T.C.
İSTANBUL MEDENİYET ÜNİVERSİTESİ
MÜHENDİSLİK VE DOĞA BİLİMLERİ FAKÜLTESİ

MEZUNİYET ÇALIŞMASI

PARKİNSON HASTALIĞININ TEŞHİSİNDE DESTEK VEKTÖR MAKİNESİ MODELİNİN MATEMATİKSEL ANALİZİ VE DİĞER MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE KARŞILAŞTIRILMASI

Elif ÇETİNZAFER-19120808019

MATEMATİK BÖLÜMÜ

Danışman

Doç. Dr. Betül HİÇDURMAZ

Temmuz, 2023
İSTANBUL

ÖNSÖZ

Parkinson hastalığı, beyinde dopamin üreten beyin hücrelerinin ölmesiyle ya da zarar görmesiyle ortaya çıkan bir beyin hastalığıdır. Hastalığa sahip kişilerin konuşurken sesleri daha titreşimli, güçsüz ve monotondur. Bu nedenle konuşma bozuklukları hastalığın belirtileri arasında önemli bir yere sahiptir. Hastalığın makine öğrenmesi algoritmaları ile teşhis edilmesinde ses verisi oldukça tercih edilen bir yöntemdir. Bu çalışmada da 31 bireyden alınan ses verisi ile hastalığın teşhis edilmesi amaçlanmıştır. Bu bir sınıflandırma problemidir. Ses verisine ait sürekli değişkenler arasında çoklu doğrusal bağlantı problemini ortadan kaldırmak için Temel Bileşen Analizi (PCA) uygulanmıştır. PCA uygulanırken Tekil Değer Ayrışımı ile öznitelik matrisinin özdeğer ve özvektörleri hesaplanmıştır. Özdeğerler aracılığıyla bileşenlerin önem sırası belirlenmiştir. PCA analizi sonucunda 5 bileşenden oluşan ve datanın %87.52' sini açıklayabilen yeni veri seti oluşturulmuştur. Parkinson hastalığının ses verisi üzerinden sınıflandırılmasında SVM, Logistik Regresyon, K-NN, RF, CART ve YSA algoritmaları ile modeller oluşturulmuştur ve karşılaştırılmıştır. Oluşturulan modeller arasında en iyi sınıflandırmayı yapan algoritmanın %90 doğruluk skoru ile SVM olduğu sonucuna varılmıştır.

İçindekiler

1 GİRİŞ

1.1 Parkinson Hastalığı Nedir?

Parkinson hastalığı beyin kaynaklı motor refleks, konuşma vb. hayati fonksiyonların yerine getirilmemesine sebep olan merkezi sinir sistemine ait nörolojik bir hastalıktır. Beyinde dopamin üreten bölgedeki hücrelerin tahrip olmasıyla veya ölmesiyle ortaya çıkan bir hastalıktır. Adını hastalığı ilk defa 1817'de titremeli felç olarak tarifleyen James Parkinson'dan almıştır (1).

Beyinde dopamin üreten bölgedeki hücrelerin kaybolmasıyla veya bu maddenin daha az salınımı sonucunda vücudun hareket kontrol merkezi etkilenir ve beyin normal fonksiyonlarını yerine getiremez. Hastalık ilk yıllarda hafif olduğundan hastalar önemli motor bozukluklarıyla karşılaşmayabilirler. Ancak ilerleyen zamanlarda hareketleri yavaşlar, kas hareketlerini kontrol edemez ve el, kol, bacak, çene gibi uzuvlarda titreme meydana gelir. Titreme genellikle istirahat halindeyken artış gösterir. Hastalığın tedavisi genellikle, görülen semptomları hafifletmek veya hastalığın ilerlemesini yavaşlatmak için yapılır. Hastalığın başlarında genellikle ilaç tedavisi kullanılarak beyindeki dopamin seviyesini artırarak hastanın hareket kontrolünü iyileştirir. Sürekli kullanılan ilaçların etkileri zamanla azalır. Hastalığın aynı zamanda cerrahi tedavileride bulunmaktadır ve bazı hastalarda derin beyin stimülasyonu etkilidir. Ayrıca hastalığın konuşma bozuklukları gibi önemli karakteristik belirtileri de bulunmaktadır. Parkinson hastalarının günlük yaşamlarını kolaylaştırmak için genellikle fizyoterapi, konuşma terapisi ve beslenme desteği gibi yöntemler tercih edilmektedir. Hastalık sürecinde psikolojik danışman veya destek grupları sayesinde yaşam kaliteleri artmaktadır.

1.2 Parkinson Hastalığının Nedenleri

Parkinson hastalığı beyinde dopamin üreten bölgedeki hücre kaybı nedeniyle veya bu maddenin az salınımı sonucu oluşur. Bu hücrelerde meydana gelen tahribata genellikle zirai ilaçlar gibi kimi kimyasallar neden olsa da genetik faktörlerin de sebep olmaktadır. Bazı genlerdeki mutasyonlar hastalığın ortaya çıkma riskini artırmaktadır. Ayrıca sigara içmek ve kahve tüketmek parkinson hastası olma riskini azalttığı düşünülmektedir..

1.3 Parkinson Hastalığının Belirtileri Nelerdir?

Parkinson hastalığı yavaş ilerleyen sinsi bir hastalıktır. Bu nedenle ilk aşamada belirtileri aniden ve şiddetli bir şekilde görünmez. Genellikle ilk aşamada kişilerin konuşmalarında veya hareket kabiliyetlerinde bozulmalar meydana gelmektedir. Kişiler daha mimiksiz, monoton bir konuşmaya sahip olurlar. Vücutları ise eğilimli bir görünüme sahip olur.

Hastalık genellikle 65 yaş ve üstü kişilerde ortaya çıktığından yaşlılık belirtisi ile karıştırılabilmektedir. Bu ancak nöroloji doktorları tarafından anlaşılmaktadır. Ayrıva hastalık genellikle kendini istirahat halindeyken göstermektedir. Hastalığın beyinde yarattığı tahribat arttıkça hastalığn belirtileri de gözle görülebilir hale gelmektedir.

Şekil 1: Parkinson hastalığını fiziksel belirtileri

1.4 Parkinson Hastalığı Tanı Yöntemleri Nelerdir?

Parkinson hastalığının tespit edilebilmesi için nöroloji hekimlerine müracaat edilmesi gerekir. Doktorlar tarafından, başvuran kişilerin tıbbi geçmişleri incelenir. Genellikle ilk aşamada

kişilere fiziksel muayene uygulanır. Bu fiziksel muayene sırasında kişilerin el, kol, bacak veya çene gibi uzuvlarında titreme, hareket kabiliyetinde azalma, denge problemi ve kol - bacak - gövde bölgesinde katılaşma gibi ana motor belirtilerinin kişide bulunup bulunmadığı kontrol edilir. Eğer kişi bu belirtilerden iki veya daha fazlasına sahip ise radyolojik görüntüleme yöntemlerine başvurarak hastalığın tespiti yapılabilir (12).

1.5 Parkinson Hastalığında Ses ve Konuşma Bozuklukları

Genellikle hastalığa sahip kişilerin el, kol, bacak ve çene gibi uzuvlarında titreme meydana gelmektedir. Parkinson hastalığı tanısı alan olguların %60-90'ında çeşitli konuşma ve ses bozuklukları görünmektedir (9). Bu bozukluklar, konuşma hızında yavaşlama, konuşurken duraklama, ses yüksekliğinde değişiklikler, konuşmanın anlaşılabilirliğinde azalma vb.'dir. Konuşma akıcılığının bozulması, anlaşılırlık düzeyi ve sosyal performansları tarafından olumsuz etkilenmektedir (11). Bu ise kişilerin hayatlarını önemli ölçüde etkilemektedir. Parkinson hastalığına sahip kişilerin konuşmaları genellikle monoton, güçsüz ve titreşimlidir. Bu ise hastalığa sahip kişilerin ses frekansındaki ve tonundaki değişimlerin azalmasıyla ilişkilendirilebilir. Kişilerin seslerinde meydana gelen değişimler kolayca tespit edilmektedir. Bu durum ise Parkinson hastalığın teşhisinde önemli bir yere sahiptir. Ayrıca ses verisinin kolay elde edilebilmesi ve uygulanabilmesi açısından da oldukça tercih edilen bir yöntemdir. Bu çalışmada da Parkinson hastalığı ses verisi ile tespit edilmeye çalışılmıştır.

1.6 Parkinson Hastalığının Ses Çıkarımı ile Teşhisine Yönelik Yapılan Çalışmalar

Teknolojinin gelişmesiyle birlikte Parkisnon hastalığının tanı ve teşhisinde doktorlara yardımcı olmak için birçok yöntem geliştirilmiştir. Bu yöntemlerden biri de ses ve konuşma bozuklukları ile teşhisidir. Ses verisinin tercih edilmesinin birçok nedeni vardır. Kolay veri elde edilmesi ve uygulanabilirliği açısından oldukça kolay olması da bu nedenlerden biridir.

Little ve arkadaşları Parkinson hastalığının belirtilerinden biri olan ses kısıklığını ölçerek sesin şiddetini derecelendirmeyi baz aldıkları bir çalışma yürütmüşlerdir. Çalışma da 23 Parkinson hastası ve 8 sağlıklı bireylere ait ses verisinden yararlanmışlardır. Parkinson hastalığının teşhisi ve hangi aşamada olduğunun tespitinde kullanılabilecek sesteki hızlı dalgalanma, titreşim, ses düzeyi ve harmoniklik gibi özelliklerin çıkartılması için konuşma testleri uygulanmış ve %91,4'e yakın bir başarı elde edilmiştir.

Das yapmış olduğu benzer bir çalışmada UCI veri tabanından elde ettiği 23'ü hasta 31 bireye ait veriler için yapay sinir ağları, karar ağaçları ve regresyon metotlarını içeren farklı sınıflandırma yöntemlerini kullanmıştır. %92,9 doğru sınıflandırma sonucunu sinir ağları uygulaması ile elde etmiştir (Das, 2010) (5).

2 Parkinson Hastalığı ve Ses Analizi

2.1 Veri Seti

Oxford Üniversite'sindeki Max Little tarafından, konuşma sinyallerini kaydeden Denver veri setini, Colorado Ulusal Ses ve Konuşma Merkezi ile işbirliği içinde oluşturulmuştur. Oluşturulan bu veri setinin amacı, durum çubuğuna bağlı olarak Parkinson hastalığına sahip olan kişilerle sahip olmayan kişileri birbirinden ayırabilmektir. Bu amaç için bireylerden ses verisi alınmış ve ses sinyalleri üzerinden çeşitli öznitelik çıkarma teknikleri kullanılarak veri seti oluşturulmuştur. Bu ölçümler 23'ü Parkinson hastası, 8'i sağlıklı olan toplam 31 bireyden alınmıştır. Kişilerin bazılarından 6 bazılarından 7 adet ses kaydı alınmıştır. Durum çubuğunda sağlıklı olduğu bilinen kişiler 0, parkinson hastası olduğu bilinen kişiler ise 1 ile gösterilmiştir.

Parkinson hastalığının ses analizi çalışmasında Oxford Üniversitesin'den alınan veri setinin kullanılmasının sebebi (2) sitesinden alınan bilgilere göre; ses analizinde değerlendirilen en basit ses, parametre, f0'dır ve temel frekanstır. Akustik ses analizinde kullanılan parametrelerin genel olarak f0 ve f0 değişiklikleri; Ötüm parametreleri (Voicing grubu); Frekans pertürbasyonu parametreleri (Jitter grubu); Genlik (Amplitüd) pertürbasyon parametreleri (Shimmer grubu); Spektral parametreler (Harmonicity grubu); GHO (gürültü-harmonik oranı) ve HGO (harmonik-gürültü oranı) olduğunu gördük.

Veri setindeki değişkenler aşağıdaki şekildedir:

- MDVP:Fo(Hz) - Ortalama vokal temel frekansı
- MDVP:Fhi(Hz) - Maksimum vokal temel frekansı
- MDVP:Flo(Hz) - Minimum vokal temel frekansı

- MDVP:Jitter(%),MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP - Birçok temel frekanstaki değişim ölçüleri
- MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA - Genlikteki çeşitli varyasyon ölçüleri
- NHR,HNR - Sesteki gürültünün ton bileşenlerine oranının iki ölçüsü
- status - Deneğin sağlık durumu (bir) - Parkinson, (sıfır) - sağlıklı
- RPDE,D2 - İki doğrusal olmayan dinamik karmaşıklık ölçüsü
- DFA - Sinyal fraktal ölçekleme üssü
- spread1,spread2,PPE - Temel frekans değişiminin üç doğrusal olmayan ölçümü

2.2 Öznitelik seçimi

Öznitelik seçimi için Temel Bileşenler yaklaşımından yararlanılmıştır. Temel bileşenler yaklaşımı özniteliklerden oluşan matrisin sütunları arasındaki lineer bağımlılık yapısını yok etme ve boyut indirgeme amaçları için kullanılmaktadır. Yüz tanıma, sınıflandırma, boyut indirgemesi ve yorumlanmasını sağlayan, çok değişkenli bir istatistik yöntemdir. Bu yaklaşım verinin içindeki en güçlü örüntüyü bulmaya çalışır.

Bu çalışmada veri setindeki yüksek korelasyona sahip sürekli değişkenlerin çoklu doğrusal bağlantı problemini ortadan kaldırmak için Temel Bileşen Analizi (PCA) uygulanmıştır. PCA dönüşümünün arka planında Tekil değer ayrışımından yararlanılmaktadır. Kısaca Tekil değer ayrışımdan bahsedelim.

Tekil Değer Ayrışımı (TDA), bir matrisin sütunları arasında lineer bağımlı olup olmadığını tespit etmek için kullanılan bir yöntemdir. Matrisi üç bileşen matrisi şeklinde ayrıştırılmasına dayanır. Tekil Değer ayrışı bir çok alanda kullanılır. Google'ın PageRank algoritmasından insan yüzü modellemeye, görüntü sıkıştırma, gürültü filtreleme, veri sıkıştırma vb. bir çok alanda kullanılır.

Denkelemlerde skaler değerli ifadeler küçük harflerle, vektörler kalın küçük harflerle, matrisler kalın büyük harflerle yazılmıştır. ^{H} işareti Hermit (karmaşık) transpozu göstermek üzere kullanılmıştır. $a \in C^t$ vektörünün L_2 normu (Euclidean normu) $||.||_2$ sembolleri ile gösterilir.

L_2 normu aşağıdaki şekilde tanımlanmıştır.

$$||a||_2 = \sqrt{|a_1|^2 + ... + |a_t|^2} = \sqrt{a^H a} \tag{2.1}$$

$\mathbf{A} \in C^{[x]}$ matrisinin spektral normu, benzer şekilde $||.||_2$ sembolleri ile gösterilir ve $A^H A$ matrisinin en büyük özdeğerlerinin karekköküne eşittir;

$||\mathbf{A}||_2$=$A^H A$ matrisinin en büyük özdeğerinin karekökü

$$||A||_2 = max_{||x||_2} \neq 0,\ \frac{||Ax||_2}{||x||_2} \tag{2.2}$$

Ortogonal $\mathbf{A} \in C^{[x]}$ matrisinin sütunlarının iç çarpımları sıfır ve sütun normları ise 1'dir. Eğer matris kare ise, üniter matris olarak adlandırılır ve $A^H A = 1$ denkliğini sağlar.

Her $\mathbf{A}\in C^{[x]}$ matrisi

$$\mathbf{A}=\mathbf{U}\ \sum V^H$$

şeklinde çarpanlarına ayrılır ve aşağıdaki maddeleri sağlar.

- $U = [u_1 u_2 .....u_I] \in C^{[x]}$ üniter bir matristir.
- $V = [v_1 v_2 .....v_I] \in C^{[x]}$ üniter bir matristir.
- $\sum \in C^{[x]}$ sözdeköşegen (pseudodiagonal) bir matristir.

$\sum$ = köşegen $(\sigma_1, \sigma_2, ...\sigma_{\min_{I,J}})$ ve sıralıdır $(\sigma_1 \geq \sigma_2 \geq .... \geq \sigma_{\min_{I,J}} \geq 0)$

$A : nxp$ boyutunda bir matris ve $rank(A) = r$' dir. $U : nxn, V : pxp$ ortogonal kare matris olmak üzere

$$A = U \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} V' \tag{2.3}$$

şeklinde çarpanlarına ayrılır. Burada, U matrisi AA' nın ve V matrisi ise $A'A$ normlanmış özvektörlerini içeren matrislerdir. D matrisi, AA' matrisinin ($A'A$ matrisinin) sıfırdan farklı $d_i (d_i > 0, i = 1, 2...., r)$ özdeğerlerinin karekökünden yani tekil değerlerden oluşan

$$D = \begin{bmatrix} \sqrt{d_1} & \dots & \\ \vdots & \ddots & \vdots \\ & \dots & \sqrt{d_r} \end{bmatrix} \tag{2.4}$$

diagonal bir matristir.

U matrisinin ilk r sütunundan oluşan matris $U_1 : n \times r (U_1'U_1) = I_rV$ matrisinin ilk r satırından oluşan matris $V_1 : r \times p (V_1V_1') = I_rV$ olmak üzere,

$$A = U_1 D V_1'$$

Tersine, bu gösterimden yukarıdakine geçmek için D matrisi sıfır matrisleri ile genişletilecek ve U_1 ile V_1 matrisleri ortogonal matrisler olacak şekilde genişletilecektir. A matrisinin yukarıdaki gibi ayrıştırılmasına tekil değer ayrışımı denir. D matrisinin elemanlarına ise A'nın tekil değerleri denir (3). Temel bileşen analizinde tekil değerler bileşenlerin hangi varyans oranı ile açıklandığını gösterir.

İleriki bölümlerde sürekli değişkenler arasındaki yüksek korelasyondan dolayı meydana gelen çoklu doğrusal bağlantı probleminin çözümü için veri setine PCA dönüşümü uygulanacaktır.

3 SINIFLANDIRMA YÖNTEMLERİ

3.1 Sınıflandırma

Sınıflandırma farklı veri sınıflarını gruplandırabilen bir denetimli öğrenme algoritmasıdır. Bağımlı değişkeninin (Y) kategorik olduğu durumlarda sınıflandırma yöntemi kullanılarak model oluşturulur. Girdi olarak verilen X bağımsız değişkenlerinin sonucunda Y bağımlı değişkeninin hangi sınıfa ait olduğu tahmin edilir. Birden çok sınıflandırma yöntemi vardır. Bunlardan bazıları aşağıdaki gibidir;

- Destek Vektör Makinaları (SVM)
- k En Yakın Komşu (k-NN)

- Random Forest (Rastgele Orman)
- Lojistik Regresyon
- Yapay Sinir Ağları (YSA)

Bu çalışmanın ileriki bölümlerinde SVM ile diğer sınıflandırma yöntemleri karşılaştırılacaktır.

3.2 Sınıflandırma Yöntemlerinin Açıklanması

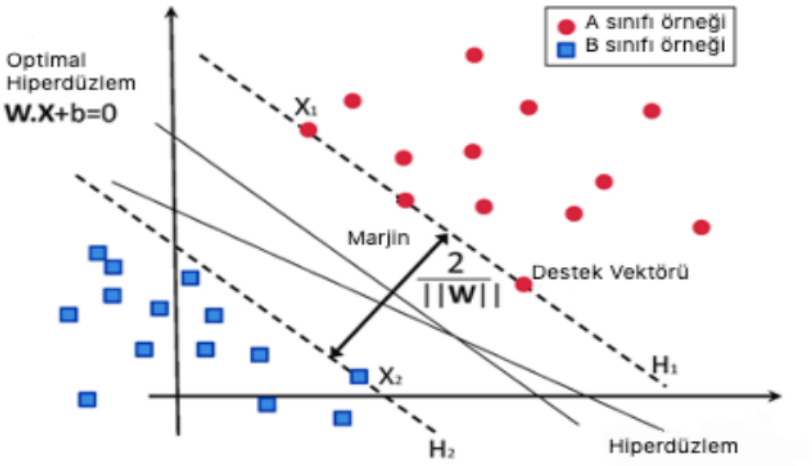
3.2.1 Destek Vektör Makineleri (SVM)

Destek vektör makineleri sınıflandırma ve regresyon analizinde kullanılır. Genellikle sınıflandırmada problemlerinde daha çok tercih edilen bir denetimli makine öğrenme yöntemidir. Daha iyi genelleme olanağı sunan ve modelin eğitimi aşamasında hatanın azaltılmasını sağlayan yapısal risk minimizasyonu ilkesini kullandığı için SVM'nin önemli avantajlara sahip olduğu bilinmektedir (13). Algoritmanın çalışması esnasında verilerin lineer olarak ayrılıp ayrılmamasına göre çekirdek fonksiyonları da kullanılabilmektedir. Bu durum SVM' in diğer sınıflandırma metotlarına göre daha çok tercih edilmesine neden olmuştur. Kullanılan çekirdek fonksiyonlarıyla doğrusal ve doğrusal olmayan sınıflandırma işlemlerini gerçekleştirebilmektedir. Eğer sınıflandırma işleminde, tam ayrıştırılabilir veriler kullanılırsa tüm verileri bir hiper düzlem ile ayırmak mümkündür. Eğer tam ayrıştırılamayan veriler kullanılırsa, sınıfları tek bir hiper düzlem ile ayırmak mümkün değildir. Bu durumda çekirdek fonksiyonu kullanarak verileri tam ayrıştırılabilir hale getirilir.

Destek vektör makine modeli ilk versiyonlarında, ikili veri kümesinin doğrusal vektörler yardımıyla sınıflara ayırmak için kullanılmaktaydı. İlerleyen yıllarda geliştirilerek destek vektör makine modellerinde çekirdek fonksiyonları da kullanılarak birden fazla sınıfı kolayca ayrılabilecek hale gelmiştir. Bu nedenle en çok tercih edilen sınıflandırma yöntemi olmuştur. Destek vektör makinesi, sınıflara ait verileri uygun olacak şekilde çekirdekler yardımıyla daha yüksek boyuta taşıyarak doğrusal olmayan bir haritalama kullanılmaktadır. Bu haritalama ile verilerin dönüştürüldüğü yeni uzayda iki sınıf arasındaki ayrımın optimum olmasını sağlayacak hiperdüzlemi bulmayı amaçlamaktayız. Marjin geometrik olarak her iki sınıfa ait en yakın gözlemler ile hiperdüzlem üzerindeki herhangi bir nokta arasındaki en kısa mesafe olarak tanımlanır. Bu veri noktalarına support vector (destek vektör) denir. SVM eğitim hatasını minimize eden tüm hiperdüzlemler arasından, iki sınıfa ait örnekleri birbirinden

ayıracak olan marjini maksimize edecek hiperdüzlemi bulmayı amaçlar.

Veri setindeki sınıf sayısına ve özniteliğine bağlı olarak kullanılacak SVM algoritması da değişmektedir. Eğer veriler lineer bir hiperdüzlem ile ayrıştırılıyorsa Hard margin, belirli bir hata ile ya da ek değişken ile lineer olarak ayrıştırılabilen veriler için Soft margin kullanılır. Eğer sınıfları bir hiperdüzlem ile ayrıştırmak mümkün değil ise kernel trick adı verilen çekirdek hilesi kullanılır. Burada çekirdek fonksiyonları kullanarak girdi uzayı daha yüksek boyutlu uzaya taşınır ve sınıflar ayrıştırılmaya çalışılır. Bu durumları ve arka planında çalışan matematiğine değinelim.



Şekil 2: SVM ile sınıflandırma

Lineer Olan SVM

Hard Margin

Lineer hiperdüzlem ile ayrılabilen veri kümelerine uygulanır. Şekil 2' de A ve B şeklinde iki sınıflı bir SVM modeli örneği bulunmaktadır. Veri setinde m adet gözlem olduğunu varsayalım. $(x_i, y_i),\ i = 1, 2, ..., m$ olacak şekilde

$$x_i^T = (x_{i1}, x_{i2}, ..., x_{id}) \in \mathbf{R^d} \tag{3.1}$$

i. gözlemin d boyutlu değişkeni şeklinde ifade edilir ve $y \in -1, +1$ şeklindedir.
Hiper düzleme en yakın olan veri noktalarına destek vektörleri denir. Destek vektör makinesinin amacı verileri maksimum marjinle ayıran optimum ayırma hiper düzlemini bulmayı amaçlar (24). Ayırma hiper düzlemi H'nin denklemi aşağıdaki gibi verilsin:

$$H(w, b) = \{x | x^t x + b = 0\} \tag{3.2}$$

w : Ağırlık değerleri (düzlemin normali)
b : Bias
x : Gözlem
$\frac{b}{||w||}$: Hiperdüzlemden orjine olan dik uzaklık
H' ye eşit uzaklıkta olan H_1 ve H_2 iki paralel hiper düzlem olacaktır. Veriler lineer olarak ayrıştırıldığından bu iki düzlem arasında veri yoktur ve düzlemlerin denklemleri aşağıdaki gibidir;

$$H_1 : (w^t x_i + b \geq 1)\ y_i = +1 \tag{3.3}$$

$$H_2 : (w^t x_i + b \leq -1)\ y_i = -1 \tag{3.4}$$

Burada x_i pozitif sınıfına aitse $y_i\ +1$ negatif sınıfa aitse $y_i,\ -1$ değerini alır. Ayrıca w, hiperdüzlemi belirleyen ağırlıkların vektörüdür, b ise sapmayı temsil eder. (3.3) ve (3.4) eşitsizliklerinden;

$$y_i(x_i.w + b) - 1 \geq 0, \forall_i \tag{3.5}$$

elde edilir.
Marjinin mümkün olabildiğince büyük olmasını amaçlamaktaydı. Bu durumda $Marjin = \frac{2}{||w||}$ uzunluğundadır ve marjini maksimize edebilmek için (3.5)' de verilen koşullar altında, w değerinin minimize edilmesi gerekir. Hesaplamalarda kolaylık sağlanması için $||w||$ yerine $\frac{1}{2}||w||^2$ alınır.

$$y_i(x_i.w + b) - 1 \geq 0, \forall_i \quad \text{koşulu altında} \qquad \min \frac{1}{2}||w||^2 \tag{3.6}$$

Verilen koşullar altında minimizasyon yapılabilmesi için, $\alpha_i \neq 0\ \forall_i$ olan kuadritip program-

lama probleminin lagrange çarpanları ile çözümünden yararlanılarak w ve b hesaplanır.

$$L_p \equiv \frac{1}{2}||w||^2 - \alpha[y_i(x_i.w + b) - 1], \forall_i \tag{3.7}$$

$$\equiv \frac{1}{2}||w||^2 - \sum_{i=1}^{L} \alpha_i[y_i(x_i w + b) - 1] \tag{3.8}$$

$$\equiv \frac{1}{2}||w||^2 - \sum_{i=1}^{L} \alpha_i y_i(x_i.w + b) + \sum_{i=1}^{L} \alpha_i \tag{3.9}$$

(3.9) eşitsizliğini minimize edecek w ve b değerleri ise maksimize edecek α değeri, L_p' nin w ve b' ye göre türevlerinin alınması ve bu türevlerin sıfıra eşitlenmesiyle elde edilir.

$$\frac{\partial L_p}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{L} \alpha_i y_i x_i \tag{3.10}$$

$$\frac{\partial L_p}{\partial b} = 0 \Rightarrow \sum_{i=1}^{L} \alpha_i y_i = 0 \tag{3.11}$$

Eşitlik (3.10) ve (3.11)'i, eşitlik (3.9)'da yerine yazarsak α değerine bağımlı ve bu α değerinin maksimize edilmesi gereken;

$$L_D \equiv \sum_{i=1}^{L} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i x_j, \quad \alpha_i \neq 0 \quad \forall_i \quad \sum_{i=1}^{L} \alpha_i y_i = 0 \tag{3.12}$$

$$\equiv \sum_{i=1}^{L} \alpha_i - \frac{1}{2}\sum_{i,j} \alpha i H_{ij} \alpha_j, H_{ij} \equiv y_i y_j x_i x_j \tag{3.13}$$

$$\equiv \sum_{i=1}^{L} \alpha_i - \frac{1}{2}\alpha^T H \alpha, \alpha_i \geq 0, \forall_i \sum_{i=1}^{L} \alpha_i y_i = 0 \tag{3.14}$$

elde edilir. Bu kısımda L_D, lagrange denkleminin ikili formudur. Dual formda her girdi vektörü x_i için dot product kullanılır.

L_p' yi minimize edecek w ve b değerleri türev işleminden sonra L_D eşitliğinde yer almamaktadır. Daha öncesinde L_P minimize edilmeye çalışılırken, şimdi L_D maksimize edilmeye çalışılacaktır.

$$\max_{\alpha} \left[\sum_{i=1}^{L} \alpha_i - \frac{1}{2} \alpha^T H \alpha \right], \quad \alpha_i \geq 0, \forall_i \quad \sum_{i=1}^{L} \alpha_i y_i = 0 \tag{3.15}$$

Konveks karesel bir optimizasyon problemi olan (3.15) eşitsizliği, karesel programlama kullanılarak çözüldüğünde α elde edilir ve bu değer kullanılarak eşitsizlik (3.10)' den w hesaplanır. Geriye ise sadece b değerinin hesaplanması kalmaktadır.

Eşitsizlik (3.11)'i sağlayan tüm x_s Destek vektörleri, $y_s(x_s.w + b) = 1$ formundadır. w değerinin eşitlik (3.10)' deki karşılığı, bu formda yerine koyulursa aşağıdaki eşitlik elde edilir.

$$y_s \left(\sum_{m \in S} a_m y m x_m . x_s + b \right) = 1 \tag{3.16}$$

Burada S, destek vektör indislerini temsil eder. Eşitlik (3.14)' de eşitliğin iki tarafı da y_s ile çarpılarak eşitlik (3.15) elde edilir. $y_s^2 = 1$ yerine koyularak;

$$y_s^2 \left(\sum_{m \in S} a_m y_m x_m . x_s + b \right) = y_s \tag{3.17}$$

$$b = y_s - \sum_{m \in S} a_m y_n x_m . x_s \tag{3.18}$$

elde edilir. Rastgele bir x_s Destek Vektörü kullanmak yerine, S boyunca tümünün ortalaması alınabilir.

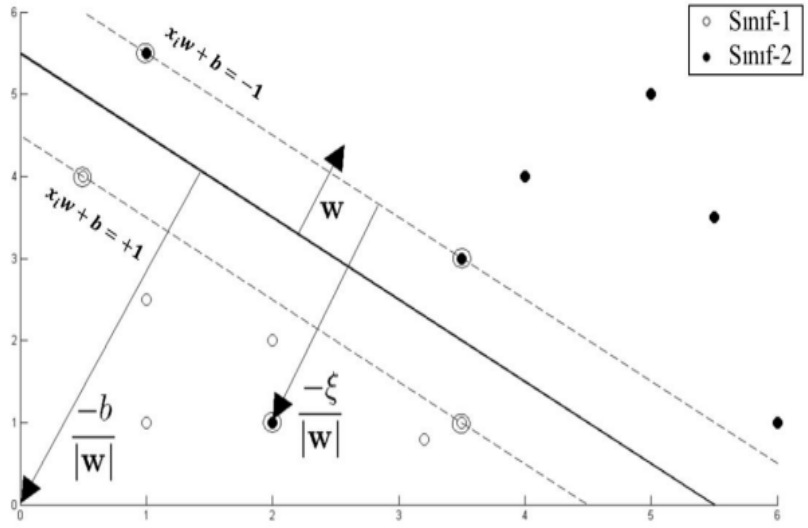
$$b = \frac{1}{N_s} \sum_{s \in S} \left(y_s - \sum_{m \in S} \alpha_m y_m x_m . x_s \right) \tag{3.19}$$

Bu adımlar sonucunda, ayırıcı hiperdüzlemin optimum yönelimini tanımlayan w ve b değerlerine ulaşılarak SVM oluşturulmuş olur. Yeni gelen herhangi bir gözlem olan x' noktasının ait olduğu sınıf y', aşağıdaki eşitlik kullanılarak elde edilir.

$$y' = sgn(w.x' + b) \tag{3.20}$$

Soft Margin

Belirli bir ek değişken ile lineer olarak ayrıştırılabilen verilerde kullanılır. Bu durumdaki SVM'in geometrik olarak gösterimi aşağıda verilmiştir.



Şekil 3: Soft margin uygulanan SVM'in geometrik yapısı

Veri kümesindeki noktalar tam olarak lineer bir şekilde ayrıştırılamadığı durumlarda verilen sınır koşulları ξ_i yapay değişkeni kullanılarak bir miktar esnetilmektedir. $\xi_i \geq 0, \forall_i$ ve $i = 1, 2, ..., L$ olmak üzere aşağıdaki denklem elde edilir.

$$x_i.w + b \geq +1 - \xi_i, y_i = +1 \quad \text{için} \tag{3.21}$$

$$x_i.w + b \leq -1 + \xi_i, y_i = -1 \quad \text{için} \tag{3.22}$$

$$y_i(x_i.w + b) - 1 + \xi_i \geq 0, \quad \xi_i \geq 0, \quad \forall_i \tag{3.23}$$

Burada aylak değişken (ξ_i):

- $\xi_i = 0$ ise veri setindeki i. gözlem marjinin doğru tarafındadır.

- $\xi_i > 0$ ise veri setindeki i. gözlem marjini ihlal etmiştir, yani $0 < \xi_i < 1$ durumunda

veri setindeki i. gözlem doğru sınıflandırılır; ancak marjinin içinde kalır.

- $\xi_i > 1$ ise i. gözlem hiperdüzlemin ters tarafındadır ve hatalı sınıflandırılmıştır.

Soft marjin kullanılan SVM'de, ceza parametresi kullanılır. Burada ceza parametresi (3.6) eşitsizliği aşağıdaki hale dönüştürülür.

$$y_i(x_i.w + b) - 1 + \xi_i \geq 0 \quad \forall_i \quad \text{koşulu altında} \quad \min \frac{1}{2}||w||^2 + C\sum_{i=1}^{L} \xi_i \tag{3.24}$$

Burada C negatif olmayan ayar parametresidir. C=0 ise $\xi_i = 0$ değerini alır ve hard marjin olur. C ile marjinin büyüklüğü arasındaki ilişkiyi ifade eder ve C azaldıkça marjin daralır, algoritma ihlallere karşı daha az toleranslı, eğitim verisine daha duyarlı olur (düşük bias, yüksek varyans). C arttıkça marjin genişler ve algoritma ihlallere daha müsait, eğitim verisine ise daha az duyarlı hale gelir. Lagrange çarpanları kullanılarak aşağıdaki eşitlik elde edilir.

$$L_p \equiv \frac{1}{2}||w||^2 + C\ \sum_{i=1}^{L} \xi_i - \sum_{i=1}^{L} \alpha_i[y_i(x_i.w + b) - 1 + \xi_i] + \sum_{i=1}^{L} \mu_i\xi_i \tag{3.25}$$

Bu eşitlikte w, b ve ξ_i değişkenlerine göre minimize, α değişkenine göre maksimize edilir. $\alpha_i \geq 0$ ve $\mu_i \geq 0$, $\forall_i$ olmak üzere, L_p'nin w, b ve ξ_i değişkenlerine göre türev alınır ve türevler sıfıra eşitlenir.

$$\frac{\partial L_p}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{L} \alpha_i y_i x_i \tag{3.26}$$

$$\frac{\partial L_p}{\partial b} = 0 \Rightarrow \sum_{i=1}^{L} \alpha_i y_i = 0 \tag{3.27}$$

$$\frac{\partial L_p}{\partial \xi_i} = 0 \Rightarrow C = \alpha_i + \mu_i \tag{3.28}$$

Bu değerler (3.25)'de yerine yazıldığında L_D elde edilir. $\mu_i \geq$, $\forall_i$ ve $\alpha_i \leq C$ olduğu göz

önünde bulundurularak, aşağıdaki eşitlik hesaplanır.

$$\max_{\alpha} \left[\sum_{i=1}^{L} \alpha_i - \frac{1}{2}\alpha^T H \alpha], \quad 0 \leq \alpha_i \leq C \; \forall_i \quad \sum_{i=1}^{L} \alpha_i y_i = 0 \tag{3.29}$$

Burada yine karesel programlama kullanılarak α elde edilir ve bu değerler kullanılarak eşitlik (3.26)'dan w ve b hesaplanır. Ancak bu kez b'yi hesaplamak için kullanılacak olan Destek Vektörler kümesi, $0 \leq \alpha_i \leq C$ aralığındaki i indisleri bulunarak elde edilir. Böylece SVM oluşturulmuş olur. Yeni gelen herhangi bir gözlem olan x' noktasının ait olduğu sınıf y', aşağıdaki eşitlik kullanılarak elde edilir.

$$y' = sgn(w.x' + b) \tag{3.30}$$

Lineer Olmayan SVM

Veri kümesi lineer hiper düzlem ile ayrıştırılamayan yani lineer olmayan veri setlerini ele alırız. Bu durumlarda sorunu gidermek için ortaya çekirdek kavramı çıkmaktadır. Bu tip sorunları ortadan kaldırmak için Wapnik ve arkadaşları konveks bir amaç fonksiyonu ve doğrusal kısıtlar kullanmışlardır. Yani Lagrange fonksiyonlarından yararlanmışlardır.
Lagrange Fonksiyonu:

$$\mathcal{L}(\beta, \beta_0; \alpha) = \frac{1}{2}\beta^T.\beta - \sum_{i \in A} \alpha_i (y_i(\beta^T x_i + \beta_0) - 1) \tag{3.31}$$

Burada ki amaç fonksiyonumuz $\frac{1}{2}\beta^T\beta$, kısıtlarımız ise ;

$$\sum_{i \in A} \alpha_i (y_i(\beta^T x_i + \beta_0) - 1) \tag{3.32}$$

Burada kısıtlı olan bir problemi kısıtsızmış gibi varsaydığımızdan amaç fonksiyonunun yanına α Lagrange çarpanı eklenmelidir. Ayrıca optimallik şartlarına göre sadece o anki $(i \in A)$ anlık şartlara bakılması gerekir. Optimallik şartlarına göre Lagrange fonksiyonunda β göre türev alınır (25). Çok boyutlu olduğundan Gradyan şeklinde $\nabla_\beta \mathcal{L}(\beta, \beta_0; \alpha)$ gösterilir. Elimizdeki ifade vektördür. Diğer yandan $\frac{1}{2}\beta^T\beta$ ifadesi ikinci dereceden olduğundan β' nın kendisi

gelmektedir. Bu durumda β ve β_0' a göre alınan türevler 0'dır.
β ve β_0' bilinmeyenlerdir ve aynı zamanda optimallik şartıdır.

$$\nabla_\beta \mathcal{L}(\beta, \beta_0; \alpha) = \beta - \sum_{i \in A} \alpha_i y_i x_i = 0 \Rightarrow \beta = \sum_{i \in A} \alpha_i y_i x_i \tag{3.33}$$

$$\frac{\partial \mathcal{L}(\beta, \beta_0; \alpha)}{\partial \beta_0} = -\sum_{i \in A} \alpha_i y_i = 0 \Rightarrow \sum_{i \in A} \alpha_i y_i = 0 \tag{3.34}$$

Eğer $\sum_{i \in A} \alpha_i y_i = 0$ ve $\sum_{i \in A} \alpha_i y_i x_i$ denklemleri yerine yazılırsa;

$$\begin{aligned} \mathcal{L}(\beta, \beta_0; \alpha) &= \frac{1}{2}\left(\sum_{i \in A} \alpha_i y_i x_i\right)^T . \left(\sum_{j \in A} \alpha_j y_j x_j - \left(\sum_{i \in A} \alpha_i y_i \left(\sum_{j \in A} \alpha_j y_j x_j\right)^T x_i\right)\right. \\ &- \sum_{i \in A} \alpha_i y_i \beta_0 + \sum_{i \in A} \alpha_i \end{aligned} \tag{3.35}$$

$$\mathcal{L}(\beta, \beta_0; \alpha) = \sum_{i \in A} \alpha_i - \frac{1}{2} \sum_{i \in A} \sum_{j \in A} \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{3.36}$$

Burada girdilerin iç çarpanlarına $(x_i^T x_j)$ bağlı bir vektör ortaya çıkmaktadır. Bu x_i ve x_j vektörlerinin çarpımı, bu vektörlerin aralarındaki açı ile alakalıdır. Burada iç çarpım işlemine göre yüksek boyuta taşıyarak iki vektör arasındaki benzerliği çekirdekler ile ölçmekteyiz. Bu da lineer bir şekilde ayrılmayan veri setlerine ayırmamıza yardımcı olur.

Farklı tipte birçok çekirdek vardır. Bunlardan bazıları şunlardır;

- Doğrusal,
- Polinom,
- Gauss,
- Laplace,
- Anova,
- Sigmoid,
- Bessel çekirdek fonksiyonlarıdır.

Çekirdek Fonksiyonları

Doğrusal Çekirdek

Doğrusal çekirdek, en basit çekirdek fonksiyonudur. $< x, y >$ iç çarpım işlemine göre aşağıdaki denklem elde edilir.

$$K(x_i, x_j) = x_i^T x_j \tag{3.37}$$

Polinom Çekirdek

Durağan olmayan bir çekirdektir. Genellikle eğitim verilerinin normalleştirildiği problemler için kullanılır. Denklemi aşağıdaki şekildedir:

$$K(x_i, x_j) = (x_i^T x_j + 1)^d \tag{3.38}$$

Burada d polinomun derecesini ifade eder.

Gauss Çekirdek

Gauss çekirdeği radyal tabanlı fonksiyondur ve genellikle doğrusal olarak ayrılamayan veri setlerinde en çok kullanılan çekirdek türlerinden biridir. Üstel bir fonksiyon olduğundan aşırı uyum sağladığında, üstel ifade doğrusallaşır. Bir doğru gibi hareket ettiğinden doğrusal olmayan yapısını yok olur. Gauss çekirdeğinin denklemi aşağıdaki şekildedir:

$$K(x_i, x_j) = e^{\gamma ||x_i - x_j||^2}, \quad \gamma > 0 \tag{3.39}$$

Dağılım genişliği gama (γ) ile kontrol edilmektedir. Gama değeri probleme uygun olacak şekilde ayarlanmalıdır.

Laplace Çekirdek

Laplace çekirdeği veri noktalarının benzerlik ölçüsünü hesaplamak için kullanılır. Ayrıca gauss çekirdeği ile eşdeğerdir. Laplace çekirdeğinin denklemi aşağıdaki gibidir:

$$K(x_i, x_j) = e^{\left(-\frac{||x_i - x_j||}{\sigma}\right)} \tag{3.40}$$

Anova Çekirdek

Anova çekirdeği, sınıflar arasındaki farklılıkları yüksek bir ağırlıkla vurgularken, aynı sınıfa ait veri noktaları arasındaki benzerlikleri düşük bir ağırlıkla dikkate alır. Bu sayede, sınıflar arasında ayrım daha iyi hale gelir (26). Anova çekirdeğinin denklemi aşağıdaki gibidir:

$$K(x_i, x_j) = \sum_{i=1}^{n} \exp(-\sigma(x_i^i - x_j^i)^2)^d \tag{3.41}$$

Sigmoid Çekirdek

Sigmoid çekirdek fonksiyonu lojistik regresyonda aktivasyon fonksiyonu olarak da kullanılır. Sigmoid çekirdek fonksiyonun pratikte iyi performans gösterdiği görülmektedir. Denklemi aşağıdaki şekildedir:

$$K(x_i, x_j) = \tanh(\alpha x_i^T x_j + c) \tag{3.42}$$

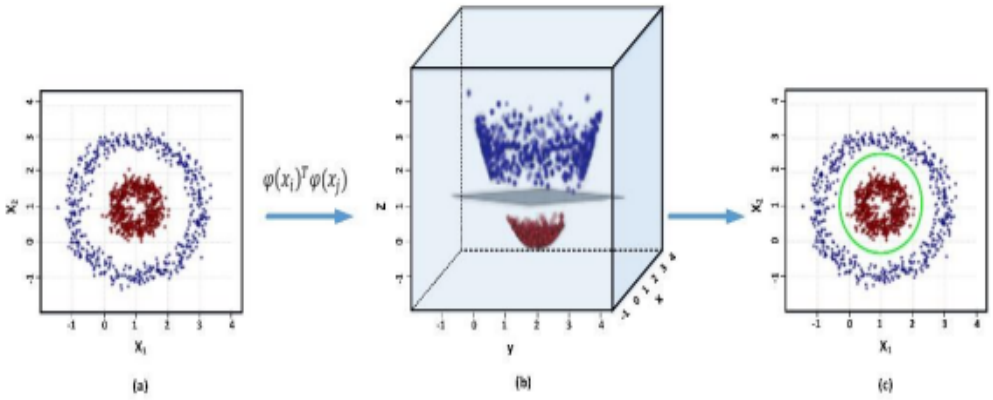
Çekirdekte iki adet ayarlabilir parametre vardır. Bunlar α ve c kesme sabitidir. Genellikle α $\frac{1}{N}$'dir ve buradaki N değeri verinin boyutunu ifade eder.

Bessel Çekirdek

Bessel çekirdeği düzgün fraksiyonel uzaylar teorisinde iyi bilinen bir çekirdektir. Denklemi aşağıdaki şekildedir:

$$K(x_i, x_j) = \frac{J_{v+1}(\sigma||x_i - x_j||)}{||x_i - x_j||^{-n(v+1)}} \tag{3.43}$$

Bessel çekirdeğindeki J ifadesi birinci türden bessel değeridir. Burada Bessel çekirdeği ile veriler iki boyutlu uzaydan, çok boyutlu uzaylara taşınarak iki boyutlu uzayda ayrılamayan veriler ayrılabilir hale getirilir.



Şekil 4: Çekirdek Fonksiyon ile Ayırıcı Düzlemi Belirleme

Burada verilen veride kırmızı ve mavi noktalar doğrusal bir şekilde ayrılmadığından bir sınıfın bir miktar yukarıya taşınmasıyla iki sınıf birbirinden ayrılabilir hale getirilmiştir.

3.2.2 K En Yakın Komşu (K-NN) Algoritması

1950'li yılların başında Evelyn Fix ve Joseph L. Hodges Jr. tarafından geliştirilmeye başlansa da 1960 yıllarında kullanılmıştır. Algoritmanın gelişiminin geç olmasının nedenlerinden biri de büyük hesaplama gücü gerektirmesi ve mevcut olan donanımların buna imkan vermemesidir (27).
k-NN algoritması seçilen bir özelliğin kendisine en yakın olan özellikle arasındaki yakınlığı kullanarak sınıflandırma işlemini gerçekleştirir. Sınıflandırma problemlerinde ve regresyon analizinde kullanılan bir yöntemdir. Küçük veri setleri için oldukça tercih edilir. Algoritmadaki k değeri komşuluğu ifade eder ve sabit bir değeri yoktur, optimal bir k değeri bulunur. Algoritma basit ve gürültülü eğitim verilerine karşı oldukça dirençli olmasından dolayı da sıklıkla tercih edilir. Algoritmanın bir çok avantajı olduğu gibi dezavantajları da vardır. Bu dezavantajlardan biri de tahmin edilecek verinin, diğer bütün verilere olan uzaklıkları hesaplanıp bellekte saklandığından oldukça depolama alanına gereksinim duyulmaktadır. Büyük veri setleri için bu yöntem genellikle tercih edilmez. K-nn algoritmasında veri setindeki tüm değişkenlerin sürekli olması halinde Öklid, değişkenlerin kategorik olması durumunda ise Manhattan uzaklık hesabından yararlanarak işlemler gerçekleştirilir.

Öklid Uzaklığı: Öklid uzaklığı x ve y gibi iki noktanın arasındaki uzaklığı ifade eder ve $d(x, y)$ olarak gösterilir. Denklemi aşağıdaki şekildedir.(28)

$$d(x,y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \tag{3.44}$$

Manhattan Uzaklığı: Manhattan uzaklığı ise, gözlem boyutları arasındaki ağırlıklı mutlak farkların tüm boyutları üzerinden toplamı şeklindedir. $P = (x_1, x_2, ..., x_n)$ ve $Q = (y_1, y_2, ..., y_n)$ olmak üzere P ve Q değişkenlerinin noktaları arasındaki uzaklık manhattan uzaklığı ile aşağıdaki denklem ile hesaplanır (29).

$$Manhattan_{PQ} = \sum_{p=1}^{p} w_p(x_{pp} - y_{pQ})^2 \tag{3.45}$$

Bir diğer ifade ile;

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + ... + |x_{ip} - x_{jp}| \tag{3.46}$$

Algoritmanın adımları;

- İlk önce k parametresi belirlenir. k parametresi tahmin edilecek veriye en yakın olan komşuların sayısıdır.
- Tahmin edilecek verinin veri setindeki bütün noktalara olan uzaklığı hesaplanır.
- İlk adımda belirtiğimiz en yakın k adet komşu ele alınır. Öznitelik değerlerine göre k komşuların sınıfına atanır.
- Seçtiğimiz bu sınıf, tahmin etmek istediğimiz verinin sınıfında olur ve veri artık etiketlenmiş olur.

.

Chebyshev Uzaklığı: İki N boyut noktasının bir boyutundaki maksimum mutlak mesafedir. Basit bir ifadeyle, iki vektör arasındaki mesafe, herhangi bir koordinat boyutu boyunca farklılıklarının en büyüğüdür. Denklemi aşağıdaki gibidir:

$$D_{Chebyshev}(x, y) = \max_i(|x_i - y_i|) \tag{3.47}$$

3.2.3 Rastgele Ormanlar

Bagging (Breiman,1996) ile Random Subspace (Ho,1998) yöntemlerinin birleşimi ile oluşturulmuştur. Rastgele ormanlar algoritması hem sınıflandırma hem de regresyon problemlerinde kullanılır. Algoritmanın temeli birbirinden bağımsız birden çok karar ağacının ürettiği tahminlerin bir araya getirilerek değerlendirilmesine dayanır.Bu nedenle rastgele orman algoritması karar ağaçlarından daha iyi performans gösterir. Ağaç sayısı arttıkça algoritmanın doğru çalışması da artmaktadır. Ayrıca karar ağacının her bir düğümünde en iyi dallara ayırıcı değişken tüm değişkenler arasından rastgele seçilen daha az sayıdaki değişken arasından seçilir. Değişkenlerin öenm sıralamasının belirlenmesi algoritma için önemli adımlardan biridir. Değişken önemliliğinin belirlenmesinde en basit ve en etkili yöntemlerden birisi permütasyona dayalı yöntemdir. Rastgele orman algoritması aşağıdaki gibidir.

1. Eğitim verisinden bootstrap (yeniden örnekleme) tekniği ile oluşturulan örneklemlerden belirlenen ağaç sayısına göre ağaçlar oluşturulur.

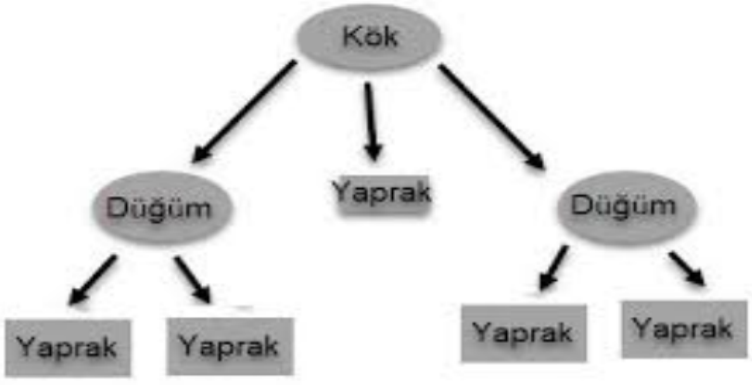
2. Her bir bootstrap örneklemi için her düğümde p adet öznitelik değişkenleri arasından rassal olarak m adet öznitelik değişkeni seçilir ve en iyi dal belirlenir.

   - sınıflandırmada $m = \sqrt{p}$
   - regresyonda $m = \frac{p}{3}$

3. n tane karar ağacının ayrı ayrı yaptığı tahminler birleştirelerek, sınıflandırma için mod , regresyon için ise ortalama alınarak model tahmini gerçekleştirilir.

3.2.4 Karar Ağaçları

Çok sayıda veri içeren bir kümeyi, bazı teknikler kullanarak alt bölümlere ayırmak için geliştirilmiş bir algoritmadır. Özellik ve hedeflere göre karar ağaçları ve yaprak düğümlerinden oluşan ağaç yapısı formunda bir model oluşturan bir sınıflandırma yöntemidir. Karar ağaçları iki adımda çalışır.

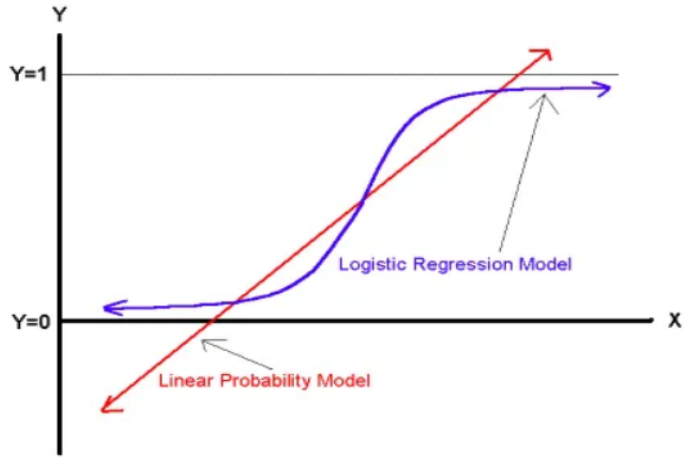


Şekil 5: Karar Ağaçlarının Genel Yapısı

Ağacın kök düğümünden başlayarak ara düğümler üzerinden gideceği yol belirlenir. Ağaç üzerinde her sınıf bir yaprak olarak gösterilmelidir. Dallanma işlemi kökten başlar ve yaprak düğüme ulaşana kadar devam eder (30).

3.2.5 Lojistik Regresyon

Regresyon anali, bir hedef değişken ile bir yada daha fazla değişken arasındaki ilişkiyi açıklamaktadır. Buradaki hedef değişken nicel ise klasik doğrusal regresyon, kategorik ise lojistik regresyon kullanılır. Ayrıca lojistik regresyon çoklu doğrusal bağlantı problemi olmaması dışında, klasik doğrusal regresyondaki varsayımlar aranmamaktadır bu durumda lojistik regresyon geniş kullanım alanına sahiptir (31).



Şekil 6: Lojistik Regresyon Eğrisi

Lojistik regresyonda bir şeyin olup olmama ihtimali üzerinde çalışılır. Örneğin müşterinin bir ürünü satın alıp almayacağı, müşterinin maaşına göre krediyi ödeyip ödeyemeyeceği üzerinde durulur. Bu algoritmada önce nitel değişken olan bağımlı değişkene logit dönüşümü yapılır ve daha sonra maksimum olabilirlik tahmini uygular. Doğrusal regresyondaki gibi bir doğrusal regresyon doğrusu yerine, hedef değişkenin kategorilerini tahmin eden bir "S" şeklinde lojistik fonksiyon oluşturulur. Burada hedef değişkenin tahmin edilmesiyle değil, değişkenin 1 ya da 0 almasıyla ilgilenilir (32). Kategorik değişkenin iki sınıftan oluşuyorsa "ikili lojistik regresyon", daha fazla kategorik sınıftan oluşuyorsa "çoklu lojistik regresyon", sıralayıcı ölçek düzeyinde ölçülmüş ise "sıralı lojistik regresyon" denir. Biz bu tezde ikili lojistik regresyon kullanacağız.
İkili lojistik regresyonda 0 değeri gerçekleşme olasılığının olmadığını, 1 ise belirli bir gerçekleşme olasılığını temsil eder. Bu iki durum arasında yer alan değerlerin olasılığı dağılım grafiğinde hangisine yakın olduğuna karar verilerek hesaplanır. Sınıflandırma işlemini yapabilmek için bir eşik değer belirlenir. Genellikle literatürde bu eşik değer 0.5'tir. Bu eşik değerin altında kalan veriler 0, eşik değerin üzerinde olan veriler 1 olarak sınıflandırılır. 0.5' e eşit olan veriler sınıflandırılmaz (33).

3.2.6 Yapay Sinir Ağları

Yapay sinir ağları (YSA) günümüzde genellikle sağlık, finans, endüstri gibi bir çok alanda uygulamaları kullanılmaktadır. YSA, biyolojik sinir hücrelerinin yapı ve işleyişlerini bilgisayara aktarmak isteyen ve doğrusal olmayan problemlerde sıklıkla kullanılan yöntemlerden

biridir.

YSA, beynin ilk duyusal işleme modellerinden esinlenerek oluşturulmuş, yapay zekanın en çok tercih edilen bir alt çalışma alanıdır (34). Ağı oluşturan her bir elemana nöron denmektedir. YSA, birbirine ağırlıklandırılmış bağlantılar üzerinden sinyal göndererek haberleşen basit nöronlardan meydana gelmektedir. YSA' nın temel işlem ünitesi olan nöran diğer komşu nöranlardan aldığı girdiyi işleyerek sonraki nörona veya çıktı katmanına iletir.

4 UYGULAMA

Bu çalışmada Parkinson hastalığının ses verisi ile analizinde SVM ve diğer sınıflandırma algoritmaları kullanılmıştır.

4.1 Çalışmanın Örneklemi

Çalışmada kullanılan veri seti UCI veri havuzunda bulunan Oxford Parkinsons'tur. İsim ve durum çubuğu olmak üzere iki kategorik değişken ve 22 adet sürekli değişkenden oluşmaktadır. Veri setinde, her kişiden farklı zamanlarda olmak üzere 6 adet ses kaydedilmiş olup toplam 197 adet veriden oluşmaktadır. Sağlıklı bireyler 0, parkinson hastalığına sahip olduğu bilinen kişiler ise 1 olarak gösterilmiştir. Veri setindeki değişkenlerin cinsi ve ağırlık değerleri aşağıdaki tabloda verilmiştir.

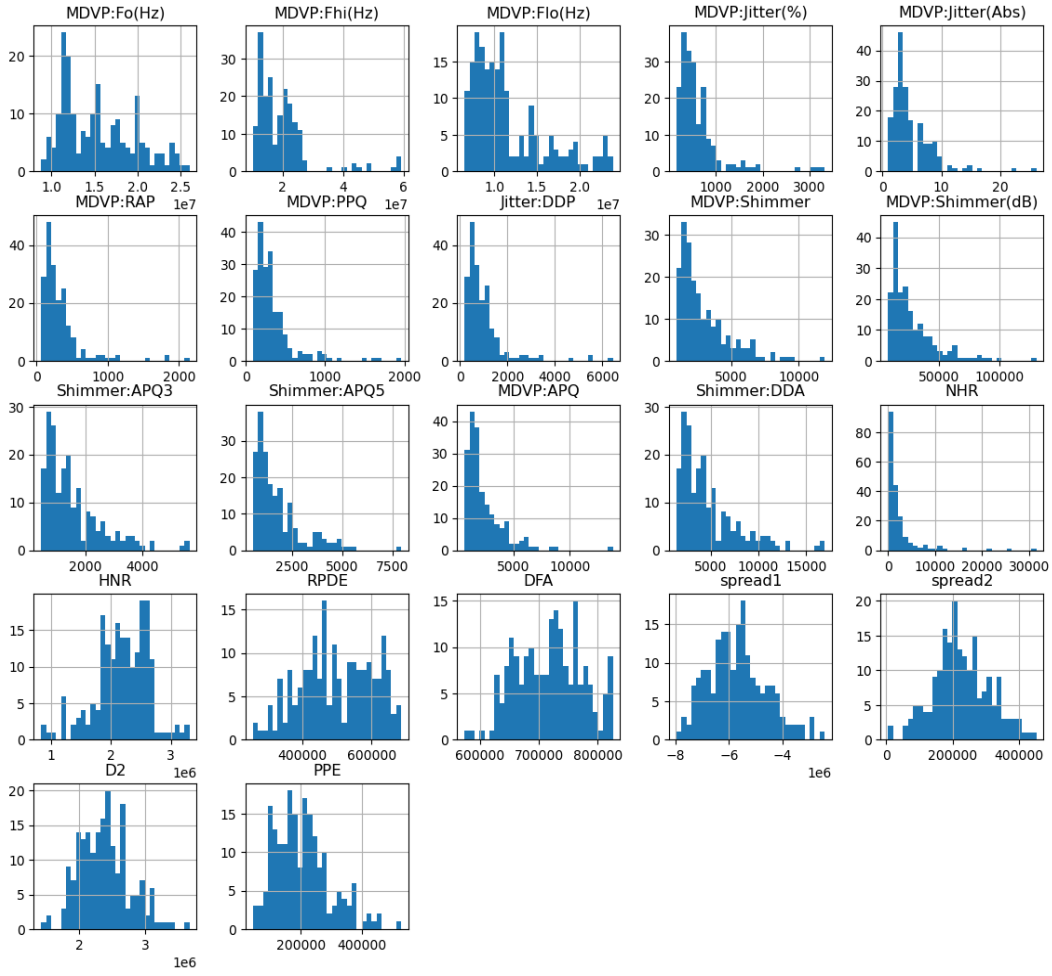
Tablo 1: Değişkenlerin cinsi ve aralık değerleri

Değişkenler	**Veri Tipi**	**Aralık Değeri**
İsim	Kategorik	-
MDVP:Fo(Hz)	Sayısal	8833300-26010500
MDVP:Fhi(Hz)	Sayısal	10214500-59203000
MDVP:Flo(Hz)	Sayısal	6547600 - 23917000
MDVP:Jitter(%)	Sayısal	168 - 3316
MDVP:Jitter(Abs)	Sayısal	1 - 26
MDVP:RAP	Sayısal	68 - 2144
MDVP:PPQ	Sayısal	92 - 1958
Jitter:DDP	Sayısal	204 - 6433
MDVP:Shimmer	Sayısal	954 - 11908
MDVP:Shimmer(dB)	Sayısal	8500 - 130200
Shimmer:APQ3	Sayısal	455 - 5647
Shimmer:APQ5	Sayısal	570 - 7940
MDVP:APQ	Sayısal	719 - 13778
Shimmer:DDA	Sayısal	1364 - 16942
NHR	Sayısal	65 - 31482
HNR	Sayısal	844100 - 3304700
status	Kategorik	0 ve 1
RPDE	Sayısal	256570 - 685151
DFA	Sayısal	574282 - 825288
spread1	Sayısal	7964984 - -2434031
spread2	Sayısal	6274 - 450493
D2	Sayısal	1423287 - 3671155
PPE	Sayısal	44539 - 527367

4.2 Veri Setine Uygulanan Ön İşlemler

Makine öğrenmesi yöntemlerinde en önemli adım veriye uygulanan ön işlemlerdir. Veri setine işlem yapmadan önce, isnull komutu ile veri setindeki değişkenlerde eksik değer olup olmadığına bakılmıştır ve eksik veri olmadığı görülmüştür. Status değişkeni string (metin) ifadelerden oluştuğundan nümerik (sayısal) hale getirilmiştir. Daha sonra sürekli değişken-

lerin normal dağılıma uygunluğu Shapiro Wilk's testi ile kontrol edilmiştir. Testin sonucuna göre spread2 değişkeninin p değeri 0.05'ten büyüktür yani normal dağılmıştır. Fakat spread2 değişkeni hariç bütün değişkenlerin normal dağılım göstermediği tespit edilmiştir. Aşağıda sürekli değişkenlerin histogram grafikleri verilmiştir.

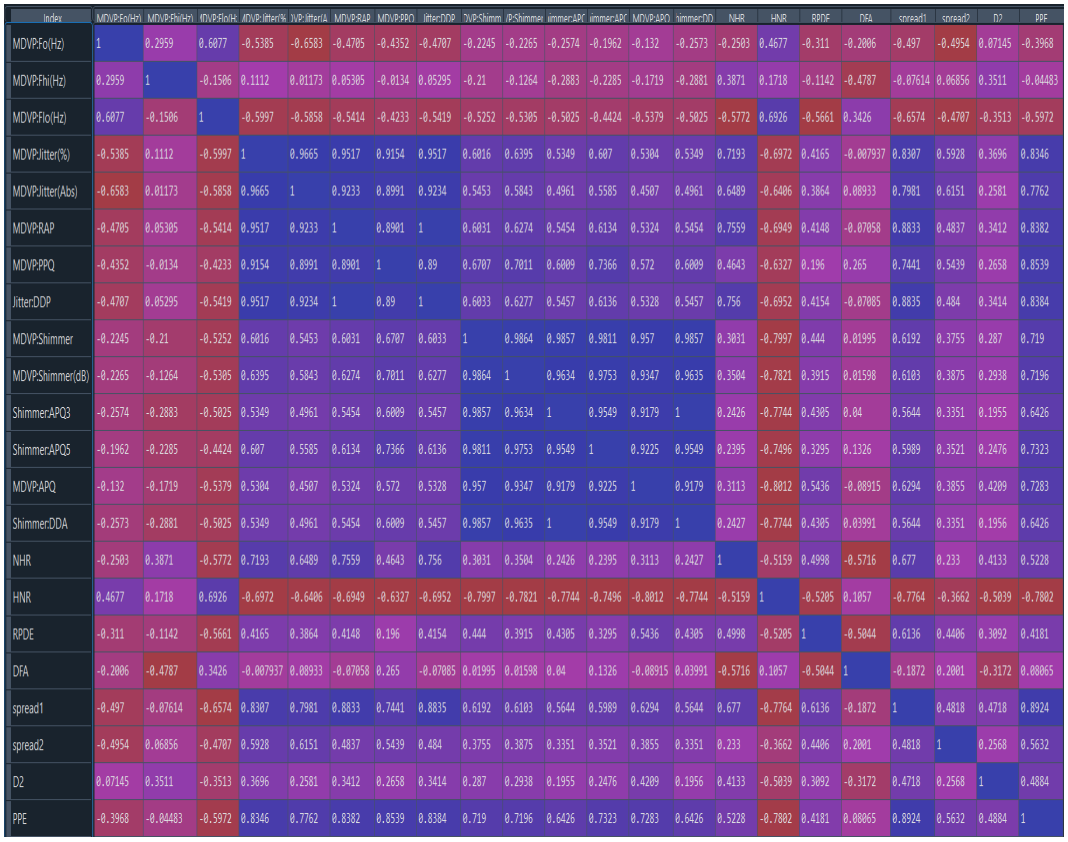


Şekil 7: Sürekli Değişkenlerin Histogram Grafiği

Şekil 7' deki histogram grafiklerinden de anlaşılacağı üzere spread2 değişkeni normal dağılıma sahiptir. Diğer sürekli değişkenleri standartscaler ile ortalama değeri 0, standart sapma değeri ise 1 olacak şekilde dönüşüm yapılmıştır ve normal dağılıma uygun hale getirilmiştir.

Veri setinin %70' i eğitim (training), %30' u ise test seti olmak üzere rastgele ikiye ayrılmıştır.

Korelasyon tablosundan yararlanılarak sürekli değişkenler arasındaki ilişki incelenmiştir.



Index	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%	MDVP:Jitter(A	MDVP:RAP	MDVP:PPQ	Jitter:DDP	MDVP:Shimm	MDVP:Shimmer	Shimmer:APQ	Shimmer:APQ	MDVP:APQ	Shimmer:DD	NHR	HNR	RPDE	DFA	spread1	spread2	D2	PPE
MDVP:Fo(Hz)	1	0.2959	0.6077	-0.5385	-0.6583	-0.4705	-0.4352	-0.4707	-0.2245	-0.2265	-0.2574	-0.1962	-0.132	-0.2573	-0.2503	0.4677	-0.311	-0.2006	-0.497	-0.4954	0.07145	-0.3968
MDVP:Fhi(Hz)	0.2959	1	-0.1506	0.1112	0.01173	0.05305	-0.0134	0.05295	-0.21	-0.1264	-0.2883	-0.2285	-0.1719	-0.2881	0.3871	0.1718	-0.1142	-0.4787	-0.07614	0.06856	0.3511	-0.04483
MDVP:Flo(Hz)	0.6077	-0.1506	1	-0.5997	-0.5858	-0.5414	-0.4233	-0.5419	-0.5252	-0.5305	-0.5025	-0.4424	-0.5379	-0.5025	-0.5772	0.6926	-0.5661	0.3426	-0.6574	-0.4707	-0.3513	-0.5972
MDVP:Jitter(%)	-0.5385	0.1112	-0.5997	1	0.9665	0.9517	0.9154	0.9517	0.6016	0.6395	0.5349	0.607	0.5304	0.5349	0.7193	-0.6972	0.4165	-0.007937	0.8307	0.5928	0.3696	0.8346
MDVP:Jitter(Abs)	-0.6583	0.01173	-0.5858	0.9665	1	0.9233	0.8991	0.9234	0.5453	0.5843	0.4961	0.5585	0.4507	0.4961	0.6489	-0.6406	0.3864	0.08933	0.7981	0.6151	0.2581	0.7762
MDVP:RAP	-0.4705	0.05305	-0.5414	0.9517	0.9233	1	0.8901	1	0.6031	0.6274	0.5454	0.6134	0.5324	0.5454	0.7559	-0.6949	0.4148	-0.07058	0.8833	0.4837	0.3412	0.8382
MDVP:PPQ	-0.4352	-0.0134	-0.4233	0.9154	0.8991	0.8901	1	0.89	0.6707	0.7011	0.6009	0.7366	0.572	0.6009	0.4643	-0.6327	0.196	0.265	0.7441	0.5439	0.2658	0.8539
Jitter:DDP	-0.4707	0.05295	-0.5419	0.9517	0.9234	1	0.89	1	0.6033	0.6277	0.5457	0.6136	0.5328	0.5457	0.756	-0.6952	0.4154	-0.07085	0.8835	0.484	0.3414	0.8384
MDVP:Shimmer	-0.2245	-0.21	-0.5252	0.6016	0.5453	0.6031	0.6707	0.6033	1	0.9864	0.9857	0.9811	0.957	0.9857	0.3031	-0.7997	0.444	0.01995	0.6192	0.3755	0.287	0.719
MDVP:Shimmer(dB)	-0.2265	-0.1264	-0.5305	0.6395	0.5843	0.6274	0.7011	0.6277	0.9864	1	0.9634	0.9753	0.9347	0.9635	0.3504	-0.7821	0.3915	0.01598	0.6103	0.3875	0.2938	0.7196
Shimmer:APQ3	-0.2574	-0.2883	-0.5025	0.5349	0.4961	0.5454	0.6009	0.5457	0.9857	0.9634	1	0.9549	0.9179	1	0.2426	-0.7744	0.4305	0.04	0.5644	0.3351	0.1955	0.6426
Shimmer:APQ5	-0.1962	-0.2285	-0.4424	0.607	0.5585	0.6134	0.7366	0.6136	0.9811	0.9753	0.9549	1	0.9225	0.9549	0.2395	-0.7496	0.3295	0.1326	0.5989	0.3521	0.2476	0.7323
MDVP:APQ	-0.132	-0.1719	-0.5379	0.5304	0.4507	0.5324	0.572	0.5328	0.957	0.9347	0.9179	0.9225	1	0.9179	0.3113	-0.8012	0.5436	-0.08915	0.6294	0.3855	0.4209	0.7283
Shimmer:DDA	-0.2573	-0.2881	-0.5025	0.5349	0.4961	0.5454	0.6009	0.5457	0.9857	0.9635	1	0.9549	0.9179	1	0.2427	-0.7744	0.4305	0.03991	0.5644	0.3351	0.1956	0.6426
NHR	-0.2503	0.3871	-0.5772	0.7193	0.6489	0.7559	0.4643	0.756	0.3031	0.3504	0.2426	0.2395	0.3113	0.2427	1	-0.5159	0.4998	-0.5716	0.677	0.233	0.4133	0.5228
HNR	0.4677	0.1718	0.6926	-0.6972	-0.6406	-0.6949	-0.6327	-0.6952	-0.7997	-0.7821	-0.7744	-0.7496	-0.8012	-0.7744	-0.5159	1	-0.5205	0.1057	-0.7764	-0.3662	-0.5039	-0.7802
RPDE	-0.311	-0.1142	-0.5661	0.4165	0.3864	0.4148	0.196	0.4154	0.444	0.3915	0.4305	0.3295	0.5436	0.4305	0.4998	-0.5205	1	-0.5044	0.6136	0.4406	0.3092	0.4181
DFA	-0.2006	-0.4787	0.3426	-0.007937	0.08933	-0.07058	0.265	-0.07085	0.01995	0.01598	0.04	0.1326	-0.08915	0.03991	-0.5716	0.1057	-0.5044	1	-0.1872	0.2001	-0.3172	0.08065
spread1	-0.497	-0.07614	-0.6574	0.8307	0.7981	0.8833	0.7441	0.8835	0.6192	0.6103	0.5644	0.5989	0.6294	0.5644	0.677	-0.7764	0.6136	-0.1872	1	0.4818	0.4718	0.8924
spread2	-0.4954	0.06856	-0.4707	0.5928	0.6151	0.4837	0.5439	0.484	0.3755	0.3875	0.3351	0.3521	0.3855	0.3351	0.233	-0.3662	0.4406	0.2001	0.4818	1	0.2568	0.5632
D2	0.07145	0.3511	-0.3513	0.3696	0.2581	0.3412	0.2658	0.3414	0.287	0.2938	0.1955	0.2476	0.4209	0.1956	0.4133	-0.5039	0.3092	-0.3172	0.4718	0.2568	1	0.4884
PPE	-0.3968	-0.04483	-0.5972	0.8346	0.7762	0.8382	0.8539	0.8384	0.719	0.7196	0.6426	0.7323	0.7283	0.6426	0.5228	-0.7802	0.4181	0.08865	0.8924	0.5632	0.4884	1

Şekil 8: Özelliklerin Korelasyon Tablosu

Korelasyon tablosundan görüldüğü üzere bazı değişkenler arasında güçlü bir ilişki bulunmaktadır. Örneğin; MDVP:Jitter(%) ile MDVP:Jitter(Abs) arasında %96.65 oranında pozitif yönlü güçlü bir ilişki bulunmaktadır. Bu durum aralarında güçlü ilişki bulunan (korelasyon değeri 0.95'den fazla olan) değişken çiftlerinin konuşma sinyalinin çok benzer yönlerini ölçtüğünü göstermektedir. Bu durum çoklu doğrusal bağlantı problemine sebep olmaktadır ve sınıflandırma aşamasında bu değişkenler yanıltıcı olabilmektedir. Temel Bileşen Analizi (PCA) dönüşümü yaparak, veri setinin boyutu küçültülür. ve çoklu doğrusal bağlantı problemi ortadan kaldırılır.

4.3 PCA

PCA dönüşümü kullanılarak veri setine boyut indirgeme işlemi yapılacaktır. Dönüşüm yapılmadan önce veri setinin standartlaştırılması gerekir. Standartlaştırma, her bir değişkenin ortalamasını 0 ve standart sapmasını 1 yaparak verileri aynı ölçekte tutmayı sağlar. Bu, değişkenler arasındaki farklılıkları dengelemeye yardımcı olur. PCA dönüşümü yapılırken aşağıdaki adımlar izlenir.

- Adım 1: Veriyi Merkezileştirme
  Merkezileştirme işlemi, değişkenlerin varyanslarına odaklanılmasına ve daha iyi bir önem sıralaması elde edilmesine yardımcı olur. Veri setindeki her değişken kendi ortalamasından çıkarılır ve veri merkezileştirilir. Böylece ortalaması 0 olan veri seti elde edilir. Bu kısımda sürekli değişkenlerin ortalaması mean() komutu ile hesaplanmıştır.

- Adım 2: Kovaryans matrisi
  Kovaryans matrisi, özdeğerler ve özvektörlerin elde edilmesi için kullanılmaktadır. Veri setinde sese ait 22 sürekli değişken olduğundan kovaryans matrisi de 22x22 boyutlu olacaktır. Kovaryans matrisi, matris ile matrisin transpozu skaler olarak çarpılır ve boyutuna bölünerek hesaplanmaktadır.

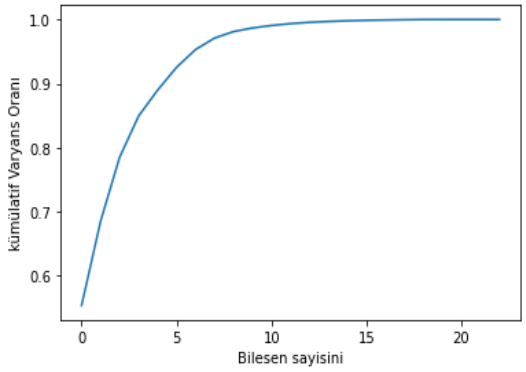
Şekil 9: Kovaryans matrisi

- Adım 3 : Özdeğerler ve Özvektörler
  Kovaryans matrisi karesel bir matristir ve özvektörler ve özdeğerler bu matristen elde edilir. Özdeğer denklemi $det(A - \lambda * I) = 0$ kaarkteristik denklemi ile ifade edilir. Bu denklemin kökleri bulunur ve bu özdeğerlere karşılık gelmektedir. Özvektörleri ise $(A - \lambda * I) * v = 0$ denkleminin çözülmesi ile elde edilir. Ayrıca özdeğer ve özvektörleri, linalg.eig() komutu ile kovaryans matrisi üzerinden de kolayca hesaplanmaktadır.

- Adım 4: Bileşenlerin Seçilmesi ve Özellik Vektörünün Oluşturulması
  Kovaryans matrisinden elde edilen özdeğer ve özvektörler yardımıyla eksenler ve özellik vektörü elde edilecektir.

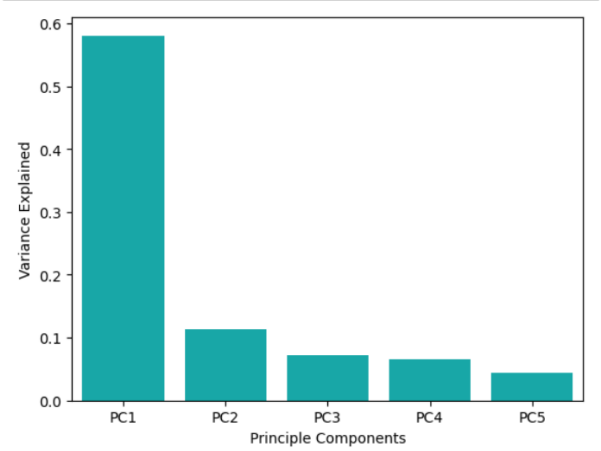
Veri setinin temel bileşenleri en yüksek özdeğerler ile özvektörlerden oluşmaktadır. Genellikle özvektörler öncelikli olarak kovaryans matristen elde edilir ve daha sonra yüksek değerden düşük değere doğru sıralanır. Amaç bileşenleri veriyi temsil etme oranına göre sıralamaktır. Böylelikle en önemli bileşenden en az öneme sahip bileşe doğru bir sıralama yapılır. Çıkan sonuca göre bileşen seçimi yapılır. Bu durumda bazı bileşenler veri setinden atılmış olur ve elde edilecek veri seti orjinal veri setinden daha az boyuta sahip olur.

Python'da PCA dönüşümünde kullanılan matematiksel adımlar (yukarıdaki 4 adım) arka planda yapılmaktadır.



Şekil 10: Özellik sayısı ile kümülatif Varyans Oranı tablosu

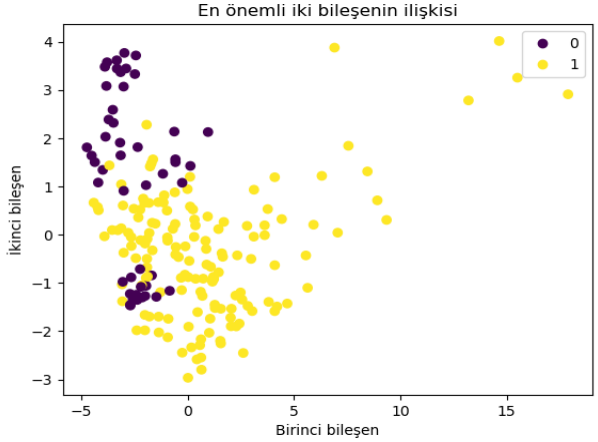
PCA dönüşümü sonucunda Şekil 10' daki grafik elde edilir. Grafiğe göre veri setine PCA dönüşümü yapıldığında bileşen sayısının 5'ten fazla oluşu veriye ihmal edilebilecek bir katkı sağlar. Bu durumda bileşen sayısının 5 seçilmesi yeterlidir.
Sonuç olarak aralarında yüksek korelasyona sahip olan değişkenlerin bulunduğu veri setine PCA dönüşümü yapılarak yüksek korelasyon sorunu olmayan, datanın %87.52' sini açıklayabilen ve 5 bileşenden oluşan yeni veri seti oluşturulmuştur. Bu yeni veri setinde korelasyon sorunu çözülerek olası çoklu doğrusal bağlantı probleminin önüne geçilmiş, değişken sayısı azaltılarak sınıflandırma aşamasında training süresi kısaltılmış ve overfitting riskinin azalması sağlanmıştır.
İlk 5 bileşeni alındığında her bir bileşenin tüm veriye katkı miktarı şu grafik ile verilir:



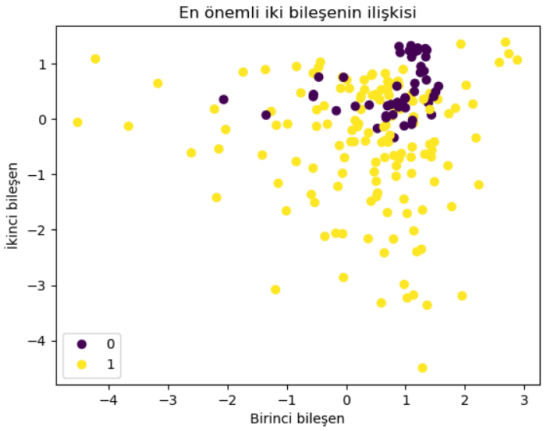
Şekil 11: Varyans Oranı Grafiği

4.4 PCA Yardımıyla Verinin Görselleştirilmesi

Orjinal veri setine PCA dönüşümü yaparak 5 bileşenden oluşacak şekilde verinin boyutu küçültülmüştür. Elde edilen bu yeni veri setindeki en önemli özelliklerin birbirlerine göre çizimlerini gerçekleştirilecektir. En önemli iki bileşen olan birinci ve ikinci bileşenin aralarındaki ilişki çizdirildiğinde aşağıdaki grafik elde edilir.



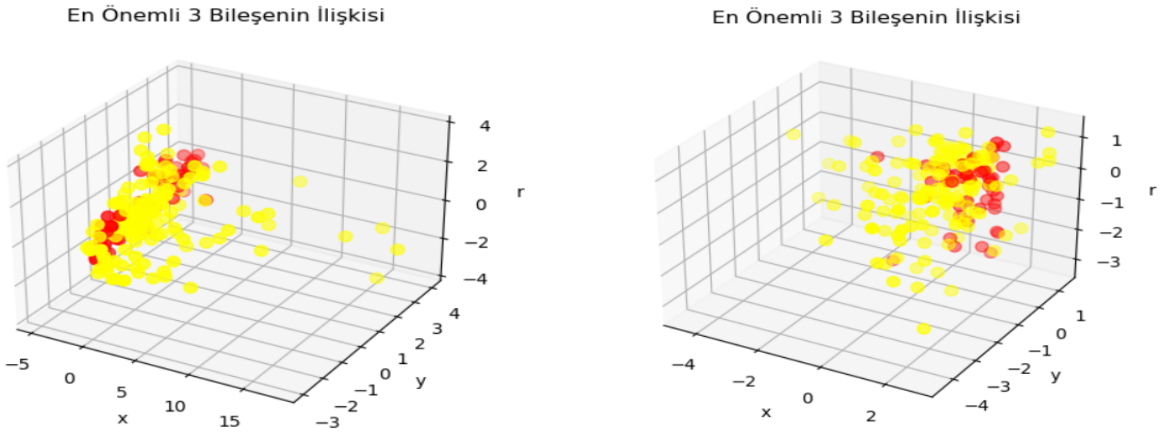
(a) En önemli iki bileşen arasındaki ilişki grafiği



(b) Bileşenlerin logaritmik dönüşüm altındaki grafiği

Birinci bileşenin yüksek değerleri için kişinin parkinson hastası olma olasılığı daha fazla iken ikinci bileşenin yüksek değerleri için sağlıklı birey olma olasılığı daha yüksektir. Ayrıca PCA dönüşümü sonucunda elde edilen en önemli iki bileşen ile kişinin parkinson hastasılığına sahip olup olmama durumu yani sınıflar arasındaki ilişki gözlenmişitr. Sınıflar çok net

olmasa da birbirinden ayrılmaya başlamıştır. Bileşenlerin logaritmik dönüşüm ile grafikleri çizildiğinde sınıflar arası sınırlar daha belirgin hale gelmiştir. Ayrıca ikinci bileşenin negatif değerleri için parkinson hastalığına sahip olma olasılığı daha fazladır.
Sınıflar arasındaki ilişkinin 3 boyutta incelenmesi aşağıdaki gibidir.



(a) En önemli üç bileşen arasındaki ilişki grafiği

(b) Bileşenlerin logaritmik dönüşüm altındaki grafiği

Birinci bileşen x, ikinci bileşen y ve üçüncü bileşen r ile gösterilmektedir. Birinci bileşenin pozitif değerleri için hastalığa sahip olma durumu daha fazla iken ikinci ve üçüncü bileşenin pozitif değerleri için hastalığa sahip olma durumu daha azdır. Ayrıca 2 boyutlu grafiğe göre sınıflar arası sınırlar daha belirgin ve birbirinden daha fazla ayrılmaktadır. Bileşenlerin logaritmik dönüşüm altındaki grafiği de çok fazla değişmediği gözlenmiştir. Sınıflar arası fark aynı şekilde tam ayrılmasa da 2 boyutlu grafiğe göre daha iyidir. Boyut arttırıldığında sınıflar arası fark daha anlaşılır hale gelmektedir. Bu da PCA analizini doğrular niteliktedir.

4.5 Modelin Kurulması

Makine öğrenmesi algoritmalarından Destek Vektör Makinesi, Lojistik Regresyon, K-En Yakın Komşu, Karar Ağaçları, Yapay Sinir Ağları yöntemleri uygulanmıştır.

Makine öğrenmesi sınıflandırma yöntemleri sonucunda ortaya çıkan modellerin doğruluğunun ölçülmesi ve model performanslarının karşılaştırılması gerekmektedir. Elde edilen sonuçlar doğrultusunda veri setine en iyi sınıflandırmayı açıklayan, en güvenilir ve doğru sonucu veren model seçilmektedir. Sınıflandırma algoritmalarının başarı oranının belirlenmesinde doğruluk (accuracy) ve hata matrisi ölçütlerinden yararlanılmıştır.

Tablo 2: Hata Matrisi Tablosu

		Tahmin Edilen Değerler	
		Pozitif	Negatif
Doğru Değerler	Pozitif	DP	YN
	Negatif	YP	DN

Hata matrisinde verilen ;
DP: Doğru sınıflandırılan pozitif birimlerin sayısı
YP: Gerçekte negatif iken pozitif olarak sınıflandırılan birim sayısı
YN: Gerçekte pozitif iken negatif olarak sınıflandırılan tahmin sayısını
YP: Gerçekte negatif iken pozitif olarak sınıflandırılan birim sayısını göstermektedir.

Doğruluk (Accuracy): Anlaşılması ve yorumlanması en basit ölçülerden biridir. Makine öğrenmesi algoritmalarında sıklıkla kullanılan bir ölçümdür. 0 ve 1 arasında değerler alır ve değerin 1'e yakın olması modelin başarılı olduğunu göstermektedir. Doğruluk değeri aşağıdaki şekilde hesaplanmaktadır:

$$\frac{DP + DN}{DP + YP + DN + YN} \tag{4.1}$$

Modeller oluşturulurken X' e sürekli değişkenleri, y' ye ise status değişkeni atanmıştır. Veri seti %30 oranında eğitim ve test seti olarak ikiye ayrılmıştır. Daha sonra X değişkeninin eğitim ve test verisine PCA uygulanmıştır. 4.3' deki PCA analizinin sonucundan bileşen sayısı 5 alınmıştır.

4.5.1 SVM

Class paketindeki "SVC" komutunu kullanarak model oluşturulmuştur. Modelde farklı çekirdek fonksiyonları ve ceza hiperparametreleri denemek için "GridSearchCV" komutu kullanılmıştır. $0.1, 1, 10, 100$ ceza değerleri ile linear, rbf (gauss), poly, sigmoid çekirdek fonksiyonlarının 10 katlı çapraz doğrulama ile 160 adet model kombinasyonu oluşturulmuştur. En başarılı modelin çekirdek fonksiyonunun gauss ve C değerinin 100 olduğu parametreler ile kurulduğu sonucuna varılmıştır. Bu değerler ile model oluşturulduğunda doğruluğu %66.10 gelmektedir. Gauss çekirdeği ile oluşturulan modele farklı gama değerleri atayarak model doğruluk değeri arttırılabilir. Gama değeri ile dağılımın genişliğini kontrol edilir ve marjin

genişliği ile doğru orantılıdır.

Tablo 3: Gauss çekirdeği için GridSearchCV için en iyi parametreleri bulma

C	0.1, 1, 10, 100
$gama$	0.1, 1, 10, 100

Modele gama parametresi eklendiğinde ve farklı ceza değerleri ile tekrar GridSearchCV komutu kullanılarak en başarılı modelin parametreleri bulunmuştur. Gama parametresi de eklendiğinde en başarılı model için ceza değeri de değişmiştir. Gauss çekirdeği için en iyi parametrelerin C'nin 0.1 ve gama değerinin 100 olduğu durumdur. Bu parametrelere göre model oluşturulduğunda doğruluk değeri %69.49' dur. Hata matrisi ise aşağıdaki şekildedir.

Tablo 4: Gauss Çekirdeği, $C = 100$ ve gamma=0.1 olduğu Modelin Hata Matrisi

		Tahmin Edilen Değerler	
		Hasta Değil	Hasta
Doğru Değerler	Hasta Değil	2	13
	Hasta	5	39

Hata matrisi incelendiğinde 2 kişiyi doğru sınıflandırarak parkinson hastalığına sahip olmayan kişiler hasta değil olarak sınıflandırılmıştır. 39 kişiyi de doğru sınıflandırılarak sağlıklı bireyleri hasta değil olarak sınıflandırmıştır. Model sağlıklı bireyleri genellikle hasta olarak sınıflandırmıştır ve model doğruluğu da düşük gelmiştir. Bu istenen bir durum olmadığı için çekirdek fonksiyonları ve ceza parametreleri kullanılarak farklı modeller oluşturulmuştur. Modellerin çekirdek fonksiyonları ve parametre değerleri için doğrulukları ve hata matrisleri incelenmiştir. Farklı C değerleri ve çekirdek fonksiyonları için oluşturulan modellerin doğruluk değerleri aşağıdaki gibidir;

Tablo 5: Farklı C ve Çekirdek Fonksiyonları için Doğruluk Değerleri

Kernel	C=0.1	C=1	C=10	C=100
Linear	72.88	67.79	67.79	67.79
Rbf	74.57	74.57	71.18	66.10
Poly	74.57	74.57	76.27	72.88
Sigmoid	74.57	72.88	74.57	71.18

Tablo 5 incelendiğinde en iyi doğruluk değerine sahip olan modelin çekirdek fonksiyonunun polinom ve ceza parametresinin 10 olduğu durumdur. Polinom çekirdeğinin de doğruluğunu

arttırmak için farklı dereceler alınabilir. Polinom çekirdeğini farklı parametre değerleri için incelenmiş ve aşağıdaki tabloya aktarılmıştır..

Tablo 6: Polinom çekirdeği için GridSearchCV için en iyi parametreleri bulma

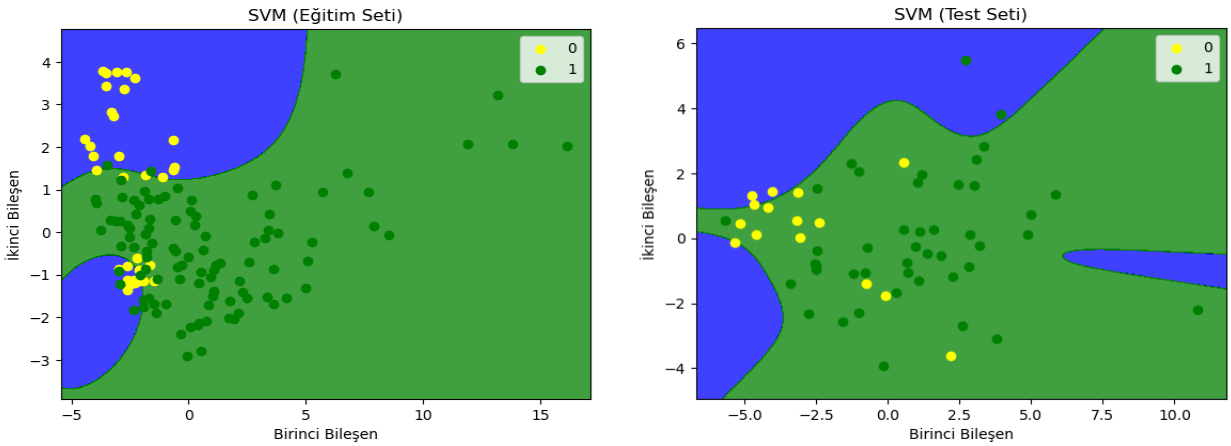
C	0.1, 1, 10, 100
dereceler	1, 2, 3, 4, 5, 6, 7, 8, 9, 10

Parametre değerleri için 10 katlı çapraz doğrulama ile 20 adet model fit edilmiştir. GridSearchCV komutunun sonucunda polinom çekirdeği için en iyi parametre değerlerinin C' nin 100 ve derecesinin 9 olduğu sonucuna varılmıştır. Fit edilen her modelin doğruluğu aşağıdaki gibidir.

Tablo 7: C=10 ve 100 için Polinom Çekirdeğinin Derecelerinin Tablosu

Derece	C=10 için Doğruluk	C=100 için Doğruluk
1	67.79	76.27
2	72.88	67.79
3	76.27	74.57
4	69.49	71.18
5	74.57	76.27
6	67.79	69.49
7	76.27	76.27
8	67.79	71.18
9	72.88	77.96
10	69.49	69.49

Polinom çekirdeği ile model oluşturulurken polinomun derecesi de etki etmektedir ve bu durumdan dolayı Tablo 7 ile GridSearchCV komutu ile elde edilen ceza parametresinin sonucu farklı çıkmıştır. En iyi parametreler ile model oluşturulduğunda eğitim verisinin doğruluk değeri %88.23 iken test verisinin doğruluk değeri %77.96 çıkmaktadır. Eğitim ve test verisinin doğruluk değerlerinin farklı çıkması modelin overfitting olduğunu göstermektedir. Model eğitim verisindeki durumları ezberlemiştir ve test verisi ile sınıflandırma yapmak istendiğinde eğitim setindeki durumları aramaktadır. Eğitim verisindeki durumları bulamadığında yanlış sınıflandırma yapmaktadır. Eğitim ve test verisinin grafikleri aşağıdaki şekildedir.

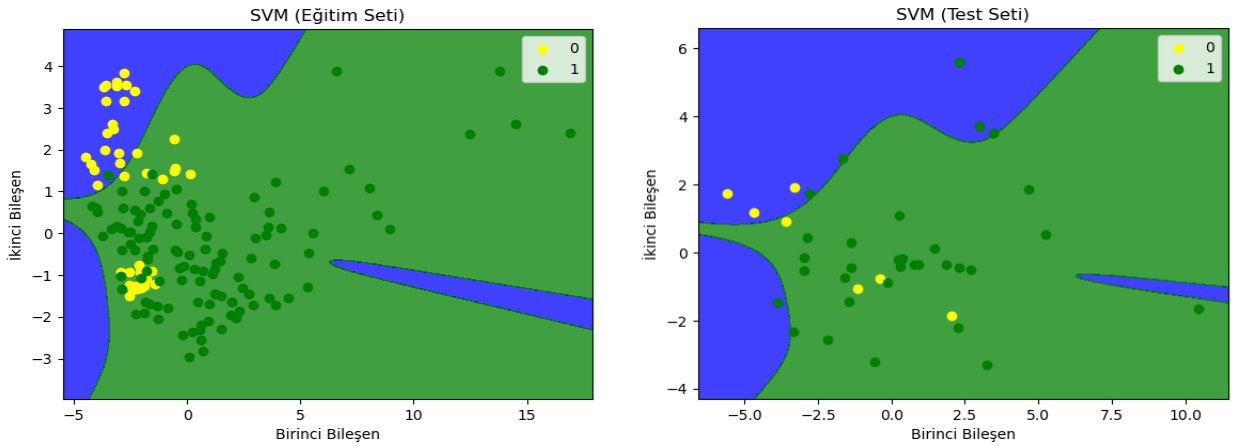


(a) Eğitim setinin grafiği

(b) Test setinin grafiği

Şekil 14: SVM için Oluşturulan Modellerin Görselleştirilmesi

Yukarıdaki grafikler incelendiğinde karar yüzeyi veriler için daha uyumludur. Fakat model test seti ile değerlendirildiğinde hasta olmayan kişileri hasta olarak sınıflandırmıştır. Bu durumun önüne geçebilmek için değişken sayısı azaltılabilir, eğitim verisinin oranını arttırılabilir, eğitim ve test verisinin doğruluğu aynı olduğunda eğitim durdurulabilir. Bu çalışmada veriyi %20 oranında ikiye ayırarak eğitim setinin verileri artırılmış, bileşen sayısı da 7 alınarak overfitting durumunun önüne geçilmeye çalışılmıştır. Veri setine yapılan işlemler sonucunda eğitim setinin model doğruluğu %87.82, test verisinin model doğruluğu ise %89.74'dir ve overfitting durumu ortadan kaldırılmıştır. Overfitting ortadan kaldırıldığında eğitim ve test verisinin grafikleri aşağıdaki gibidir.



(a) Eğitim setinin grafiği

(b) Test setinin grafiği

Şekil 15: SVM için Oluşturulan Modellerin Görselleştirilmesi

Sonuç olarak polinom çekirdeği için C'nin 100 ve derecesinin 9 olduğu parametre değerleri için model oluşturulmuştur. Modelin hata matrisi aşağıdaki tabloda verilmiştir.

Tablo 8: Polinom Çekirdeği, $C = 100$ ve derece=9 olduğu Modelin Hata Matrisi

		Tahmin Edilen Değerler	
		Hasta Değil	Hasta
Doğru Değerler	Hasta Değil	4	3
	Hasta	1	31

Sağlıklı olan 7 kişinin 4'ü doğru, 3' ünü ise parkinson hastası olarak sınıflandırmıştır. 32 parkinson hastasının ise 1'ini yanlış sınıflandırarak kişi parkinson hastası olduğu halde sağlıklı olarak sınıflandırmıştır. 31 parkinson hastasını ise doğru sınıflandırmıştır.
Sınıflandırma raporu aşağıdaki şekildedir:

Tablo 9: Sınıflandırma Raporu

	Precision	Recall	f1-score	support
Hasta değil	0.80	0.57	0.67	7
Hasta	0.91	0.97	0.94	32
accuracy			0.90	39
macro avg	0.86	0.77	0.80	39
weighted avg	0.89	0.90	0.89	39

Precision (kesinlik), bir sınıflandırma modelinin ne kadar doğru pozitif tahminler yaptığını ölçen bir değerdir. Precision, doğru pozitif tahminlerin toplam pozitif tahminlere oranını temsil eder. Tablo 9' a göre sağlıklı kişilerin %80 nini, parkinson hastası olan kişilerin ise %91' ini doğru sınıflandırmıştır.

Recall (duyarlılık), bir sınıflandırma modelinin ne kadar doğru pozitif örneği tespit ettiğini ölçen bir değerdir. Recall, gerçek pozitiflerin (TP) toplam pozitif örneklerin (TP + FN) oranını temsil eder. Örneğin sağlıklı bireyler için tahmin edilmesi gereken değerlerin ne kadarını doğru tahmin ettiğini gösterir.

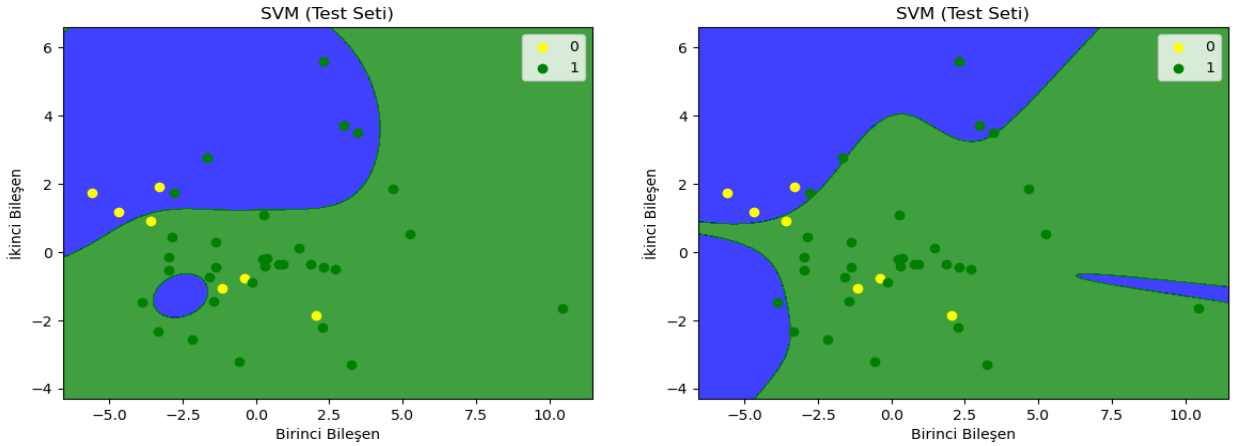
F1 skoru, bir sınıflandırma modelinin performansını ölçen bir metriktir. Precision (kesinlik) ve recall (duyarlılık) değerlerinin harmonik ortalamasını temsil eder.

Support, belirli bir sınıfa ait örneklerin toplam sayısını temsil eder.

Sınıflandırma raporu incelendiğinde modelin parkinson hastası olan kişileri, sağlıklı bireylere göre daha iyi sınıflandırdığı görünmektedir.

Sonuç olarak gauss ve polinom çekirdeği için oluşturulan modellerin doğruluk değerleri ve hata matrisleri karşılaştırıldığında polinom çekirdeği ile oluşturulan model parkinson hastası olan kişileri daha iyi sınıflandırmaktadır.

PCA uygulanmış eğitim verisinin ilk iki bileşeni ile geri kalan 3 bileşene göre oluşturulan modelleri görselleştirilmesi aşağıdaki gibidir;



(a) Gauss Çekirdeği ve $C = 100$, $gamma = 0.1$ (b) Polinom Çekirdeği, $C = 100$ ve $Degree = 9$

Şekil 16: SVM için Oluşturulan Modellerin Görselleştirilmesi

Yeşiller parkinson hastasılığına sahip kişileri, sarılar ise sağlıklı bireyleri göstermek üzere PCA dönüşümü yapılmış eğitim verisinin ilk iki bileşeni ile geri kalan 3 bileşene göre kurulan modelleri görselleştirdiğimizde elde edilen sonuclar, model doğruluğu ve hata matrislerinden elde ettiğimiz sonuçları doğrular niteliktedir. Gözlemler sonucunda SVM algoritması için en iyi modelin parametrelerinin polinom çekirdeği, ceza parametresinin 10 ve derecesinin 9 olduğu görünmektedir.

Modellerdeki overfitting sorununu gidermek için %20 oranında eğitim ve test verisi olarak ikiye ayrılmıştır. Bileşen sayısı ise 7 olarak alınmıştır.

4.5.2 Lojistik Regresyon

Class paketindeki "LogisticRegression" komutunu kullanarak model oluşturulmuştur. Bu kısımda veri setini eğitim ve test seti olarak ayrılmış daha sonra PCA uygulanarak model oluşturulmuştur. Modelin katsayıları 0.57573284, -0.73600302, -0.35230083, -0.37910126, -0.7883077, 0.09372298, -0.71737375 şeklindedir ve doğruluğu %79.48 'dir. Hata matrisi ise aşağıdaki gibidir.

Tablo 10: Lojistik Regresyon için Hata Matrisi

		Tahmin Edilen Değerler	
		Hasta Değil	Hasta
Doğru Değerler	Hasta Değil	4	3
	Hasta	5	27

SVM algoritması ile karşılaştırıldığında hastalığa sahip olan kişilerin sınıflandırılmasında fark vardır. Bu durumda SVM algoritması hastalığa sahip olan kişileri daha iyi sınıflandırıyor denebilir.

4.5.3 K-En yakın Komşu

Class sınıfından "KNeighborsClassifier" komutunu kullanarak model oluşturulmuştur. GridSearchCV fonksiyonu ile Euclidean, Manhattan, Minkowski ve Chebyshev metrikleri ve komşuluk sayısı için 0-50 arasındaki değerler alınarak en iyi parametre değerleri bulunmuştur. Komşuluk sayısı 11 alındığında ve Chebyshev metriği kullanıldığında modelin doğruluk skoru %84.6' dır. Hata matrisi ise aşağıdaki şekildedir.

Tablo 11: K-En Yakın Komşu Algoritması için Hata Matrisi

		Tahmin Edilen Değerler	
		Hasta Değil	Hasta
Doğru Değerler	Hasta Değil	3	4
	Hasta	2	30

7 sağlıklı kişinin 3'ü doğru, 4' ü ise yanlış sınıflandırılmıştır. Test setindeki 32 parkinson hastasının 30'unu doğru sınıflandırmıştır.

4.5.4 Karar Ağaçları

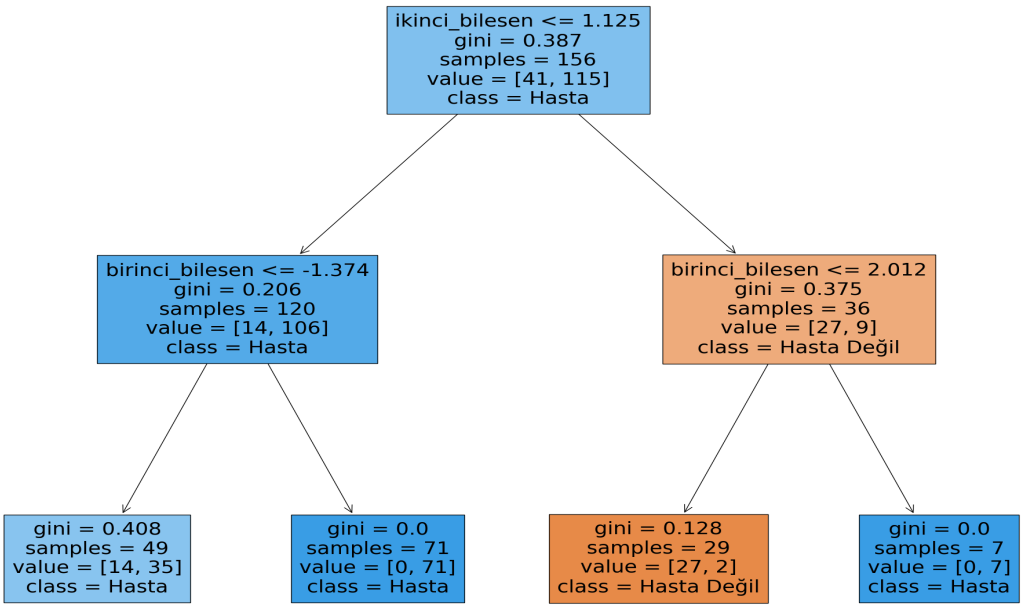
Class sınıfının "DecisionTreeClassifier" fonksiyonunu kullanarak model oluşturulmuştur. Karar ağaçlarından Classification and Regression Trees (CART) algoritması kullanılmıştır. Algoritmanın max depth ve min samples split olmak üzere iki adet ayar parametresi bulunmaktadır. Max depth maksimum derinlik demektir ve kök düğümden yaprak düğüme arasındaki en uzun yolu ifade eder. Min samples split ise bir düğümün tekrardan bölünmeden önce sahip olması gereken minimum verinin sayısıdır. max depth= [1,3,5,8,10] ve min samples split=[2,3,5,10,20,50] değerlerini vererek GridSearchCV fonksiyonu kullanıldığında en iyi parametre değerlerinin max depth için 2 ve min samples split için 10 olduğu görülmüştür. Bu

parametreler ile oluşturulan modelin doğruluk skoru %84.6 ve hata matrisi ise aşağıdaki şekilde bulunmuştur:

Tablo 12: CART Algoritması için Hata Matrisi

		Tahmin Edilen Değerler	
		Hasta Değil	Hasta
Doğru Değerler	Hasta Değil	3	4
	Hasta	6	26

Hastalığa sahip olmayan 7 kişi alındığında 4 kişiyi parkinson hastası olarak yanlış sınıflandırılmıştır. Aynı zamanda parkinson hastası olduğu bilinen 32 kişi alındığında 26 kişiyi doğru sınıflandırmıştır.



Şekil 17: Karar Ağacı

Eğitim setine PCA dönüşümü yapıldığında ilk bileşenin veriye katkısı (tablo 11'den) diğer bileşenlere oranla daha yüksektir. Karar ağacı oluşturulurken iki bileşenden yararlanılmıştır. Karar ağacı eğitim seti üzerinden oluşturulur. İkinci bileşen, birinci bileşenin değerlerine göre 2 adet IF şartı içerir. Örneğin eğer birinci bileşen -1.374 değerine eşit veya küçük ise alınan

14'ü sağlıklı, 106'sı parkinson hastası olan toplam 120 kişinin birinci bileşenin 2. IF şartına göre 71'i parkinson hastasıdır.

4.5.5 Rastgele Orman

Class sınıfındaki "RandomForestClassifier" komutu kullanılarak model oluşturulmuştur. Algoritmanın ayar parametreleri;
n estimators: Kaç adet karar ağacı kullanılacağının ayarlandığı parametredir.
max features: Bir düğümü bölerken dikkate alınması gereken özelliklerin rastgele alt kümelerinin boyutudur.
min samples split: Bir düğümün tekrardan bölünmeden önce sahip olması gereken minimum verinin sayısıdır.
n estimators=[100,200,500,1000], max features=[3,5,7,8] ve min samples split=[2,5,10,20] parametreleri için GridSearchCV komutu ile en iyi parametrelerin n estimators için 400, max features için 3 ve min samples split için 30 olduğu görülmüştür. Bu parametre değerleri ile oluşturulan modelin doğruluk skoru %84.6 ve hata matrisi ise aşağıdaki şekildedir:

Tablo 13: Rastgele Orman Algoritması için Hata Matrisi

		Tahmin Edilen Değerler	
		Hasta Değil	Hasta
Doğru Değerler	Hasta Değil	3	4
	Hasta	2	30

Sağlıklı olduğu bilinen 7 kişinin 3' ünü doğru sınıflandırmıştır. Parkinson hastası olduğu bilinen 32 kişinin ise 30' unu doğru sınıflandırmıştır.

4.5.6 Yapay Sinir Ağları

Class sınıfındaki "MLPClassifier" komutu ile model oluşturulmuştur. Modelin parametreleri alpha ve hidden layer sizes' dır. Alpha değeri L2 cezası (düzenleme süresi) parametresidir ve momentum algoritmasında momentumun ağırlığını belirleyen bir faktördür. Hidden layer sizes ise gizli katmanların her birinin içerdiği nöron sayısını belirtir. Modelde alpha değeri için 1 ve hidden layer sizes için $(20, 5)$ değerleri alınmıştır. Ayrıca aktivasyon fonksiyonları için sigmoid, tanh, relu, ve identity fonksiyonları kullanılmıştır (identity fonksiyonu doğrusaldır ve lineer olarak ayrılabilen veri setleri için, diğer fonksiyonlar ise doğrusal olarak ayrılamayan veri setleri için kullanılır.). Aktivasyon fonksiyonu için ise "sigmoid" seçilmiştir. Bu aktivasyon fonksiyonu ve parametre değerleri için modelin doğruluk skoru % 87.1' dir. Hata matrisi

ise aşağıdaki şekildedir.

Tablo 14: Yapay Sinir Ağları Algoritması için Hata Matrisi

		Tahmin Edilen Değerler	
		Hasta Değil	Hasta
Doğru Değerler	Hasta Değil	4	3
	Hasta	2	30

7 sağlıklı olduğu bilinen kişinin 4' ünü, parkinson hastası olduğu bilinen 32 kişinin ise 30' unu doğru sınıflandırmıştır.

4.5.7 Modelin Denenmesi

Bu kısımda parkinson hastası olmadığı bilinen 8 kişiden izinleri doğrultusunda sesleri kayıt edilmiştir. Kaydedilen bu sesleri, Praat yazılımından yararlanarak jitter, HNR gibi parametreleri ölçülmüş ve çıkan bu sonuçlar ile yeni bir veri seti oluşturulmuştur. Analizlerin sonucunu modelde predict komutu ile yerine koyularak kişilerin parkinson hastalığına sahip olup olmama durumu belirlenmeye çalışılmıştır.

PRAAT

Praat, konuşma ve ses dosyalarını analiz etme, düzenleme imkanı sunan popüler bir ses işleme yazılımıdır. Praat, kullanıcılarına bilgisayar ile erişme ve oluşturulan ses örnekleri üzerinde çeşitli işlemler yapma imkanı sunar. Yazılım özellikle dilbilimciler, konuşma terapistleri ve akademik çalışmalarda oldukça sık kullanılmaktadır.

Praat' ın özelliklerinden biri de ses dosyalarının analizi yapılırken grafikler ile görselleştirilir. Bu sayede kullanıcılar kaydedilen sesleri dalga biçiminde görebilir, frekans analizi, formant analizi, spektrogramlar ve pitch analizi gibi araçlar yer alır ve bu sesler üzerinde işlemler yapılabilir. Praat özellikle fonatik incelmelere yönelik bir programdır ve konuşmanın jitter, shimmer, perde, HNR gibi parametreleri ölçmektedir.

Praat yazılımı, dilbilimcilerin sesleri analiz etmelerine imkan sunmak için geliştirilmiştir.

Sesin frekansını, şiddetini, tonlama, duraklar ve vokal kalitesi gibi özelliklerinin analizleri için kullanılmaktadır. Bu özellikler sayesinde dilbilimciler tarafından, konuşmanın doğru bir şekilde analizi gerçekleştirilir. Aynı zamanda konuşma bozukluklarının teşhis ve tedavisinde de Praat yazılımdan yararlanılmaktadır. Konuşma terapistleri, Praat yazılımından elde ettikleri analiz sonucunda hastalarının konuşma bozukluklarına yönelik tedaviler geliştirebilirler.

Sonuç olarak Praat ses analizleri, dilbilim ve müzik araştırmaları için oldukça kullanışlı bir yazılımdır. Tercih edilmesinin bir diğer sebebi ise herkes tarafından kolaylıkla erişilebilen açık kaynak kodlu olmasıdır.

Praat yazılımı ile kişilerden alınan ses kayıtları işlenmiş ve sonuçları excel dosyasına kaydedilmiştir. Kişilerin 2' sinden 10 ve 30 saniyelik, ikisinden sadece 30 saniyelik ve 4'ünden de sadece 10 saniyelik sesler olmak üzere 10 adet ses verisi kaydedilmiş ve işlenmiştir. Praat ile elde edilen analizdeki parametreler ile orjinal veri setimizdeki değişkenlerden ortak olanları seçilmiştir. Ortak olan değişkenler sırasıyla; Jitter:DDP, Shimmer:APQ3, Shimmer:APQ5 ,Shimmer:DDA, NHR, HNR' dir. Daha önceki adımlarda oluşturulan SVM modeli tekrar bu değişkenler üzerinden oluşturulmuştur. Yeni veri setindeki kişilerin ses verilerini analizi sonucunda elde edilen değerleri predict komutundan yararlanarak parkinson hastalığına sahip olup olamama durumları tahmin edilmiştir. Bulunan sonuçlar aşağıdaki tablodaki gibidir.

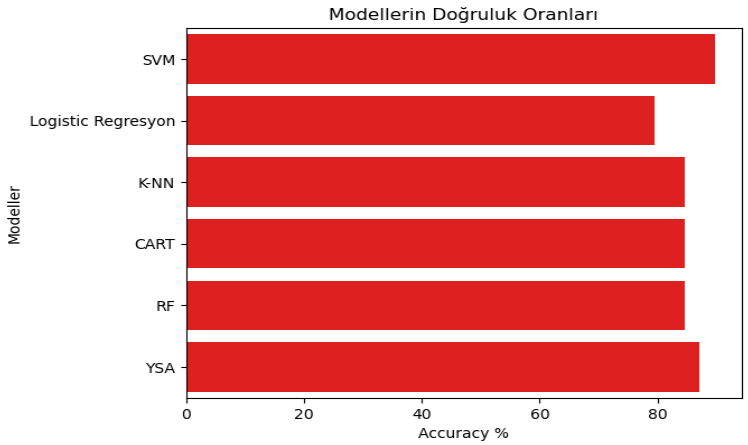
Kişiler	Status
Tahsin (10 sn)	1
Kerem (10 sn)	0
Melike (10 sn)	1
Nihal (10 sn)	0
Beyza (10 sn)	1
Havva (10 sn)	1
Melike (30 sn)	0
Elif (30 sn)	0
Zeynep (30 sn)	0
Havva (30 sn)	0

Elde edilen sonuçlar doğrultusunda modelde denenen 10 farklı ses ile yapılan sınıflandırmanın 6' sını doğru olduğu görünmektedir. Modellerin hata matrisleri ve sınıflandırma

raporları incelendiğinde sağlıklı bireyleri sınıflandırmada başarı düşüktü fakat model denendiğinde sağlıklı bireyleri de sınıflandırma da başarılı olduğu görünmektedir. Ayrıca modelde denenen seslerin uzunluğu da sonucun doğruluğunu etkilemektedir. Aynı zamanda aynı kişiler (Melike ve Havva) tarafından elde edilen 10 ve 30 saniyelik seslerden elde edilen sonuçlar değerlendirildiğinde 30 saniyeden elde edilen sonucun doğru olduğu görünmektedir.

5 SONUÇ

Bu çalışmada denetimli makine öğrenmesi algoritmaları kullanarak Parkinson hastalığının ses verisi ile teşhis edilmesi amaçlanmaktaydı. Kullanılan veri seti hasta ve hasta olmayan kişilerin 6 adet ses kayıtlarından oluşmaktaydı. Farklı algoritmalar ile oluşturulan modellerin doğruluk skorları aşağıdaki gibidir:



Şekil 18: Modellerin Doğruluk Oranı

Kurulan modellerde Parkinson hastalığının ses verisi ile teşhisinde doğruluk skorları ve hata matrisleri göz önüne alındığında en iyi sınıflandırma algoritmasının (Şekil 18' den) doğruluk skoru %89.74 olan SVM algoritmasının olduğu görünmektedir. SVM algoritmasında çekirdek kullanarak modelin performansını arttırmak mümkündür. Burada seçilen çekirdeğin türü kullanılan veri setinin yapısına, boyutuna ve problemin doğasına göre seçilmelidir. Yanlış seçilen çekirdek modelin performansını olumsuz etkiler ve hatalı sınıflandırmaya yol açabilir. Örneğin bu çalışmada gauss çekirdeği ile oluşturulan modelin performansı oldukça düşmüştür ve test verisini tek bir sınıftan oluşuyormuş gibi sınıflandırmıştır. Fakat polinom çekirdeği kullanıldığında modelin performansı artmış ve iki sınıfı birbirinden daha iyi ayırt

edebilir hale gelmiştir. Ayrıca çekirdek türlerine ait parametreler de modelin performansını etkilemektedir. Tablo 7 incelendiğinde de tek dereceler için modelin daha iyi performans gösterdiği görünmektedir.

Oluşturulan modeller performanslarına göre sıralanacak olursa SVM'den sonra en iyi sınıflandırmayı yapan algoritma YSA' dır. KNN,CART ve RF algoritmaları aynı performansı göstermektedir. En kötü performansı ise Lojistik Regresyon göstermektedir.

Çalışmada kullanılan modeller parkinson hastalığına sahip kişileri genellikle doğru tahmin etmektedir. Ancak hastalığa sahip olan bireyleri sınıflandırmadaki başarısı, hastalığa sahip olmayan bireyleri sınıflandırmadaki başarısına göre oldukça yüksektir. Hastalık sınıflandırma problemlerinde önemli olan nokta hastalığa sahip olma ihtimali yüksek olan kişileri tespit etmek ve bir hekime yönlendirmek olduğundan istenilen amaca ulaşılmıştır.

Modelin denenmesi kısmında Parkinson hastası olmadığı bilinen 10 kişinin 6'sı doğru sınıflandırılmıştır. Burada önemli olan nokta kişilerin seslerinin uzunluğudur. Aynı kişiden, aynı zamanda alınan 10 ve 30 saniyelik ses kayıtları ile sınıflandırma yapıldığında 30 saniyelik ses analizi ile doğru sınıflandırma yapıldığı,10 saniyelik ses ile yanlış sınıflandırma yapıldığı görünmektedir. Modelin denendiği kısımda elde edilen sonuçlara göre algoritmanın Parkinson hastası olmadığı bilinen kişilerde de performansının oldukça yüksek olduğu söylenebilir.

Kısaca Parkinson hastalığının ses verisi ile teşhisinde modeller oluşturulmuştur. En iyi performansa sahip modelin SVM olduğu belirlenmiştir. SVM modeli Parkinson hastalığına sahip ve sahip olmayan kişileri birbirinden ayırt edebilmektedir. Hastalığa sahip olma şüphesi olan kişiler tarafından kolayca kullanılabilir ve çıkan sonuca göre hastanelerin nöroloji bölümüne başvurabilirler.

Çalışmaya ait Python kodları aşağıdaki web sitesinde yer almaktadır.
`https://github.com/Cetinzafer/parkinson-hastaliginin-ses-verisi-ile-teshisi`

Kaynaklar

[1] 0159894.pdf (trakya.edu.tr)

[2] https://www.sonerturudu.com/akustik-ses-analizinde-kullanilan-parametreler-ve-yazilimlar/

[3] Temel Bileşenler Analizi (zafercomert.com)

[4] Suitability of dysphonia measurements for telemonitoring of Parkinson's disease (nature.com)

[5] http://dspace.trakya.edu.tr/xmlui/bitstream/handle/trakya/3095/0159894.pdf?sequence=1

[6] https://acikerisim.medipol.edu.tr/xmlui/bitstream/handle/20.500.12511/7903/Pence-Kadriye-2020.pdf?sequence=1&isAllowed=y

[7] Pagonabarraga J, Kulisevsky J. Cognitive impairment and dementia in Parkinson's disease. Neurobiology of Disease. Academic Press. 46;590–6, 2012

[8] Walsh B, Smith A. Linguistic complexity, speech production, and comprehension in Parkinson's disease: Behavioral and physiological indices. Journal of Speech, Language, and Hearing Research, 2011

[9] Lechien JR, Blecic S, Ghosez Y, Huet K, Harmegnies B, Saussez S. Voice Quality and Orofacial Strength as Outcome of Levodopa Effectiveness in Patients with Early Idiopathic Parkinson Disease: A Preliminary Report. J Voice. Sep 1;33(5):716–20, 2019

[10] Miller N. Communication changes in Parkinson's disease. Practical Neurology, 17(4), 266-274, 2007.

[11] Kara MS, Caplan DN. Communication impairment in Parkinson's disease: impact of motor and cognitive symptoms on speech and language. Brain and language 185;38–46, 2018.

[12] https://www.cerrahi.com.tr/parkinson-nedir-neden-olur-parkinson-belirtileri-tedavisi

[13] Awad, M., & Khanna, R. (2015). Support vectors machines for classification. In Efficient learning machines (36-99)

[14] Yücelbaş, Ş., & Yücelbaş, C. (2019). Temel Bileşen Analizi Yöntemleri Kullanarak Parkinson Hastalığının Otomatik Teşhisi. Avrupa Bilim ve Teknoloji Dergisi, (16), 294-300.

[15] GÜNDÜZ, H. (2019). Parkinson Hastalığı Tespitinde Farklı Boyutsallık İndirgeme Yöntemlerinin Karşılaştırılması. Avrupa Bilim ve Teknoloji Dergisi, (17), 1164-1172.

[16] BADEM, H. (2019). Parkinson Hastaliğinin Ses Sinyalleri Üzerinden Makine Öğrenmesi Teknikleri ile Tanimlanmasi. Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi, 8(2), 630-637.

[17] Esmer, S., Uçar, M. K., İbrahim, Ç. İ. L., & Bozkurt, M. R. (2020). Parkinson hastalığı teşhisi için makine öğrenmesi tabanlı yeni bir yöntem. Düzce Üniversitesi Bilim ve Teknoloji Dergisi, 8(3), 1877-1893.

[18] Mei, J., Desrosiers, C., & Frasnelli, J. (2021). Machine learning for the diagnosis of Parkinson's disease: a review of literature. Frontiers in aging neuroscience, 13, 633752.

[19] Senturk, Z. K. (2020). Early diagnosis of Parkinson's disease using machine learning algorithms. Medical hypotheses, 138, 109603.

[20] Akkaya, E. M. Ü. Ü. (2019). Parkinson hastalığında ses ve konuşmanın akustik analizi (Doctoral dissertation, Anadolu University (Turkey)).

[21] SEVİNDİK, S., & ŞİRAY, G. Ü. (2018). DİSKRİMİNANT ANALİZİ VE BAZI ALTERNATİF REGRESYON ANALİZLERİ.

[22] Ersungur, Ş. M., KIZILTAN, A., & Polat, Ö. (2007). Türkiye'de Bölgelerin Sosyo-Ekonomik Gelişmişlik Siralamasi: Temel Bileşenler Analizi. Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi, 21(2), 55-66.

[23] Albayrak, A. S. (2005). Çoklu doğrusal bağlanti halİnde enküçük kareler teknİğİnİn alternatİfİ yanli tahmİn teknİklerİ ve bİr uygulama. Uluslararası Yönetim İktisat ve İşletme Dergisi, 1(1), 105-126.

[24] Erdem, A. (2023). Destek Vektör Makinesi Sınıflandırma Algoritması için Genişletilmiş Gauss Çekirdeği ve Uygulaması, Yüksek Lisans Tezi, Medeniyet Üniversitesi, Uygulamalı Matematik ve Hesaplamalı Bilimleri Enstitüsü, İstanbul

[25] Demirhan, M.E.(2016).Sentetik Açıklıklı Radar Görüntülerinde Otomatik Hedef Tanıma, Yüksek Lisans Tezi, Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Ankara

[26] http://abakus.inonu.edu.tr/xmlui/bitstream/handle/11616/7944/Tez

[27] https://miracozturk.com/python-ile-siniflandirma-analizleri-knn-k-nearest-neighbours-k-en-yakin-komsu-algoritmasi/

[28] Dekhtyar A. CSC 466: Knowledge Discovery from Data-Distance/Similarity Measures. 2009.

[29] Kresse W, Danko DM. Springer handbook of geographic information: Springer; 2012.

[30] Utgoff, P.E.(1989). Incremental induction of decision trees. Machine Learning, 4(2),161-186.

[31] Hosmer, D. W., JR., S.Lemeshow, & R. X. Sturdıvant, (2013). Applied Logistic Regression, John Wiley & Sons, Canada

[32] JavaTpoint,(2022), Logistic Regression in Machine Learning

[33] Theobald, O. (2017). Machine Learning for Absolute Beginners, Independently published.

[34] Krogh, A. (2008). What are artificial neural networks? Nature Biotechnology, 26(2),195-197. https://doi.org/10.1038/nbt1386

[35] Yalçın, M., & Aksu-Koç, A. (2018). Praat programı ile Türkçe'de tonların frekans analizi. Dil ve Konuşma Terapisi Araştırmaları Dergisi, 4(1), 1-8.

[36] Arslan, S. (2015). Praat programı ve kullanımı. Turkish Studies, 10(3), 177-190.

[37] Korkmaz, B., & Özçelik, Ö. (2016). Praat programı kullanılarak ses kaydı ve analizi. Dil ve Dilbilim Araştırmaları Dergisi, 12(2), 113-126.

[38] Özbay, M. (2017). Dilbilim araştırmalarında Praat programının kullanımı. Dil Dergisi, 173, 61-67.