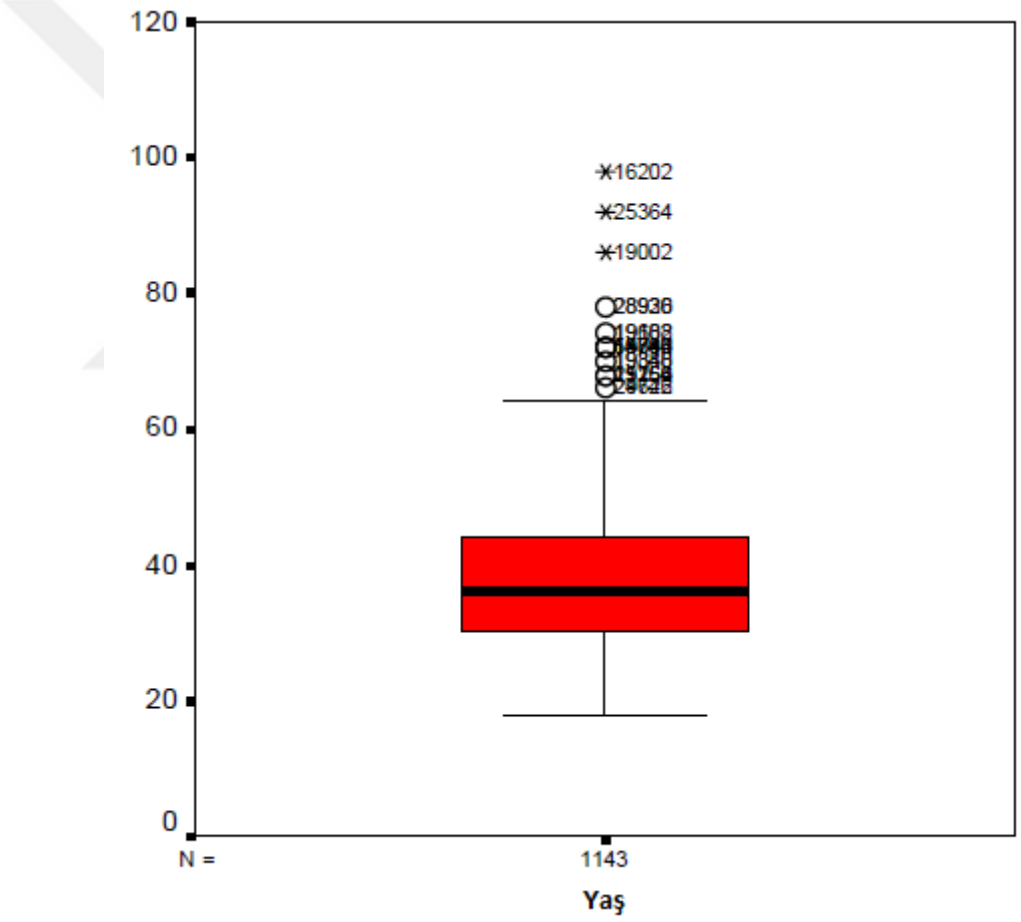
4.7 Değişim Analizi

Veri değişimi analizi, davranışı zamanla değişen nesneler için düzenleri veya eğilimleri tanımlar ve modeller. Bu, karakterizasyon, ayırt etme, ilişkilendirme, sınıflandırma ya da zamana bağlı verilerin kümelenmesini içerebilse de, böyle bir analizin farklı özellikleri arasında zaman dizisi veri analizi, dizi ya da periyodik model eşleştirmesi ve benzerliğe dayalı veri analizi yer alır.

4.8 Görselleştirme

Görselleştirme, basitçe verilerin grafiksel sunumudur. Bazı durumlarda veriler en iyi grafiklerle anlaşılabilir. Örneğin, görselleştirme teknikleri Şekil 4.1 'de görüldüğü gibi aykırı değerleri kolayca gösterebilir.

Veriyi grafiksel olarak temsil etme süreci, günümüzde çoğu sorgu aracında kullanılmaktadır. Görselleştirme, iki boyutlu grafikler ve haritalardan çok daha fazlası anlamına gelebilir. Aykırı değerlerin saptanması için bir örnek aşağıda verilmiştir, KDD CUP 2000 örnek verisindeki bir kutu grafiği çizgilerini kolayca saptayabilmektedir. Bu örnekteki aykırı değerler, verilerdeki kirliliğe işaret edebilir.



Şekil 4.1: SPSS İstatistik Paketi ile Bir Outlier Analizi

Bölüm 5

Veri Madenciliği Algoritmaları

5.1 Karar Ağaçları

Karar ağaçları, sınıfları önceden bilinen bir veriden tümevarım yöntemiyle öğrenilen ağaç yapılı bir karar yöntemi çeşididir. Bir karar ağacı, basit karar verme adımları uygulanarak, büyük miktarlardaki kayıtları, çok küçük kayıt gruplarına böler ve her başarılı bölme işlemiyle, sonuç gruplarının temsilcileri bir diğeriyle çok daha benzer hale gelir. Büyük veri tabanlarının kullanıldığı pek çok sınıflama probleminde ve karmaşık yada hata veri içeren durumlarda karar ağaçları kullanışlı bir çözüm olmaktadır. Tahmin edici ve tanımlayıcı özelliklere sahip olan karar ağaçları, Veri madenciliğinde yapıların kolay olması, yorumlanmalarının kolay olması, veri tabanı sistemlerine kolayca entegre edilebilmeleri, güvenilirliklerinin sağlam olması nedenleri ile sınıflama modelleri içerisinde en yaygın kullanıma sahip olan bir yöntemdir [7], [8], [9], [10].

Karar ağacı yaklaşımlarının en büyük avantajı anlaşılabilir olmasıdır. Bununla birlikte, karar ağacı yaklaşımını kullanarak verileri başarılı bir şekilde modellemek için, birkaç bölme gerekebilir. Ağaç, yüksekliğe göre verileri alt bölümlere ayırır. Örneğin, belirli bir yaşın üzerindeki daha ağır insanların yüksek tansiyon eğilimi olduğunu öğrenmek için yaş ve kilo bazında daha fazla alt bölümlere ayırmak gerekebilir.

Aşağıda, bir karar ağacı oluşturmanın basitleştirilmiş, aşamalı bir örneği bulunmaktadır. Günümüzde kullanılan karar ağaçları için pek çok yaklaşım söz konusudur. İstatistiksel bir yaklaşım olan CART bu yaklaşımın en iyi örneğidir. Her terminal olmayan düğümden

tam olarak iki dalın bulunduğu istatistiksel tahmin kullanır. Başka bir yaklaşım, terminal olmayan düğümden çıkan dalların sayısının, kategori sayısına eşit olduğu yerdir.

Bütün karar ağacı algoritmaları benzer bir süreçten geçerken, farklı değişkenlerin öneminin nasıl gruplandırılacağı ve sıralanacağını belirlemek için farklı matematiksel algoritmalar kullanılır. Örneğin, C4.5 Makine Öğrenimi Programlarında, Quinlan, bölmenin oluşturduğu bilginin yararlı olan oranını ifade eden bir kazanım oranı algoritmasını tartışır. Karar ağacı algoritmalarındaki ana adımlar aşağıdaki gibidir.

Adım 1: Bir veri kaynağından değişkenler seçilir ve veri kaynağında sunulan değişkenlerden, bağımlı değişken bir kullanıcı tarafından seçilir.

Adım 2: Bir sonucu etkileyen her değişken incelenir. Bu değişkenlerin her birinde bulunan değerler üzerinde yinelenen gruplama değerleri bir arada gerçekleştirilir.

Adım 3: Her bir değişken için gruplamalar hesaplandıktan sonra, bir değişken bağımlı değişken için en öngörücü olarak kabul edilir ve ağacın yaprak düğümlerini oluşturmak için kullanılır.

Çoğu insan karar ağaçlarını sezgisel bir yöntem olduğunu düşünür. Karar ağaçlarının olumsuz tarafı, verilerinin karmaşıklığı arttıkça yönetilmesi zorlaşır. Bunun nedeni ağaçtaki artan dal ve düğüm sayısıdır. Eksik verilerin işlenmesiyle ilgili bir sorun da vardır, çünkü bir veri öğesi olmadan, bu verilere bağlı olarak bir ağaç düğümü hesaplanamaz [6].

5.2 Genetik Algoritmalar

Genetik algoritmalar, biyolojik evrimdeki süreçlere dayalı bir optimizasyon yöntemidir. Temel fikir, zamanla, evrimin "en uygun türler"i seçmesidir. Bu düşüncenin veri madenciliğine uygulanması, genellikle "en uygun" modeller elde etmek için genetik yöntemler kullanılarak bir veri modelinin optimize edilmesini içerir. Genetik algoritmalar, çoğunlukla, model verisine yönelik sinir ağları ile bağlantılı olarak kullanılmıştır. Genetik algoritmalar aynı zamanda verileri kümelemekte iyidir. Örneğin, bir veri kümesini üç gruba bölmek veya kümelendirmek istendiğinde, yapılacak işlemler aşağıda verilmiştir.

Adım 1: Genetik bir algoritma için, rastgele bir veri gruplamasıyla başlanır. Bir organizma olarak oluşturulacak üç kümenin her biri düşünülür. Genetik algoritma, bir veri

kümesinin üç "organizma" ya da kümeden biri için bir eşleşme olup olmadığını belirleyen bir uyum fonksiyonu olarak adlandırılacaktır. Bu uyum fonksiyonu, bazı veri kümelerini diğerlerine göre daha iyi uyarlayan bir fonksiyon olabilir. Veri kümeleri okunduğunda, bir kümedeki diğer veri öğelerinin ne kadar iyi bir ilişki içinde olduklarını görmek için bu işlevler uyum fonksiyonu işlevi tarafından değerlendirilebilir. Örneğimizde, bir uyum fonksiyonu, bir grup içindeki veri setleri arasındaki benzerlik düzeyini belirlemek için bir işlev olabilir.

Adım 2: Genetik algoritmalar, veri gruplarının açıklamalarının kopyalanmasına ve değiştirilmesine izin veren operatörlere sahiptir. Bu operatörler, hayatın ürettiği, eşleştiği ve mutasyona uğradığı doğada bulunan işlevi taklit eder. Veri kümesindeki bir veri satırı, uyum fonksiyonu açısından uygunsa, o zaman hayatta kalır ve bir kümeye kopyalanır. Bir veri satırı uygun değilse, başka bir gruba geçebilir veya başka bir deyişle, daha iyi uyum sağlamak için diğer kümelerle eşleştirilebilir. Yeni veri kümelerine geçildiğinde küme, kendini en uygun hale getirmek için değiştirecektir.

Genetik algoritmalar, diğer teknolojilerin zor zamanlarda karşılaştığı karmaşık problemleri çözmüştür; Bununla birlikte, genetik algoritmalar, yaklaşımların içerisinde en az "açık" olarak bilindiğinden en az anlaşılabilir olmaktadır. Örneğin, uyum fonksiyonları çok çeşitlilik gösterebilir. Temel gereksinim, bir uyum fonksiyonunun, en az hataya yakınlaşmaya izin veren belirli özelliklere sahip olması gerektiğidir.

Genetik algoritmalar, daha yüksek düzeyde bir model anlayışı sağlamak için sıklıkla sinir ağları ile birlikte kullanılmıştır. Sinir ağlarının sıklıkla "kara kutular" olduğu söylenirken, sinir ağları ile bağlantılı genetik algoritmalar, her bir sinir ağ modelinin daha ayrıntılı bir şekilde belgelenmesini sağlayan bir veri tabanını doğrudan bir veri tabanına etki eden girdi değişkenleri gruplarını kaydedebilir. Çeşitli modellerle deneyler yaptıktan sonra, önceki modelin değişken kümelerinden birini okuyarak son bir model oluşturulabilir [6].

5.3 Sinir Ağları

Sinir ağları, iş dünyasında yardımcı modeller olarak yaygın olarak kullanılmaktadır. Özellikle, finansal hizmet sektörü kredi kartlarında ve mali işlemlerde sahteciliği modellemek için yaygın olarak sinir ağlarını kullanmaktadır.

Sinir ağları, bir insan beyni içindeki bir nöronu taklit etmeye çalışır, her bir bağlantı bir işlem elemanı (PE) olarak tarif edilir. Sinirsel ağlar deneyimlerden öğrenir ve bir dizi girdi verisi ile bir sonuç arasındaki bilinmeyen ilişkilerin tespitinde kullanılır. Diğer yaklaşımlar gibi, sinir ağları da verideki modelleri tespit eder ve genelleme yapar.

Verilerde bulunan ilişkiler ve sonuçları tahmin eder. Sinir ağları, özellikle karmaşık süreçleri tahmin etme yetenekleri bakımından dikkat çekmiştir.

Bir işlem elemanı, verileri bir dizi matematiksel işlevi kullanarak özetleyip dönüştürerek işler. Bir işlem elemanının kabiliyetleri sınırlıdır, ancak bir sistem oluşturmak için bağlandığında, nöronlar veya işlem elemanları akıllı bir model oluşturur. İşlem elemanları herhangi bir yolla birbiriyle bağlantılıdır ve modellemeye çalıştıkları verilere daha fazla uymaları için yüzlerce veya binlerce yineleme üzerinden yeniden eğitilebilirler.

İşlem elemanları, girişlere ve çıkışlara bağlanır. Ağın eğitim süreci, girişlerden çıkışa olan bağlantıların mukavemetini veya ağırlığını değiştirmeyi içerir. Bir bağlantının gücündeki artış veya azalmalar, doğru sonucu üretmenin önemine dayanır. Bir bağlantının gücü, bir deneme-yanılma işlemi sırasında aldığı ağırlığa bağlıdır. Bu süreç, ağırlıkların ayarlanması için bir matematiksel yöntem kullanır ve bir öğrenme kuralı olarak adlandırılır.

Tekrar veya tekrarlı olarak eğitim, bir sinir ağını tarihsel verilerin örneklerine maruz bırakır. İşlem elemanları verileri özetler, dönüştürür ve aralarındaki bağlantıları farklı ağırlıklar ile hesaplar. Yani bir ağ, her bir örnek için çıktı değişkenini tahmin etmek için çeşitli formüller dener.

Bir nöral ağ, belirli bir doğruluk seviyesinde bilinen sonuç değerleriyle eşleşen sonuç değerlerini veya diğer bazı durdurma kriterlerini karşılayana kadar eğitimini sürdürür. İşlem elamanlarının her biri çok sayıda girdi alır ve girişlerin ağırlıklı toplamının doğrusal olmayan bir fonksiyonu olan bir çıktı üretir. Girdilerin her birine atanan ağırlıklar, ağ tarafından üretilen çıktıların hedef çıktılarla karşılaştırıldığı bir eğitim süreci (genellikle geri yayılım) sırasında elde edilir. Ağın üretmesi istenilen cevaplar, üretilen çıktılarla karşılaştırılır ve bunlar arasındaki sapma, ağırlıkları ayarlamak için geri besleme olarak kullanılır [6].

Yeniden ayarlama ağırlıkları süreci, bir modelin doğruluğunu arttırmak için önemlidir. Gizli düğümlerin sayısı ayarlanabilir ve aslında yalnızca konuları karıştırmak için birden fazla gizli düğüm noktası olabilir. Girdi sayısı, gizli düğümler, çıkışlar ve düğümler

arasındaki bağlantılar için ağırlıklandırma algoritmaları, bir sinir ağının karmaşıklığını belirler. Genel olarak, bir sinir ağının karmaşıklığı, doğruluğu ve sinir ağı modelini oluşturmak için gereken zaman arasında bir ilişki vardır. Gizli düğümlerin ve ağırlıkların konfigürasyonu sinir ağları için çok kritik olduğundan, gizli düğümlerin doğru sayısını ve yeniden ayarlanan ağırlıkları bulmak için birçok yaklaşım vardır.

Bu açıklamalar sinir ağlarının giriş anlamında bir görünümüdür ve nasıl çalıştığını anlamak için bir başlangıç noktasıdır.

Sinir ağlarının en büyük gücü, karmaşık problemlerin sonuçlarını doğru bir şekilde tahmin etme yetenekleridir. Sinir ağları, finansal piyasalarda ve üretimde popüler olan tahmin veya sürekli sayısal çıktıları gerçekleştirmede tercih edilen bir tekniktir [6].

Sinir ağlarına ait bazı eleştiriler de vardır. Bunlardan birincisi, tahmin için yararlı oldukları, ancak her zaman bir modeli anlamadıklarıdır ve Sinir ağlarının erken uygulamalarının "kara kutu" tahmin motorları olarak eleştirildiğidir ancak; bugün piyasadaki yeni araçlarla, bu eleştiri tartışmalıdır.

İkincisi, sinir ağları aşırı eğitime yatkındır. Öğrenme kapasitesi yüksek bir ağ, bu kapasiteyi desteklemek için çok az veri örneği kullanılarak eğitilirse, ağ ilk önce verilerin genel eğilimlerini öğrenmeye başlar. Bu arzu edilir, ancak daha sonra ağ, genellikle istenmeyen olan eğitim verilerinin çok spesifik özelliklerini öğrenmeye devam eder. Bu ağların eğitim verilerini ezberlediği ve genelleme yeteneğinden yoksun olduğu söylenir. Ticari dereceli sinir ağları bugün, önyükleme tutma (test) numuneleri ile aşırı yüklenmeyi ve test hatalarına karşı test hatalarını etkin bir şekilde ortadan kaldırmıştır.

Aşırı eğitim, test veri setinin sonuçlarını düzenli olarak kontrol ederek ölçülebilir. Bir çalışma sürecinin erken aşamaları hem eğitim hem de test verisinde daha düşük hata ölçümleri sağlar. Bu, ağ kapasitesi gereksinim duyulmadıkça veya eğitim dosyasında çok az veri kümesi olmadıkça devam eder. Öğrenme sırasında eğitim verilerinin daha iyi sonuçlar üretmeye devam etmesine rağmen bir noktada test verileri daha kötü sonuçlar üretirse, aşırı eğitim gerçekleşmiştir.

Sinir ağları ile ilgili bir başka konu da eğitim hızıdır. Sinir ağları inşa etmek birçok aşama gerektirir. Bu, en doğru modeli oluşturmanın çok zaman alıcı olabileceği anlamına gelir. Tüm regresyon tekniklerinin yakınsama zamanı gerektirdiğinden bahsetmek gerekir; ve

geri yayılım yavaş olsa da, sinir ağlarını eğitmek "conjugate gradient" gibi yöntemlerle seri bir şekilde hızlandırılabilir.

5.4 İstatistik

İstatistiksel yaklaşımların güçlü yanları, sadece bu yaklaşımların doğru olmaması değil, iyi anlaşılmış ve yaygın olarak kullanılmalarıdır. İstatistiksel yaklaşımlar, çoğu kişi tarafından "en doğru" veri madenciliği biçimi olarak görülmekte ve aslında birçok veri madenciliği tekniği, uzun yıllardır var olan istatistiksel teknikleri kullanmaktadır. Popüler bir karar ağacı yaklaşımı olan CHAID, Chi Square metriğini kullanmaktadır.

İlişkilendirme algoritması, istatistiksel destek ve güven ölçütlerini kullanır ve kümeleme teknikleri K-Means algoritması gibi istatistiksel ölçümleri kullanır. Bayes Ağları, Bayes Teorisi Olasılığı'nı kullanır.

İstatistiğin en büyük eleştirisi her zaman onu etkili bir şekilde kullanmanın zorluğu olmuştur. Birçok iş uzmanı istatistikte kullanılan terminoloji ile karıştırılmaktadır [6].

Bölüm 6

Veri Madenciliği Uygulama Alanları

Çok yaygın kullanım alanına sahip veri madenciliğinin başlıca kullanıldığı alanlar aşağıdaki gibidir;

6.1 Bilimsel ve Mühendislik Verileri

Günümüzde laboratuvar veya bilgisayar ortamında sistemlerin benzetimi ve analizi sürecinde yüksek miktarda bilimsel veri üretilmektedir. Elde edilen bu verilerin anlamlandırılması için veri madenciliği çok uygun bir platform sağlamaktadır [11].

6.2 Sağlık Verileri

Sağlık ve tıp alanı veri madenciliğinin en yaygın kullanıldığı alanlardan biridir. Özellikle tarama testlerinden elde edilen veriler ile pek çok kanserlerin ön tanısı, kalp verileri kullanılarak kalp krizi riski tespiti, acil servislerdeki hasta belirtilerine göre risk ve önceliklerin tespiti gibi çok geniş bir uygulama alanı söz konusudur [11].

6.3 İş Verileri

İş süreçleri sırasında büyük miktarda veri üretilir. Bu veriyi karar verme süreçlerinde ve müşteri veri tabanlarının analizi ile reklam ve promosyon ile ilgili pek çok faydalı bilgiye ulaşmak için kullanmak mümkündür [11].

6.4 Alışveriş Verileri

Bu alanda en çok başvurulan veri madenciliği yaklaşımı Birliktelik Kuralı yöntemlerinden pazar sepet analizidir. Pazar sepet analizinde amaç satın alınan ürünler arasındaki ilişkileri bulmaktır [1]. Bu ilişkilerin bilinmesi işletmenin pazarlama, reklam stratejilerini belirlemede yol gösterici olacaktır.

6.5 Bankacılık ve Finans Verileri

Bankacılık sektöründe kredi ve kredi kartı sahtekarlığı tahminlerinde, risk değerlendirmelerinde, müşteri eğilim analizlerinde, kar analizi gibi alanlarda yaygın şekilde veri madenciliği kullanılmaktadır [1], [11].

6.6 Eğitim Alanı Verileri

Öğrenci işlerinde veriler analiz edilerek öğrencilerin başarı ve başarısızlık sebepleri, başarı oranının arttırılması için hangi konulara ağırlıklı olarak çalışılması gerektiği, üniversite giriş puanları ile okul başarısı arasında bir ilişkinin olup olmadığı gibi pek çok sorunun cevabı bulunarak eğitim kalitesi ve performansı arttırılabilir [1].

6.7 İnternet Verileri

İnternet ve web üzerindeki veriler her geçen gün artmaktadır. Web madenciliği kısaca internetten faydalı bilginin keşfedilmesi olarak tanımlanabilir. Kaynakların otomatik tarama sistemleri, bilgi alma için kullanılan sistemler ve web siteleri veya online veri tabanlarından seçilmesi web içerik madenciliği konusuna girerken; web sunucularından veya online servislerden kullanıcı erişim desenlerinin analiz ve keşfi web kullanım madenciliği konusuna girmektedir [12].

6.8 Doküman Verileri

Dokümanlar üzerinde uygulanan veri madenciliği çalışmalarında ana amaç; dokümanlar arasında elle tasnif gerekmeksizin benzerlik hesaplayabilmektir. Bu genelde otomatik olarak çıkarılan anahtar sözcüklerin tekrar sayısı sayesinde yapılır. Polis kayıtlarında mevcut rapora benzer kaç adet ve hangi raporlar var. Ürün tasarım dokümanları ve internet dokümanları arasında mevcut tasarım için kullanılabilecek ne tür dosyalar var gibi sorulara cevap bulunabilir [1].

6.9 Askeri Veriler

Hedef tanıma ve askeri takip sistemlerinden elde edilen veriler ile veri madenciliği çalışmaları yapılabilmektedir. Bunun yanı sıra sensörlerin ve simülasyon modellerin performans analizleri ve görüntü verileri ile veri madenciliği çalışmaları yapılabilmektedir [13].

6.10 Sosyal Ağ Verileri

Günümüzde internet kullanımının yaygınlaşmasıyla son yıllarda ortaya çıkan yeni pek çok uygulama iletişim ve eğlence amacı ile kullanılmaya başlanmıştır. Sosyal medya olarak tanımladığımız bu uygulamalar ile kişiler ve geniş kitleler hakkında büyük miktardaki verilere internet üzerinden rahatlıkla ulaşılabilmektedir.

Tezin bundan sonraki bölümlerinde veri madenciliği yaklaşımının sosyal ağlar üzerindeki etkisi incelenecektir.

Bölüm 7

Sosyal Ağlar

Son yıllarda internet erişiminin yaygınlaşması ile pek çok uygulama geliştirilmiştir. Geliştirilen bu uygulamalar insanlara sadece iletişim ve bilgi paylaşımı olanağı değil eğlence ve iyi vakit geçirme seçeneği de sunmaktadır. Genel olarak sosyal ağlar olarak tanımlanan bu uygulamalar kişiler ve geniş kitleler hakkında büyük miktarda veriye internet üzerinden rahatlıkla erişim imkanı sunmaktadır.

Sosyal ağ uygulamaları genellikle belli bir kitleyi kendine hedef edinmektedir. Günümüzde en popüler ve ülkemizden erişilebilen sosyal ağ uygulamaları aşağıdaki gibidir;

Facebook: Gerçek hayatta tanımış olduğunuz kişilerle etkileşime geçmenizi sağlar.

Instagram: Fotoğraf ve kısa video paylaşımlarında bulunabilmenizi sağlar.

Google+: Google tarafından meydana getirilmiştir. Tanıdığınız ve ilgi duyduğunuz kişilerle etkileşimde bulunabilmenizi sağlar.

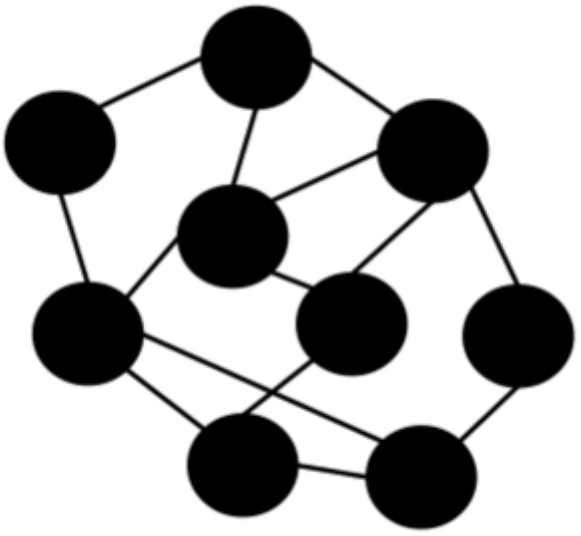
Twitter: Sadece 140 karakter ile fikir paylaşımında bulunabilmenizi sağlar.

YouTube: Sosyal video paylaşım ve etkileşim platformudur.

Linked-in: Profesyonel iş ağı olarak kullanılmaktadır.

7.1 Çizge Teorisi Yaklaşımı

Sosyal ağların bilimsel olarak çok farklı ifade ve modellemeleri olmakla birlikte, literatürde en çok kabul görmüş gösterim şekli; Çizge Teoremi (graph theory)'dir. Bireyler veya varlıklar birer düğüm (node), ilişkiler ise birer kenar (edge) olarak tasvir edilir [14]. Şekil 7.1 'de Sosyal Ağların Çizge Teoremi ile temsili bir gösterimi yer almaktadır.



Şekil 7.1: Sosyal Ağların Çizge Teoremi ile Temsili Gösterimi

Sosyal ağ kavramının çıkışından bu yana kullanılan ve en çok kabul gören ve ilk yöntemlerden birisi olan çizge teorisi günümüzde de en önemli sosyal ağ analizi yöntemi olarak kabul edilmektedir. Genellikle bir ağdaki varlıkları ve bu varlıklar arasındaki ilişkileri göstermek için kullanılır. Özellikle çok büyük ölçekli veri tabanlarında çizge teorisinin kullanımı çok büyük bir öneme sahiptir [14].

7.2 Sosyal Ağların Genel Özellikleri

Sosyal ağların pek çoğu kullanıcı odaklıdır ve kullanıcıların daha fazla vakit geçirebilmeleri için onlara ekstra uygulamalar sunmaktadır. Sundukları uygulamaların çoğunluğu ücretsizdir. Sunulan uygulamalar anlık mesajlaşma, video, elektronik posta, oyun, dosya ve fotoğraf paylaşımı gibi çeşitli hizmetleri içerir. Bu hizmetleri sağlamaktaki amaç, kullanıcıların etkileşimini kolaylaştırmaktır. Sosyal ağ kullanıcıları; kendilerine ait bilgilerinin tutulduğu veri tabanı sayesinde rahatlıkla arkadaşlarının paylaşımlarını

ve ilgilendikleri konuları takip edebilirler. Kullanıcılar aynı zamanda kullandıkları sosyal ağın özelliklerine göre kendi profillerini oluşturabilir, profillerinde paylaştıkları bilgilerin ve beğenilerin üzerinde gizlilik ayarları yapabilirler [1].

7.3 Sosyal Ağ Uygulamalarında İletişim

Sosyal ağ uygulamalarında iletişim kullanıcıların istek göndermesi ve karşı tarafın onaylaması şeklinde iki yönlü olabileceği gibi yalnızca bir kullanıcının diğerini takip etmesi şeklinde tek yönlü de gerçekleşebilmektedir. Aynı zamanda öneri sistemlerinin geliştirilmesi ile bir kullanıcının davranış ve beğeni özelliklerine benzer kullanıcıların önerilmesi ile de sosyal ağ uygulamalarında etkileşim başlayabilmektedir.

Sosyal ağ uygulamalarındaki arkadaşlık terimi yanlış algıya sebebiyet verebilmektedir. Sosyal ağlarda kullanılan arkadaşlık kavramı bireyler arası bağlantıyı temsil etmektedir. Bu bağlantı gerçek hayat ile ilişkisi olmadan sadece sosyal ağlarda varlığını sürdürebilmektedir [1], [14].

7.4 Sosyal Ağ Uygulamaları

Literatürde sosyal ağ uygulamalarının çevrim içi sosyal ağ uygulamaları ve kurum içi sosyal ağ uygulamaları olarak iki ayrı grupta incelendiği görülmektedir.

Çevrimiçi Sosyal Ağ Uygulamaları; Kullanıcıların topluma açık veya gizli profiller oluşturmasına imkan sağlayan, ilişkiye sahip olduğu diğer kullanıcıların bağlantı listelerini görmesine müsaade eden web tabanlı uygulamalardır.

Kurum İçi Sosyal Ağ Uygulamaları; kuruluşlara özel olarak hazırlanan intranete dayalı olarak sosyal ağ oluşturma araçları ile kuruluşların kendi iç ağları içerisinde oluşturup kullanabildikleri ve sadece kuruluş çalışanlarının katılabileceği uygulamalardır.

Çevrimiçi sosyal ağlar internet üzerinden herkesin erişebileceği bir ortam sunarken; kurum içi sosyal ağlarda yalnızca kurum çalışanları bu ağa erişebilmektedir [15].

Bölüm 8

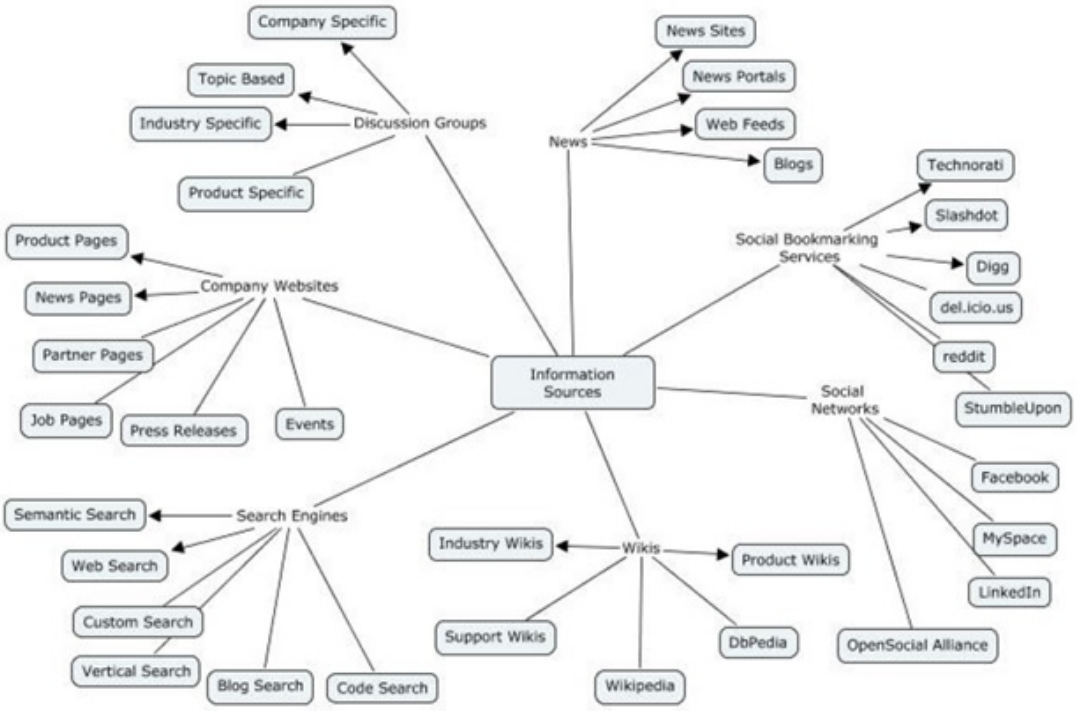
Sosyal Ağlarda Veri Madenciliği

Sosyal ağlar, içerdikleri sosyal aktörlerin ilişkilerine ait oldukça faydalı bilgiler barındırmaktadırlar. Bu yapı ve ilişkilerin analiz edilmesi aracılığıyla benzerlikler, yakınlıklar, eğilimler ve etkileşimler gibi verilere ulaşılarak sosyal ağlardaki ilişkiler hakkında çeşitli çıkarımlara veya tahminlere varılabilir. Özellikle internet teknolojisinde yaşanan teknik ve kültürel ilerlemeler ile sosyal ağlara ilişkin veriler ölçülebilir hale gelmiştir. İnternet üzerinde gerçekleşen sosyal iletişim sonucu oluşan veri yığını çok büyük boyutlara ulaşmaktadır. Söz konusu devasa büyüklükteki veri yığınları içerisinden anlamlı bilginin çıkarılabilmesi için Veri Madenciliği türlerinden biri olan "Web Madenciliği" teknikleri kullanılmaktadır. Sosyal ağların web madenciliği teknikleri kullanılarak analiz edilmesi ve bu sayede akademik, ticari, sosyolojik ve pekçok alanda anlamlı verilere ulaşılması önemli bir çalışma konusu haline gelmiştir [16].

8.1 Web Madenciliği

Web madenciliği; çeşitli yapıdaki web sayfalarını, dokümanlarını ve kayıt bilgilerini incelemek ve bunlardaki anlamlı kalıpları keşfetmek için veri madenciliği tekniklerinin kullanılması olarak tanımlanmaktadır [17].

Web madenciliğinde kullanılan veriler, web üzerinde çok geniş bir alandan toplanmaktadır [17]. Şekil 8.1 'de web madenciliği veri kaynakları gösterilmektedir.



Şekil 8.1: Web Madenciliği Veri Kaynakları

8.2 Sosyal Ağlarda Web Madenciliği

Sosyal ağlardan elde edilen devasa büyüklükte ve karmaşıklıktaki verilerin etkin bir biçimde analiz edilebilmesi için web madenciliğinde kullanılan tüm yöntemler 4 ana işlem adımından oluşmaktadır [16].

- Kaynak Bulma,
- Bilgi çıkarımı ve ön işleme,
- Genelleştirme
- Çözümleme

8.2.1 Kaynak Bulma

Kaynak bulma özetle verinin elde edilmesi kısmıdır. Çeşitli verilerin bir veri ambarında toplanması ile yapılır. Web Madenciliği için veri kaynakları Şekil 8.1'de verilmiştir.

8.2.2 Bilgi Çıkarımı ve Ön İşleme

Veri kaynaklarından elde edilen verilerin, veri ambarında toplanması sonrasında verilerin işlenmesi veya işlenecek hale getirilmesi işlemidir.

8.2.3 Genelleştirme

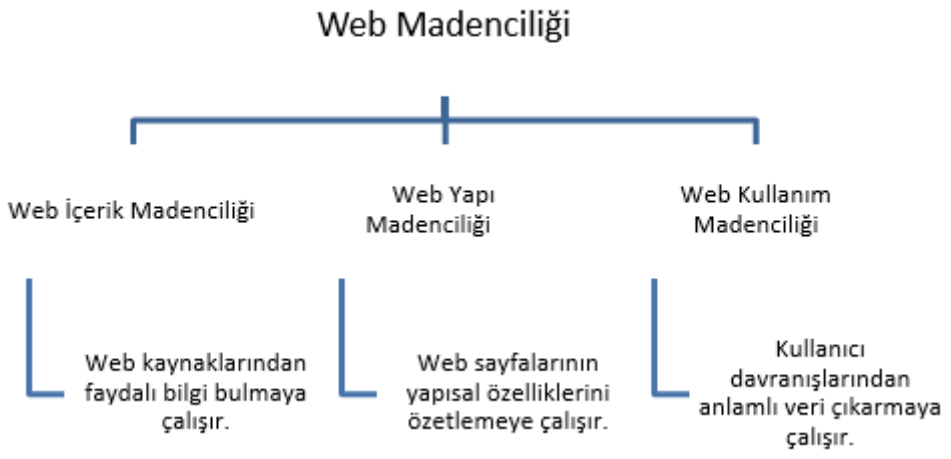
Daha önceden kazanılan tecrübenin veriler üzerinde uygulanması ile genel geçer kuralların üretilmesi işlemidir.

8.2.4 Çözümleme (Analiz)

Çıkarılan kurallar neticesinde eldeki verilerden anlamlı bilginin keşfedilmesi aşamasıdır. Bu aşama sonucunda geleceğe yönelik tahmin ve değerlendirmelerde bulunabilme yeteneği kazanılmış olur.

8.3 Web Madenciliği Yöntemleri

Web Madenciliği, veri madenciliği yaklaşımının bir alt dalıdır. 3 başlık altında incelenir. Bu başlık Şekil 8.2'de gösterilmektedir.



Şekil 8.2: Web Madenciliği Yöntemleri

8.3.1 Web İçerik Madenciliği

Web sitelerinin içeriğine odaklanır. Yapay zeka, doğal dil işleme veya resim işleme ve benzeri yöntemler kullanılarak web kaynaklarının içeriklerinden yararlı bilginin elde edilmesi işlemidir [16].

Web sayfasında kullanılan dili tespit etmek, kullanılan kelimlerin sıklıklarını hesaplamak, anahtar kelime tespitinde bulunmak web içerik madenciliği adına yapılan çalışmalara örnek olarak verilebilir [14].

8.3.2 Web Yapı Madenciliği

Web Yapı Madenciliğinde amaç, web sayfaları arasındaki linkleri takip ederek bilgi elde etmektir [17].

Web Yapı Madenciliği daha önceki bölümlerde bahsedilen Çizge Teorisi yaklaşımına dayanır. Web siteleri ve web sayfaları arasındaki bağlantılar incelenir. Çizge yaklaşımına göre web sayfaları birer düğümdür ve web yapı madenciliğinde düğümler arası bağlantılara odaklanılır.

Hangi web sitelerin hangi web sitelerine bağlantı verdiği bir grafik ile gösterilebilir. Bu grafik yardımı ile en çok bağlantı alan veya veren siteleri tespit etmek mümkündür [14], [16].

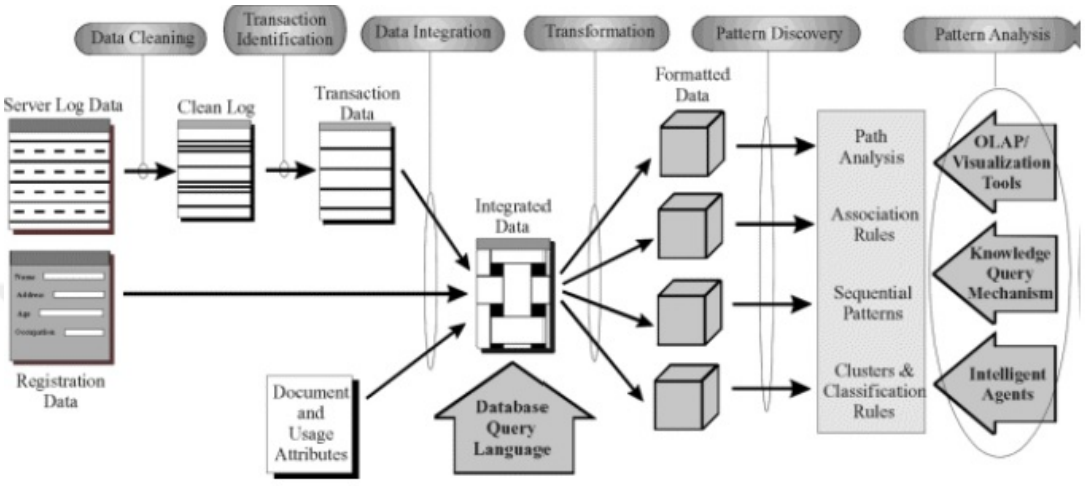
8.3.3 Web Kullanım Madenciliği

Bu yöntemde ihtiyaç duyulan veriler internet mecrası üzerindeki çeşitli sunucularda kayıt altına alınmış olan kullanıcılara ait kullanıcı erişim hareketlerinin yer aldığı çeşitli log dosyalarından elde edilir.

Web Kullanım Madenciliğinde amaç, kullanıcıların web sayfaları ile olan ilişkilerini; geride bıraktıkları erişim kayıtlarından elde etmektir. Log dosyalarının incelenmesi ile kullanıcılar hakkında detaylı bilgi ve tahminlere ulaşılır. Örneğin kullanıcıların tıklama alışkanlıkları ve sıklıkları, dolaştıkları web siteleri, hangi web sitesine hangi web sitesinden sonra girdikleri, en çok hangi reklamlara tıkladıkları ve üzerinde durma süreleri, resim içerikli mi yazı içerikli mi yoksa video içerikli mi web sitelerine daha çok tıkladıkları gibi

pek çok bilgi elde edilir. Bu bilgi kazanımının ardından kullanıcıya özel içerik sunma, teklif verme, kullanıcıya özel reklam oluşturma gibi pazarlama davranışları geliştirilebilir; kullanıcının eğilimleri belirlenerek ona yönelik tahminler yürütülebilir. Bu sayede ticari firma sahipleri avantaj elde edebilir [14], [16], [17].

Şekil 8.3'de Web Kullanım Madenciliğinin genel mimarisi gösterilmektedir.



Şekil 8.3: Web Kullanım Madenciliği Mimarisi

Mimari incelendiğinde Web Kullanım Madenciliğinin iki ana aşamadan oluştuğu gözlemlenmektedir.

İlk Aşama web sitelerinden verilerin elde edilmesi, temizlenmesi ve dönüştürülmesi adımlarını kapsar ve bu adımlarda web sitelerine bağımlı olarak işlemler yapılır.

İkinci Aşama ise web sitelerinden bağımsız olarak veri madenciliği yöntemlerinin uygulandığı kısımdır.

8.4 Fikir Madenciliği

Fikir Madenciliği veya Duygu Analizi; insanların görüş, tutum ve duygularının bir varlığı ifade ettiği metin madenciliği çalışmalarıdır. Bu varlık, bireyleri, grupları, olayları veya konuları ifade edebilir [18].

Fikir madenciliği ve Duygu Analizi literatürde çoğunlukla birbiri yerine kullanılmaktadır. Ancak bir grup araştırmacı bu kavramların kısmen de olsa farklı olduğunu savunmaktadır. Fikir Madenciliği insanların herhangi bir varlık hakkındaki fikirlerini analiz edip,

ortaya çıkarırken; Duygu Analizi bir metindeki duygusal ifadeleri ortaya çıkarır. Bu açıdan bakıldığında anlam olarak birbirlerini tam olarak karşılamadıkları anlaşılmaktadır [18].

Fikir Madenciliğinde üç temel sınıflandırma seviyesi bulunur.

- Doküman Seviyesi
- Cümle Seviyesi
- Görüş Seviyesi

Doküman seviyesi bir dokümanın, pozitif yada negatif bir fikir/duygu ifade edip etmemesine göre sınıflandırılır. Tüm bir dokümanı tek bir fikir bilgisi olarak görür.

Cümle seviyesi bir dokümanda yer alan her bir cümle için fikir/duygu ifade edip etmediğini kontrol eder. Cümlenin subjektif veya objektif olmasına göre durumu belirlenir. Şayet cümle subjektif ise cümlenin pozitif veya negatif bir fikir/duygu ifade etme durumuna karar verilir [19].

Doküman ve cümle seviyelerindeki sınıflandırmalarda yorumların detaylandırılması zorunluluğu bulunmamaktadır. Ancak Görüş seviyesinde yorumlar detaylandırılır. Görüş seviyesi, duyguyu/fikri varlıkların belirli yönlerine göre segmente etmeyi amaçlar. İlk adım olarak, varlıkların ve özelliklerini belirlemek gerekir. Yorumcular aynı varlığın farklı özellikleri için farklı yorumlar da bulunabilirler. "Tabletin ses kalitesi hiç iyi değil fakat şarj ömrü oldukça uzun" gibi bir cümle bu duruma örnek olarak gösterilebilir [18].

Duygu sınıflandırma teknikleri Şekil 8.4'te verilmiştir [19].

Şekil 8.4'te de görüldüğü üzere duygu analizi çalışmaları makine öğrenmesi, veri sözlüğü tabanlı yaklaşım gibi yaklaşımları kullanır. Makine öğrenmesi, verilen bir problemi ortamdan edindiği bilgiye göre modelleyen Yapay Zeka disiplininin alt dalıdır. Makine öğrenmesi teknikleri denetimli ve denetimsiz öğrenme metotlarından meydana gelir. Daha önceki bölümlerde de bahsediliği üzere Denetimli öğrenme, önceden gözlemlenmiş ve sonuçları bilinen verileri kullanarak bu verileri ve sonuçlarını kapsayan bir fonksiyon oluşturmayı amaçlayan makine öğrenimi yöntemidir. Denetimsiz öğrenme ise, etiketlenmemiş verideki gizli yapıyı bulmayı amaçlar. Özetle, veriler arasında var olan ama gözle görülmeyen bağıntıyı açığa çıkarır [20], [21].

Şekil 8.4: Duygu Sınıflandırma Teknikleri

8.5 İlgili Çalışmalar

Twitter verileri kullanılarak duygu analizi alanında yapılan çalışmalar aşağıda özetlenmiştir.

Ikoro, Sharmina, Malik and Navarro 2018 yılında, denetimli öğrenme söz dizimine dayalı bir yaklaşımı optimize etmeye odaklanmışlardır. Enerji tüketicilerinden enerji sağlayıcılarına gönderilen tweetleri toplayarak yeni bir sosyal medya analizi uygulaması sunmuşlardır. Yapılan çalışmada elde edilen veriler pozitif, negatif ve nötr olarak 3 sınıfa ayrılmıştır. Sektöre yeni giren Üç enerji sağlayıcıya karşı Big Six (İngiltere'nin en büyük ve en eski gaz ve elektrik tedarikçileri) ile etkileşim halindeki tüketici tweetleri üzerindeki duygu analizi sonuçlarını karşılaştırmışlardır. Sonuçlar değerlendirildiğinde genel olarak, yeni giren enerji tüketicilerinin duygularının, Big Six'in tüketicilerinden gelenlerden daha olumlu olduğunu göstermişlerdir [22].

Siddiqua, Ahsan, and Chy 2016 yılında kural tabanlı bir sınıflandırıcıyı zayıf denetimli

Naive- Bayes sınıflandırıcıyla birleştirerek twitter üzerindeki duygu analizi için bir yaklaşım önermişlerdir. Tweetlerden elde edilen veriler ile duygu sınıflandırması için, kural tabanlı sınıflandırıcılar ve Naive-Bayes sınıflandırıcısını uygulamışlardır. Deneyleri Stanford sentiment140 veri setine dayandırarak gerçekleştirmişlerdir. Deney sonuçları, geri çekme, kesinlik, F1 skoru ve doğruluk açısından önerilen metodun taban çizgisi üzerindeki etkinliğini göstermiştir [23].

Kim, Lee and Kyeong tarafından 2013 yılında Amerika Birleşik Devletleri'nde (ABD) toplanan Twitter verilerini kullanarak Eyaletler arası sosyal konulara dayalı coğrafi kümeleme analizi çalışması yapmışlardır. ABD eyaletleri boyunca bir dizi konu kelimeyi zaman serileri ile ilişkilendirerek coğrafi topluluklar bulmuşlardır. Coğrafi kümelenme sonucunda Google Fusion Tablosu kullanarak görselleştirmişlerdir. Önerilen yöntem, gerçek zamanlı veri akışını analiz etmek ve coğrafi toplulukları bulmak için basit ama kullanışlı bir yaklaşım sunmaktadır [24].

Dedhia ve Ramteke 2017 yılında Ensemble modelinin güçlü bir temel öğrenci, yani SVM ile birleştirilmesine odaklanmışlardır. Ensemble sınıflandırma yöntemleri ile Pozitif ve Negatif etiketlerden verileri sınıflandırmak ve Precision, Recall ve F1 skorlarını tahmin etmek için uygulamışlardır [25].

Joshi, Simon ve Murumkar 2018 yılında bir işletme markasına yönelik bir kampanya yürütmek için akıllı bir yöntem önermektedir; bu sayede işletme sahibi pazardaki pozisyonunu belirlemektedir ve işinin ne kadar iyi (ya da kötü) olduğunu, veri madenciliğini yaparak ve çıkarımlarda bulunarak ortaya çıkarmaktadır. Böylece, işletme sahiplerine işlerine değer katma ve rekabet avantajı sağlama yeteneği kazandırmayı hedeflemektedirler. [26].

Chauhan, Sutaria ve Doshi 2018 yılında, semiyotiklerin, iletişim için sınırlı sözleri olan sosyal medyadaki duyguları ifade etmek için yaygın olarak kullanıldığını göstermişlerdir. Mevcut çalışmalarda semiyolojinin göz ardı edildiğini ya da gösterge bilim kabul edildiğinde duygu puanının belirlenmesinde kullanılmadığını öne sürmüşlerdir. Önceden, gösterge puanı belirlemesi gösterge bilim ve metin için ayrı ayrı yapıldığını ancak her ikisini birleştirerek ve daha doğru duygu puanını belirlemek için bir yöntem olması gerektiğini, Tweet'in duygularını da etkileyeceği için tweet'de kullanılan semiyotiklerin sıklığını göz önünde bulundurulması gerektiğini vurgulamışlardır [27].

Barnaghi, Breslin ve Ghaffari 2016 yılında metin sınıflandırma için iyi bilinen bir makine öğrenme metodu kullanarak twitter duyguları ve meydana gelen olaylar arasında bir ilişki aramışlardır. Twitter mesajlarında olumlu veya olumsuz bir duygu ortaya çıkarmaya çalışmışlar ve bu amaçla eğitimli bir yöntem oluşturmak için elle işaretlenmiş (pozitif / negatif) tweet'leri kullanmışlardır. Eğitimli model Bayesistik Lojistik Regresyon (BLR) sınıflandırma yöntemine dayanmaktadır. Subjektif veya objektif tweetleri tespit etmek için harici sözcükler kullanılmıştır. FIFA Dünya Kupası 2014'ü örnek olay incelemesi olarak kullanarak, kamuoyunun beklenmedik olaylara karşı yansımasını analiz etmek için Twitter Akış API'sini ve resmi dünya kupası hashtag'lerini madencilik, filtreleme ve tweet işlemleri kullanılmıştır [28].

Jose ve Chooralil 2016 yılında, sınıflandırıcı topluluk yaklaşımını kullanarak gerçek zamanlı bir twitter duygu analizörü uygulamışlardır. Makine öğrenme sınıflandırıcılarını lexicon tabanlı sınıflandırıcı ile birleştirmişler, politik verilerin doğru sınıflandırılması için SentiWordNet sınıflandırıcı, naif bayes sınıflandırıcı ve gizli markov model sınıflandırıcı gibi üç sınıflandırıcının avantajlarını kullanmışlardır. Böylece, gerçek zamanlı tweetlerden politik duyguları bulmak için yeni bir doğru duyarlılık sınıflandırıcısı geliştirmişlerdir. Daha sonra iki siyasetçiye karşı politik duyarlılık hesaplanmıştır. Aynı zamanda bu sınıflandırıcı twitter verilerindeki duygu analizi ile iki yeni yayınlanan filmi karşılaştırmak için kullanılmıştır [29].

Sharma ve Moh, 2016 yılında genel devlet seçimleri için kampanya döneminde, Hindistan'da beş ulusal siyasi partiye başvuruda bulunan bir aylık süre boyunca toplanan 42.235 tweet üzerine veri (metin) madenciliği gerçekleştirmiştir. Hem denetlenen hem de denetlenmeyen yaklaşımlardan yararlanmışlardır. Sınıflandırıcıyı oluşturmak ve test verilerini pozitif, negatif ve nötr olarak sınıflandırmak için Sözlük Tabanlı, Naive Bayes ve SVM algoritmasını kullanmışlardır. Twitter kullanıcılarının düşündükleri her bir Hint politik partisine karşı sınıflandırıcı çalıştırılmış ve Naive Bayes ve SVM algoritmalarının seçim sonuçlarını doğru tahmin ettiği seçim sonrası kanıtlanmıştır [30].

Linares, Herrera, Cuadros ve Alfaro 2015 yılında, kullanıcıların turist trafiğinin öngörücü bir aracını oluşturmak için Peru'ya seyahat etme arzusunu ortaya koyan tweet'lerin kullanımını kapsamaktadır. Bu çalışmayı yapmak için, web taraması kullanarak tweet'lerin otomatik olarak toplanması yapılmış ve duygu analizinin bir parçası olarak tweet'leri sıralamak için Naive Bayes algoritması kullanılmıştır [31].

Tripathi, Vishwakarma ve Lala 2015 yılında, insanların Twitter'da paylaştığı görüşlere ilişkin duygu analizi yapmak için sınıflandırma amaçlı veri madenciliği teknikleri kullanmıştır. Veri kümesini, yani doğal dilde olan twitter tweet'leri ve metin madenciliği teknikleri ile mutlu, hüzünlü ve nötr duyguları tahmin edebilen duyarlılık sınıflandırıcısı oluşturmuşlardır. Bu amaçla RapidMiner aracı kullanılmıştır [32].

Hodeghatta 2013 yılında Twitter'daki Hollywood filmlerinde ifade edilen duyguları analiz etmeyi amaçlamıştır. Bu sayede seyircilerin fikirleri, alışkanlıkları ve tercihleri daha iyi seyirci deneyimi ve piyasa davranışını anlamak için analiz edilmiş ve kullanılmıştır [33].

Garg, Garg ve Ranga 2017 yılında, 18 Eylül 2016'da gerçekleşen güvenlik güçlerinin bir grup terörist tarafından saldırıya uğradığı olay ile ilgili Twitter tweetlerini çıkararak tweet sonrası terör saldırısının duygularını ve hayatta kalma durumunu incelemiştir. Twitter'da yayınlanan verilerin bilgi akışını incelemek için son retweet, retweet sayısı, sık kullanılanların sayısı gibi faktörler kullanılmıştır. Retweetlerin sayısı arttıkça, erişimin de o kadar yüksek olduğu düşünülmüştür. Sonuç olarak, tweetlerin hayatta kalma durumunu yansıttığı düşünülmüştür [34].

Mishra, Rajnish ve Kumar 2016 yılında, Modi ji'nin Dijital Hindistan Kampanyası hakkında fikir ifade eden Twitter veri setinin duyarlılık analizini yapmaya çalıştılar. Çalışmada duygular toplandı ve Pozitif, Negatif veya Nötr olma durumlarına göre sınıflandırıldı. Twitter verileri, Twitter API kullanılarak analiz için toplandı. Duygu analizi, Makine Öğrenme ve Sözlük Tabanlı yaklaşım için yaygın olarak kullanılan iki yaklaşımdan farklı olarak, farklı kullanıcılar tarafından yayınlanan verileri analiz etmek için Sözlük Tabanlı yaklaşım kullanıldı. Daha sonra bu verinin polarite sınıflandırması yapıldı [35].

Abdullah ve Hadzikadic 2017 yılında, ABD cumhurbaşkanı Donald Trump hakkındaki Twitter tartışmalarını incelediler. Ayrıca, tartışmaların ardından tweetlerin adayı destekleyip desteklemediğini tespit edip edemeyeceklerini görmek için insanların Trump ile ilgili duygularını araştırdılar. Bu çalışmanın en önemli bulgularından biri tweetlerdeki negatif veya pozitif kutupların bu adayı desteklemenin iyi bir göstergesi olmadığını kabul etmek oldu [36].

Bilgin ve Şentürk 2017 yılında, Doc2Vec kullanarak Türkçe ve İngilizce Twitter mesajlarında duygu analizinin yapılmasını amaçlamışlardır. Doc2Vec algoritması, Semi-Supervised

öğrenme metodu kullanılarak Pozitif, Negatif ve Nötr olarak etiketlenmiş veriler üzerinde çalıştırılmış ve sonuçları kaydedilmiştir [37].

Parveen ve Pandey 2016 yılında, İş zekası hakkında bazı öngörülerde bulunmaya yardımcı olan tweet'lerle ilgili duygu analizi yapmıştır. Twitter web sitesinde bulunan yorumlar, geri bildirim ve yorumlar biçimindeki film veri kümesini işlemek için Hadoop Framework kullanılmıştır. Twitter verilerindeki duygu analizi sonuçları, olumlu, olumsuz ve tarafsız görüşler sunan farklı bölümler olarak gösterilmiştir [38].

Windasari, Uzzi ve Satoto 2017 yılında, özellikle GoJek (GoJek çevrimiçi ulaşım, özellikle toplu taşıma araçlarına ulaşımın zor olduğu veya trafik sıkışıklığının olduğu alanlarda birçok kullanıcı tarafından tercih edilmektedir.) çevrimiçi ulaşım hizmetleriyle ilgili Twitter gönderisine dayanan kamu duygularını tespit eden bir sistem önermişlerdir. Sistem tweetleri toplayıp, SVM algoritması kullanarak tweetleri analiz ederek bunları olumlu ve olumsuz duygulara ayırmaktadır [39].

Rezaei ve Jalali 2017 yılında, McDiarmid ağaç algoritmasını önermişlerdir. Hoeffding ağaç algoritması, madencilik veri akışlarında kullanılan en popüler araçtır. Bu ağaç algoritması için, Hoeffding'in sınırı, bir ayrıştırma özniteliğini seçmek için bir düğümde gerekli en küçük miktarda örneği bulmaktır. MacDiarmid'i Hoeffding ağaç algoritmasına bağlı olarak değiştirerek twitter'daki duygu analizi için McDiarmid ağacından gelen doğruluk oranı Hoeffding ağacından daha yüksek olması sağlanmıştır; bununla birlikte, işlem süresi önemli ölçüde azalmıştır [40].

Rane ve Kumar 2018 yılında, 6 büyük Amerikan Havayolu Şirketinin tweetlerinden oluşan bir veri kümesi üzerinde çalışmış ve çok sınıflı duygu analizi gerçekleştirmişlerdir. Analiz 7 farklı sınıflandırma stratejisi kullanılarak gerçekleştirilmiştir: Karar Ağacı, Rastgele Orman, SVM, K-En Yakın Komşu, Lojistik Regresyon, Gauss Naive Bayes ve AdaBoost. Sınıflandırıcılar, verilerin % 80'i kullanılarak eğitilmiş ve kalan % 20'lik veriler test edilmiştir. Test setinin sonucu tweet hissi pozitif, negatif, nötr şeklinde olmuştur. Elde edilen sonuçlara göre, her bir sınıflandırma yaklaşımı arasında bir karşılaştırma yapmak için doğruluk değerleri hesaplanmış ve altı havayolu şirketinin tümünü birleştiren toplam duygu sayımı görselleştirilmiştir [41].

Jain ve Katkar 2015 yılında, veri madenciliği sınıflandırıcılarını kullanarak twitter kullanıcılarının duygu analizi üzerinde çalışmıştır. K en yakın komşu, Random Forest,

BaysNet ve Naive Bayes algoritmaları kullanmışlardır. Elde edilen sonuçlar, k-en yakın komşu sınıflandırıcısının çok yüksek tahmin doğruluğu verdiğini göstermiştir. Sonuç ayrıca, tek sınıflandırıcıların sınıflandırıcı onayı grubundan daha iyi performans gösterdiğini ortaya koymuştur [42].

Subramaniam, Ranjitha, Aswini ve Kumar 2017 yılında, anket yolu ile elde edilen twitter verilerinin analizi üzerinde çalışmıştır. Twitter'da yaygın olarak kullanılan bilgiler, anketin uygulandığı web sitesinde trend olan bir konu olarak sunucuda güncellenmiştir. Twitter'daki veriler analiz edilmiş ve her bir tweet tarafından verilen reaksiyonlar incelenmiştir. Her tweetde verilen farklı reaksiyon türlerini bulmak için duygusal analiz kullanılmıştır. Çalışmanın ana amacı, trend olan konuların görüntülendiği bir website ortaya koymaktır [43].

Nirmala, Roopa ve Kumar 2015 yılında, twitter sitesinde yayınlanan bilgileri işlenmemiş etiketler kullanarak işlemeye odaklanmışlardır. Önerilen sistem, geniş bir yelpazede mevcut API'lar aracılığıyla Twitter'daki veri madenciliği desteğiyle Twitter'dan veri almayı kolaylaştıran R dili kullanılarak gerçekleştirilmiştir. Her tweet puanlarını atamak için metin sözlüğü kullanılmıştır. Önerilen bu yöntem ile işsizlik oranının analizi gerçekleştirilmiştir [44].

Shital ve Anil Phand 2017 yılında, hizmet sağlayıcı firmaların veya endüstrilerin, müşterilerinin ürünlerine, hizmetlerine ya da teklifleri hakkında görüşlerini bulmalarına yardımcı olan yaklaşımı üzerinde çalışmıştır. Araştırmanın sonucu, ürün veya hizmetin ilgili müşteriler, tarafından nasıl algılandığı ortaya koymuştur [45].

Hao ve diğerleri 2011 yılında, Twitter ve Twitter analizlerini, gerçek zamanlı Twitter veri akışlarını araştırmak için coğrafi ve zamana dayalı etkileşimli görselleştirmelerle birleştiren Twitter zaman çizelgelerinin görsel analizi üzerinde çalışmışlardır. Bugünün görsel analiz araçlarının (örn. SAS JMP, Vivisimo, Polyanalyst vb) esas olarak, evet / hayır soruları, sayısal derecelendirmeleri ve doğrudan yorumlarını kullanarak incelemeler hakkında geri bildirim sağlamışlardır [46].

Alkalbani ve diğerleri 2017 yılında, tüketicilerin bulut hizmetleriyle ilgili deneyimlerini yansıtan bulut tüketicileri tarafından yapılan incelemeleri araştırmıştır. Her bir gözden geçirmenin tutumunu belirlemek ve ifade edilen görüşün olumlu, olumsuz ya da tarafsız olduğunu belirlemek için yaklaşık 6.000 bulut hizmeti kullanıcısı tarafından yapılan

incelemeler, duygu analizi kullanılarak analiz edilmiştir. Analizde iki veri madenciliği aracı, KNIME ve RapidMiner kullanılmıştır ve sonuçlar karşılaştırılmıştır [47]

Baydoğan ve Alataş 2018 yılında, KNIME ile twitter verileri üzerine çalışmışlardır. Aslında bir sınıflandırma çalışması olan duygu analizi çalışması, Twitter verilerinde makine öğrenme algoritmaları kullanılarak gerçekleştirilmiştir. Çalışmanın sonuçları bir doğruluk analizi yapılarak yorumlanmıştır. Zengin görselleştirme araçlarına sahip olan KNIME'nin kullanımının, bu çalışmaları hem daha kolay hem de daha güvenilir hale getirmek için duyarlılık analiz çalışmalarında yaygınlaşacağı öngörülmektedir [48].

Bölüm 9

Uygulama

Bu bölümde Twitter mesajları üzerinde makine öğrenmesi yöntemlerine dayalı veri madenciliği algoritmaları ile duygu analizi yapılması amaçlanmıştır.

9.1 Veri Seti

Uygulama gerçekleştirilirken günümüzün en popüler ağlarından biri olan twitter tercih edilmiştir. Twitter belli başlı konular hakkında güncel bilgilere ulaşılmasını sağlayan gerçek zamanlı bir bilgi ağıdır. Düşünce paylaşımı olarak düşünüldüğünden her bir tweet en fazla 140 karaktere sahip olabilmektedir. Sadece kişisel kullanıcılar değil, kurumsal şirketlerde bu ağın kullanıcısı olabilmektedir. Şirketlerin amacı, müşterileri ile iletişim kurarak ürün veya hizmetleri hakkında güncel bilgi ve yeniliklerini, faaliyetlerini paylaşabilmektir. Aldıkları geri bildirimler aracılığıyla müşteri memnuniyet oranlarını arttırabilmektir.

Twitter'da paylaşılan bilgiler herkese açık olabileceği gibi sadece takipçilere özel olacak şekilde de ayarlanabilmektedir. Twitter, herkese açık olan verilere twitter.api uygulamasıyla ulaşılmasına izin vermektedir. API aracılığıyla elde edilen Keys ve Tokens kullanılarak (Consumer API Keys, Access token and Access token secret) istenilen filtreler ve ayarlar ile herkese açık hesapların tweetlerine erişilebilmektedir.

9.2 Uygulamada Kullanılan Program

Uygulamada günümüzde hızla yaygınlaşan ve kolay kullanıma sahip bir platform olan KNIME kullanılmıştır. KNIME, 2004 yılında Konstanz Üniversitesi'nde Java dilinde geliştirilmiştir [49].

Knime, veri madenciliği yapan bir yazılımdır. Akış mantığı ile çalışır ve tamamen ücretsizdir. Temel amaç bir veri kaynağından bir hedefe/bilgiye akış sağlamaktır. Müşteri analizi, kampanya analizi, tahminler, farklı kaynaklardan gelen verilerin birleştirilmesi vb. gibi pek çok amaçla kullanılabilir. Knime tüm bunları karşılayabilen ücretsiz bir veri madenciliği programıdır [50].

Açık bir platform olması sebebiyle geliştiricilerin yeni bileşenler üretmesine olanak sağlamaktadır. Aynı zamanda kullanıcılarına Weka, R Project, LIBSVM, ImageJ gibi diğer açık kaynaklı programların olanaklarına KNIME içerisinden erişim imkanı sunmaktadır [49].

KNIME çalışma ortamı, veri madenciliği çalışmaları için gerekli olan tüm temel gereksinimleri karşılar. KNIME'da yapılacak tüm veri madenciliği çalışmaları bir iş akışı ile tanımlanır ve her bir iş akışı birbiri ile ilişkilendirilmiş düğümlerden oluşur. Düğümler arası ilişkiler kurularak iş akışı oluşturulur ve her bir iş akışı birbirinden bağımsızdır.

KNIME aynı zamanda dünyada en yaygın kullanılan veri madenciliği araçlarından biri olan WEKA'nın algoritmalarını küçük bir eklenti yükleme işlemi ile kullanabilmektedir [49].

KNIME'da çalışmak için yeni bir workflow penceresi açılması gerekir. Yeni açılan workflow ismi ve yeri bu pencerede belirlenir. Aktarılacak dosya tipine ve yapılması planlanan işlemlere göre node repository bölümünden uygun düğümler seçilip sürüklenerek workflow penceresine bırakılır. Seçilen düğüme sağ tıklayarak configure bölümünden gerekli ayarlamalar yapılır. Kullanılması gereken düğümler node repository bölümünden sırayla seçilip workflow'a aktarıldıktan sonra her biri için configure bölümünden gerekli ayarlamalar yapılır ve ardından sistem çalıştırılır.

Düğümlerin altındaki yanan kırmızı, sonra sarı sonra yeşil ışıkların anlamları vardır. Kırmızı, problem olduğunu gösterir ve örneğin verilen veride ya da configure'de bir problem

olabilir ve bunları değiştirmekle giderilebilir. Sarı, bekleme durumudur yani çalıştırılmayı beklemektedir. Yeşil, herhangi bir problem yok başarılı biçimde çalıştı demektir. Output Table ise sistem çalıştırıldıktan sonra oluşan sonucu göstermektedir [50].

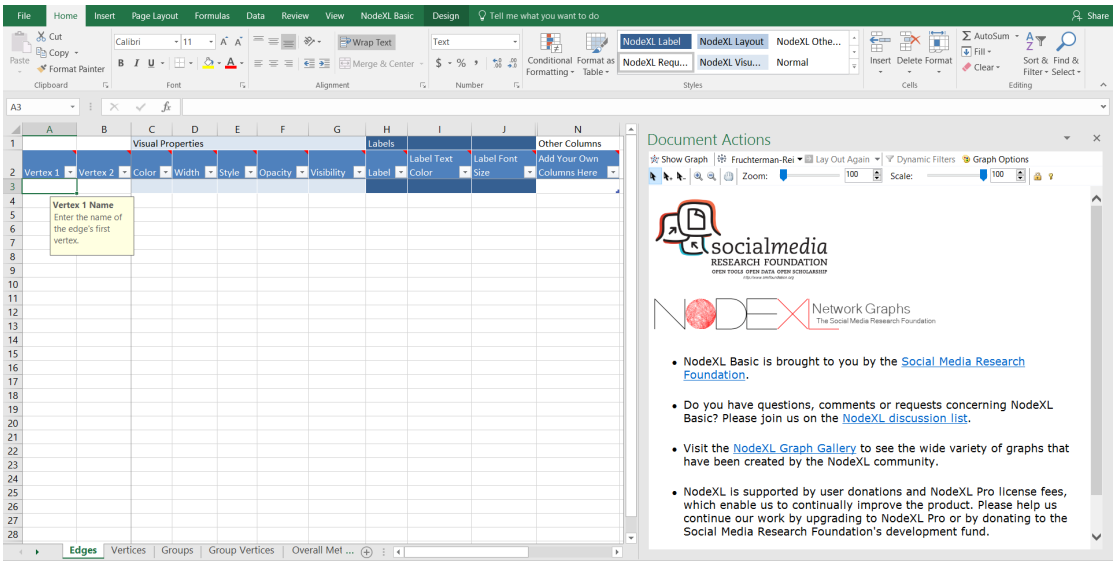
9.3 Yapılan Çalışma

Tez kapsamında uygulama için yapılan ilk adım sosyal ağ verilerinin yani Twitter mesajlarının elde edilmesi işlemi olmuştur.

Bir sonraki adım twitterdan ne tür verilerin alınacağına karar verilmesi olmuştur. Çalışmada şirketlerin sosyal ağlardaki itibarını ölçmek amaçlanmıştır. Bu nedenle 5 Türk Savunma Sanayi Şirketi hakkında yazılan tweet mesajları 01.01.2018-31.01.2019 tarih aralığı için incelenmiştir.

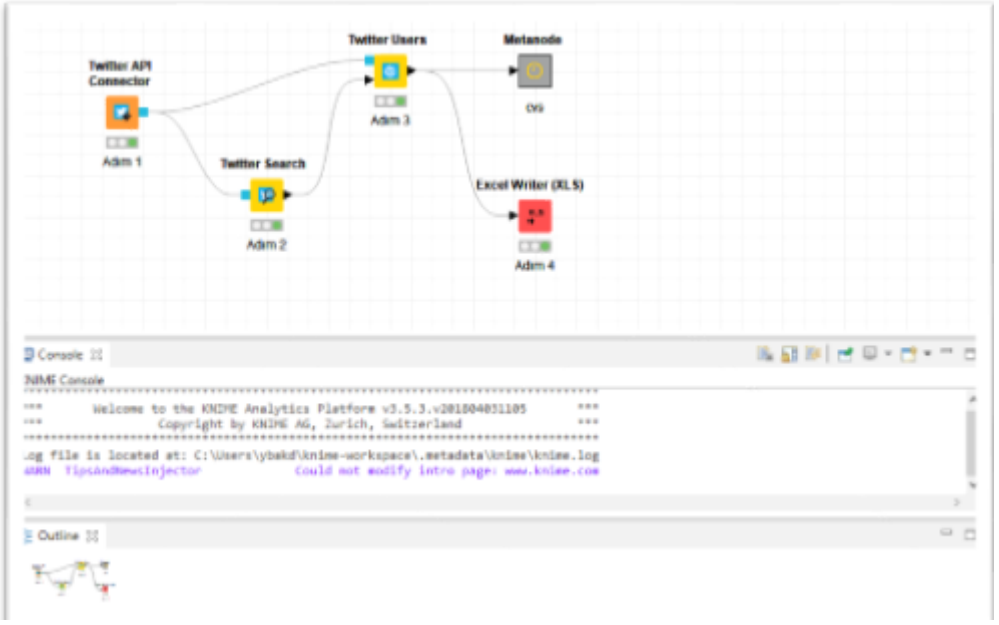
İncelemesi yapılan konu spesifik bir bir başlık olduğundan elde edilen tweet mesajları sayısını arttırmak amacıyla, analizler için kullanılacak olan KNIME programına ek olarak verilerin elde edilmesi aşamasında NodeXL ve Java Programlama Dilinde yazılmış bir kod kullanılmıştır. Kullanılan kod, Ek A'da verilmiştir.

NodeXL (Network Overview Discovery and Exploration add-in for Excel 2007, 2010, 2013) Microsoft Excel'in Office 2007 ve sonrası sürümleri üzerinde bir eklenti olarak çalışan bir sosyal ağ analizi aracıdır. Dünyanın önde gelen üniversitelerindeki (Microsoft Research, University of Maryland, Cornell University, Stanford University, Oxford Internet Institute) akademisyen ve araştırmacıların bir araya gelerek ücretsiz ve açık kaynak kodlu kullanım konseptiyle tasarlamış olduğu ve sosyal ağ uygulamaları için özelleştirilmiş bir araçtır. Node XL ile Facebook, Twitter, Flickr ve YouTube üzerinden veri çekebilmek mümkün olsa da; en başarılı performansı Twitter verileri üzerinden vermektedir. NodeXL ile sosyal ağlar aracılığıyla elde edilen veriler, çeşitli algoritmalarla görselleştirilebilmekte ve gerekli istatistikler hesaplanarak ağlar analiz edilebilmektedir. Bütün bunlar doğrudan Microsoft Excel ekranında yapılabildiği için programlama bilgisi olmayan kullanıcılara da kullanım kolaylığı sağlamaktadır. NodeXL'in arayüzü Şekil 9.1 'de verilmiştir.



Şekil 9.1: NodeXL'in Arayüzü

Verilerin elde edilmesi işlemi için KNIME Programı üzerinde oluşturulan akış Diyagramı ise Şekil 9.2'deki gibidir:

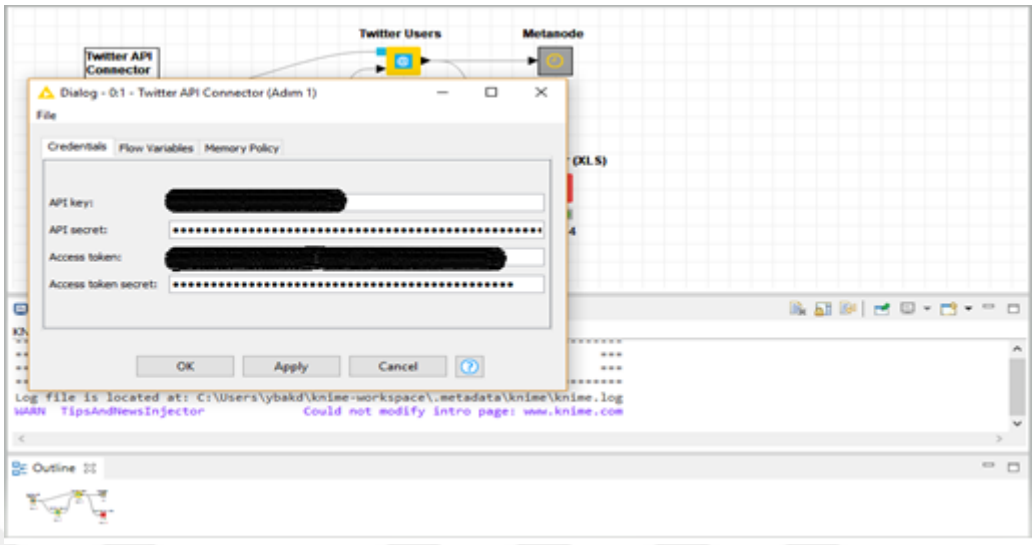


Şekil 9.2: KNIME Analytics Platform v3.5.3 Twitter Verilerinin Alınması için Oluşturulan Akış Diyagramı

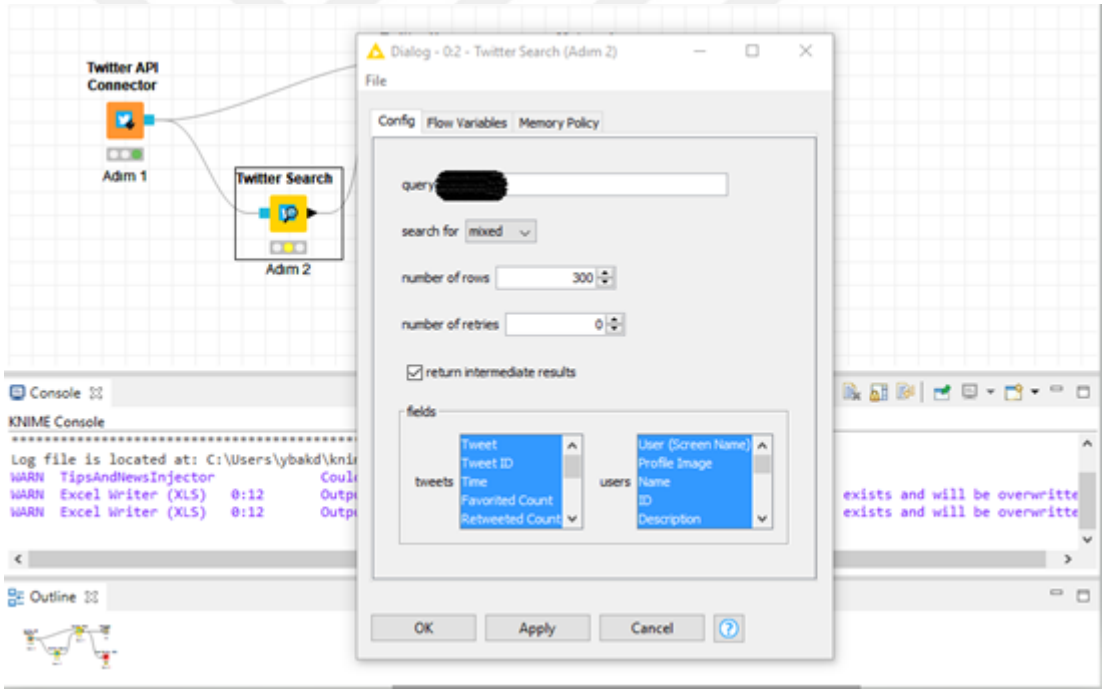
Twitter verilerinin alınması için oluşturulan akış diyagramında "Twitter API Connector", "Twitter Search", "Twitter Users" ve "Excel Writer" düğümleri (operatörleri) kullanılmıştır.

Oluşturulan akış diyagramını çalıştırmak için Twitter API aracılığıyla oluşturulan API Key, API secret, Access Token ve Access Token Secret değerleri Twitter API Connector

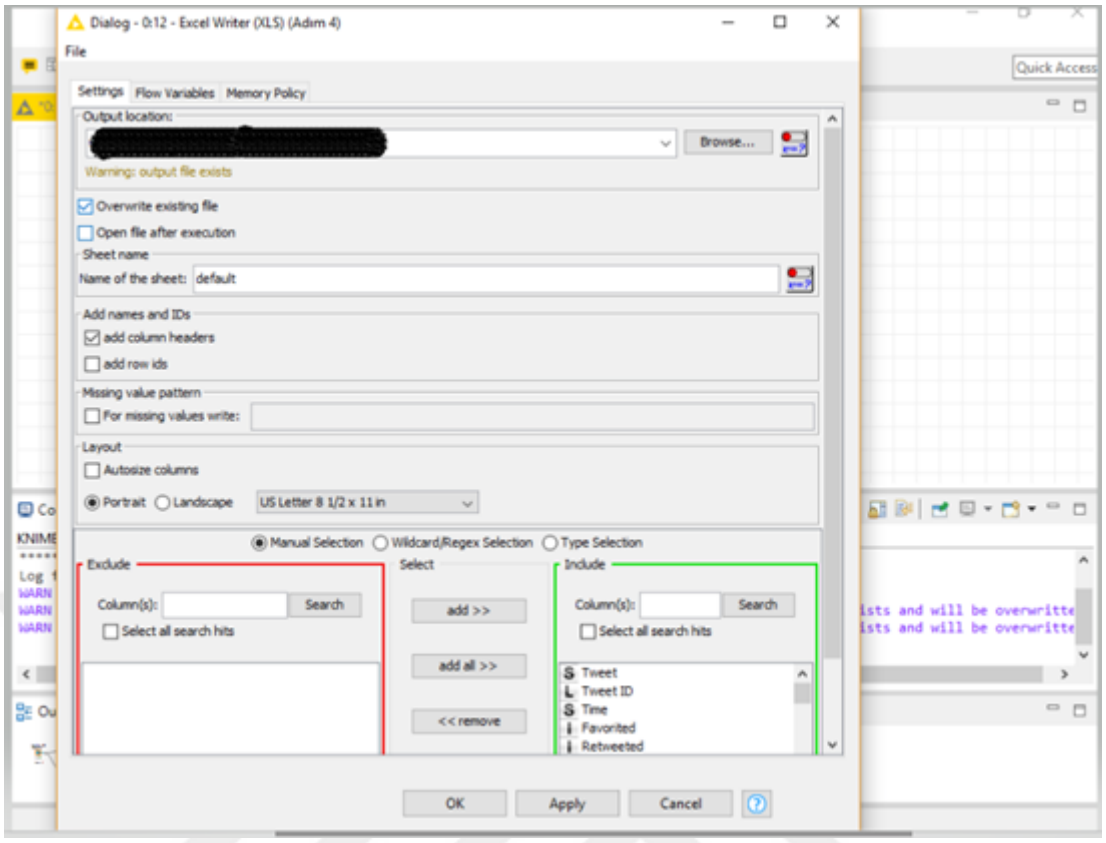
düğümünün konfigürasyon kısmında uygulanmıştır.



Şekil 9.3: KNIME Analytics Platform v3.5.3 Twitter API Connector Ayarı



Şekil 9.4: KNIME Analytics Platform v3.5.3 Twitter Verilerinin Alınması



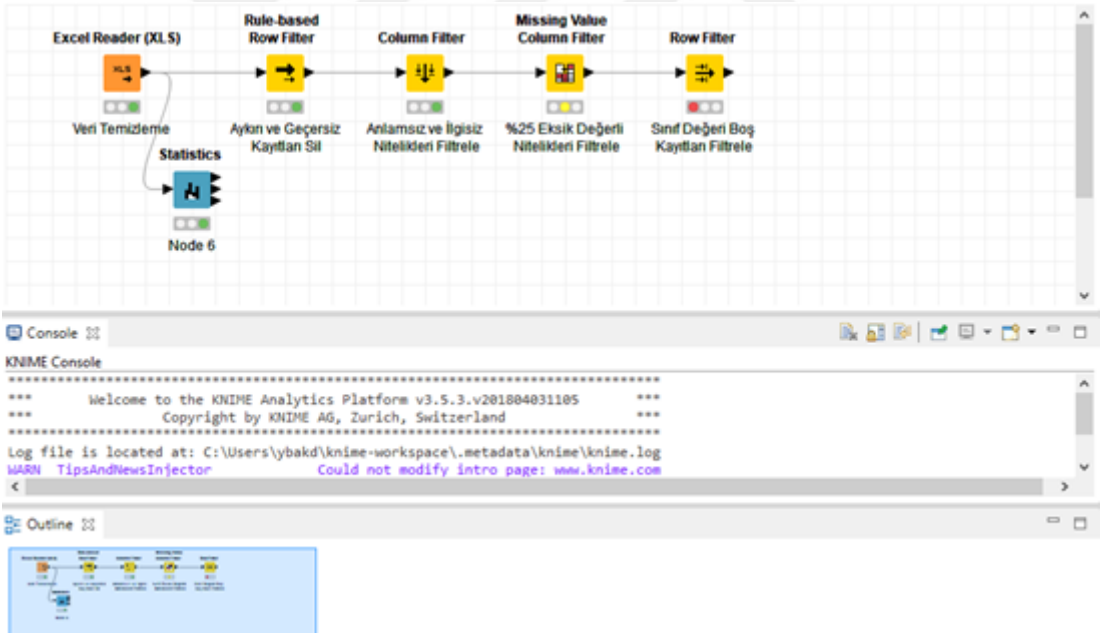
Şekil 9.5: KNIME Analytics Platform v3.5.3 Twitter Verilerinin Kaydedilmesi

Kullanılan programlardan (KNIME, NodeXL, Java Kodu) her bir şirket için tweet mesajlarının elde edilmesinin ardından yaklaşık 15.000 tweet mesajı excell olarak kaydedilmiştir. Yazılan tweet mesajlarına ek olarak kullanıcılara ait aşağıdaki bilgiler de kayıt altına alınmıştır.

- Kullanıcı Adı
- Tarih
- Cinsiyet

Her bir şirket için veri setinin hazırlanmasının ardından verilerin sınıf etiketleri oluşturulmuştur. Savunma Sanayi şirketlerinin sosyal ağlardaki durumu Olumlu ve Olumsuz olmak üzere iki sınıf etiketi üzerinden değerlendirilmiştir. Kullanıcıların tweet mesajları içerisinde, mutluluk ifadesi barındıran kelimelerin varlığı durumunda tweet pozitif olarak etiketlenirken; üzgünlük ifadesi barındıran kelimelerin varlığı durumunda tweet negatif olarak etiketlenmiştir. Mutluluk ve üzgünlük belirten kelimeler, "Sosyal Psikolojide Duygusal Durumlar" [51] isimli kaynaktan yararlanılarak oluşturulmuştur. Kelime seti içinden 100 adet kelime/kelime grubu mutluluk ve üzgünlük belirten kelime olarak seçilmiştir. 48 adet kelime mutluluk, 52 adet kelime üzgünlük bildiren kelime olarak Ek B'deki Tablo B.1'de gösterilmektedir.

Bir sonraki aşamada etiketlenmiş verilere, önişleme adımları uygulanmıştır. Veriyi önişlemden geçirmek, yani filtrelemek eksik satırlardan, kirli yada gürültülü verileri temizlemek anlamına gelir. Temizlemedeki amaç makine öğrenme algoritmalarının başarısını arttırmaktır. Veri Önişleme işlemi için oluşturulan akış diyagramı ise Şekil 9.6'teki gibidir.



Şekil 9.6: KNIME Analytics Platform v3.5.3 Veri Önişleme İşlemi için Akış Diyagramı

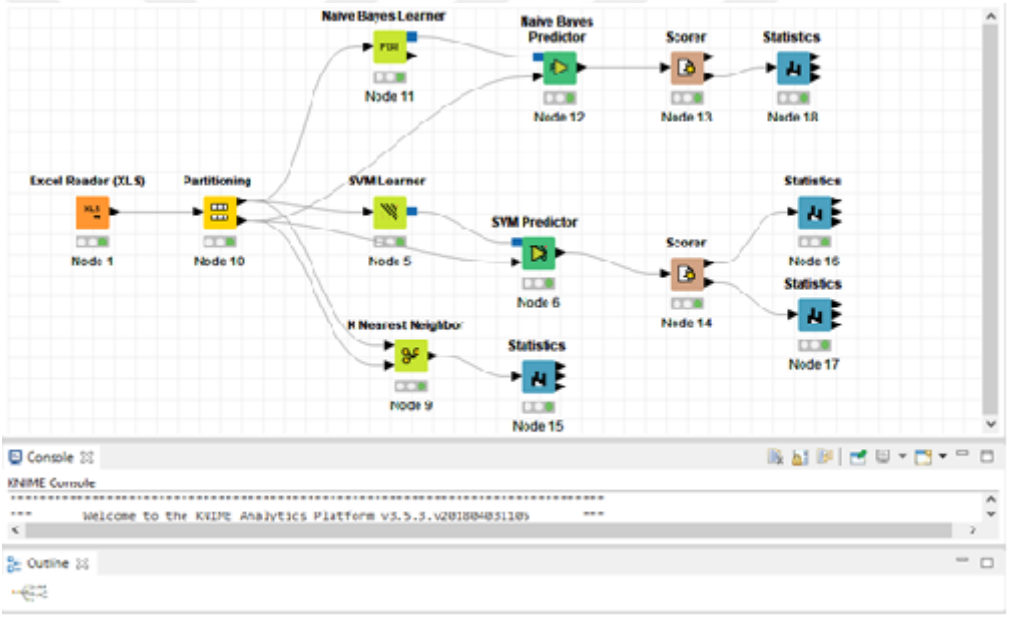
Veri önişleme için kullanılan düğümlerin (operatörlerin) açıklamaları aşağıdaki gibidir.

Rule-based Row Filter: İleri düzey satır filtreleme anlamına gelir. Row Filter'dan farkı kural yazılmasına imkan veriyor olmasıdır. Bu düğüm sayesinde veri seti içindeki aykırı ve geçersiz kayıtlar silinmiştir.

Column Filter: Veri setinde herhangi bir sebepten dolayı bir veya birden fazla kolonun kullanılmayacağı durumda column filter ile filtreleme yapılır. İlgisiz kolonlar çıkarılmıştır. Column Filter, seçilen kolonun komple silinmesine imkan tanıdığı gibi configure edildiği zaman silinmesi istenilen seçeneklerin elenmesine de olanak vermektedir.

Row Filter: Seçilen kolondaki eksik verilerin filtrelenmesi veya alt ve üst sınırların dışında kalan verilerin filtrelenmesi amacıyla kullanılır.

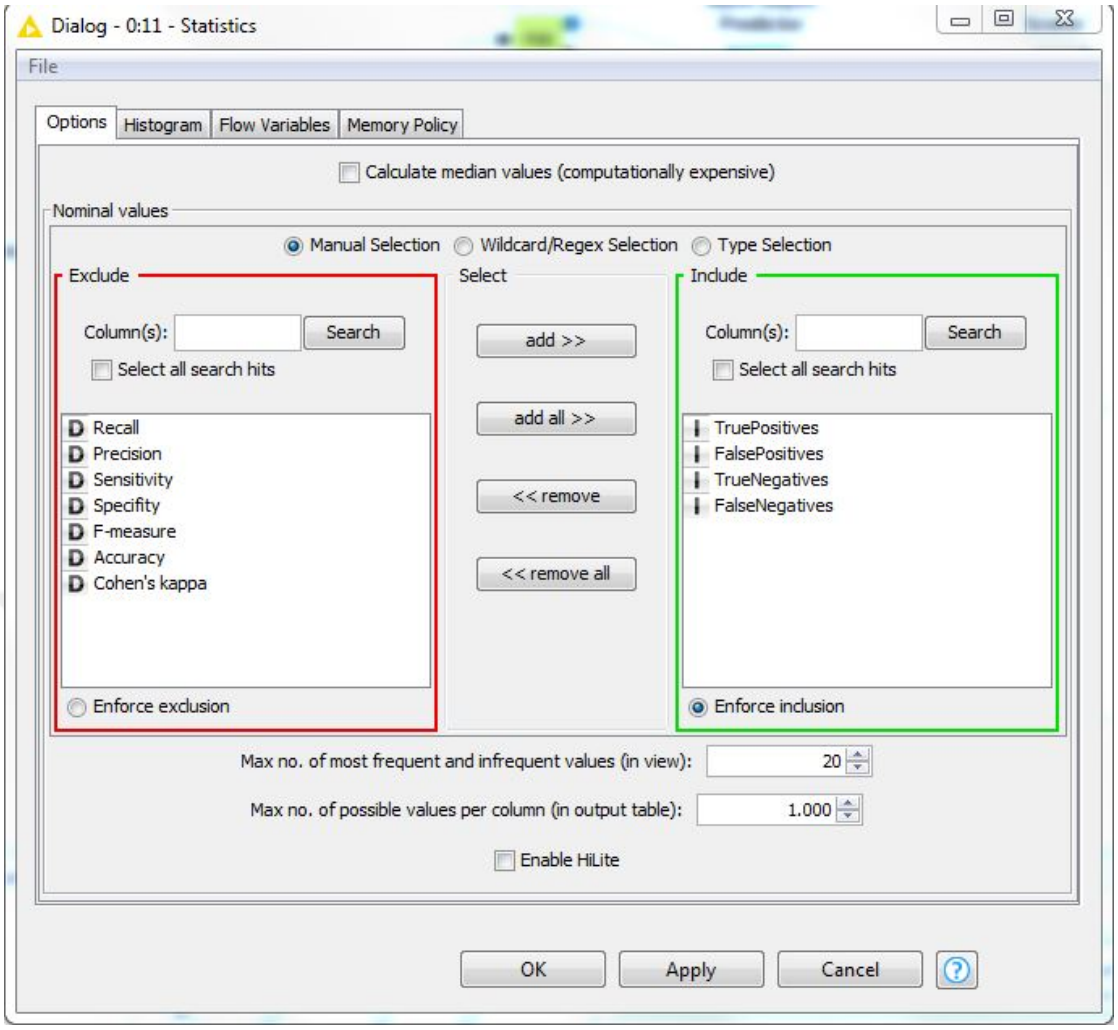
Önişleme işlemlerinin tamamlanmasından sonraki aşamada Sınıflandırma Algoritmalarına geçilmiştir. 10.780 adet tweet mesajı için uygulamada sınıflandırmada en çok tercih edilen Sade Bayes (Naive Bayes), Destek Vektör Makineleri (Support Vector Machine) ve K En Yakın Komşu (K Nearest Neighbor) sınıflandırma algoritmaları kullanılmıştır. Sınıflandırma için oluşturulan akış diyagramı Şekil 9.7'daki gibidir.



Şekil 9.7: KNIME Analytics Platform v3.5.3 Sınıflandırma Algoritmaları için Akış Diyagramı

Sınıflandırma için oluşturulan akış diyagramında kullanılan düğümlerin açıklamaları ve kullandıkları parametreler aşağıdaki gibidir:

Statistics: İstatistik düğümü, bir veri kümesine ait temel istatistik ölçülerini (ortalama, maks., min., varyans, orta değer vb.) görüntüleye yarar. İstatistik düğümünün (operatörünün) konfigürayon penceresi Şekil 9.8'deki gibidir.



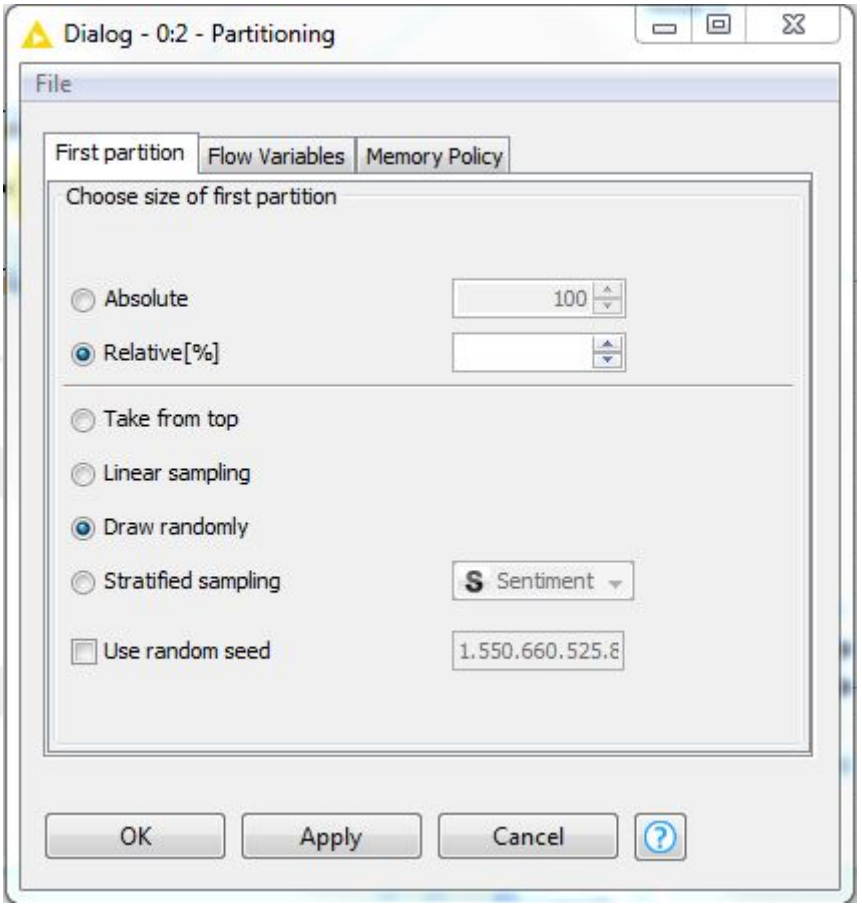
Şekil 9.8: KNIME Analytics Platform v3.5.3 Statistics Configure

Partitioning: Bölümleme düğümü giriş portuna aktarılan veriyi, ayarlanan oranlarda (60/40, 70/30 vb.) iki parçaya bölmek için kullanılır. Bu parçalardan biri eğitim; diğeri ise eğitilen modelin testi için kullanılır. Bu bölümlemenin sadece oransal durumu değil, her bir parçaya düşecek olan verilerin nasıl seçileceği de bu düğümde ayarlanabilmektedir.

- Üstten Almak (Take frop top): Bu mod, en üstteki satırların belirlenen oranını eğitim tablosuna, kalan oranı ise test tablosuna koyar.
- Doğrusal Örnekleme (Linear Sampling): Bu mod her zaman ilk ve son satırı içerir ve kalan satırları tüm tablo boyunca doğrusal olarak seçer (örneğin, her üçüncü satır). Minimum ve maksimum değerleri korurken sıralı bir sütunu tekrar örneklemek için kullanışlıdır.
- Rastgele Dizilim (Draw randomly): Tüm satırların rastgele örneklemesi.

- Tabakalı Dizilim (Stratified sampling): Tabakalı örnekleme yapmak için bu mod kullanılır.

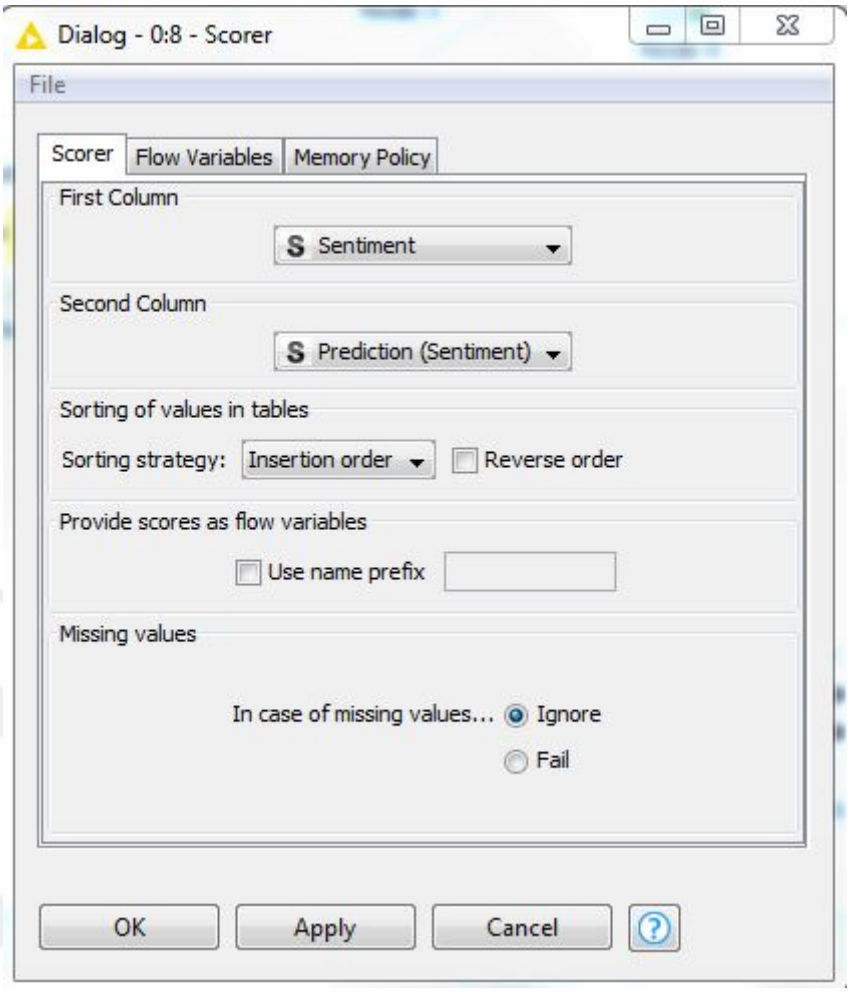
Partitioning düğümünün (operatörünün) konfigürayon penceresi Şekil 9.9'daki gibidir.



Şekil 9.9: KNIME Analytics Platform v3.5.3 Partitioning Configure

Scorer: Değerlendirici düğümü, kendisine aktarılan verideki iki sınıf değerini (biri gerçek, diğeri tahmin edilen) karşılaştırarak tahminlerin doğruluğunu hesaplar. Çıkış potlarından ilki doğruluk matrisini, diğeri ise doğruluk tablosunda yer alan istatistikleri dışarıya aktarır.

Scorer düğümünün (operatörünün) konfigürayon penceresi Şekil 9.10'daki gibidir.

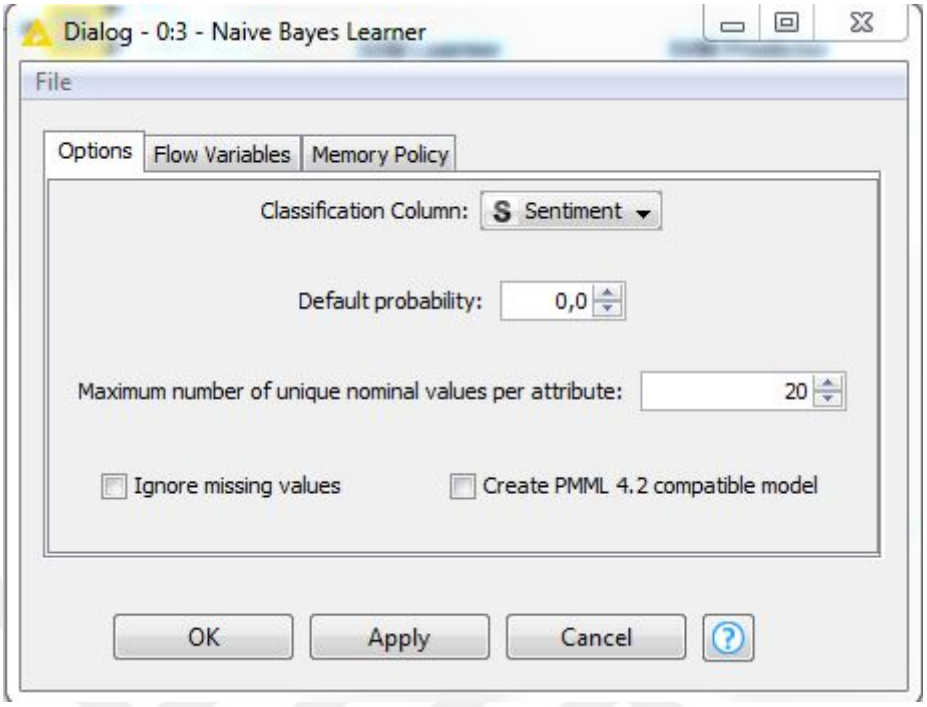


Şekil 9.10: KNIME Analytics Platform v3.5.3 Scorer Configure

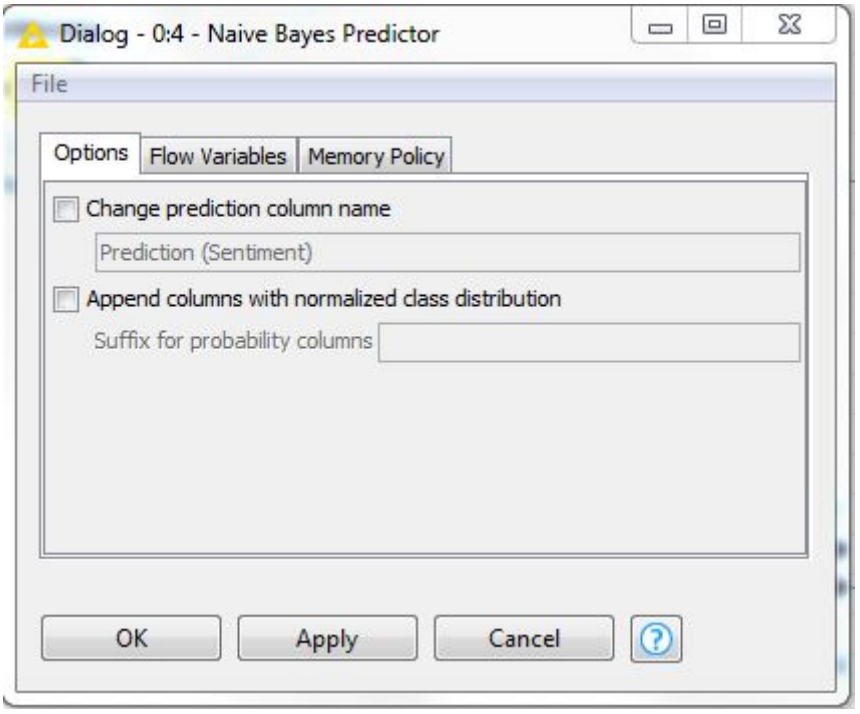
Naive Bayes Sınıflandırıcısı: Bayes teorisinin veri kümesindeki herbir niteliğin sınıf niteliği ile ilişkisinin diğer niteliklerden bağımsız olduğunu varsayarak uygulanmasına Sade (Naif) Bayes Sınıflandırıcı adı verilir. Sade Bayes Sınıflandırıcısı nominal değerlerde daha başarılıdır [49].

Uygulamada kullanılan Naive Bayes Düğümü, verilen eğitim verilerinden bir Bayesian modeli oluşturur. Oluşturulan model, sınıflandırılmamış verilerin sınıf üyeliğini tahmin etmek için saf Bayes belirleyici kullanır. Desteklenmeyen veri türlerinden dolayı herhangi bir sütun göz ardı edilirse, düğüm bir uyarı mesajı görüntüler.

Naive Bayes düğümünün (operatörünün) konfigürayon pencereleri Şekil 9.11 ve 9.12'deki gibidir.



Şekil 9.11: KNIME Analytics Platform v3.5.3 Naive Bayes Learner Configure



Şekil 9.12: KNIME Analytics Platform v3.5.3 Naive Bayes Predictor Configure

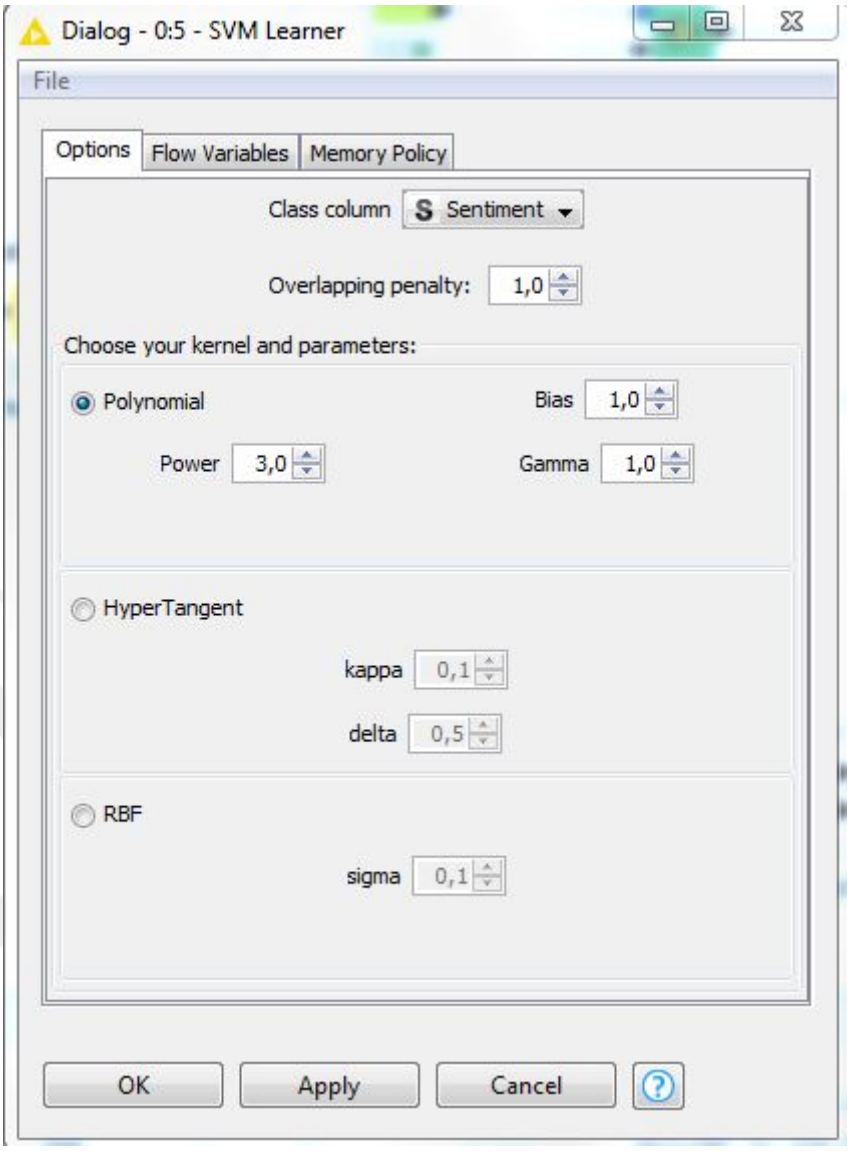
Naive Bayes Learner düğümü için konfigürasyon açıklamaları şu şekildedir:

- Classification Column, Sınıflandırma yapılması istenen kolon belirlenir.
- Maximum number of unique nominal values per attribute, Sınıflandırma için kullanılması planlanan maksimum nominal değer içeren kolon sayısı belirlenir.
- Default probability, Varsayılan olasılık düzeltme olmadığı durumlarda sıfır olarak ayarlanır.
- Ignore missing values, Düğüm tahmin sonucunu iyileştirmek için eksik değer bilgilerinin yok sayılmasıdır.
- Create PMML 4.2 compatible model, PMML 4.2 standardıyla uyumlu bir model oluşturmak için bu seçenek kullanılır. PMML 4.2 standardı, eksik değerleri yok sayar ve bit vektörlerini desteklemez. Bu nedenle, bit vektör sütunları ve eksik değerler, eğer bu seçenek seçiliyse öğrenme ve tahmin sırasında göz ardı edilir.

Support Vector Machine (SVM) Sınıflandırıcısı: Destek vektör makineleri, sınıflandırma ve regresyon analizi için kullanılan veriyi analiz eden ilişkili öğrenme algoritmalarıyla çalışan denetimli öğrenme modelleridir [52]

Uygulamada kullanılan SVM düğümü giriş verisi üzerinde bir destek vektörü makinesi eğitir. Bir dizi farklı çekirdeği (HyperTangent, Polynomial ve RBF) destekler. SVM Learner, birden fazla sınıf problemini de destekler (her sınıf ve diğer sınıflar arasındaki hiper düzlemi hesaplayarak), ancak bu çalışma süresini arttırır.

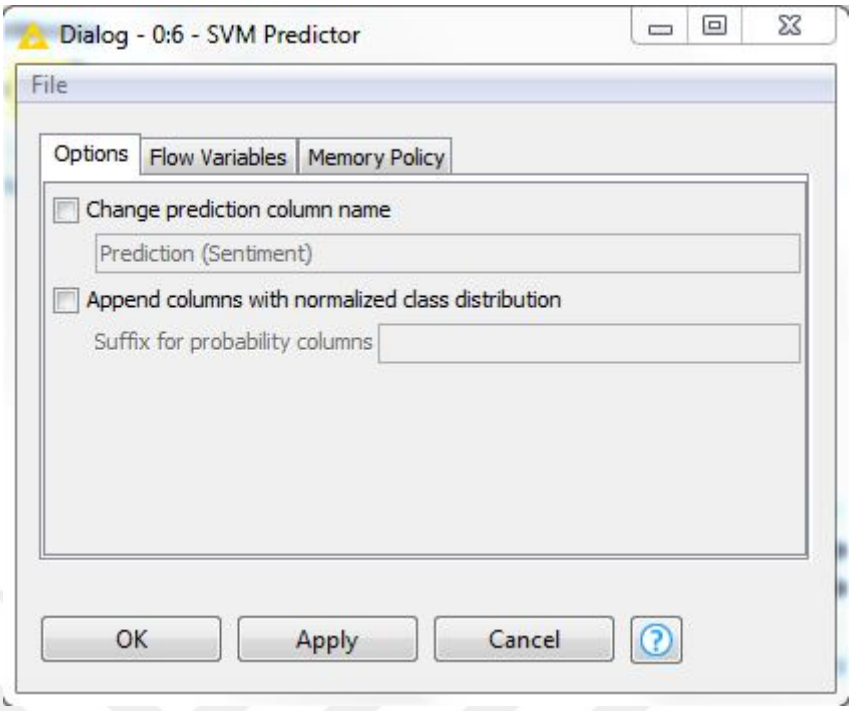
SVM düğümünün (operatörünün) konfigürayon pencereleri Şekil 9.13 ve 9.14'deki gibidir.



Şekil 9.13: KNIME Analytics Platform v3.5.3 SVM Learner Configure

SVM Learner düğümü için konfigürayon açıklamaları şu şekildedir:

- Class column, Nominal hedef değişkenini içeren sınıf seçilir.
- Overlapping penalty, Giriş verilerinin ayrılmaması durumunda faydalıdır. Yanlış sınıflandırılan her bir noktaya ne kadar ceza verileceğini belirler. Bunun için 1 iyi bir değerdir.



Şekil 9.14: KNIME Analytics Platform v3.5.3 SVM Predictor Configure

- Kernel type, Kernel ve parametre seçimi yapılır.Her çekirdeğin kendine ait parametreleri bulunmaktadır.

  Polynomial Kernel: Makine öğreniminde, destek vektör makinelerinde (SVM) polinom çekirdeği yaygın olarak kullanılan bir çekirdek işlevidir. Polinomiyal çekirdek işlemi doğrusal olmayan bir çekirdektir. Polinomiyal çekirdeği, tüm eğitim verilerinin normalleştirildiği problemler için daha uygundur. İlgili çekirdek fonksiyonu aşağıdaki formül ile ifade edilir [53]:

$$k(x, y) = (ax^T y + c)^d \tag{9.1}$$

  alfa eğimi, c sabit terimi ve d polinom derecesini ifade eder.

  HyperTangent Kernel: Hiper Tanjant (Sigmoid) Çekirdeği, Yapay sinir ağlarında aktivasyon fonksiyonu olarak da kullanılır. İlgili çekirdek fonksiyonu aşağıdaki formül ile ifade edilir [53]:

$$k(x, y) = tanh(ax^T y + c) \tag{9.2}$$

RBF Kernel: Makine öğreniminde RBF çekirdeği, çeşitli çekirdek öğrenim algoritmalarında kullanılan popüler bir çekirdek işlevidir. Özellikle, SVM Sınıflandırmasında yaygın olarak kullanılır. İlgili çekirdek fonksiyonu aşağıdaki formül ile ifade edilir [53]:

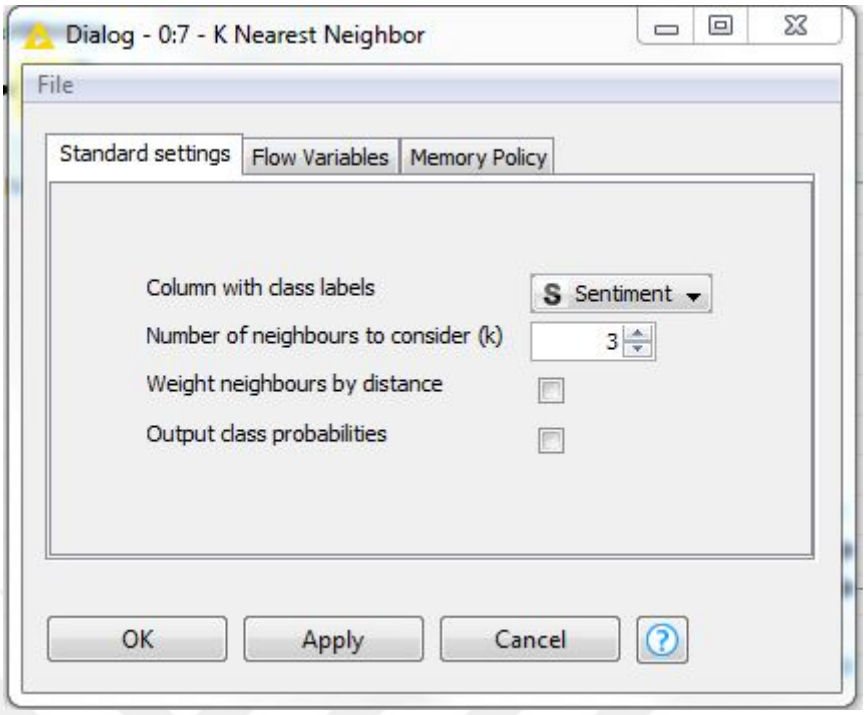
$$k(x, y) = exp(\frac{-|x - y|^2}{2a^2}) \tag{9.3}$$

Çekirdek (Kernel) yöntemi, makine öğrenimini yüksek oranda arttırmaktadır. En çok kullanılan çekirdek yöntemi; Polynomial Kernel ve Gaussian RBF (Radial Basis Function) Kernel'dir. "A Comparison Study of Kernel Functions in the Support Vector Machine and Its Application for Termite Detection"da isimli kaynakta [54] bu çekirdeklerin karşılaştırması yapılmış ve Polynomial çekirdeğinin 0,9188 AUC ile en iyi sınıflandırma doğruluğunu sağladığı görülmüştür. Bu sebeple yapılan çalışmada da Polynomial çekirdeği kullanılmıştır.

K Nearest Neighbor Sınıflandırıcısı: K En Yakın Komşu algoritması, basit ve etkili sınıflandırma yöntemlerinden biridir ve makine öğrenme algoritmaları arasında popüler olarak kullanılmaktadır. Önerilen veri noktasının bulunduğu sınıfın ve en yakın komşusunun, k değerine göre belirlediği bir sınıflandırma yöntemidir [55]

Uygulamada yer alan KNN düğümü eğitim verilerini kullanarak En Yakın Komşu algoritmasına göre bir test verisi kümesini sınıflandırır. Bu tip sınıflandırıcı sadece birkaç bin ila on bin eğitim örneği için uygundur. Tüm (ve sadece) sayısal sütunlar ve Öklid mesafesi bu uygulamada kullanılmaktadır. Test verilerindeki diğer tüm sütunlar (sayısal olmayan türde) çıktıya olduğu gibi iletilir.

KNN düğümünün (operatörünün) konfigürayon penceresi Şekil 9.15'deki gibidir.



Şekil 9.15: KNIME Analytics Platform v3.5.3 KNN Configure

KNN düğümü için konfigürayon açıklamaları şu şekildedir:

- Column with class labels, Sınıflandırma niteliği olarak kullanılacak sütun seçilir.
- Number of neighbours to consider (k), Yeni bir örneği sınıflandırmak için kullanılacak en yakın komşu sayısı seçilir. Tek sayı olması önerilir.
- Weight neighbours by distance, Bu seçeneğin seçilmesi durumunda, sorgu paterninin depolanan eğitim paternlerine olan mesafesini sınıflandırmaya dahil eder. Daha yakın komşular, ortaya çıkan sınıf üzerinde, uzaktakilerden daha büyük etkiye sahiptir. (Ancak yine de sadece k sayıda komşu dikkate alınacaktır)
- Output class probabilities, Bu seçenek etkinse, sınıf olasılıklarını içeren ek sütunlar çıktı tablosuna eklenir.

Algoritmanın çalışması için bir k değeri belirlenir. k değerinin anlamı sınıflandırma için bakılacak eleman sayısıdır. Bir değer geldiğinde en yakın k kadar eleman alınarak gelen değer arasındaki uzaklık hesaplanır. Uzaklık hesaplama işleminde genellikle Öklid fonksiyonu kullanılır. Öklid fonksiyonuna ek olarak Manhattan, Minkowski ve Hamming fonksiyonları da kullanılabilir. Uzaklık hesaplandıktan sonra sıralanır ve gelen değerin

uygun sınıfa ataması yapılır. Yapılan çalışmada k değeri için 1'den başlayarak denemeler yapılmıştır. k=3'ten itibaren başarı sabitlendiğii için k, 3 olarak belirlenmiştir.

9.4 Değerlendirme Ölçütleri

Sınıflandırma algoritmaların sonuçlarının doğruluğunu ölçmek için çeşitli metrikler vardır. Bunlardan en önemlileri aşağıdaki gibidir:

9.4.1 Doğruluk

Doğruluk bir karar vericinin verdiği kararın gerçekte olması gerekenlerle karşılaştırarak aşağıdaki formülde gösterildiği şekilde hesaplanan bir ölçüttür [49].

$$Dogruluk = \frac{TP + TN}{TP + TN + FP + FN} \tag{9.4}$$

TABLO 9.1: Karışıklık Tablosu

		Pozitif	Negatif
Tahmin Edilen Değer	Pozitif	TP	FP
Tahmin Edilen Değer	Negatif	FN	TN

Karışıklık tablosundan ve doğruluk denkleminden anlaşılacağı üzere bir sınıflandırıcının doğruluğu, pozitif ve negatif olan durumları doğru tahmin ettiği örnek sayısının, toplam örnek sayısına bölümü ile elde edilmektedir [49].

9.4.2 Hassasiyet

Doğru Pozitif Oranı sınıflandırıcının gerçekte pozitif olan örnekleri doğru şekilde etiketlemede ne kadar başarılı olduğunu gösterir. 9.5 Hassasiyet Denklemi aşağıdaki gibidir:

$$TPR = \frac{TP}{TP + FN} \tag{9.5}$$

9.4.3 Kesinlik

Doğru Negatif Oranı sınıflandırıcının gerçekte negatif olan örnekleri doğru şekilde etiketlemede ne kadar başarılı olduğunu gösterir [49]. 9.6 Kesinlik Denklemi aşağıdaki gibidir:

$$TNR = \frac{TN}{TN + FP} \tag{9.6}$$

9.4.4 Hatalı Pozitif Oranı

Kesinlik değerinin tersi ve sınıflandırıcının hatalı pozitif tespit etme oranı ile ilgilenir [49]. 9.7 Hatalı Pozitif Oranı Denlemi aşağıdaki gibidir:

$$FPR = \frac{FP}{FP + TN} \tag{9.7}$$

FPR değeri aynı zamanda 1-TNR değerine eşit olacaktır. İyi bir sınıflandırıcının hem hassasiyetinin hem de kesinliğinin yüksek dolayısıyla FPR değerinin düşük olması beklenir.

9.4.5 Eğri Altı Alan

ROC eğrisi altında kalan alan (AUC), sınıflandırma algoritmalarının başarımlarının değerlendirilmesinde kullanılan bir diğer ölçüttür. ROC eğrisi altında kalan alan, [0-1] aralığında değer alır ve yüksek değerler alması, sınıflandırma algoritmasının tahmin etme başarısının daha yüksek olduğunu gösterir [56]

9.5 Analiz Sonuçları

Seçilen sınıflandırma algoritmaları her bir savunma sanayi kuruluşu verileri için ayrı ayrı çalıştırılmıştır ve sonuçları aşağıdaki gibidir:

1. Savunma Sanayi Kuruluşu Analiz Sonuçları

Tablo 9.2: 1. Kuruluş, Naive Bayes Algoritması Karmaşıklık Matrisi

	Pozitif	Negatif
Pozitif	1105	0
Negatif	0	2875

Tablo 9.3: 1. Kuruluş, SVM Algoritması Karmaşıklık Matrisi

	Pozitif	Negatif
Pozitif	1105	0
Negatif	0	2875

Tablo 9.4: 1. Kuruluş, KNN Algoritması Karmaşıklık Matrisi

	Pozitif	Negatif
Pozitif	944	161
Negatif	0	2875

Karmaşıklık matrislerinden anlaşılacağı üzere Naive Bayes ve SVM sınıflandırma algoritmaları 1. Kuruluş için aynı değerleri üretirken, KNN sınıflandırma algoritması daha düşük performans göstermiştir.

Değerlendirme Ölçütlerine göre 1. kuruluş verilerinin sonuçları aşağıdaki gibidir:

Tablo 9.5: 1. Kuruluş, Naive Bayes ve SVM Algoritmaları Değerlendirme Ölçütü Sonuçları

	min	mean	max	std. dev.
True Pozitives	1105	1990	2875	1.251,579
False Pozitives	0	0	0	0
True Negatives	1105	1990	2875	1.251,579
False Negatives	0	0	0	0
Recall	1	1	1	0
Precision	1	1	1	0
Sensitivity	1	1	1	0
Specifity	1	1	1	0
F-measure	1	1	1	0
Accuracy	1	1	1	0
Cohen's Kappa	1	1	1	0

TABLO 9.6: 1. Kuruluş, KNN Algoritması Değerlendirme Ölçütü Sonuçları

	min	mean	max	std. dev.
True Pozitives	944	1909,5	2875	1.365,4232
False Pozitives	0	80,5	161	113,8442
True Negatives	944	1909,5	2875	1.365,4232
False Negatives	0	80,5	161	113,8442
Recall	0,8543	0,9271	1	0,103
Precision	0,947	0,9735	1	0,0375
Sensitivity	0,8543	0,9271	1	0,103
Specifity	0,8543	0,9271	1	0,103
F-measure	0,9214	0,9471	0,9728	0,0363
Accuracy	0,9595	0,9595	0,9595	0
Cohen's Kappa	0,8944	0,8944	0,8944	0

2. Savunma Sanayi Kuruluşu Analiz Sonuçları

TABLO 9.7: 2. Kuruluş, Naive Bayes Algoritması Karmaşıklık Matrisi

	Pozitif	Negatif
Pozitif	870	0
Negatif	0	401

TABLO 9.8: 2. Kuruluş, SVM Algoritması Karmaşıklık Matrisi

	Pozitif	Negatif
Pozitif	870	0
Negatif	10	391

TABLO 9.9: 2. Kuruluş, KNN Algoritması Karmaşıklık Matrisi

	Pozitif	Negatif
Pozitif	870	0
Negatif	49	352

Karmaşıklık matrislerinden anlaşılacağı üzere her bir sınıflandırma algoritması 2. Kuruluş için farklı dğerleri üretmiştir. En başarılı sınıflandırmayı Naive Bayes algoritması sağlamıştır.

Tablo 9.10: 2. Kuruluş, Naive Bayes Algoritması Değerlendirme Ölçütü Sonuçları

	min	mean	max	std. dev.
True Pozitives	401	635,5	870	331,6331
False Pozitives	0	0	0	0
True Negatives	401	635,5	870	331,6331
False Negatives	0	0	0	0
Recall	1	1	1	0
Precision	1	1	1	0
Sensitivity	1	1	1	0
Specifity	1	1	1	0
F-measure	1	1	1	0
Accuracy	1	1	1	0
Cohen's Kappa	1	1	1	0

Tablo 9.11: 2. Kuruluş, SVM Algoritması Değerlendirme Ölçütü Sonuçları

	min	mean	max	std. dev.
True Pozitives	391	630,5	870	338,7041
False Pozitives	0	5	10	7,0711
True Negatives	391	630,5	870	338,7041
False Negatives	0	5	10	7,0711
Recall	0,9751	0,9875	1	0,0176
Precision	0,9886	0,9943	1	0,008
Sensitivity	0,9751	0,9875	1	0,0176
Specifity	0,9751	0,9875	1	0,0176
F-measure	0,9874	0,9908	0,9943	0,0049
Accuracy	0,9921	0,9921	0,9921	0
Cohen's Kappa	0,9817	0,9817	0,9817	0

Tablo 9.12: 2. Kuruluş, KNN Algoritması Değerlendirme Ölçütü Sonuçları

	min	mean	max	std. dev.
True Pozitives	352	611	870	366,2813
False Pozitives	0	24,5	49	34,6482
True Negatives	352	611	870	366,2813
False Negatives	0	24,5	49	34,6482
Recall	0,8778	0,9389	1	0,0864
Precision	0,9467	0,9733	1	0,0377
Sensitivity	0,8778	0,9389	1	0,0864
Specifity	0,8778	0,9389	1	0,0864
F-measure	0,9349	0,9538	0,9726	0,0266
Accuracy	0,9614	0,9614	0,9614	0
Cohen's Kappa	0,9077	0,9077	0,9077	0

2. Kuruluş verileri Değerlendirme Ölçütü sonuçlarına göre de en yüksek doğruluk değerini Naive Bayes sınıflandırma algoritmasının sağladığı görülmektedir.

3. Savunma Sanayi Kuruluşu Analiz Sonuçları

Tablo 9.13: 3. Kuruluş, Naive Bayes Algoritması Karmaşıklık Matrisi

	Pozitif	Negatif
Pozitif	266	0
Negatif	0	1236

Tablo 9.14: 3. Kuruluş, SVM Algoritması Karmaşıklık Matrisi

	Pozitif	Negatif
Pozitif	266	0
Negatif	0	1236

Tablo 9.15: 3. Kuruluş, KNN Algoritması Karmaşıklık Matrisi

	Pozitif	Negatif
Pozitif	108	158
Negatif	0	1236

Karmaşıklık matrisi sonuçlarına göre Naive Bayes ve SVM algoritmaları aynı değerleri üretirken, KNN algoritması daha düşük performans göstererek diğer algoritmaların gerisinde kalmıştır.

Değerlendirme Ölçütlerine göre 3. kuruluş verilerinin sonuçları aşağıdaki gibidir:

Tablo 9.16: 3. Kuruluş, Naive Bayes ve SVM Algoritmaları Değerlendirme Ölçütü Sonuçları

	min	mean	max	std. dev.
True Pozitives	266	751	1236	685,8936
False Pozitives	0	0	0	0
True Negatives	266	751	1236	685,8936
False Negatives	0	0	0	0
Recall	1	1	1	0
Precision	1	1	1	0
Sensitivity	1	1	1	0
Specifity	1	1	1	0
F-measure	1	1	1	0
Accuracy	1	1	1	0
Cohen's Kappa	1	1	1	0

Tablo 9.17: 3. Kuruluş, KNN Algoritması Değerlendirme Ölçütü Sonuçları

	min	mean	max	std. dev.
True Pozitives	108	672	1236	797,6164
False Pozitives	0	79	158	111,7229
True Negatives	108	672	1236	797,6164
False Negatives	0	79	158	111,7229
Recall	0,406	0,703	1	0,42
Precision	0,8867	0,9433	1	0,0801
Sensitivity	0,406	0,703	1	0,42
Specifity	0,406	0,703	1	0,42
F-measure	0,5775	0,7587	0,9399	0,2562
Accuracy	0,8948	0,8948	0,8948	0
Cohen's Kappa	0,5294	0,5294	0,5294	0,5294

3. Kuruluş Değerlendirme Ölçütü sonuçlarına göre Naive Bayes ve SVM sınıflandırma algoritmaları başarı ölçütleri bakımından daha doğru sonuçlar üretmişlerdir.

4. Savunma Sanayi Kuruluşu Analiz Sonuçları

Tablo 9.18: 4. Kuruluş, Naive Bayes Algoritması Karmaşıklık Matrisi

	Pozitif	Negatif
Pozitif	932	0
Negatif	0	382

Tablo 9.19: 4. Kuruluş, SVM Algoritması Karmaşıklık Matrisi

	Pozitif	Negatif
Pozitif	932	0
Negatif	6	376

Tablo 9.20: 4. Kuruluş, KNN Algoritması Karmaşıklık Matrisi

	Pozitif	Negatif
Pozitif	932	0
Negatif	119	263

Karmaşıklık matrislerinden anlaşılacağı üzere 4. Kuruluş verileri için en iyi sınıflandırmayı Naive Bayes sınıflandırma algoritması sağlamıştır. Naive Bayes algoritmasından sonra SVM algoritması da KNN algoritmasına göre daha yüksek başarı göstermiştir.

Değerlendirme Ölçütlerine göre değerlendirmeleri aşağıdaki gibidir.

Tablo 9.21: 4. Kuruluş, Naive Bayes Algoritması Değerlendirme Ölçütü Sonuçları

	min	mean	max	std. dev.
True Pozitives	382	657	932	388,9087
False Pozitives	0	0	0	0
True Negatives	382	657	932	388,9087
False Negatives	0	0	0	0
Recall	1	1	1	0
Precision	1	1	1	0
Sensitivity	1	1	1	0
Specifity	1	1	1	0
F-measure	1	1	1	0
Accuracy	1	1	1	0
Cohen's Kappa	1	1	1	0

Tablo 9.22: 4. Kuruluş, SVM Algoritması Değerlendirme Ölçütü Sonuçları

	min	mean	max	std. dev.
True Pozitives	376	654	932	393,1514
False Pozitives	0	3	6	4,2426
True Negatives	376	654	932	393,1514
False Negatives	0	3	6	4,2426
Recall	0,9843	0,9921	1	0,0111
Precision	0,9936	0,9968	1	0,0045
Sensitivity	0,9843	0,9921	1	0,0111
Specifity	0,9843	0,9921	1	0,0111
F-measure	0,9921	0,9944	0,9968	0,0033
Accuracy	0,9954	0,9954	0,9954	0
Cohen's Kappa	,9889	0,9889	0,9889	0

Tablo 9.23: 4. Kuruluş, KNN Algoritması Değerlendirme Ölçütü Sonuçları

	min	mean	max	std. dev.
True Pozitives	263	597,5	932	473,0544
False Pozitives	0	59,5	119	84,1457
True Negatives	263	597,5	932	473,0544
False Negatives	0	59,5	119	84,1457
Recall	0,6885	0,8442	1	0,2203
Precision	0,8868	0,9434	1	0,0801
Sensitivity	0,6885	0,8442	1	0,2203
Specifity	0,6885	0,8442	1	0,2203
F-measure	0,8155	0,8777	0,94	0,088
Accuracy	0,9094	0,9094	0,9094	0
Cohen's Kappa	0,7582	0,7582	0,7582	0

4. Kuruluş Değerlendirme Ölçütü tablolarına göre Naive Bayes sınıflandırma algoritması diğer iki algoritmadan daha başarılı sonuçlar üretmiştir.

5. Savunma Sanayi Kuruluşu Analiz Sonuçları

Tablo 9.24: 5. Kuruluş, Naive Bayes Algoritması Karmaşıklık Matrisi

	Pozitif	Negatif
Pozitif	575	0
Negatif	0	1060

Tablo 9.25: 5. Kuruluş, SVM Algoritması Karmaşıklık Matrisi

	Pozitif	Negatif
Pozitif	575	0
Negatif	0	1060

Tablo 9.26: 5. Kuruluş, KNN Algoritması Karmaşıklık Matrisi

	Pozitif	Negatif
Pozitif	513	62
Negatif	0	1060

Karmaşıklık Matrislerinden anlaşılacağı üzere 5. Kuruluş verileri için en iyi sınıflandırmayı Naive Bayes ve SVM Algoritmaları sağlamıştır. KNN algoritması, Naive Bayes ve SVM Algoritmalarının sağladığı doğruluk değerine ulaşamamışlardır.

Değerlendirme Ölçütlerine göre değerlendirmeleri aşağıdaki gibidir.

Tablo 9.27: 5. Kuruluş, Naive Bayes ve SVM Algoritmaları Değerlendirme Ölçütü Sonuçları

	min	mean	max	std. dev.
True Pozitives	575	817,5	1060	342,9468
False Pozitives	0	0	0	0
True Negatives	575	817,5	1060	342,9468
False Negatives	0	0	0	0
Recall	1	1	1	0
Precision	1	1	1	0
Sensitivity	1	1	1	0
Specifity	1	1	1	0
F-measure	1	1	1	0
Accuracy	1	1	1	0
Cohen's Kappa	1	1	1	0

Tablo 9.28: 5. Kuruluş, KNN Algoritması Değerlendirme Ölçütü Sonuçları

	min	mean	max	std. dev.
True Pozitives	513	786,5	1060	386,7874
False Pozitives	0	31	62	43,8406
True Negatives	513	786,5	1060	386,7874
False Negatives	0	31	62	43,8406
Recall	0,8922	0,9461	1	0,0762
Precision	0,9447	0,9724	1	0,031
Sensitivity	0,8922	0,9461	1	0,0762
Specifity	0,8922	0,9461	1	0,0762
F-measure	0,943	0,9573	0,9716	0,0202
Accuracy	0,9621	0,9621	0,9621	0
Cohen's Kappa	0,9147	0,9147	0,9147	0

5. Kuruluş Değerlendirme Ölçütü tablolarına göre Naive Bayes ve SVM sınıflandırma algoritmaları KNN algoritmasından daha başarılı sonuçlar üretmişlerdir.

Tüm analiz sonuçları aşağıdaki gibi özetlenmiştir.

Tablo 9.29: Sınıflandırma Algoritmaları Genel Karşılaştırma Tablosu

Doğruluk Değerleri	Naive Bayes	SVM	KNN
1. Kuruluş	100%	100%	95,955%
2. Kuruluş	99,875%	99,213%	96,145%
3. Kuruluş	100%	100%	89,481%
4. Kuruluş	99,981%	99,543%	90,944%
5. Kuruluş	100%	100%	96,208%

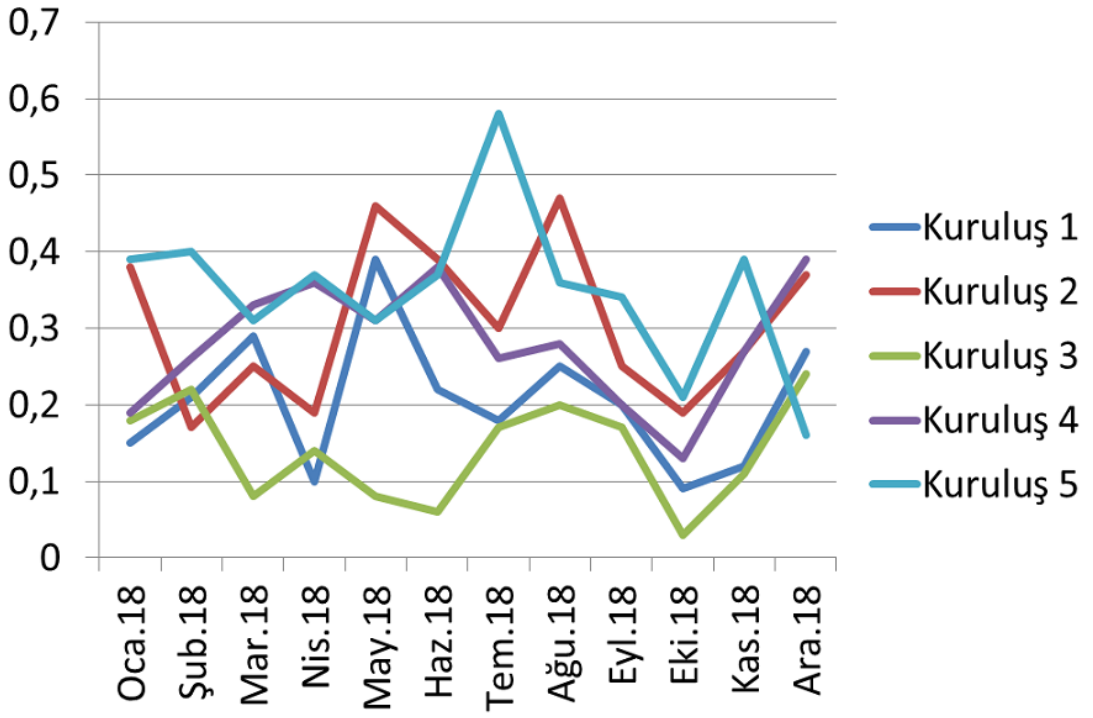
Kullanılan sınıflandırma algoritmaları arasında her beş veri seti için de en yüksek başarım Naive Sınıflandırma algoritmasında olmuştur; ancak SVM algoirtmasının da başarı değerleri Naive Bayes algoritmasına oldukça yakındır.

Tablo 9.30: Kuruluşlar Arası Genel Karşılaştırma Tablosu

	1. Kuruluş	2. Kuruluş	3. Kuruluş	4. Kuruluş	5. Kuruluş
Pozitif Yorum Oranı	72,45%	68,03%	82,49%	70,95%	64,42%

Kuruluşlar arası genel karşılaştırmaya bakıldığında ise en yüksek pozitif yoruma sahip olan kuruluş 3 kuruluş olmuştur.

Kuruluşlara ait negatif yorumların hangi dönemlerde nasıl değiştiği ise Şekil 9.16'daki grafik ile gösterilmektedir.



Şekil 9.16: Kuruluşların 2018 Yılına Ait Negatif Tweet Oranları

Grafik incelendiğinde kuruluşların genel olarak Mayıs-Ağustos ayları arasında twitter üzerinde aldıkları negatif yorumların arttığı gözlemlenmiştir. Özellikle 5. Kuruluş için Temmuz 2018 dönemi; 2. Kuruluş için Mayıs ve Ağustos 2018 dönemlerinde yaşanan olaylara geriye dönülüp bakıldığında, hangi olayların sosyal ağlar üzerinde etkili olarak kuruluşların itibarına dair negatif etkide bulunduğu değerlendirilebilmektedir.

TABLO 9.31: Literatürdeki Benzer Duygu Analizi Çalışmalarının Doğruluk Değerleri

Referanslar	Yıl	Kullanılan Algoritma	Veri Kapsamı	Doğruluk Değeri
Jose,Chooralil	2016	SWN, Naive Bayes, HMM	Delhi Election Days	71.48
Sharm, Moh	2016	Naive Bayes, SVM	Indian Election	78.4
Hodeghatta	2013	Naive Bayes	Hollywood Movies	79
Rane, Kumar	2018	RF,LR,DT,SVM,GNB,KNN	US Airline	86.5
Jain,Katkar	2015	KNN, NB, Random Forest	Political Leaders	99.64
Nizam, Akin	2014	Naive Bayes, SMO, KNN	Gıda Firmaları	72.33
Akdemir	2019	Naive Bayes, SVM, KNN	Savunma Sanayi Firmaları	100

Tablo 9.31'den de anlaşılacağı üzere twitter verileri üzerinde uygulanan duygu analizi çalışmalarında en sık kullanılan algoritmalar Naive Bayes, SVM ve KNN olmuştur. Yapılan çalışmada bu sebeple bu algoritmalar kullanılarak yapılan sınıflandırmaların doğruluk değerleri karşılaştırılmıştır. Sonuçlar karşılaştırıldığında literatürde yapılan benzer duygu analizi çalışmalarına kıyasla bu çalışmada elde edilen doğruluk değerleri daha yüksek seviyeye ulaşmıştır.

Doğruluk Oranı genel olarak, sınıflandırıcının ne sıklıkta doğru tahmin ettiğinin bir ölçüsü olduğundan çalışmalar esnasında eğitim veri seti ile test veri setinin ayrımını doğru şekilde yapmak doğruluk oranının belirlenmesinde büyük bir öneme sahiptir.

Sınıflandırıcının doğru sınıflandırma işlemini gerçekleştirmesinde ikinci önemli aşama eğitim setinin doğru etiketleme işlemi yaptığının kontrolüdür. Eğitim seti ne kadar doğru etiketlenmişse sınıflandırıcının doğru tahmin etme oranı da bir o kadar artacaktır.

Yapılan çalışmada veri setlerinin bölüntülenmesinde %66 oran ve rastgele dağılım tercih edilmiştir. Aynı zamanda veri etiketleme işleminin doğrulama işlemi gerçekleştirilmiştir. Bu işlemler sayesinde sınıflandırıcıların doğruluk oranları daha yüksek seviyelere ulaşmıştır.

Bölüm 10

Sonuç

Twitter uygulaması, insanların güncel konular ile ilgili duygu ve düşüncelerini bildirdikleri güncel ve popüler bir sosyal ağdır. Twitter, araştırmacı ve uygulayıcılar için önemli bilgiler sunan önemli bir veri kaynağıdır.

Bu çalışma kapsamında, Türkçe Twitter mesajları üzerinde, makine öğrenmesine dayalı sınıflandırıcılar kullanılarak, duygu analizi gerçekleştirilmiştir. Çalışma kapsamında Türkçe Twitter mesajlarının sınıflandırılmasında, üç temel makine öğrenmesine dayalı sınıflandırıcı (Naive Bayes algoritması, destek vektör makineleri ve k en yakın komşu) kullanılmıştır. Yapılan çalışmalarda, en yüksek başarım sınıflandırma algoritması olarak Naive Bayes algoritması kullanıldığında elde edilmektedir. Savunma Sanayi alanında belirlenen 5 kuruluş hakkında yazılan twitter mesajlarının duygusal polaritesi incelendiğinde ise en yüksek başarım 3. Kuruluş'ta elde edilmiştir. Yapılan bu çalışmanın, twitter verileri üzerinde uygulanan diğer duygu analizi çalışmalarından farkı; daha önce farklı çalışma konularında, farklı zamanlarda kullanılan Naive Bayes, Destek Vektör Makineleri ve K En Yakın Komşu algoritmalarının şirket itibarı konusunda ilk defa bu çalışma ile bir arada kullanılmış olması ve önceki kullanımlarına göre daha yüksek doğruluk değerlerine ulaşmış olmasıdır. Aynı zamanda tek bir veri seti yerine 5 farklı veri seti kullanılarak ve kullanılan sınıflandırma algoritma sayısı da arttırılarak hem algoritmalar arası karşılaştırma hem de veri setleri arası karşılaştırma imkanı sunulmuştur.

Gelecek çalışmalarda farklı makine öğrenme algoritmaları kullanılarak farklı veri setleri üzerinde duyarlılık analizi çalışmalarının gerçekleştirilmesi amaçlanmıştır.

Ek A

Java Kodu

```
package twitterinfo;

import java.io.FileWriter;

import java.io.IOException;

import java.util.List;

import twitter4j.*;

import twitter4j.conf.ConfigurationBuilder; public class TwitterInfo public static void main(String[]args) throws TwitterException, IOException

ConfigurationBuilder configurationBuilder= new ConfigurationBuilder();

configurationBuilder.setDebugEnabled(true)

.setOAuthConsumerKey("rm3dwqzG0cUIAyh8tnQgks70M")

.setOAuthConsumerSecret("02c1V8vwDOVlZpPDBJ2fhXl02t49GSBjwjcIyaSwGu8
pStwTVs")

.setOAuthAccessToken("211935066-5xM43IYy0bSHYHjQ0KhX2e9sgyEQVnkjHXKM
BAn3")

.setOAuthAccessTokenSecret("OcasiYLEGf5iCAdbMGfgGA0mEIoqf7XRsa17PUUI
31OcB");

TwitterFactory tf= new TwitterFactory(configurationBuilder.build());
```

```
twitter4j.Twitter twitter= tf.getInstance();

List<Status> status= twitter.getHomeTimeline();

QueryResult queryResult = twitter.search(new Query("text"));

// for (Status s:status)

// System.out.println(s.getUser().getName()+" "+s.getText()); //

try(FileWriter fw = new FileWriter("result.txt"))

for (Status tweet : queryResult.getTweets())

fw.write(s.get.user+tweet.getText());

fw.write("n");
```

Ek B

Mutluluk/Üzgünlük Bildiren Karakter ve Kelimeler

Mutluluk karakterleri ":)", "(:", ":D", " :d", ";)", ":>", "=)"

Üzgünlük karakterleri ":(", "):", ">:(", ":o", ":'(", ":<", ";("

Tablo B.1: Mutluluk ve Üzgünlük Kelime Tablosu

mutlu	mutluluk	başarı	güç
tebrik	şükran	teşekkür	gurur
keyif	heves	sevinç	sevinmek
sevindirici haber	değerli	önemli	hayranlık
fırsat	ödül	katkı	güçlü
güvenli	güvenilir	gurur kaynağı	memnuniyet
coşku	yetenek	nezaket	harika
neşe kaynağı	mutluluk kaynağı	ilgi	beğenmek
mutluluktan uçmak	bayram etmek	can atmak	yetenekli
güzel	kaza	zarar	kriz
terör	acı	hata	heba olmak
üzücü olay	üzüntü	hoşnutsuzluk	hayıflanma
kuşku	ümidini yitirme	umudunu kesme	ümitsizlik
çile	çile	kırılmak	dava
şüphe	örgüt	kumpas	düş kırıklığı
yas tutma	buhran	kasvet	çaresizlik
sıkıntı	hayal kırıklığı	şanssızlık	dövmek
endişe verici	tepki	sopa	gözaltı
virüs	can çekişmek	ümidini boşa çıkarma	özlemini çekme
müteessir olmak	kalbini kırmak	daralmak	üzgün
oyalanma	çöküntü	kıvranma	komik
gülünç	gülünçlük	dedikodu	utanma
utanç	neşelilik	şenlik	birlik
gerginlik	gerilmek	gerilemek	verimlilik
etkinlilik	atılım	kazandırmak	kazanç

Kaynaklar

[1] M. U. Şimşek. Sosyal ağlarda veri madenciliği Üzerine bir uygulama. Master's thesis, Gazi Üniversitesi, Ankara, Turkiye, 2012.

[2] O. O. Unal. Data mining applications on web usage analysis and user profiling. Master's thesis, 2003.

[3] J. Han. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

[4] S. Öğüt. Veri madenciliği kavramı ve gelişim süreci. Master's thesis, Görsel İletişim Tasarımı Bölümü, İletişim Fakültesi Yeditepe Üniversitesi.

[5] S.Aydin. Veri madenciligi ve anadolu universitesi uzaktan egitim sisteminde bir uygulama. Master's thesis, 2007.

[6] R. Groth. *Data mining: building competitive advantage*. Prentice Hall PTR, 2000.

[7] SPSS. Answertree algorithm summary. 1999.

[8] J. Sun and H. Liu. Data mining method for listed companies' financial distress prediction. *Knowledge-Based Systems*, 21(1):1, 5, 2008.

[9] M. Ture ve F. Tokatli ve İ. Kurt. Using kaplan meirer analysis together with decision tree methods (c rt, chaid, quest, c4.5 and id3) in determining recurrence-free survival of breast cancer patients. Master's thesis, 2008. Expert Systems With Applications, Article in Pres, 2008.

[10] A. Vahaplar. Bir coğrafi veri madenciliği uygulaması. Master's thesis, Ege Üniversitesi Fen Bilimler Enstitüsü, İzmir, 2003.

[11] M. Altun. Veri madenciliği ve uygulama alanları. 2017.

[12] H. Takçı and İ. Soğukpınar. Kütüphane kullanıcılarının erişim Örüntülerinin keşfi, bilgi dünyası. 2002.

[13] Ş. Z. Erdoğan. Veri madenciliği ve veri madenciliğinde kullanılan k-means algoritmasının Öğrenci veri tabanında uygulanması p.82. 2004.

[14] S. E. Şeker. Ybs ansiklopedi sosyal ağlarda veri madenciliği (data mining on social networks) cilt, vol. 2, no. 2. 2015.

[15] H. Kuduğ. Sosyal ag analizi olcutlerini is aglarina uyarlanmasi. Master's thesis, 2011.

[16] A. Karci ve O. Boy. Sosyal aglarin web madenciligi teknikleri ile benzerlik tahmini, elektrik bilgi sempozyumu pp 154-161. 2011.

[17] A. Baykal and C. Coskun. Web madenciligi teknikleri. Master's thesis, 2009.

[18] U. Can and B. Alatas. Duygu analizi ve fikir madenciligi algoritmalarininn incelenmesi. Master's thesis, 2017.

[19] W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications a survey. 2014.

[20] H. Nizam and S. S. Akin. Sosyal medyada makine ogrenmesi ile duygu analizinde dengeli ve dengesiz veri setlerinin performanslarinin karsilastirilmasi.

[21] B. Liu. Analysis of comparative opinions. *Sentiment Analysis*, pages 202,217.

[22] V. Ikoro, M. Sharmina, K. Malik, and R. Batista-Navarro. Analyzing sentiments expressed on twitter by uk energy company consumers. *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2018.

[23] U. A. Siddiqua, T. Ahsan, and A. N. Chy. Combining a rule-based classifier with weakly supervised learning for twitter sentiment analysis. *2016 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, 2016.

[24] H.G. Kim, S. Lee, and S. Kyeong. Discovering hot topics using twitter streaming data. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM 13*, 2013.

[25] C. Dedhia and J. Ramteke. Ensemble model for twitter sentiment analysis. *2017 International Conference on Inventive Systems and Control (ICISC)*, 2017.

[26] P. A. Joshi, G. Simon, and Y. P. Murumkar. Generation of brand/product reputation using twitter data. *2018 International Conference on Information , Communication, Engineering and Technology (ICICET)*, 2018.

[27] D. Chauhan, K. Sutaria, and R. Doshi. Impact of semiotics on multidimensional sentiment analysis on twitter: A survey. *2018 Second International Conference on Computing Methodologies and Communication (ICCMC)*, 2018.

[28] P. Barnaghi, P. Ghaffari, and J. G. Breslin. Opinion mining and sentiment polarity on twitter and correlation between events and sentiment. *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*, 2016.

[29] R. Jose and V. S. Chooralil. Prediction of election result by enhanced sentiment analysis on twitter data using classifier ensemble approach. *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, 2016.

[30] P. Sharma and T.-S. Moh. Prediction of indian election using sentiment analysis on hindi twitter. *2016 IEEE International Conference on Big Data (Big Data)*, 2016.

[31] R. Linares, J. Herrera, A. Cuadros, and L. Alfaro. Prediction of tourist traffic to peru by using sentiment analysis in twitter social network. *2015 Latin American Computing Conference (CLEI)*, 2015.

[32] P. Tripathi, S. Kr. Vishwakarma, and A. Lala. Sentiment analysis of english tweets using rapid miner. *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*, 2015.

[33] U. R. Hodeghatta. Sentiment analysis of hollywood movies on twitter. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM 13*, 2013.

[34] P. Garg, H. Garg, and V. Ranga. Sentiment analysis of the uri terror attack using twitter. *2017 International Conference on Computing, Communication and Automation (ICCCA)*, 2017.

[35] P. Mishra, R. Rajnish, and P. Kumar. Sentiment analysis of twitter data: Case study on digital india. *2016 International Conference on Information Technology*

(InCITe) - The Next Generation IT Summit on the Theme - Internet of Things: Connect your Worlds, 2016.

[36] M. Abdullah and M. Hadzikadic. Sentiment analysis of twitter data: Emotions revealed regarding donald trump during the 2015-16 primary debates. *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, 2017.

[37] M. Bilgin and I. F. Senturk. Sentiment analysis on twitter data with semi-supervised doc2vec. *2017 International Conference on Computer Science and Engineering (UBMK)*, 2017.

[38] H. Parveen and S. Pandey. Sentiment analysis on twitter data-set using naive bayes algorithm. *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, 2016.

[39] I. P. Windasari, F. N. Uzzi, and K. I. Satoto. Sentiment analysis on twitter posts: An analysis of positive or negative opinion on gojek. *2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, 2017.

[40] Z. Rezaei and M. Jalali. Sentiment analysis on twitter using mcdiarmid tree algorithm. *2017 7th International Conference on Computer and Knowledge Engineering (ICCKE)*, 2017.

[41] A. Rane and A. Kumar. Sentiment classification system of twitter data for us airline service analysis. *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, 2018.

[42] A. P. Jain and V. D. Katkar. Sentiments analysis of twitter data using data mining. *2015 International Conference on Information Processing (ICIP)*, 2015.

[43] G. Subramaniam, R. Aswini, M. Ranjitha, and Praveen Kumar Rajendran. Survey on user emotion analysis using twitter data. *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 2017.

[44] C. R. Nirmala, G. M. Roopa, and K R Naveen Kumar. Twitter data analysis for unemployment crisis. *2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, 2015.

[45] S. Anil Phand and J. Anil Phand. Twitter sentiment classification using stanford nlp. *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*, 2017.

[46] M. Hao, C. Rohrdantz, H. Janetzko, U. Dayal, D. A. Keim, L.-E. Haug, and M. Hsu. Visual sentiment analysis on twitter data streams. *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2011.

[47] A. M. Alkalbani, L. Gadhvi, B. P., F. K. Hussain, A. M. Ghamry, and O. K. Hussain. Analysing cloud services reviews using opining mining. *2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA)*, 2017.

[48] C. Baydogan and B. Alatas. Sentiment analysis using konstanz information miner in social networks. *2018 6th International Symposium on Digital Forensic and Security (ISDFS)*, 2018.

[49] I. Kose. Veri madenciligi teori uygulama ve felsefesi, nisan. 2018.

[50] Ş. E. Şeker ve D. Erdogan. *KNIME ile Uçtan Uca Veri Bilimi*. Demet Erdogan, 2018.

[51] W. G. Parrott. *Emotions in social psychology: essential readings*. Psychology Press, 2001.

[52] S. Beser. Destek vektor makineleri, temmuz 2017. URL `https://veribilimcisi.com/2017/07/19/destek-vektor-makineleri-support-vector-machine/`.

[53] E.Güldoğan. Çeşitli Çekirdek fonksiyonlari ile oluşturulan destek vektör makinesi modellerinin performansının İncelenmesi: Bir klinik uygulama, 2017.

[54] D. Nandika M.A. Nanda, K. B. Seminar and A. Maddu. A comparison study of kernel functions in the support vector machine and its application for termite detection. 2018.

[55] H. A. Zengin. K en yakın komşu methodu, kasım 2017. URL `https://yazilimagiris.com/k-en-yakin-komsu-methodu-k-nearest-neighborhood/`.

[56] M. Based and M. L. Methods. Yonetim bilisim sistemleri dergisi pp. 1,14. 2017.