T.C.
KARABÜK ÜNİVERSİTESİ
MÜHENDİSLİK FAKÜLTESİ BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

MAKİNE ÖĞRENMESİ YÖNTEMLERİ KULLANARAK PROSPEKTÜS-İLAÇ SINIFLANDIRMASI

BİLGİSAYAR MÜHENDİSLİĞİ LİSANS

BİTİRME PROJESİ

Proje Danışmanı: Dr.Öğr.Üyesi Ferhat ATASOY

Gürkan KIRMACI

2020

ÖNSÖZ

Bu tez çalışmasının planlanması, tasarımı ve uygulama olarak hayata geçirilmesinde ilgi ve desteğini esirgemeyen, engin bilgi birikimi ve tecrübelerinden faydalandığım, gerekli yönlendirmeleriyle çalışmamı bilimsel temeller ışığında şekillendiren sayın hocam Dr. Öğr. Üyesi FERHAT ATASOY'a teşekkürlerimi sunarım...

İÇİNDEKİLER

Sayfa

ŞEKİLLERİN LİSTESİ

Şekil **Sayfa**

1.ÖZET

Teknoloji hızla ilerlerken, her geçen gün bilginin hacmi de artmaktadır. Bu doğrultuda her şey birbirine bağlı olarak büyüyerek ilerlemektedir. Bilgi arttıkça, saklama kapasitesi de artmakta dolayısıyla veri kaydı yapılan alanlara da fazlasıyla ihtiyaç duyulmaktadır. Bundan dolayı eldeki verilerin analizi, sonucu ve bu verilerin anlamlandırılmasının yöntemlerinin önemi makinalar için gittikçe artmaktadır. Bilgisayarlarca üretilen veriler tek başlarına değersizdir, anlamsızdır. Çünkü çıplak gözle bakıldığında bir anlam ifade etmezler. Bu veriler belli bir amaç doğrultusunda işlendiği zaman anlamlı hale gelmeye başlarlar. Bu nedenle çok çok büyük verileri işleyebilen teknikleri kullanabilmek büyük önem kazanmaktadır. Bu ham veriyi bilgiye veya anlamlı hale dönüştürme işlemleri veri madenciliği ile yapılabilmektedir. Veri madenciliği yöntemleri ile eldeki veriler sınıflandırılarak, gruplandırılarak ya da veriler arasında ilişkiler, bağıntılar, istatistiksel sonuçlar oluşturularak modeller oluşturulur. Oluşturulan model, oluşturulduğu veri kümesinde olmayan yeni bir veri geldiğinde yeni gelen veri hakkında tahmin yapma imkanı verir. Yapılan tahminlerin doğruluk yüzdesi, oluşturulmuş olan veri modelin, veri üzerindeki başarı oranını göstermektedir. Dolayısı ile bir veri madenciliği uygulamasında kullanılan algoritma ile ortaya çıkan sonuç, kullanılan algoritmaya göre değişmektedir. En az sonuç kadar kullanılan bu algoritma önemlidir ve direk olarak sonuca etki etmektedir. Daha sonra çıkan bu sonuç, başka üretilen algoritmalarla kıyaslanarak ortaya çıkan tüm bu sonuçlar test edilebilir. Makine öğrenmesini özetlemek gerekirse, geçmiş tecrübelerden geleceği tahmin etmektir. Bu hususta ilgili önemli görülen unsurları şu şekilde sıralanabilir

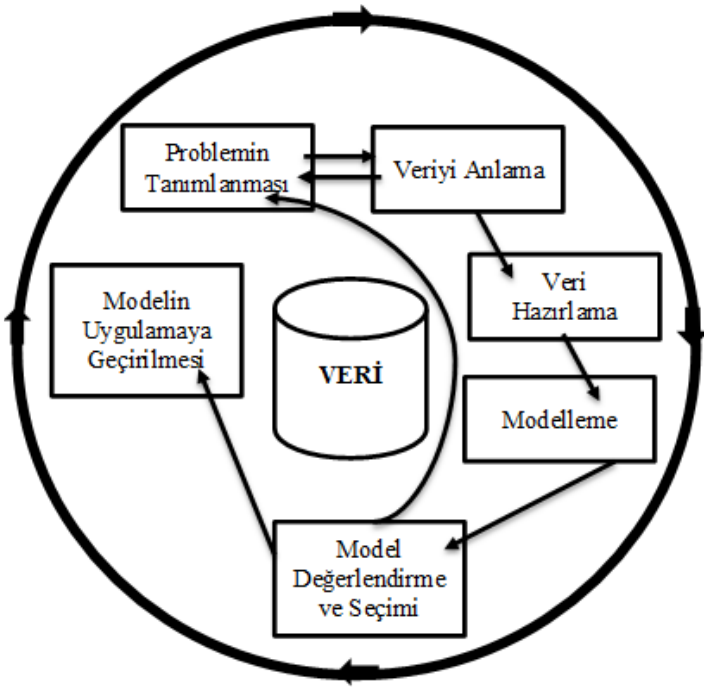
- Makine öğrenmesinde verilerden öğrenmeyi gerçekleştiren makine bilgisayarlardır

- Tecrübe, teknik ve performans tanımları makine öğrenmesi için kritik noktalardır.
- Öğrenmenin başarısı ölçülebilirdir. Başarı yüzdesi için belirlenen ölçütler sayesinde oluşturulan modelin performansı değerlendirilmektedir.

- Makinanın öğrenmesi için oluşturulan modele ait parametreler varsa bu parametreler değiştirilmektedir.

Makine öğrenmesi, optik karakter algılama, yüz tanıma, spam e-posta filtrelemesi, konuşma dili anlama, tıbbi teşhis, müşteri segmentasyonu, sahtekarlık tespiti hava durumu tahmini gibi birçok farklı problemin çözümünde kullanılmaktadır.

Bu projede ise büyük bir veri yığını içinden bir ilacın prospektüsüne göre hangi türe ait olduğunu makine öğrenmesi yöntemlerini kullanarak sınıflandırılacak. Bu doğrultuda yapılan tez çalışmasındaki sınıflandırmaya dayalı makine öğrenmesi uygulaması gerçekleştirilirken şekil 1.a da ki CRISP modeli adımları izlenecektir.



ŞEKİL 1.a : CRİSP MODELİ

1.1.Problem

Doğru ve zamanında karar almanın hasta sağlığı üzerindeki etkisi tartışmasız çok önemlidir. Fakat yığın halindeki bu verilerden hastanın faydalanması pek olanaklı değildir. Anlamsız görünen verinin anlamlı, kullanılır hale dönüşmesi gerekmektedir.

1.2.Amaç

Bu noktada veri madenciliği kayıt altına alınan yığın halindeki verileri, anlam kazandırmak için gerekli işlemlerden geçirir ve bilgiye dönüştürür. Anlamsız görünen verileri hastanın ihtiyacına göre prospektüsünden kategorize ederek, hastanın aradığı sınıfta ilaca en hızlı şekilde ulaşmasını sağlanır.

1.3.Kullanılacak Yöntem ve Teknikler

Python, veri madenciliği, Karar Ağacı Algoritması, Rassal Orman Algoritması, Naive Bayes Algoritması

2.VERİYİ ANLAMA

2.1.Problemin Tanımlanması

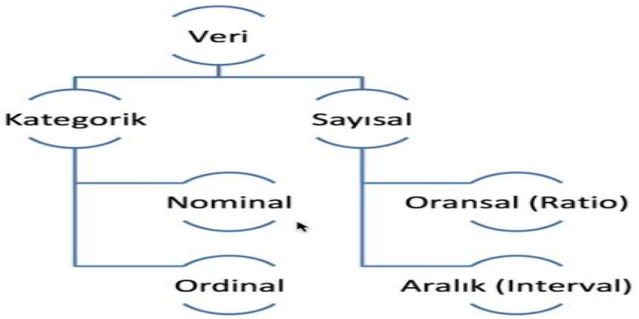
Doğru ve zamanında karar almanın hasta sağlığı üzerindeki etkisi tartışmasız çok önemlidir. Fakat yığın halindeki bu verilerden hastanın faydalanması pek olanaklı değildir. Anlamsız görünen verinin anlamlı, kullanılır hale dönüşmesi gerekmektedir. (“T.Mitchell ,Machine Learning”) Problemin çözümünde görev(G),performans(P) ve deneyim (D) belirlenerek problem adım adım çözülebilir. Bu prospektüs-ilaç sınıflandırma problemi için:

- Görev(G) : İlaçların prospektüsüne göre sınıflara ayrılması
- Performans(P) : İlaçların doğru sınıflandırma yüzdeleri
- Deneyin(D) :Belli sınıflardaki ilaçlardan alınan örnekler ve gözlemler

2.2. Anlama

Makinanın deneyim için yararlanacağı veri, probleme uygun bir şekilde toplanır. Bu aşamada problemin çok iyi anlaşılması gerekmektedir. Problem net bir şekilde anlaşıldığı takdirde veri de daha iyi tanınır. Zaman zaman bu aşama uzun ve zahmetli olabilmektedir. İnternet üzerinde probleme yönelik veri setleri bulmakta mevcuttur.

Bu veri setleri eksik ve yanlış bilgi içerdiği de görülebilir. Problemin iyi anlaşılamaması bu veri setlerinin de iyi analiz edilememesine yol açabilir. (Lichman,2013) Veriler:



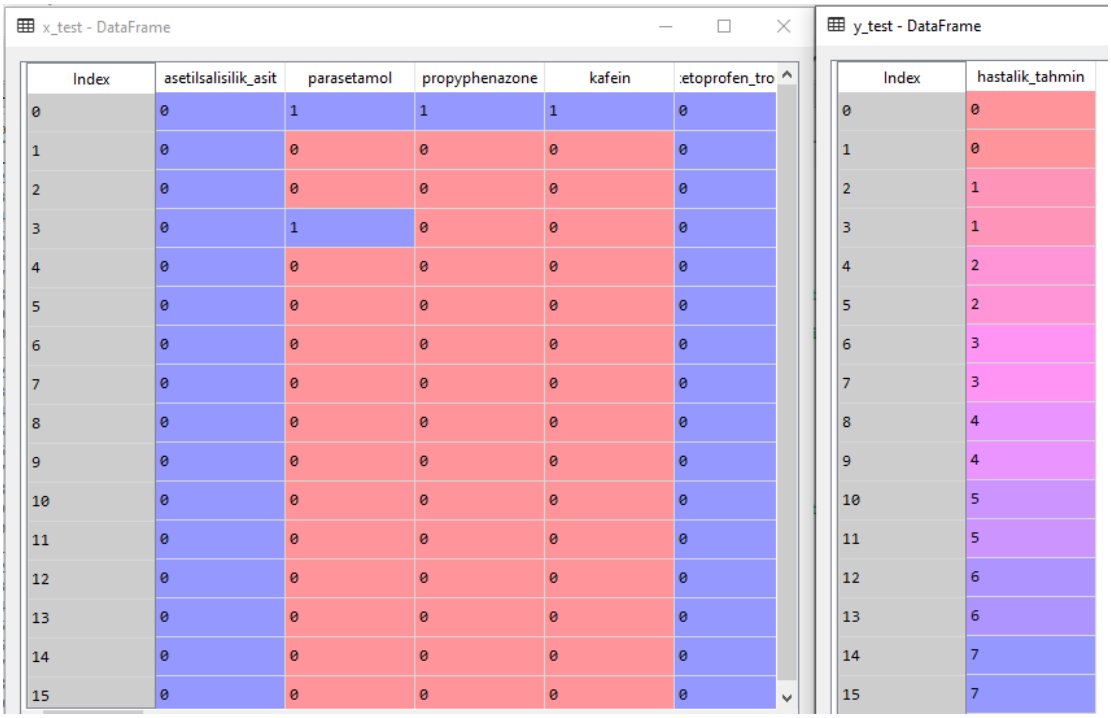
Şekil 2.a : Veri Tanımı

Şeklinde iki grupta veriyi inceleyebiliriz. Veri noktasındaki ayrım problem tiplerini de ayırmaya yardımcı olmaktadır. Kategorik veriler üzerinde tahmin yapılması sınıflandırma (classification), sayısal veriler üzerinde tahmin yapılması ise tahmin (prediction) olarak tanımlanır. Kategorik verilerin alt kolu olan nominal verilerde sıralama imkanı olmazken (araba markası vb.) ,ordinal verilerde ise sıralama imkanı olmaktadır.(Plaka numaraları gibi).Sayısal verilerin alt kolu olan oransal (ratio) veriler birbirine göre çarpılabilen orantı kurulabilen değerler, aralık (interval) ise çarpılamayan değerlerdir. (Oda sıcaklığı gibi)

3.VERİYİ HAZIRLAMA

Veri hazırlama bir veri madenciliği tekniğidir ve ham veriyi anlaşılabilir hale dönüştürmeyi gerektirir. Ham veriler eksik, tutarsız ve birçok hata içermesi gibi olasılıkları mevcuttur. Veri hazırlama bir diğer adıyla veri ön işleme bu anlamda kanıtlanmış bir yöntemdir. Bu işlemin bir standardı olmayıp kullanılacak veri setine göre değişmektedir. Bu projede bir ilaç prospektüs sitesinden veriler çekilerek işlenmeye hazır hale getirilmesi sağlanacaktır. Bu verilerin çekilmesi python "requests" ve "beautifulsoup" kütüphanesi kullanılarak yapılacaktır. Requests modülü python da bir http kütüphanesidir ve veri çekmek istenilen sitelere istekte bulunarak,ilgili sitenin html kodlarını döndürülmesini sağlar.("python-requests.org").

Daha sonra beautifulsoup kütüphanesi ile veriler işlenerek , istenilen kısımlar bir metin şeklinde elde edilir.Veri hazırlanması aşamasında örnek bir python kütüphane oluşturarak bu kodlarla uygulanacak veriye göre değişiklikler yaparak verinin hazırlanması sağlanacaktır. Csv formatında bir excel dosyasına veriler aktarılacak, etken maddeler sütun değerleri, ilaç grupları ise satır değerlerine denk gelecek şekilde her bir ilaç için skorlanacak ve bu küme yüzde %20 test , %80 eğitim kümesi olarak ayrılarak makine öğrenmesi yöntemleri uygulanacaktır.



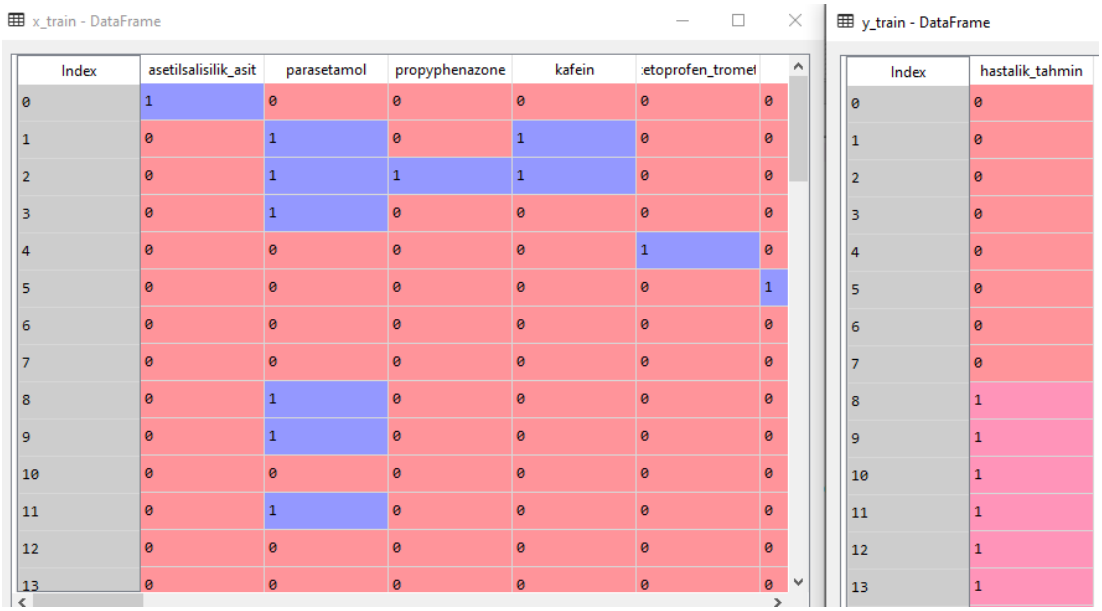
x_test - DataFrame

Index	asetilsalisilik_asit	parasetamol	propyphenazone	kafein	:etoprofen_tro
0	0	1	1	1	0
1	0	0	0	0	0
2	0	0	0	0	0
3	0	1	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0
11	0	0	0	0	0
12	0	0	0	0	0
13	0	0	0	0	0
14	0	0	0	0	0
15	0	0	0	0	0

y_test - DataFrame

Index	hastalik_tahmin
0	0
1	0
2	1
3	1
4	2
5	2
6	3
7	3
8	4
9	4
10	5
11	5
12	6
13	6
14	7
15	7

Şekil 3.a : x_test ve y_test veri seti



x_train - DataFrame

Index	asetilsalisilik_asit	parasetamol	propyphenazone	kafein	:etoprofen_tromel	
0	1	0	0	0	0	0
1	0	1	0	1	0	0
2	0	1	1	1	0	0
3	0	1	0	0	0	0
4	0	0	0	0	1	0
5	0	0	0	0	0	1
6	0	0	0	0	0	0
7	0	0	0	0	0	0
8	0	1	0	0	0	0
9	0	1	0	0	0	0
10	0	0	0	0	0	0
11	0	1	0	0	0	0
12	0	0	0	0	0	0
13	0	0	0	0	0	0

y_train - DataFrame

Index	hastalik_tahmin
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	1
9	1
10	1
11	1
12	1
13	1

Şekil 3.b: x_train ve y_train veri seti

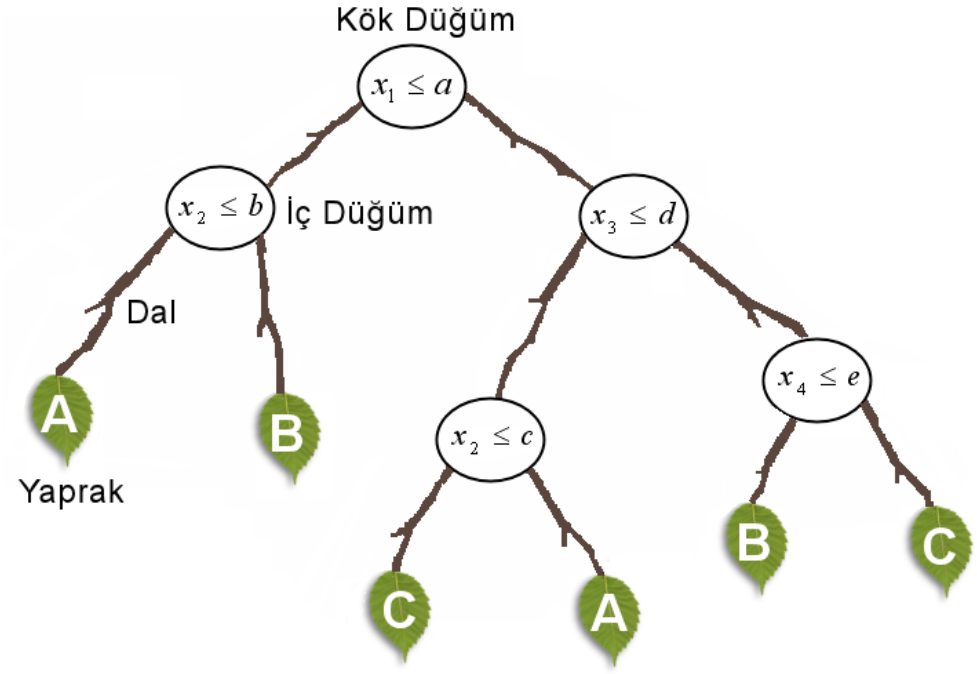
4.MODELLEME

Bu aşamaya kadar yukarıda anlatılan aşamalar geçilerek hazır edilen veri artık kullanılacak yöntemlerle anlaşılır hale getirilebilir. Modellemedeki asıl amaç verinin içindeki anlaşılır olmayan kurguyu anlamlı hale getirerek en yalın sonuçları gruplayarak ortaya çıkarmaktır. Tüm modellerin yer aldığı bir M kümesinden amaç , en iyi sonucu veren model oluşturmak ise (Piyush Rai) :

- Aynı öğrenme modeli farklı biçimlerde şekilde denenebilir
- Birbirinden farklı öğrenme modeli kullanılabilir

4.1.Karar Ağacı Algoritması

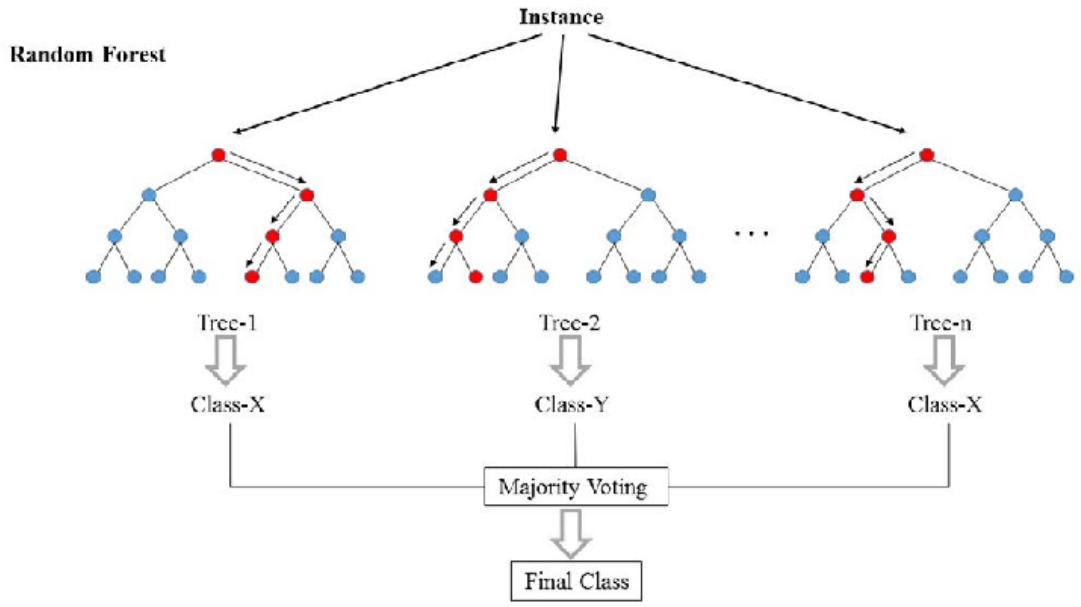
Karar ağacının yapısında düğüm, yaprak ve dal olmak üzere 3 bileşen bulunmaktadır. Ağacın atası ilk düğüm, kök olarak adlandırılmaktadır. Bir düğümün en uç düğümleri, çocukları ise yaprak olarak adlandırılmaktadır. Ara bileşenler ise dallarıdır. Bir düğüm o ağacın özniteliklerini temsil eder. Bir sınıflandırma algoritması tasarlanırken kök düğümden başlanarak sorgu işlemlerine göre ağacın derinliklerine doğru ilerlenir. Sorma işlemleri ağacın yapraklarına ulaşılana dek devam etmektedir. Makina öğrenmesinde bu algoritma uygulanırken oluşturulan eğitim seti kök düğümden başlanarak uygulanır. Alınan sonuçlara göre alt düğümlere doğru ilerlenir. Yaprak düğümlerine ulaşılana dek bu işlem devam etmektedir



Şekil 4.a : Karar Ağacı Algoritması

4.2.Rassal Orman Algoritması

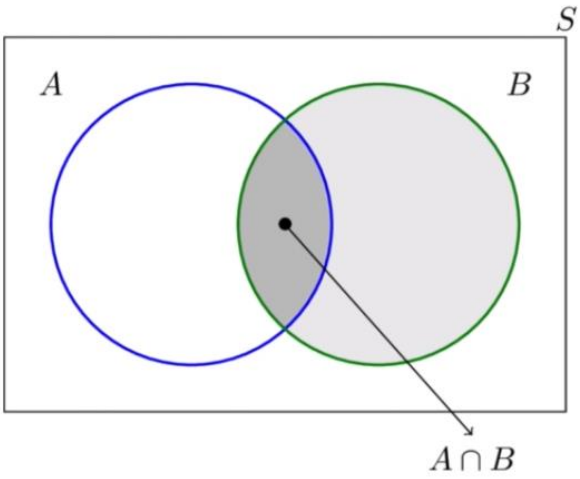
Toplu bir sınıflandırma algoritması olan rassal orman algoritması temelde bir karar ağacı algoritmasıdır. Birden fazla karar ağacının bir araya gelmesiyle isminden de anlaşıldığı gibi bir ormandan oluşur. Bu orman veriyi birçok parçaya bölerek birden fazla karar ağacı oluşturur. Oluşturulan birden fazla karar ağaçları ise çoğunluğun kararıyla (majority voiting) sınıflandırma işlemini gerçekleştirir. Temel bileşenleri karar ağacı gibidir. (bknz. 4.1 Karar Ağacı Algoritması) Topluluğa çok ağaç eklenmesi test kümesinde hatalı tahmin sayısını azaltmaktadır. Fakat ağaç sayısının artması da tekrarlı eğitime ve algoritmanın yavaşlamasına sebebiyet verebilmektedir.



Şekil 4.b : Rassal Orman Algoritması

4.3.Naive Bayes Algoritması

Naive Bayes algoritmasının temeli , Bayes teoremine dayanmaktadır. Birbirinden bağımsız gerçekleşen A ve B olaylarının aynı anda gerçekleşirken,B olayınında gerçekleşmesi durumudur..Bu durum koşullu olasılık olarak ifade edilmektedir.



$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

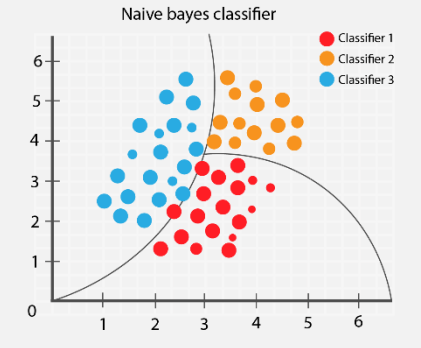
Şekil 4.c :Bayes Teoremi

Naive Bayes algoritmasını da yine A ve B kümeleri üzerinden değerlendirecek olursak B durumu gerçekleştiğinde A durumununda gerçekleşme olasılığı hedef durum olarak açıklanabilir. Bu ihtimalin hesaplanması yapılırken A durumu gerçekleştiğinde B durumununda gerçekleşme ihtimali ile A durumunun gerçekleşme ihtimali çarpılarak çıkan sonucun, B durumunun gerçekleşme ihtimaline bölünür, Naive Bayes algoritması hedef durum için hesaplanmış olur.

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

using Bayesian probability terminology, the above equation can be written as

$$\text{Posterior} = \frac{\text{prior x likelihood}}{\text{evidence}}$$



Şekil 4.d : Naive Bayes Algoritması

Python'ın sklearn kütüphanesi içinde birden fazla Naive Bayes metodu tanımlanmıştır. En genel kullanılan yöntemler Gaussian , Multinominal ve Bernoulli Naive Bayes yöntemleridir.Gaussian Naive Bayes sürekli değerler için kullanılır. Büyükten küçüğe doğru uzaylar arası geçiş yapılan kümeler için kullanımı idealdir. Bu proje kapsamında oluşturulan veri setinde, tahmin edilmek istenen hem bağımlı hem de bağımsız değişkenler reel sayılara dönüştürüldüğünden bu yöntem kullanılabilir.

Multinominal Naive Bayes'de, oluşturulan nominal değerler içeren veri setinde sınıflandırma yaparken her bir eleman için sayısal bir değer atanır, sınıflandırma bu değerler üzerinden yapılır. Bu proje kapsamında bu yöntem ekstra bir yük oluşturacağı düşünüldüğünden model olarak tercih edilmemiştir.

Bernoulli Naves Bayes yöntemi ise ikili tahmin yöntemi olarak adlandırılabilir. Yine nominal değerler üzerinde kullanılır. Bu değerler boole cebiri mantığıyla sınıflandırılır ve içerikler; doğru(true) =1 , yanlış(false) = 0 durumlarına göre kategorize edilir. Veri seti başka türde veriler içerirse bu yöntem tercih edilemez. Hedef değişkeninin iki sonuçtan birine ait olması, tüm değişkenlerin sıfır ve bir değerleriyle eşleştirilmesi

durumu en uygun örnek kullanım senaryosudur. Bu projede bağımlı değişkenlerin 2'den fazla oluşu nedeniyle bu yöntemde bu projede kullanılamamaktadır.

5.MODEL DEĞERLENDİRME VE SEÇİMİ

Projede Python dili ve ağırlıklı olarak Python sklearn kütüphaneleri, görsel form içinde tkinter kütüphanesi kullanıldı. Nedeni, Python'ın makine öğrenmesinde en çok kullanılan iki dilden birisi olması ve zengin fonksiyonlar ve içerik içermesidir. Birçok dökümana da ulaşmak bu sayede mümkün olmaktadır. Sklearn kütüphanesi ise, ihtiyaç duyulan temel yöntemlerin birçoğunu bünyesinde barındırması ve veri analitiği uygulamalarını problemsiz yürütülmesine olanak sağlamaktadır. İçerdiği modüllerle zaman zaman karşılaşılan eksik verileri sklearn doldurmakla beraber, öznitelik seçimi, çapraz doğrulama işlemleri yaparak sonuçları değerlendirme olanağı da sağlamaktadır.

6.MODELİN UYGULAMAYA GEÇİRİLMESİ

Bir ilaç prospektüs sitesinden hazırlanan Python kodlarıyla veri çekilerek bir veri seti oluşturuldu ve makinanın öğrenmesi için hazırlanarak .csv formatında kaydedildi.
Veri analizi, tahmin, sınıflandırma aşamaları uzun süren çalışmalar sonucu her biri kendi içindeki tüm sınıflarla tek tek incelenerek örnek veriler üzerinde denemeler yapıldı ve ana problemin çözümü için uygulamaya geçişi tasarlandı. Ürünü görselleştirmek, grafiksel kullanıcı arayüzü (GUI) tasarlamak için tkinter kütüphanesi kullanılarak bir form nesnesi oluşturularak girdi ve çıktıları bu form nesnesi üzerinden görüntülenmiştir

Şekil 6.a: Uygulama Formu

Uygulamayı çalıştırdıktan sonra şekil 6.a da ki uygulama formu ekrana gelmektedir. Uygulamada her bir ilaç için beş etken madde girilmesi yeterli görüldüğünden, beş etken madde etiketi eklenmiştir. Her bir etken maddenin karşısında bir seçim menüsü yer almaktadır. Bu menüde ilaç prospektüsünde yer alan etken maddeler tek tek girilmektedir. Bir ağrı kesici olan Parol Plus için şekil 6.b de etken maddeler ilaç prospektüsündeki içeriğe uygun girilmiştir.

Şekil 6.b: Parol Plus etken maddelerinin seçim menüsüne eklenmesi

Uygulama formunun sağ tarafında, kullanılan her bir algoritma için bir buton eklenmiştir. Bir ilaç prospektüsünde yer alan etken maddeleri tek tek girildikten sonra algoritma butonları sırasıyla çalıştırılır. Her bir algoritmanın çıktısı alt orta kısımda bir metin içinde yer almaktadır. Sol alt kısımda algoritma etiketleri buton sırasına göre yazılmıştır. Algoritma etiketlerinin karşılarında yer alan metinde, her bir algoritmayı çalıştırdıktan sonra hangi ilaç türüne ait ise uygulama tarafından ilgili algoritma ile sınıflandırılıp çıktısı metin içine yazılmaktadır.

Şekil 6.c: Algoritma çıktıları

Algoritma çalıştıktan sonra şekil 6.c de görüldüğü gibi algoritmalar başarılı bir şekilde sınıflandırma yapmaktadır. Parol Plus prospektüsünde yer alan etken maddeler girildiğinde üç algoritmada ağrı kesici olarak sonuçları vermiştir.

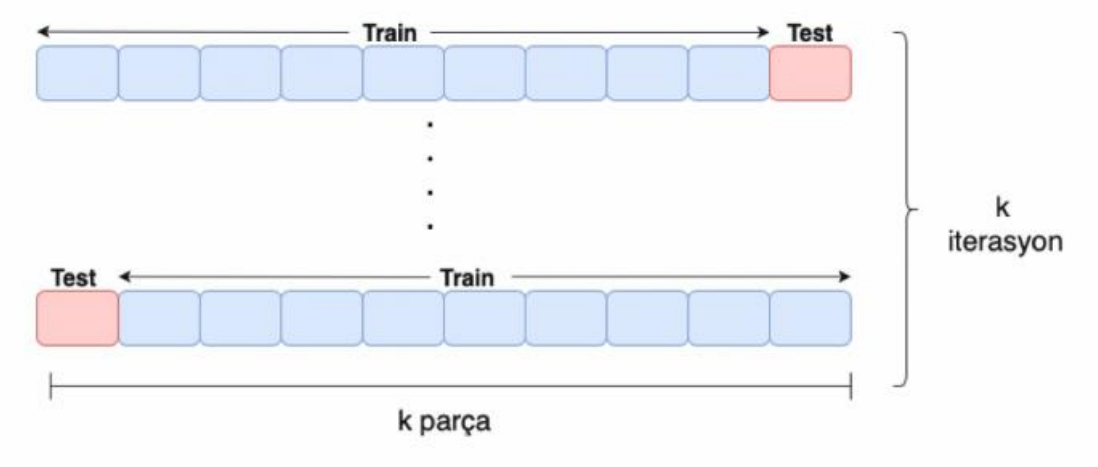
7.WEKA'DA VERİ SETİ DEĞERLENDİRME

Makine öğrenmesi ile alakalı birçok kütüphaneyi içinde bulunduran, hazır paket programlardan biri olan WEKA, Waikato Üniversitesi tarafından geliştirilmiş açık kaynak kodlu bir yazılımdır. Yüzde seksen eğitim, yüzde yirmi test olarak bölünen veri seti, birleştirilerek WEKA'da Rassal Ağaç ve Naif Bayes algoritmalarında sonuçlar gözlemlenmiştir.

7.1.Çapraz Doğrulama (Cross Validation)

Makine öğrenmesinde başarının değerlendirilmesi için sıklıkla kullanılan bir yöntem olan çapraz doğrulama yöntemi literatürde k katlamalı çapraz doğrulama olarak tanımlanır. Bu yöntem WEKA uygulamalarında sıkça kullanılmaktadır. Öncelikle ana veri seti her defasında biri test, kalan kısımlar ise eğitim kümesi olarak bölünür ve seçilen algoritma uygulanır. Çıkan sonuç ilk başarı yüzdesi olarak alınır. Daha sonra bölünen ikinci test, eğitim kümesi için aynı işlemler yapılır ve k. sayıda bölünen veri seti, k sayıda iterasyon yapılana kadar bu işlem sürer. En son çıkan tüm sonuçların ortalaması alınarak başarı yüzdesi hesaplanır. Genel olarak çapraz doğrulama değeri

’10’ alındığı için bu uygulamada da çapraz doğrulama değeri ’10’ olarak kullanılmıştır.



Şekil 7.a: k katlamalı çapraz doğrulama

7.2 Karışıklık Matrisi(Confusion Matrix)

WEKA’da makine öğrenmesi algoritmaları uygulanırken sonuçların doğruluğu ve başarısını ölçmek için kullanılan ölçütlerden biri de karışıklık matrisidir. Uygulamada n sayıda bağımlı değişken varsa, n*n boyutunda bir matris oluşturarak bu matris içerisinde tahmin değerlerinin doğruluğunun gösterilmesinde kullanılan ölçütlerden birisidir .

		Tahmin Edilen Sınıf	
		Pozitif (P)	Negatif(N)
Gerçek Sınıf	Pozitif (P)	TP (True positive)	FN (False negative)
	Negatif(N)	FP (False positive)	TN (True negative)

Şekil.7.b: Karışıklık Matrisi(Confusion Matrix)

Burada;

‘True positive’ : Doğru tahmin edilen pozitif sınıf değeri

‘False negative’ : Yanlış tahmin edilen negatif sınıf değeri

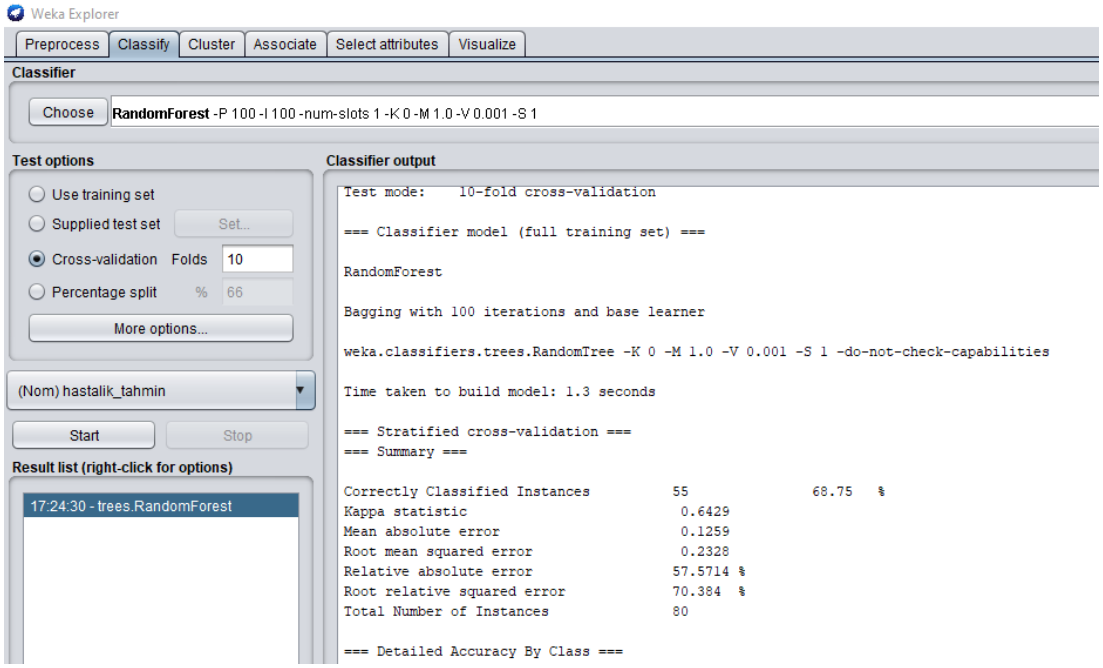
‘False positive’ : Yanlış tahmin edilen pozitif sınıf değeri

‘True negative’ : Doğru tahmin edilen negatif sınıf değeri

olarak tanımlanır. Bu uygulamada da her bir ilaç için karışıklık matrisi oluşturulmuştur. Hangi tahminin, ne ölçüde doğru, ne ölçüde yanlış yapıldığı bu matris tarafından açık bir şekilde gözlemlenmektedir.

7.3.Rassal Orman Algoritması

WEKA uygulamasında sınıflandırma algoritmalarını içinde barındıran 'explorer' uygulaması seçildi. Daha sonra 'Random Forest Algoritması' seçilerek çapraz doğrulama değeri '10' olarak girildi. Başarı oranı %68.75 olarak gözlemlendi.



Şekil 7.c: WEKA'da Rassal Orman Algoritması

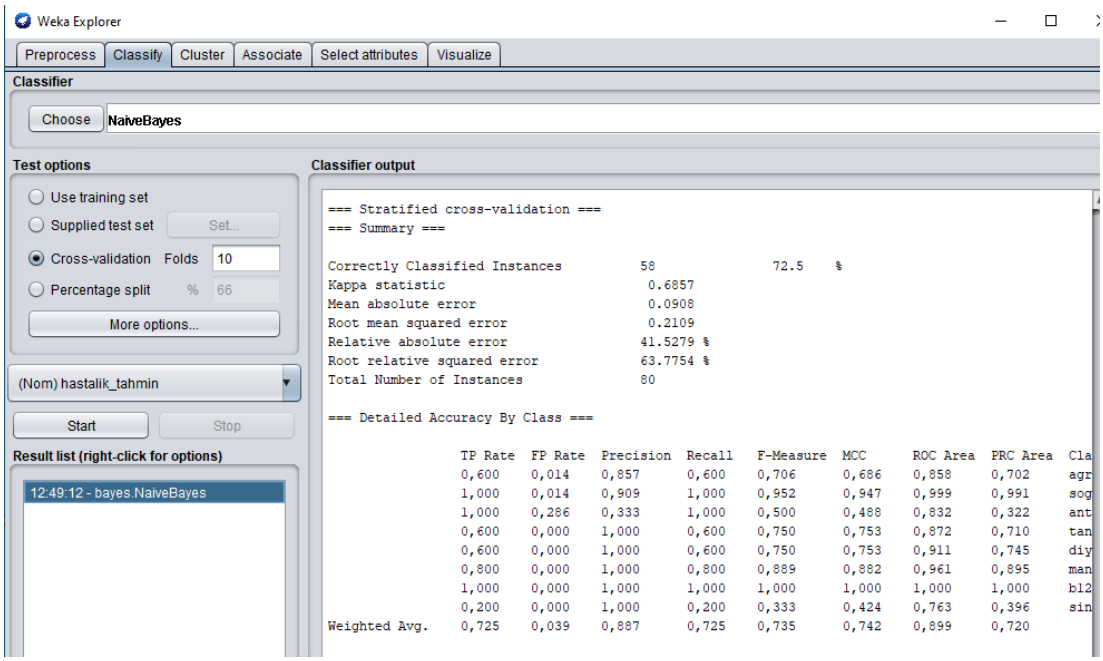
```
=== Confusion Matrix ===

 a b c d e f g h   <-- classified as
 6 1 2 0 1 0 0 0 | a = agri_kesici
 0 9 0 0 1 0 0 0 | b = soguk_alginligi
 0 0 9 0 1 0 0 0 | c = antidepresan
 0 0 4 6 0 0 0 0 | d = tansiyon
 0 0 4 0 6 0 0 0 | e = diyabet
 0 0 2 0 0 8 0 0 | f = mantar
 0 0 1 0 0 0 9 0 | g = b12_vitamini
 1 0 7 0 0 0 0 2 | h = sindirim_sistemi
```

Şekil 7.d: Karışıklık Matrisi Sonuçları

Bu tez çalışmasında sekiz farklı ilaç türü veri setinde hazırlandığı için 8*8'lık bir matris oluşmuştur. Burada her bir satır değeri, satırda belirtilen ilaç türünü ait tahmin sayısını, sütun değerleri ise kaç tahminin hangi ilaç türü olarak sonuçlandığını göstermektedir.

7.4.Naive Bayes Algoritması

WEKA uygulamasında sınıflandırma algoritmalarını içinde barındıran 'explorer' uygulaması seçildi. Daha sonra 'Naive Bayes Algoritması' seçilerek çapraz doğrulama değeri '10' olarak girildi. Başarı oranı %72.50 olarak gözlemlendi.



Şekil 7.e: WEKA'da Naive Bayes Algoritması

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h   <-- classified as
  6  1  3  0  0  0  0  0 |  a = agri_kesici
  0 10  0  0  0  0  0  0 |  b = soguk_alginligi
  0  0 10  0  0  0  0  0 |  c = antidepresan
  0  0  4  6  0  0  0  0 |  d = tansiyon
  0  0  4  0  6  0  0  0 |  e = diyabet
  0  0  2  0  0  8  0  0 |  f = mantar
  0  0  0  0  0  0 10  0 |  g = b12_vitamini
  1  0  7  0  0  0  0  2 |  h = sindirim_sistemi
```

Şekil 7.f: Karışıklık Matrisi sonuçları

Bu tez çalışmasında sekiz farklı ilaç türü veri setinde hazırlandığı için 8*8'lık bir matris oluşmuştur. Burada her bir satır değeri, satırda belirtilen ilaç türünü ait tahmin sayısını, sütun değerleri ise kaç tahminin hangi ilaç türü olarak sonuçlandığını göstermektedir

8.SONUÇLAR VE ÖNERİLER

Makine öğrenmesi yöntemleri, Python'ın zengin kütüphanelerinin içeriklerinden faydalanılarak, sınıflandırma işlemlerinde yaygın olarak kullanılan Karar Ağacı, Rassal Orman, Naif Bayes algoritmaları ile gerçekleştirildi. Bu uygulamada üç farklı sınıflandırma algoritma kullanılma sebebi, bu algoritmaların yüzde yüzlük bir başarı ile çalışmadığı için en az iki algoritmanın verdiği çıktı doğru kabul edilmektedir. Genel olarak yapılan gözlemlerde de bu tez desteklenmektedir. WEKA içinde veri seti hazırlanarak başarı yüzdeleri bu algoritmalar için gözlenmiştir. Yapılan tez çalışması kapsamındaki başarı gözlemleri, WEKA uygulamasına oranla daha iyi sonuçlar vermektedir.

Makine öğrenmesi ve yapay zeka uygulamaları artık çağımızın önemli alanları arasına girmiştir. Büyük verilerin anlamlandırılması, işlenmesi, anlaşılır hale getirilmesi büyük önem arz etmektedir. Özellikle kullanıcıların ilgi alanlarına, isteklerine göre en doğru ürünlerin optimize edilerek kişilere sunulması her yönüyle tasarruf haline gelmektedir. Bu noktada makine öğrenmesinin etkili kullanılması şirketleri hatta ülkeleri birçok alanda ön plana çıkarmakta ve fayda sağlamaktadır. Bu alanda çalışmaya başlayacak kişiler, veriyi hazırlama kısmının da modelleme kadar önemli olduğunu bilmelidirler. Çünkü veri hazırlama kısmı çok zahmetli, dikkat gerektiren bir bölümdür. Bazen bu süreç aylar sürebilmektedir. Dolayısıyla bu alanda en güncel teknolojileri takip etmek, büyük veri kümesinde çalışıyormuş gibi en baştan tasarım yapmak, ilerde daha büyük veri setleri üzerinde çalışırken sorun yaşamamak adına en önemli noktadır. Mevcut algoritmalar üzerinden zayıf yönler tespit edilerek, yeni algoritmalar üretmek ve bunları makine öğrenmesi ile sınıflandırma yaparken uygulamak, bu alandaki bir diğer önemli noktadır. Sonuç olarak bu iki kısmın başarılı bir şekilde gerçekleştirilmesi hem bilime hem de insanlığa büyük faydalar sağlayacağı çıkarılacak en önemli sonuçtur.

9.LİTERATÜR TARAMASI

•Alpaydın E. (2004) “Introduction to Machine Learning”, The MIT Press, 3-6

•Machine Learning in Medicine- a Complete Overview- Ton J.Cleophas,Aeilko H.Zwinderman

•An Introduction to Machine Learning / Miroslov Kubat

•Principles of Data Mining / Max Bramer

•Data Mining : Concepts and techniques,Han,Kamper,Pei 2001

•http://www.bilkav.com

•Shannon : model of communication, 1948

•Data minnig: Practical machine learning tools and techniques, Lan H.witten,Eibe Frank,Mark Hall,Christoper J.Pal

10.YAPILAN BENZER PROJELER

Makine öğrenmesi yöntemlerinin polisomnografik verilere uygulanması : (http://dspace.trakya.edu.tr/xmlui/bitstream/handle/1/1667/127.pdf?sequence=1&isAllowed=y)

Makina öğrenmesi yöntemleri ile glokom hastalığının teşhisi:

(http://acikerisimarsiv.selcuk.edu.tr:8080/xmlui/handle/123456789/1256)