T.C.
TRAKYA ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
EKONOMETRİ ANABİLİM DALI
EKONOMETİRİ BİLİM DALI
YÜKSEK LİSANS TEZİ



MAKİNE ÖĞRENMESİ ALGORİTMALARINI KULLANARAK BİTCOİN FİYAT TAHMİNİ

ZEYNEP ŞENGÜL

TEZ DANIŞMANI

PROF.DR. NURCAN METİN

EDİRNE, 2022

Çok kıymetli; Babam Ethem, Annem Zümran

Ve Kardeşlerim Nefise ve Nisa'ya,

ÖNSÖZ

Öncelikli olarak, değeli tez danışmanım Prof. Dr. Nurcan METİN'e verdiği destek ve zaman için teşekkürü borç bilirim.

Yüksek lisans dönemimde beni makine öğrenmesine yönlendiren Doç. Dr. Mehmet Kenan TERZİOĞLU'na teşekkür ederim

Lisans yıllarımda özellikle son sınıfta dersime girerek bölümümü çok sevmeme sebep olan hocalarıma sonsuz teşekkürlerimi sunarım.

İÇİNDEKİLER

ÖNSÖZ ..... **İİİ**

İÇİNDEKİLER ..... **İV**

KISALTMALAR ..... **Vİ**

SEMBOLLER ..... **Vİİ**

ÇİZELGE LİSTESİ ..... **Vİİİ**

ŞEKİL LİSTESİ ..... **İX**

ÖZET ..... **Xİ**

SUMMARY ..... **Xİİ**

1. GİRİŞ **1**

2.KAVRAMLAR **3**

2.1. Para **3**

2.1.1. İtibari Para 4

2.1.2. Elektronik Para 5

2.1.3. Dijital Para 5

2.1.4. Kripto Para 5

2.2. Makine Öğrenmesi **8**

2.2.1. Denetimli Makine Öğrenmesi 10

2.2.1.1.Regresyon Analizi 11

2.2.1.2. Sınıflandırma 12

2.2.2.Yarı Denetimli Makine Öğrenmesi 12

2.2.3. Denetimsiz Makine Öğrenmesi 13

2.2.4. Takviyeli Makine Öğrenmesi 15

3. METOTLAR **17**

3.1. Destek Vektör Makineleri **17**

3.1.1. Doğrusal Olarak Ayrılabilen Veriler 18

3.1.2. Doğrusal Olarak Ayrılamayan Veriler 22

3.1.3. Destek Vektör Doğrusal Regresyonu 27

3.1.4 Doğrusal Olmayan Destek Vektör Regresyonu 29

3.2. Rastgele Orman **32**

3.3. K-En Yakın Komşular **36**

4. MODEL PERFORMANS METRİKLERİ **42**

5. UYGULAMA **43**

6. SONUÇ 51

KAYNAKÇA 53

KISALTMALAR

ARCH : Otoregresif Koşullu Değişen Varyans
DVM : Destek Vektör Makinesi
DVR : Destek Vektör Regresyonu
ECB : European Central Bank
KA : Karar Ağaçları
KNN : K-En Yakın Komşular
MSE : Ortalama Kare Hata
OOB : Out Of Bag
RBF : Radyal Tabanlı Fonksiyon
RMSE : Kök Ortalama Kare Hata
RO : Rastgele Orman
YSA : Yapay Sinir Ağları

SEMBOLLER

$\boldsymbol{C}$: Ceza parametresi
$\boldsymbol{\theta_k}$: Eğitim verisine dair rastgele vektör
$\boldsymbol{\xi}$: Esneklik değişkeni
$\boldsymbol{\gamma}$: Gama
$\in$: Hata toleransı

ÇİZELGE LİSTESİ

Çizelge 1: DVR tabanlı tahmin modeli için kullanılan parametreler ve değer aralıkları**.....31**
Çizelge 2: KNN örnek veri seti**.....37**
Çizelge 3: Bitcoin'e ait tanımlayıcı istatistikler**.....44**
Çizelge 4: Makine öğrenmesi algoritmalarının tahmin değerleri**.....48**
Çizelge 5: Makine öğrenmesi tahmin algoritmalarına ait hata istatistikleri**.....49**
Çizelge 6: Makine öğrenmesi tahmin algoritmalarına ait R^2 değerleri**.....50**

ŞEKİL LİSTESİ

Şekil 1: Dünya Çapında Nakit Para Akışı **4**

Şekil 2: Bitcoin Fiyatı **7**

Şekil 3: Makine Öğrenmesi Adımları..... **9**

Şekil 4: Yaygın Kullanılan Makine Öğrenmesi Modelleri..... **10**

Şekil 5: Denetimli Öğrenme Modeli..... **11**

Şekil 6: Yarı Denetimli Öğrenme Modeli **12**

Şekil 7: Denetimsiz Öğrenme..... **13**

Şekil 8: Karadeniz Bölgesindeki İllerin Denize Kıyısı Olup Olmamasına Göre Kümelenmesi..... **14**

Şekil 9: Takviyeli Öğrenme..... **15**

Şekil 10: Takviyeli Öğrenme ile Deneyerek Çıkışı Bulma **15**

Şekil 11: Takviyeli Öğrenme İle Kısa Yolun Belirlenmesi..... **16**

Şekil 12: En İyi Ayırıcı Düzlem **17**

Şekil 13: En İyi Ayırıcı Hiper Düzlem Ve Destek Vektörler..... **19**

Şekil 14: Destek Vektör Ağ Yapısı **20**

Şekil 15: Doğrusal Olarak Belirli Bir Hata İle Ayrılabilen DVM..... **21**

Şekil 16: Doğrusal Olarak Ayrılamayan Veri Örneği..... **22**

Şekil 17: Doğrusal Çekirdek..... **24**

Şekil 18: Polinom Çekirdek Kullanımıyla 2 Boyutlu Uzaydan 3 Boyutlu Uzaya Taşınması **25**

Şekil 19: Radyal Tabanlı Çekirdeğin γ Değerine Göre Sınıflandırılması **26**

Şekil 20: Sigmoid Fonksiyonu..... **27**

Şekil 21: Kayıp Fonksiyonu..... **28**

Şekil 22: Doğrusal DVR..... **28**

Şekil 23: Doğrusal Doğrusal Fonksiyonu İçin ϵ Tüpü Ve Esneklik Değişkeni (DVR) **29**

Şekil 24: Çekirdek Yaklaşımı **30**

Şekil 25: DVR'de Kayıp Fonksiyonu **30**

Şekil 26: C' nin Belli Değeri İle Sahip Olduğu Marjin Aralığı **31**

Şekil 27: Rastgele Orman Algoritması İle Veri Setinin Oluşturulması **33**

Şekil 28: Rastgele Orman Algoritması İle Oob İşleyiş Süreci **34**

Şekil 29: KNN İki Boyutlu Düzlemde Örnek Çalışması **38**

Şekil 30: İki Boyutlu Düzleme Yeni Bir Gözlemin Eklenmesi **39**

Şekil 31: k=3 Değerine Göre Komşu Belirleme İşlemi **39**

Şekil 32: Yeni Gelen Öğrencinin Sınıfa Atanması **40**

Şekil 33: Bitcoin Kapanış Verileri **43**

Şekil 34: Doğrusal DVR **44**

Şekil 35: RBF -DVR **45**

Şekil 36: Polinominal -DVR **45**

Şekil 37: Sigmoid-DVR **46**

Şekil 38: Rastgele Orman Regresyonu **47**

Şekil 39: KNN Regresyonu **47**

MAKİNE ÖĞRENMESİ ALGORİTMALARINI KULLANARAK BİTCOİN FİYAT TAHMİNİ

ÖZET

Yatırımcılar için finansal piyasaların gelecek değerlerinin tahminini bilmek önemlidir. Bunun sebebi yatırımcıların yapmış oldukları yatırımların kendilerine maksimum kar sağlamasını istemesi ve zarara uğramak istememesidir. Ancak finansal zaman serilerinin özelliği olan aşırı oynaklık tahminin doğruluk yüzdesini azaltmaktadır.

Çalışmada 1826 gözlemden oluşan Bitcoin'e ait günlük kapanış değerleri kullanılarak makine öğrenmesi algoritmaları ile 1827. gün kapanış tahmini gerçekleştirilmektedir. Tahmin için belirlenen algoritmalar Destek Vektör Regresyonu, Rastgele Orman Regresyonu ve k-En Yakın Komşular Regresyonu olmaktadır. Tahmin algoritmalarının performans ve doğruluklarını değerlendirmek için MSE, RMSE ve R^2 değerleri dikkate alınmaktadır.

Tüm tahmin grafikleri ve model performans metrikleri incelendiğinde en iyi tahmini Rastgele Orman Regresyonu gerçekleştirmektedir. Destek Vektör Regresyonları içinde serinin polinominal artış göstermesinden dolayı en iyi tahmini verdiği ancak doğrusal, radyal tabanlı ve sigmoid çekirdek fonksiyonlu tahmin algoritmalarının etkin çalışmadığı görülmektedir. k- En Yakın Komşular Regresyonunun ise Rastgele Orman tahmini kadar etkin tahmin yaptığı görülmektedir.

Sonuç olarak belirlenen makine öğrenmesi tahmin algoritmaları içinden en az hata ve en yüksek R^2 değerine sahip olarak Rastgele Orman regresyonu olmaktadır.

Bitcoin'e ait veriler www.investing.com 'dan alınmıştır. Tüm kodlar Python'da yazılmış olup, model tahminleri için scikit-learn kütüphanesinden faydalanılmıştır.

BITCOIN PRICE PRESICTION USING MACHINE LEARNING ALGORITHMS

SUMMARY

It is important for investors to know the forecast of the future values of financial markets. This is because investors want their investments to provide them with maximum profit and do not want to be damaged. However, the characteristic of the financial time series, excessive volatility reduces the percentage of accuracy of the estimate.

With machine learning algorithms using the daily closing values of Bitcoin, which consist of 1826 observations in the study 1827. The day closing forecast is being carried out. The algorithms determined for estimation are the Support Vector Regression, Random Forest Regression, and k-Nearest Neighbors Regression. MSE, RMSE, and R^2 values are taken into account to assess the performance and accuracy of prediction algorithms.

When all forecasting graphics and model performance metrics are reviewed, the best estimate is Random Forest Regression. The Support Vector Regressions show that the series provides the best estimate due to the polynominal increase, but the prediction algorithms with linear, radial-based and sigmoid core function do not function effectively. the k-Nearest Neighbors Regression is seen to be predicting as much effective as the Random Forest forecast.

The resulting machine learning has a Random Forest regression of at least error within the estimation algorithms and with the highest value of R^2.

Bitcoin data is derived from www.investing.com. All the codes are written in Python, and the scikit-learn library is used for model estimates.

1. GİRİŞ

Bireyler ve kurumlar, yatırım yaptıkları finansal zaman serilerinde maksimum kar yapmak istemektedirler. Bunun için gelecek tahmini yapmak yatırımlarını zarara uğramadan, maksimum kara taşımak önemli olmaktadır. Finansal zaman serileri için gelecek tahmini yapmak oldukça zordur. Bunun sebebi piyasa hareketlerinin belirsiz olmasıdır. Belirsiz olmasının nedeni finansal zaman serilerinin siyasi, ekonomik, sosyal birçok olaydan etkilenmesinden kaynaklanmaktadır. Bu sebeple gelecek öngörüsü yapıldığında gerçekte o an var olmayan olayların ortaya çıkması ile piyasada ani iniş ve çıkış yaşanabilmektedir. Bu ani gelişen olaylar, yapılan tahminin tersine farklı bir değer sergileyebilmektedir.

Akademik çalışmalara göre fiyat hareketleri doğrusal olmayan, dinamik yapıya sahiptir. Piyasada yer alan finansal araçların fiyat haraketliliklerinin yönünün tahmin edilmesi ilgili piyasaya yapılan yatırımların, yatırım sahiplerine kar elde etmesi için önemli bilgi sağlamaktadır. Piyasada işlem yapmak isteyen yatırımcılar için fiyat değerlerinin hangi yönde ilerleyeceği beklentisi son derece önemli ve risklidir. Finansal serilerde hareket yönünün tahmin edilmesi yatırım yapılan finansal değerde daha fazla kazanç sağlamaktadır (Chen ve ark., 2003).

Finansal piyasalardaki gelecek fiyat tahmini; ekonometri, ekonomi, finans, bilgisayar bilimi, tarih ve matematik alanları başta olmak üzere birçok alanda, çok fazla araştırmacının dikkatini çekmiş olması ile beraber birçok araştırma yapılmıştır. Araştırmalara göre piyasadaki fiyat hareketlilikleri rassal değildir. Yüksek oynaklığa sahip finansal serileri basit zaman serisi ya da regresyon tekniklerinin uygulanmasını zorlaştırmasından ötürü yüksek kazanç elde etmek amacıyla portföy yöneticileri farklı işlem modelleri geliştirilmiştir. Fakat bu geliştirdikleri modeller fazla kazanç sağlamayarak ortalamanın altında gelir elde etmişlerdir. Finansal hareketliliklerin yönündeki değişkenlik doğrusal tekniklerin kullanımını optimumum olmaktan çıkararak ARCH gibi doğrusal olmayan modeller daha düşük hatalara sahip tahmin sağlamaktadır (Zhang ve ark., 2009).

Araştırmacılar finansal zaman serisi tahminlerinde son zamanlarda yapay zeka, derin öğrenme ve makine öğrenmesine odaklanmışlardır.

2008 yılında yaşanan küresel krizde büyük finansal kayıplar meydana gelmesi ile beraber finansal piyasalara ve ulusların mali güvenine olan kayıp sonucunda sanal paraya olan talep artmıştır. Bu sebeple kripto paraların başında gelen Bitcoin için gelecek tahmini yapılması önem arz etmektedir.

Bu çalışma Bitcoin'e ait kapanış değerleri dikkate alınarak bir sonraki gün tahmini yapılması hedeflenmektedir. Amaç, belirlenen Makine Öğrenmesi algoritmaları arasından Destek Vektör regresyonu, Rastgele Orman regresyonu ve K-En Yakın Komşu regresyonu uygulanarak zaman serisi tahmini için en iyi öngörü elde eden algoritmayı belirlemektir.

2. **KAVRAMLAR**

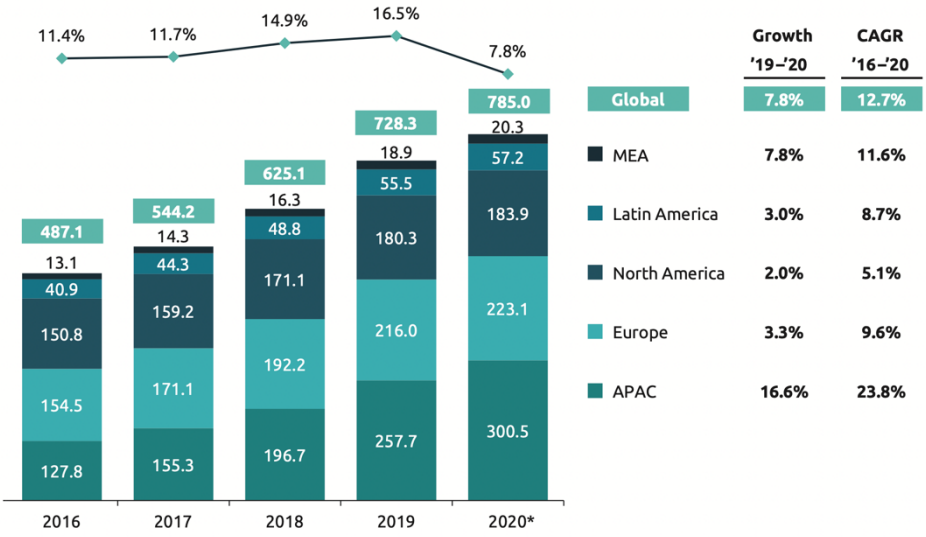
2.1. Para

Para, günümüzde insanların temel ihtiyaçlarını dahi karşılamak için mecburen sahip olunması gerekilen sosyal araçtır. Küreselleşen dünyaya bakıldığında tam anlamıyla para ile ikame ettiği görülmektedir. Para bir değişim aracı olması ile beraber aynı zaman da değer sahibidir. Tarih incelendiğinde ilk takas aracı olarak kıymetli madenler olan altın, gümüş kullanıldığı görülmektedir.

Altın para, M.Ö. 7. yüzyılda Lidyalılar tarafından kullanılmıştır (Yıldırım, 2015). Lidya kralı olan II. Alyottes altın madeni paralar bastırarak, günümüzde paranın mübadele, transfer aracı ve değer muhafaza olma özelliklerine sahip olmasını sağlamıştır. Zamanla madenin kıt olması, üretiminin maliyetli olması ve madenin aşınıp değer kaybetmesinden dolayı başka temsili paralar kullanılmıştır (Alpago, 2018). Temsili para olarak kağıt para seçilmektedir. Bu kağıtların özellikleri taklit edilmesinin zor olmasıdır. Değeri yüksek olan altın, gümüş gibi madenlerin taşınması, çalınma riskini arttırması nedeni ile insanlar sarraflara paralarını emanet etmeye başlayarak karşılığında emanetin değeri yazılı kağıtlar almışlardır. Ticaretin ilerlemesi ile emanet karşılığında alınan kağıtların yerini makbuzlar almış oldu. Böylece zamanla altın ve gümüşlerin yerini makbuz alarak temsili paraya dönüşmüş oldu. Ancak bu süreçte kıymetli madenlerin emanet edildiği sahafların altınları başkalarına vermesi veya makbuzların karşılığının çıkmamasından dolayı devlet bu duruma çözüm getirerek altın karşılığı olan kağıtların kendisini basarak bu paraları 1964 yılına kadar ticarette kullandırılmıştır (İnci ve Alpen, 2018).

Birçok klasik iktisatçı ve "Ekonominin Babası" olarak anılan Adam Smith parayı "ulusların zenginliği" olarak ifade etmektedir. Para ile dünya üzerinde çalışan var olan her mesleğin birbirine dönüşmesini ifade etmektedir. Daha geniş bir anlam ile açıklanacak olursa "çiftçinin işini berberin işine çevirme"nin yolu olarak ifade edilebilmektedir. Bireylerin paraya sahip olmak istemelerinin sebebi, değerli başka bir kıymet ile değiştirme imkânı sağlıyor olmasıdır (Mcleay ve ark., 2014).

Zamanla internetin varoluşu ve bunun sonuncunda geliştirilen mobil bankacılık, e-ticaret ve sanal işlemler parayı elde tutmaya ihtiyaç olmadan değişim yapmayı mümkün kılmaktadır. Bu sanal gelişimin göstergesi olarak dünya çapında yıllara göre nakit dışı para akışı Şekil 1'de yer almaktadır.



*2020 yılı tahmin değeridir.

Şekil 1: Dünya Çapında Nakit Para Akışı (World Payments Report, 2021)

Dünya çapında yıllara göre nakit dışı para transferinin yer aldığı Şekil 1 incelendiğinde yıllar geçtikçe nakit dışı para akışının arttığı görülmektedir. Bireyler çalışmaktan, günlük işlerden, ve e-ticarete özel indirimlerden dolayı internet üzerinden alışverişi tercih etmektedir.

2.1.1. İtibari Para

Günümüzde tüm dünya ülkelerinin sahip olduğu ekonomilerde tamamıyla kağıt paralar kullanılmaktadır. Hükümet basmayı planladığı parayı ödeme aracı olarak ilan etmesi gerekmektedir. İtibari paralar değişim aracı olarak kullanılmasına rağmen gerçekte hakiki değerleri bulunmamaktadır (Ünsal, 2017). Değeri olmayan bu kağıt paraların kullanılmasının tek nedeni devletin bu kağıtları yasal bir ödeme aracı olarak ilan etmesinden kaynaklanmaktadır.

Merkez bankası para basmadan önce ekonomik amaçlara erişmek için yapmış olduğu araştırma geliştirme çalışmaları neticesinde hedeflediği amaç ve tahminlere binaen para basımı gerçekleştirmektedir.

2.1.2. Elektronik Para

Günümüz dünyasında dijital teknolojinin gelişmesi ile tüketici davranışları büyük oranda değişiklikler sergilemektedir. Yaşamış olduğumuz zaman dahilinde dijital teknolojinin tüketicilere kolaylık sağladığı çok açık şekilde görülmektedir.

Elektronik para, bir değişim yaparken işletmeye elektronik ortamda ödenmesi sağlanan paradır. Daha açık şekilde ifade etmek gerekirse itibari paranın dijital ağa aktarılarak nakit para yerine kullanılmasıdır.

Elektronik para, transferi hızlı olan, elektronik ortamda saklanabilen alışverişi sağlayan paralardır. Bireylerin banka hesaplarında bulundurdukları elektronik paralar itibari paranın temsilidir. Kişilerin bankalara istedikleri yerde ulaşabilmesi aynı zamanda tercih olarak elektronik paraların kullanılması elektronik paranın itibari para ile arasındaki farkı azaltmaktadır (İnci ve Alpen, 2018).

2.1.3. Dijital Para

Dijital para, internet aracılığıyla kişinin mal veya hizmet karşılığında ödeme yapmak için bir şirketten aldığı para birimidir. Dijital para diğer adı ile sanal para olarak ifade edilebilmektedir. Bunun anlamı kağıt para, altın, gümüş gibi elle somut şekilde tutulamamaktadır. Bu para türü özellikle sanal ortamlarda, çeşitli ürünleri satın almak, değiş tokuş aracı olarak genellikle video oyunlarında kullanılmaktadır. Gerçek hayatta bir ürün almak için kullanılmamaktadır.

2.1.4. Kripto Para

Kripto para kelimesi, sanal olarak üretilmiş bir para biriminin kriptografi ile bu değeri güvence altına alındığını ifade eden gerçeklikten türetilmektedir (Corsello, 2018). Kripro para sanal para olarak adlandırılmaktadır.

Sanal para, sisteme geliştiriciler tarafından enjekte edilen para birimi olmasıyla beraber, sanal topluluğu oluşturan bireyler tarafından kullanılan ve kabul gören bir değerin dijital karşılığı olarak adlandırılmaktadır (European Central Bank, 2012). Sanal para ile dijital para aynı düşünülebilmektedir fakat farklı paralardır. Dünya genelinde, toplam para biriminin %95'inden daha fazlasını dijital para oluşturmaktadır. Sanal para gerçek dünyada somut olarak var olmayan ancak oyun siteleri gibi alanlarda değiş tokuş aracı olarak kullanılan para birimiyken gelişen dünyada sanal para birimi gerçek dünyaya taşmaya başladı. Bu sebeple sanal para ile dijital para arasındaki mesafe azalmış oldu (Partners, 2015). Günümüzde son derece popüler bir konu olması ile beraber, itibari para gibi devletin kamuoyuna ilanı ile bildirmesiyle değil; devlet dışındaki kişiler, firmalar tarafından üretilmektedir. Kripto para ile internet üzerinden takas yapılabilmekte ve ürün, hizmet alımı sağlanabilmektedir. Kripto paralar arz-talep içinde fiyatlandırılmaktadır.

Sanal para olarak en popüler Bitcoin olmaktadır. Bitcoin, ilk sanal para birimi hem de yüksek işlem hacmine sahip olması sebebiyle örnek olarak seçilmiştir.

Diffie ve Hellman (1976), ortak ve özel anahtarların kullanıldığı makale yayınlayarak kriptografik anahtar değişiminde kullanılan algoritma geliştirmişlerdir. Bu anahtar değişimi ile güven duygusu olmayan ortamda, ortak gizli anahtar üretilmesine imkan sağlamaktadır (Diffie ve Hellman, 1976).

Kripto paranın ilk çıkışı 1983 yılında Amerikan kriptograf David Chaum tarafından ortaya atılmaktadır (Chaum, 1982). Chaum tarafından geliştirilen Chaumian Blinding adını verdiği ilkel şifreleme algoritması ortaya çıkmaktadır. Kör imza adını aldığı algoritma işleyişi, mesajı imzalayan kişiye gönderici mesajı teslim etmektedir. Bu algoritma yüksek performans sergileyerek gizlilik sağlamıştır (Şeker, 2009).

Küresel krizin yaşandığı 2008 yılından sonra Satoshi Nakamoto adı altında Bitcoin ifadesini "Kullanıcıdan kullanıcıya elektronik nakit sistemi" adlı makalede yayınladılar. Bu makale sonrasında 2009 yılında ilk Bitcoin üretilmiş oldu. Bitcoin, 7/24 internet üzerinden işlemi yapılabilmektedir. Bitcoin'in diğer özelliklerinden biri hızlı transferi sağlanabilmektedir. Bitcoin, zaman geçtikçe daha çok dikkat çekmeye

başlamış ve birçok insan tarafından takip edilmekte olsa da gerçekte takibi zor olan kripto para birimidir. Bitcoin'e olan talep arttıkça değeri de katlanarak artmaya devam etmektedir. Bu sebepten ötürü birçok yatırımcının ilgisini çekmektedir (Halaburda ve Sarvay, 2016).

Bitcoin, eşler arası işlem için aracılık faaliyeti görmektedir. Bu faaliyeti sürdürürken herhangi bir birimden yardım almamaktır. Bu sebeple Bitcoin kuru piyasada talep ve arz miktarına belirlenmektedir. Sanal paranın çıkışı ile beraber bu parayı itibari para ile takas edecek platformlar kuruldu. Bitcoin'i talep eden kullanıcılar Bitcoin kazanmak için çeşitli mal ve hizmet satarak buna ulaşmaktadır. Açık ve özel anahtarları içinde barındıran bir dosyaya kullanıcılar sahip olmaktadır. Bu dosyanın herhangi bir yol ile bilgisayardan silinmesi, yok olması durumunda Bitcoin'inde kaybolmasını ifade etmektedir. Bitcoin transferi hızlı olan bir para birimidir, transfer işlemini yapabilmek için göndericiye ait açık anahtarın bilinmesi gerekmektedir. Bitcoin'i transfer edecek birey daha önce yapılmış işlemlerin blok zincirdeki kodu ve alıcı kişinin açık anahtarını imzalamak zorundadır. Transfer bu şekilde gerçekleşmiş olacaktır. Her transferde bu tekrarlanır ve bloğa kaydedilir (Yüksel, 2015).

Bitcoin'inin ilk çıkışından bu yana fiyatında ve işlem hacmi artışında önemli rol oynayan etkenler bulunmaktadır.



Şekil 2: Bitcoin Fiyatı (Blockchain, 2022)

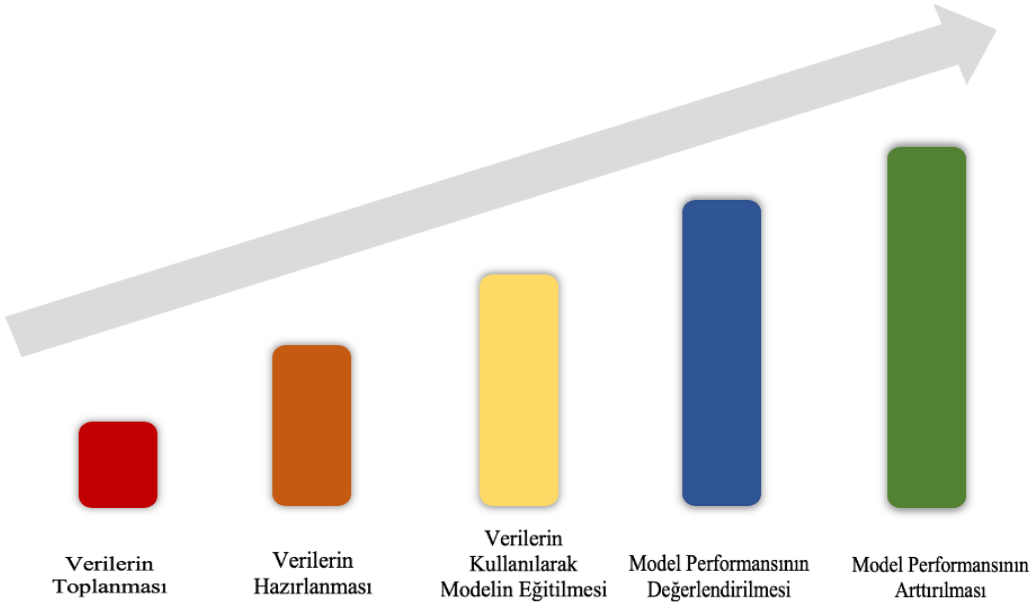
Şekil 2'de yer alan Bitcoin'e ait fiyat grafiği incelendiğinde 2016 Kasım'dan bu yana sürekli artış gözükse de ciddi iniş çıkış oynaklıkların olduğu aşikardır. Bitcoin'e olan talep arttıkça fiyatında artış yaşanmaktadır.

Küreselleşen dünyada teknolojinin hızla gelişmesi, itibari paranın varlığını sürdürmesini zorlaştırmaktadır. Bu durum ekonomik ilişkilerin daha hızlı ve kolay olmasını sağlayacak arayışlar meydana getirmektedir. Mevcut arayışın neticesinde gerçekleştirilmesi uygun görülen değişiklikler devletin otorite kararlarını belirleyen kurumların yeni bir sürece girmesine neden olacaktır (Atış, 2014).

2.2. Makine Öğrenmesi

Makine öğrenmesi son zamanlarda yaygın olarak kullanılan istatistiksel ve bilgisayar biliminden faydalanan yapay zeka algoritmasıdır. Bu algoritmalar mevcut verileri kullanarak kümeleme, sınıflandırma ve regresyon tahmini yapabilmektedir (Baran, 2020). Daha genel tanımıyla makine öğrenmesi geçmiş deneyimlerden faydalanarak elde edilen öğrenmelerden model oluşturarak geleceğe yönelik tahmin yapmaktadır. Bilgisayar mevcut geçmiş verilerden yola çıkarak bir model oluşturur ve bu modeli yeni verilerde kullanarak karar vermektedir (Öğücü, 2006).

Yapay zekanın alt birimi olan makine öğrenmesi, öğrenme algoritmaları ile bu algoritmaların performansını ölçmektedir. Bilgisayarlar algoritmalarının insanların düşünme şeklinden yola çıkarak onlar gibi öğrenmesini sağlayan bilimsel çalışma alanını oluşturmaktadır (Alpaydın, 2020).



Şekil 3: Makine Öğrenmesi Adımları

Makine öğrenmesi hakkında ilk çalışmayı ortaya koyan Alan Turing, 1950 yılında "Computing Machinery and Intelligence" adlı makalesinde günümüzde Turing testi olarak bilinen kavram ile ifade etmektedir. 1952 yılında, yapay zekanın öncüsü Arthur Samuel bir bilgisayar oyunu yazarak bu bilgisayar oyunundaki tüm hamleleri, stratejileri kendi programına dahil etmeyi başarılı şekilde gerçekleştirmiştir. 1957'de Frank Rosenblatt, insan beyninin düşünce yapısından etkilenerek bu düşünce sürecini simüle eden yapay sinir ağlarını ortaya koymuştur (Keskenler ve Keskenler, 2017).

Büyük gözlemlerin oluşturduğu veri seti içerisinde fark edilmesi zor olan yapıları makine öğrenmesi algoritmaları ile ortaya çıkarılarak problem çözülebilmektedir (Alpaydın, 2020).

Makine öğrenmesi verilerin farklılıkları, problemlerin çeşitliliklerinden dolayı farklı yöntemleri ortaya çıkarmak durumunda kalmıştır. Bu yöntemler;

- Denetimli makine öğrenmesi
- Denetimsiz makine öğrenmesi
- Yarı Denetimli makine öğrenmesi

- Takviyeli makine öğrenmesi

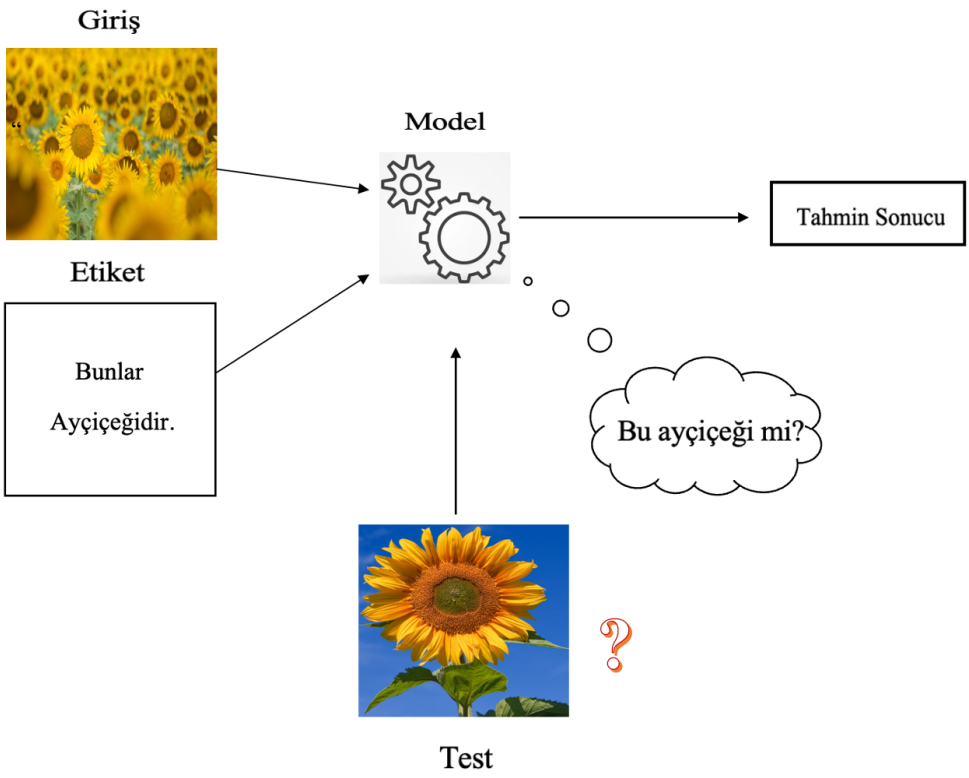


Şekil 4: Yaygın Kullanılan Makine Öğrenmesi Modelleri

2.2.1. Denetimli Makine Öğrenmesi

Denetimli makine öğrenmesi, algoritma ile analizi yapılmak istenen verilerin, eğitim ve test verisi olarak ayrılarak elde edilen çıktının arasındaki bağlantıyı kurmaya ve bir sonuca varmaya dayanmaktadır.

Denetimli makine öğrenmesi elde edilen eğitim verilerini analiz ederek gözlemleri kullanabilmek için haritalama yapmaktadır. Denetimli makine öğrenmesinde iki aşama mevcuttur. Bu aşamaları eğitim ve test verilerinin oluşturulmasını kapsamaktadır. Eğitim aşamasında kullanılacak olan gözlemler algoritmada incelenip bir tahmin oluşturmaktadır. Eğitim aşamasında yapılan tahminlerin doğruluğunu test etmek için verilerin belli bir kısmı test aşaması için ayrılmaktadır (Baran, 2020).



Şekil 5: Denetimli Öğrenme Modeli

Denetimli öğrenmede eğitim verileri kullanılarak oluşturulan model tüm veriler hakkında bilgi verebilmektedir. Algoritmanın öğrenmesi için konuya dair örnek sayısı ne kadar fazla olursa sistemin tahmin performansı da o kadar yüksek olacaktır.

2.2.1.1. Regresyon Analizi

Regresyon analizi, bir ve daha fazla bağımsız değişken ile bağımlı değişken arasındaki ilişkiyi matematiksel olarak ifade edilmesidir. Örnek olarak, bağımlı değişkeni iktisadi büyüme, bağımsız değişkeni ise işsizlik olan bir regresyon analizinde geçmiş verilerden yararlanılarak işsizliğin iktisadi büyümeyi ne oranda etkilediği tahmini yapılabilmektedir.

$$Y_i = \beta_0 + \beta_1 X_1 + \in \qquad (1)$$

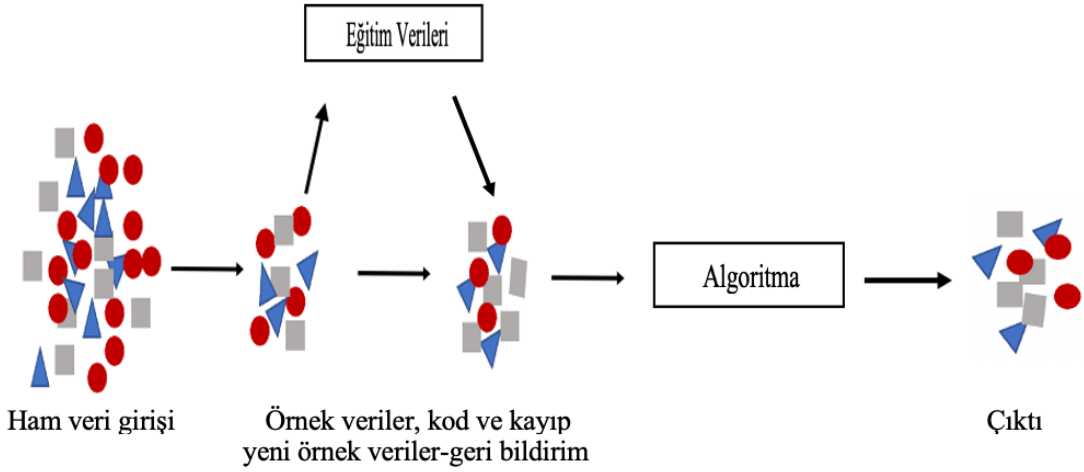
Eşitlik 1'de yer alan denklemde Y_i bağımlı değişkeni, X_1 bağımsız değişkeni ifade etmektedir. β_0 ve β_1 regresyon katsayılarını göstermektedir. β_0 sabit katsayı terimidir. $\in$ değeri ise hatayı ifade etmektedir.

2.2.1.2. Sınıflandırma

Sınıflandırma, makine öğrenmesi ve istatiksel analizlerinde problemin çözümüne uygun olduğu takdirde seçilmektedir. Sınıflandırma, yeni gözlem dahil olduğunda hangi kategoriye ait olacağını, mevcut veri ve kategorilerden yola çıkarak karar vermektedir. Sınıflandırmaya örnek olarak, giriş yapılmak istenen uygulamaya "robot", "robot değil" diye sınıflandırılması verilebilmektedir.

2.2.2.Yarı Denetimli Makine Öğrenmesi

Yarı denetimli makine öğrenmesi veri setinde az sayıda işaretlenmiş ve çok sayıda işaretlenmemiş veri olduğu zaman kullanımı tercih edilmektedir. Bu yöntem ile az sayıdaki işaretlenmiş verilerden çok sayıdaki işaretlenmemiş veriyi tahmin ve sınıflandırma yapmaktadır. Yarı denetimli makine öğrenmesi ne kadar deneme yapabildiğini ve hatayı öğretmektedir. Birçok kez deneme yapmasının sebebi geçmiş deneyimlerden öğrenek en iyi performansı elde etmesidir (Kızılkaya ve Oğuzlar, 2018).

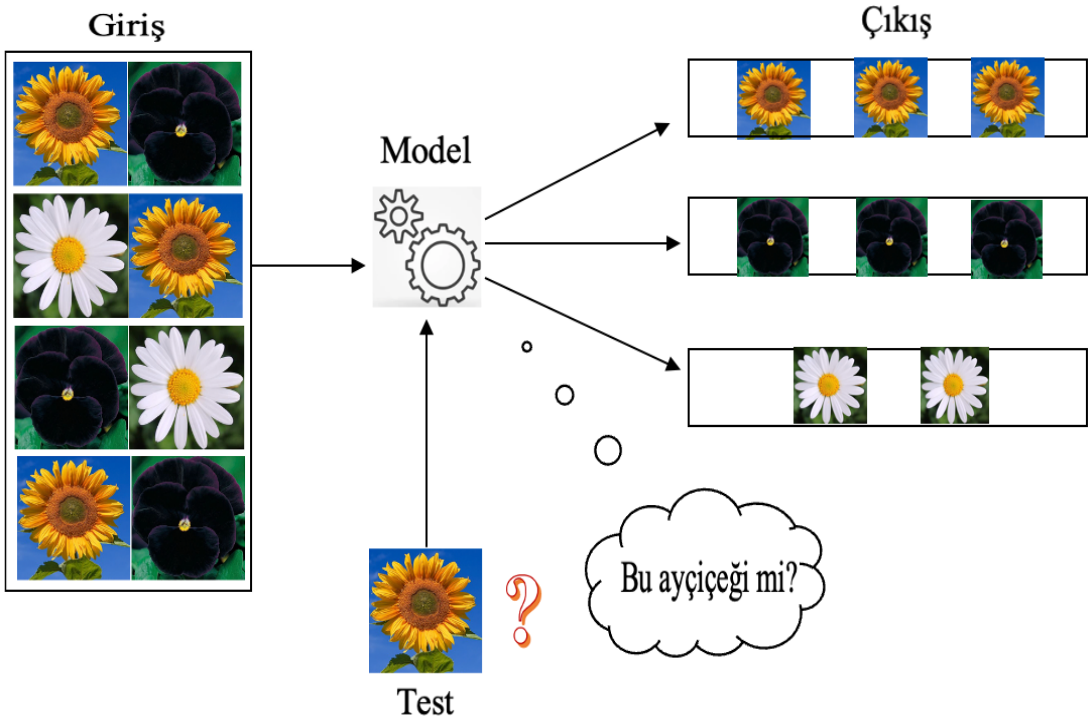


Şekil 6: Yarı Denetimli Öğrenme Modeli

2.2.3. Denetimsiz Makine Öğrenmesi

Denetimsiz makine öğrenmesi, sistemin eğitilmesi için etiketsiz veri kullanılarak öğrenmesini sağlanmaktadır. Bu yöntemin amacı gözlemlerin çıktıları bilinmediği için

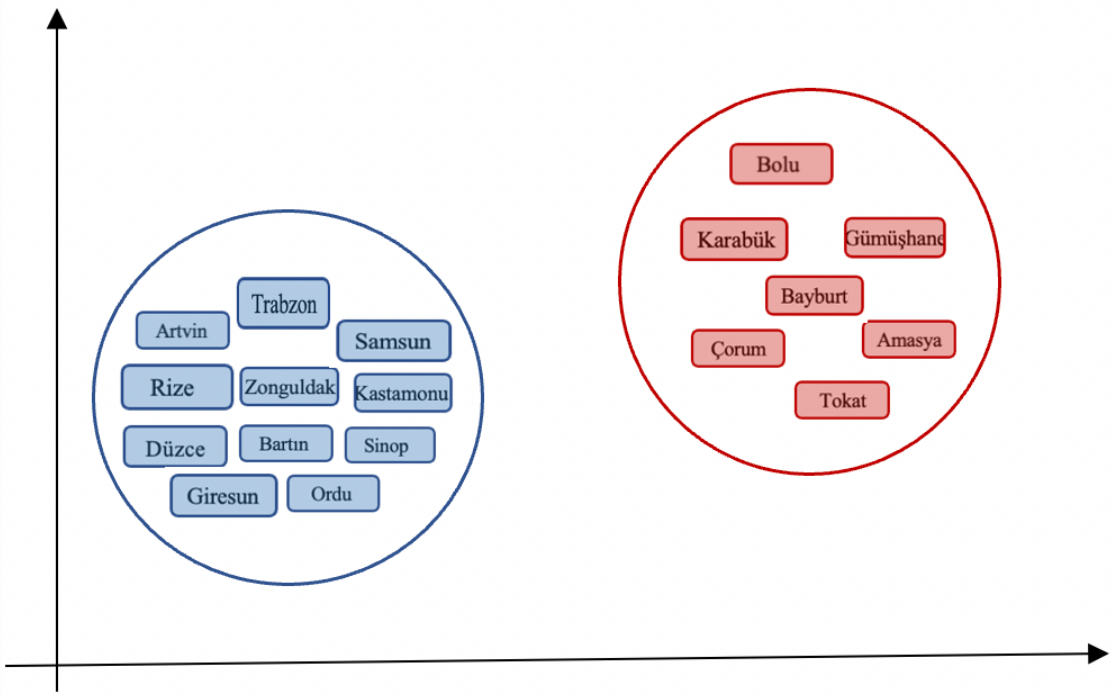
sınıflandırma yapmak değildir (Chao, 2011). Bu yöntemle veri setindeki gözlemler arasındaki benzerlikler bulunarak kümelere ayrılmaktadır (prowmes.com).



Şekil 7: Denetimsiz Öğrenme

Denetimsiz öğrenme modelinde girdi verileri bilinmektedir. Ancak sahip olduğu kategori bilinmemektedir bu sebeple büyük veri çalışmalarında denetimsiz öğrenme algoritmaları kullanılmaktadır.

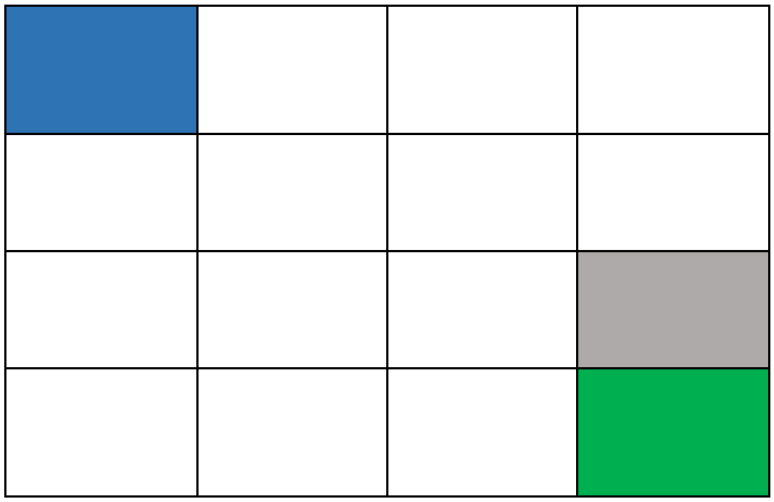
Kümeleme yöntemi, denetimsiz makine öğrenmesi analizlerinden biridir. Çok değişkenli analiz yöntemlerinden biri olan kümeleme analizi, veri setinde yer alan gözlemleri benzerleriyle aynı gruba taşıma işlemini gerçekleştirmektedir. Temel amaç, gözlemlerin sahip olduğu karakteristik özelliklere göre gruplamaktır. Şekil 5'de Kümeleme Analizi yer almaktadır. Karadeniz Bölgesinde yer alan şehirlerin denize kıyısı olup olmamasına göre gruplanmaktadır.



Şekil 8: Karadeniz Bölgesindeki İllerin Denize Kıyısı Olup Olmamasına Göre Kümelenmesi

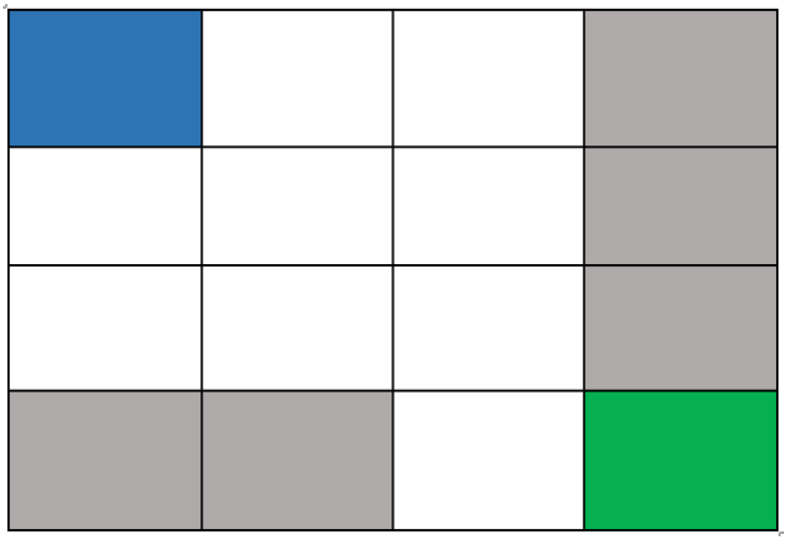
2.2.4. Takviyeli Makine Öğrenmesi

Takviyeli makine öğrenmesinde, öğrenme evresinin başında en uygun çözümün ne olduğu bilinmemektedir. Bundan dolayı çözümler yinelemeli olarak nasıl öğrenmesi gerektiğini göstermektedir. Öğrenme aşamasının sonunda çıktının iyi ve kötü olduğuna dair bilgi vermektedir. (Wenzel ve ark., 2019). Takviyeli makine öğrenmesi, bir aracın hedefine ulaşması için doğru yolu nasıl seçeceğini gösterir. Bu yöntem yaygın olarak oyun programlarında kullanılmaktadır. Örnek olarak bir robotun 4x4 matris üzerinde sol üst köşede mavi kutudan başlayarak sağ alt köşede bulunan yeşil kutuya ulaşılması istensin.



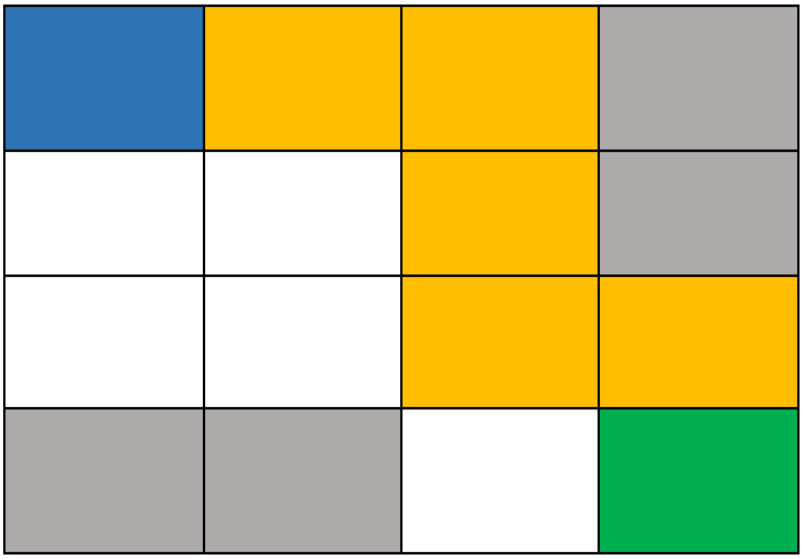
Şekil 9: Takviyeli Öğrenme İle Deneyerek Çıkışı Bulma-1

Şekil 9 incelendiğinde robot rastgele tahminlerle ilerleyerek en kısa yol ile çıkışa ulaşmayı hedeflemektedir. Bu sebeple matrisi keşfetmeye başlar.



Şekil 10: Takviyeli Öğrenme ile Deneyerek Çıkışı Bulma-2

Şekil 10'da robot matrisi keşfetmeye devam ederken her denemede bir hücreyi işaretlemektedir.



Şekil 11: Takviyeli Öğrenme İle Kısa Yolun Belirlenmesi

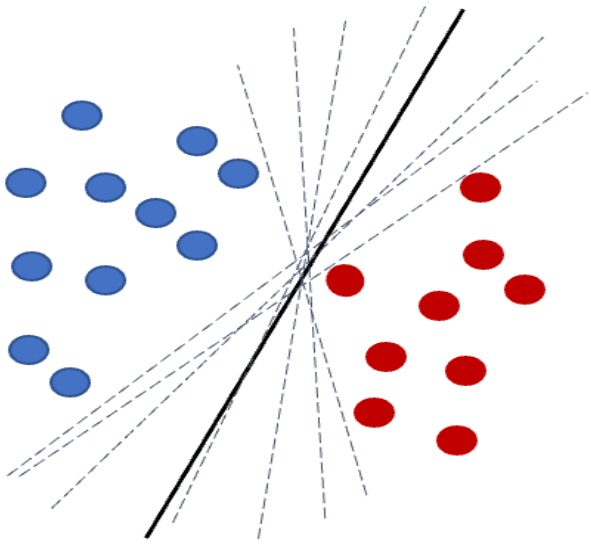
Tüm denemeler bittikten sonra en kısa yolu Şekil 10'da yer aldığı gibi sarı renkle belirlemektedir.

3. METOTLAR

3.1. Destek Vektör Makineleri

Destek Vektör Makineleri(DVM), denetimli makine öğrenmesi yöntemlerinden biri olarak Vladimir Vapnik ve arkadaşları tarafından geliştirilmiştir (Vapnik, 1999). DVM veriyi sadece sınıflandırmak için değil problemin çeşitliliğine göre tahmin yapmak için de kullanılmaktadır.

DVM temelde veriyi iki sınıfa doğrusal olarak ayırmak için kullanılan sınıflandırma algoritmasıdır. Problemlerin çeşitliliği ve büyüklüğü farklı oldukça ikiden daha fazla sınıfa ayırma ihtiyacı olmaktadır. Zamanla geliştirilerek veriyi ikiden fazla sınıfa ayırmayı başarmıştır (Wang ve Xue, 2013). DVM çalışma disiplini doğrusal olarak ayrılabilen veriler arasında en ideal ayırma hiper düzlemini oluşturmaktadır. DVM'de amaç iki sınıfı birbirinden ayırırken en uygun fonksiyonu bulmaktır. Şekil 12'de görüldüğü üzere iki sınıfın ayrımında sonsuz tane doğru çizilebilmektedir. Fakat DVM'de amaç veriyi en iyi ayıran hiper düzlemi bulmaktır. Bu iki sınıfı ayıran en uygun hiper düzleme marjin adı verilmektedir.



Şekil 12: En İyi Ayırıcı Düzlem

DVM'de marjini maksimize etmesinin sebebi hatayı minimize etmeye çalışmasıdır. Hatayı en aza düşürmeye çalışmasından dolayı uygulamalarda tercih edilmesine sebep olmaktadır. DVM hem doğrusal olarak ayrılabilen hem de doğrusal olarak ayrılamayan veri setleri için kullanılabilmektedir.

3.1.1. Doğrusal Olarak Ayrılabilen Veriler

Doğrusal olarak ayrılabilen DVM'de amaç iki sınıf veri arasında maksimum marjini bulmaktır.

Z veri kümesinin i elemanlı veri seti olduğunu varsayalım.

$$Z = \{x_i, y_i\}, \quad x_i \in R_n, \qquad y_i \in \{-1, +1\}$$

Burada x_i girdi vektörünü, y_i örneklerin hangi sınıfa düştüğünü temsil etmektedir. Doğrusal olarak ayrılabilme durumunda sonsuz tane hiper düzlem ile ayrılabilme söz konusudur. DVM hiper düzlemin iki gruba da eşit olmasını sağlayarak en uygun düzlemi seçmesini sağlamaktadır.

En uygun hiper düzlem,

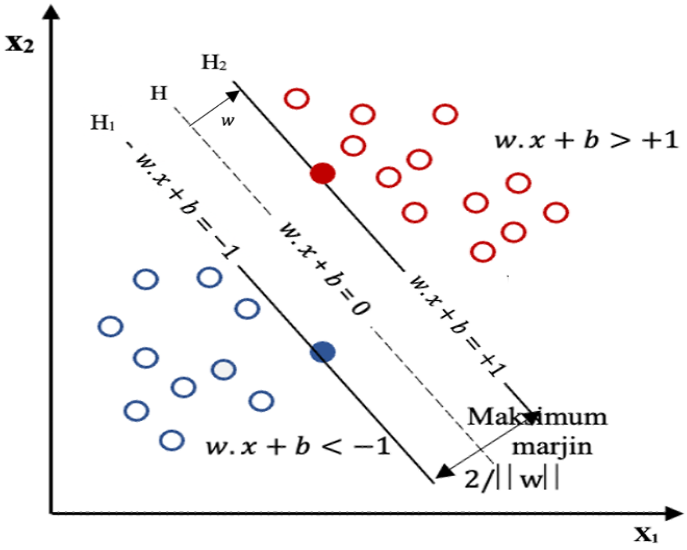
$$w.\,x + b = 0 \tag{2}$$

ifade edilmektedir. Denklemde w, ağırlık vektörünü, b sabit olarak tanımlanmaktadır. En uygun hiper düzlemin bulunabilmesi için veri setindeki örneklerin düştüğü sınıflar,

$$w.xi+b \geq +1 \quad \forall\, yi = +1 \tag{3}$$

$$w.xi+b \leq -1 \quad \forall\, yi = -1 \tag{4}$$

şeklinde ifade edilmektedir. 3'de yer alan eşitlik hiper düzlem üzerinde yer alan örneklerin +1 sınıfına düştüğünü, 4'teki denklem ise hiper düzlemin altına düşen örneklerin -1 sınıfına dahil olduğunu ifade etmektedir.



Şekil 13: En İyi Ayırıcı Hiper Düzlem Ve Destek Vektörler

Şekil 13 incelendiğinde -1 sınıfına ait olan verilerin $w.x + b < -1$ denklemi ile ifade edildiğini +1 sınıfına düşen verilerin ise $w.x + b < +1$ denklemi ile ifade edildiği görülmektedir. H1 ve H2 hiper düzlemi üzerine düşen veriler destek vektör makinelerini ifade etmektedir ve bu hiper düzlemler arasındaki mesafe $2/\|w\|$ kadardır. En uygun hiper düzlemi bulmak için $2/\|w\|$ ifadesinin maksimum olması gerekmektedir. Bu durumda $\|w\|$ ifadesinin en küçük olması gerekmektedir.

$$min\left(\frac{1}{2}||w||^2\right), \quad yi(w.xi + b) - 1 \qquad \forall \tag{5}$$

5'de ki eşitlikte optimizasyon sorununun ana problemi yer almaktadır. Problemi çözmek için Lagrange çarpanları kullanılmaktadır. Lagrange fonksiyonu,

$$L(w, b, \alpha) = \frac{1}{2}\left||w|\right|^2 - \sum_{i=1}^{N} \alpha_i \, y_i(x_i.w + b) + \sum_{i=1}^{N} \alpha_i \tag{6}$$

6'daki eşitlikte verilmektedir. Burada $L(w, b, \alpha)$ Lagrange fonksiyonunu, α_i Lagrange çarpanlarını ifade etmektedir. Denklemde $\alpha_i \geq 0 \quad \forall$ koşulunu sağladığı varsayılmaktadır. Denklemi minimize etmek için w ve b'ye göre türev alınmaktadır.

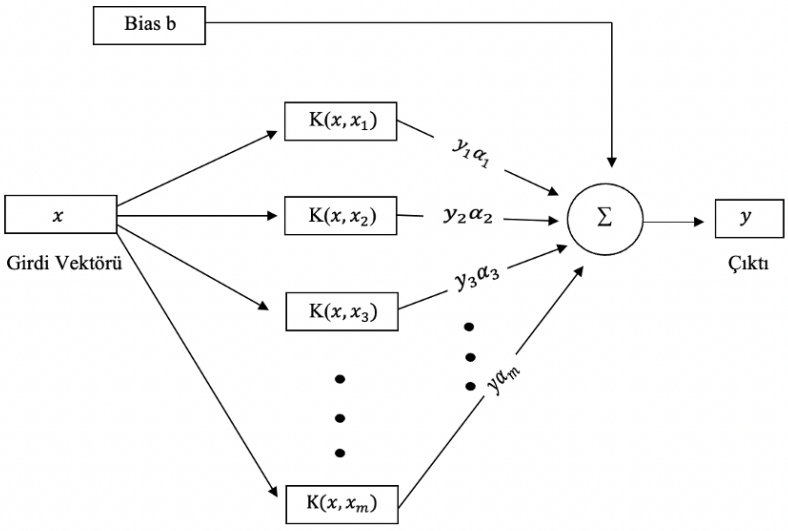
$$\frac{\partial L}{\partial w} = 0 \quad \rightarrow w = \sum_{i=1}^{N} \alpha_i \, y_i x_i \tag{7}$$

$$\frac{\partial L}{\partial b} = 0 \quad \rightarrow \quad b = \sum_{i=1}^{N} \alpha_i \, y_i \tag{8}$$

Türevden elde edilen 7 ve 8'deki denklemleri 1.8'de yer alan Lagrange fonksiyonunda yerine yazıldığında 9'daki denklem elde edilmektedir.

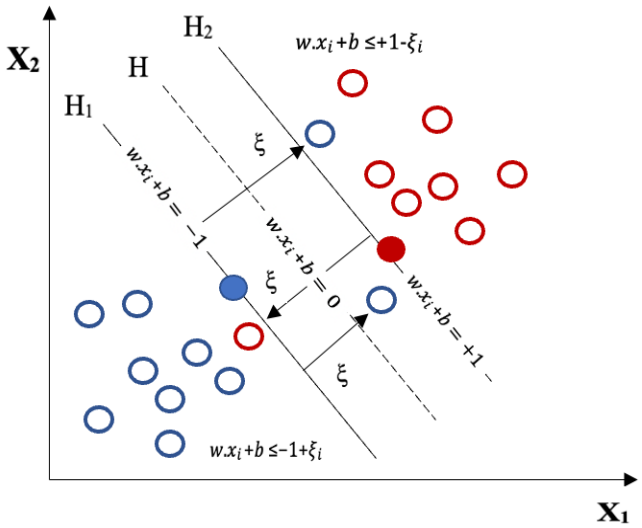
$$L_D(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=0}^{N} \alpha_i \alpha_j \, y_i \, y_j x_i . x_j \qquad \alpha_i \geq 0 \;\; \forall_i , \quad \sum_{i=1}^{N} \alpha_i \, y_i = 0 \tag{9}$$

Eşitlik 9'da dual form yer almaktadır. Problemin ana formunu minimize etmek eşitlikte yer alan dual formu maksimize etmektir. Şekil 14'de DVM'ye ait ağ yapısı yer almaktadır.



Şekil 14: Destek Vektör Ağ Yapısı

Problemlerin çeşitliliği ile beraber doğrusal sınıflandırma tamamen mümkün olmamaktadır. Uygun hiper düzlem belirlendiğinde diğer sınıfa ait veri yanlış sınıfta yer alabilmekte veya marjin aralığına düşebilmektedir. Burada meydana gelen hata oranını azaltmak için esneklik değişkeni (ξ) kullanılmaktadır. Şekil 15'da doğrusal olarak belirli hata ile ayrılabilen DVM yer almaktadır.



Şekil 15: Doğrusal Olarak Belirli Bir Hata İle Ayrılabilen DVM

Şekil 15 incelendiğinde hatalı sınıflandırmalar olduğu görülmektedir. Bu yanlış sınıflandırmanın çözümü için esneklik değişkeni (ξ) kullanılmaktadır. Esneklik değişkeni (ξ) de görüldüğü üzere marjinden sapmayı ifade etmektedir. Esneklik değişkeni, $\xi = 0$ ise veri setinin doğrusal olarak ayırabilen bir hiper düzlemin olduğunu ifade etmektedir. $0 < \xi < 1$ olduğu taktirde verinin marjin aralığına düştüğünü ifade etmektedir. $\xi < 1$ ise verinin yanlış sınıfta yer aldığını ifade etmektedir.

Esneklik değişkeni (ξ) fonksiyona eklediğimizde,

$$w.xi + b \geq +1 - \xi i \quad \forall \; yi = +1 \quad \xi i \geq 0 \tag{10}$$

$$w.xi + b \leq -1 + \xi i \quad \forall \; yi = -1 \quad \xi i \geq 0 \tag{11}$$

denklemleri elde edilmektedir. Tek bir denklemde toplanarak 12'deki eşitlik elde edilmektedir.

$$y_i(w.x_i + b) - 1 + \xi_i \geq 0 \qquad (12)$$

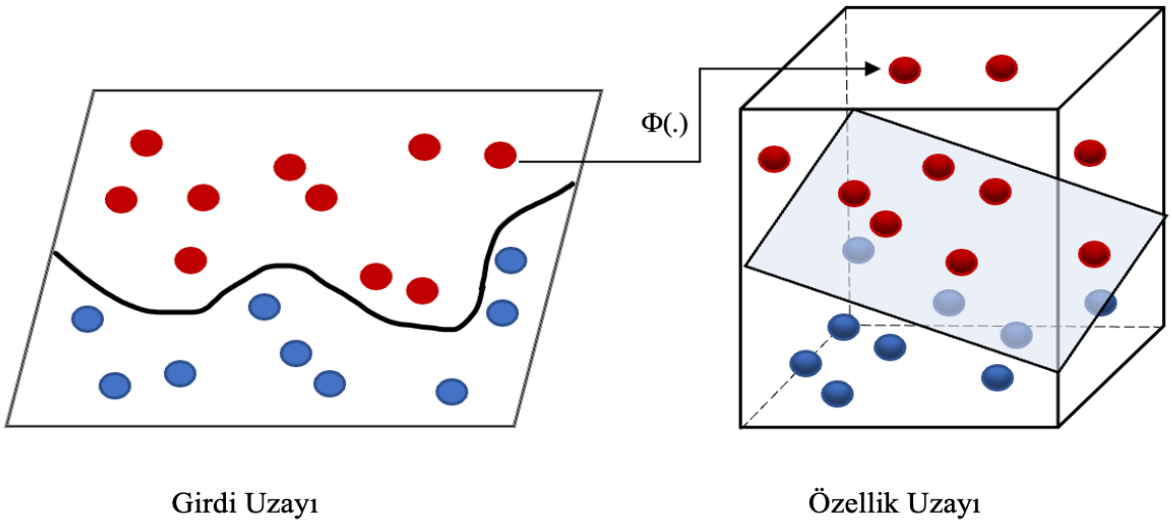
Yanlış sınıflandırma hatasının minimize etmek için 5'te yer alan denkleme esneklik değişkeni eklenirse;

$$min\left(\frac{1}{2}||w||^2 + C\sum_{i=1}^{L}\xi_i\right), \qquad y_i(w.x_i + b) - 1 + \xi_i \geq 0 \qquad (13)$$

Denklem 13'teki C ifadesi ceza parametresidir. Ceza parametresi marjini maksimize ederken, hatalı sınıflandırmayı minimize etmektedir. C parametresinin küçük olması hatanın küçük olduğunu ifade etmektedir (Burges, 1998).

3.1.2. Doğrusal Olarak Ayrılamayan Veriler

Problemlerin çeşitliliği ile beraber verilerin doğrusal olarak ayrılabilmesi pek mümkün olmamaktadır. Doğrusal olarak veriyi belli bir hataya kadar ayırmak mümkün fakat hatanın çok fazla olması uygun hiper düzlemin belirlenememesine yol açmaktadır. Bunun önüne geçmek için doğrusal olmayan DVM kullanılmaktadır.



Şekil 16: Doğrusal Olarak Ayrılamayan Veri Örneği

Şekil 16'da doğrusal olarak ayrılamayan veri setinin çekirdek fonksiyonu kullanılarak verilerin özellik uzayına taşındığını göstermektedir. Doğrusal olarak ayrılamayan DVM'ye ait denklem 14'de yer alan eşitlikte ifade edilmektedir.

$$w.\phi(x) + b = 0 \tag{14}$$

Denklem incelendiğinde doğrusal olarak ayrılabilen modellere ait fonksiyonda x yerine $\phi(x)$ yazılmaktadır.

Doğrusal olarak ayrılamayan veri setlerinde, doğrusal olarak sınıflandırılabilmeleri için daha yüksek boyutlu uzaya taşınmaktadır. Daha yüksek boyutlu uzaya taşınmak için çekirdek yaklaşımı kullanılmaktadır.

Çekirdek yaklaşımı regresyon ve sınıflandırma problemlerinde kullanılmaktadır. DVM'de esneklik kazandırmasından dolayı performans açısından etkili sonuçlar vermektedir (Güldoğan, 2017).

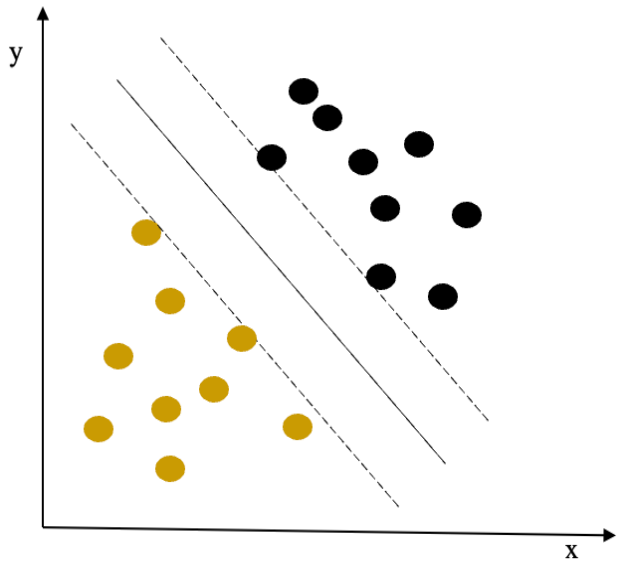
En çok kullanılan çekirdek fonksiyonları; aşağıda yer almaktadır.

- Doğrusal Çekirdek Fonksiyonu
- Polinom Çekirdek Fonksiyonu
- Radyal Tabanlı Çekirdek Fonksiyonu
- Sigmoid Çekirdek Fonksiyonu

Doğrusal çekirdek fonksiyonu, en basit çekirdek fonksiyonudur. Doğrusal çekirdek fonksiyonu 15'teki Eşitlikte yer almaktadır.

$$K(x, y) = x.y \tag{15}$$

Denklemde $x.y$ iç çarpımı ifade etmektedir.



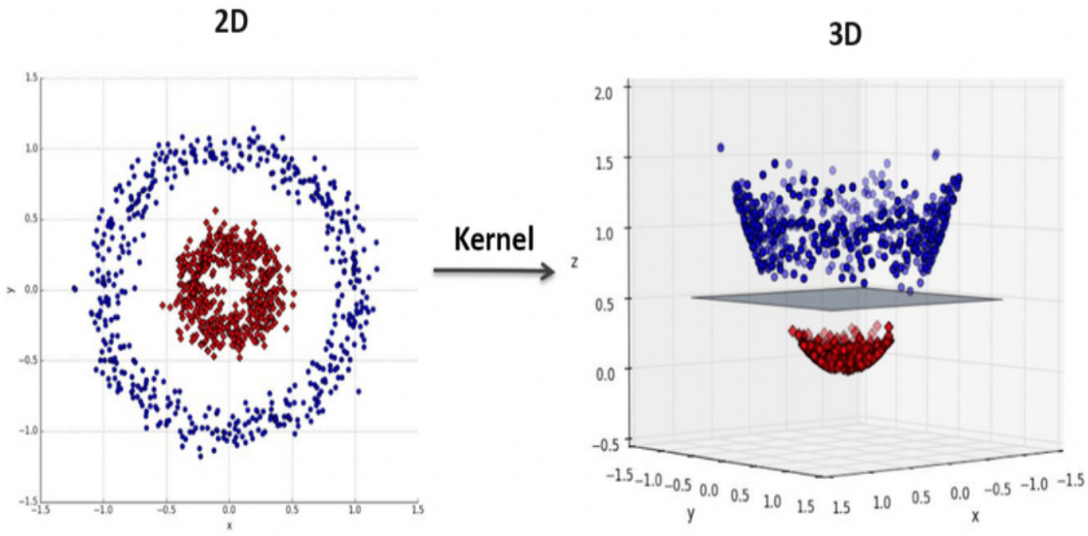
Şekil 17: Doğrusal Çekirdek

Doğrusal çekirdek kullanımı ile sınıflandırmayı gösteren Şekil 17 incelendiğinde verinin iki sınıfa doğrusal olarak ayrıldığı görülmektedir.

Polinom çekirdek fonksiyonu, durağan değildir. Bu çekirdek fonksiyonu genellikle eğitim verisini normalleştirdiği problemler için kullanılmaktadır.

$$K(x, y) = (1 + x.y)^d \tag{16}$$

Polinom çekirdek fonksiyonunda doğrusal çekirdek fonksiyonuna göre daha fazla parametre mevcuttur. 1.18'de yer alan denklemde 1 sabit terimi ifade etmektedir. d ise polinom derecesini göstermektedir.



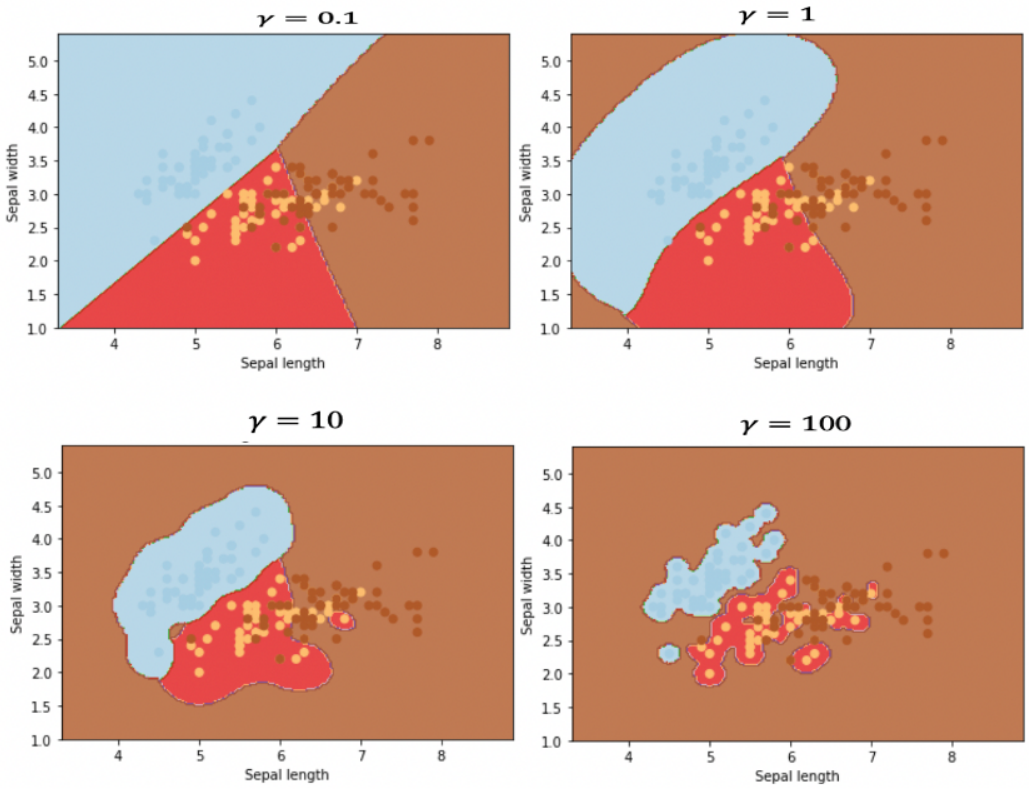
Şekil 18: Polinom Çekirdek Kullanımıyla 2 Boyutlu Uzaydan 3 Boyutlu Uzaya Taşınması (Akça, 2020)

2 boyutluda yer alan veriyi doğrusal olarak sınıflandırmamız mümkün olmamaktadır. Bu sebeple veriyi 3 boyutlu uzaya taşıyarak yani veriyi yukarıya kaldırarak sınıflandırma işlemi gerçekleştirilir. Bu işlem Şekil 18'da gösterilmektedir.

Radyal tabanlı çekirdek fonksiyonu, uygulamaya göre dikkatle seçilmelidir. Aşırı tahminin olması durumunda, üstel ifade doğrusallaştırılmalıdır. Böyle olduğu takdirde bir doğru gibi performans göstereceğinden doğrusal olmayan özelliğini kaybetmektir.

$$K(x,y) = (e)^{-\gamma||x-y||^2}, \gamma > 0 \qquad (17)$$

Fonksiyonda yer alan gama (γ) parametresi ayarlanabilmektedir. Bu parametrenin doğru değer alması tahmin algoritmasında önemli rol oynamaktadır. Aşırı tahmin söz konusu olduğunda polinom çekirdek fonksiyonunda olduğu gibi üstel ifade doğrusallaştırılarak, doğrusal olmayan yapısını kaybetmeye başlamaktadır.



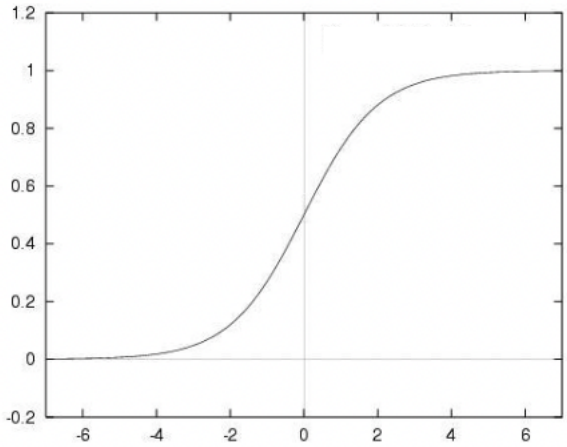
Şekil 19: Radyal Tabanlı Çekirdeğin γ Değerine Göre Sınıflandırılması (Akça, 2020)

Radyal tabanlı çekirdek, sonsuz boyutta yer alan DVM'lerini bulur ve her noktaya bakarak birbirilerine ne kadar çok benzediğini bulmak için normal dağılımla hesaplayarak bu hesaba göre sınıflandırma yapmaktadır. γ değeri ne kadar küçük olursa dağılım o kadar büyük olmaktadır. Eğer modelde aşırı tahmin söz konusu ise γ değerini düşürülürken, yetersiz tahmin olması durumunda γ değerini yükseltilmesi gerekmektedir (Akça, 2020).

Sigmoid çekirdek fonksiyonu birçok çalışmada iyi performans sergilediği ifade edilmektedir.

$$K(x, y) = \tan(\alpha x. y + c) \tag{18}$$

Sigmoid çekirdek fonksiyonunda yer alan α eğim, c kesme sabit terimini ifade etmektedir. α için $1/N$ ortak değeri kullanılabilmektedir. Burada N ifadesi veri boyutudur.



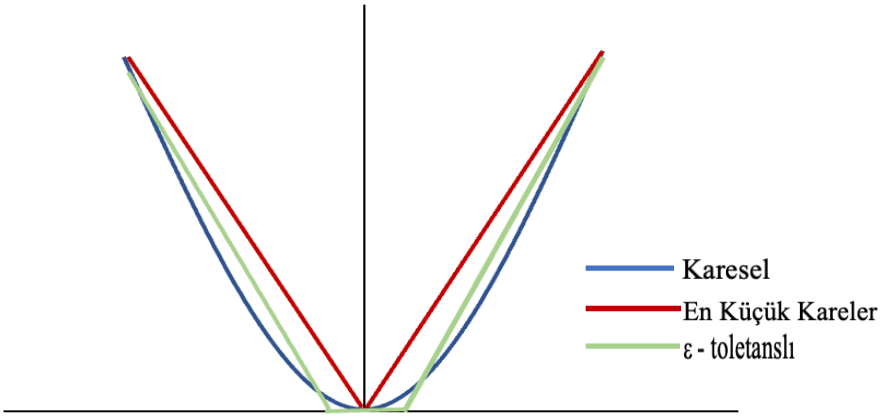
Şekil 20: Sigmoid Fonksiyonu

Sigmoid çekirdeğe ait grafiğin yer aldığı Şekil 20 incelendiğinde bu fonksiyonun her zaman 0 ile 1 aralığında değer aldığını yani aktivasyon değerinin aşırı değer almadığını göstermektedir. Bu özelliği iyi sınıflandırıcı olarak kullanılabildiğini ifade etmektedir.

3.1.3. Destek Vektör Doğrusal Regresyonu

Belirtildiği üzere DVM problemlerin çeşitliliği üzerine hem sınıflandırma hem de regresyon tahmini için kullanılabilmektedir. Regresyon tahmininde girdi (x) ve çıktı (y) verilerimiz bulunmaktadır. Regresyon tahmini ile girdinin çıktıyı ne kadar etkilediği tahmin edilmektedir.

Destek vektör regresyonu (DVR), DVM'nin temel prensiplerine benzer şekilde çalışmaktadır. DVM'ye kayıp fonksiyonun eklenmesiyle regresyon problemlerinde kullanılabilmektedir.



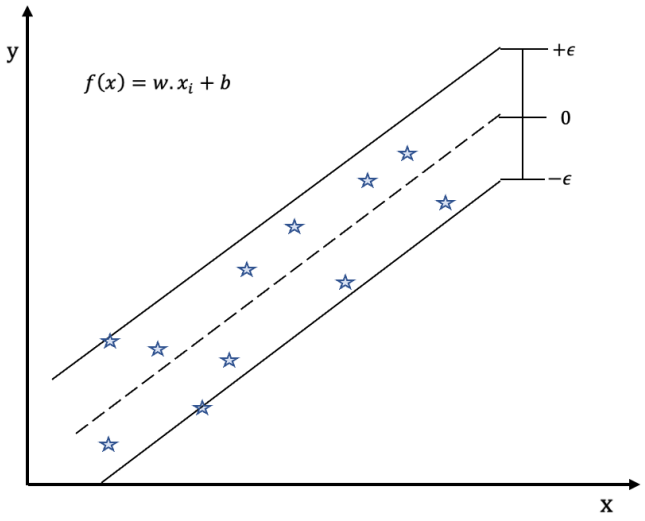
Şekil 21: Kayıp Fonksiyonu

DVR'de tercih edilen kayıp fonksiyon modeli ε – toleranslı model olmaktadır. Amaç ε değerinden daha küçük olan sapmaya sahip eğitim verilerinden $f(x)$ fonksiyonu oluşturmaktadır. ε'den daha küçük değerler dikkate alınmamaktadır. f doğrusal fonksiyonu 19 'daki Eşitlikte ifade edilmektedir.

$$(y_1, x_1) \dots, \dots, \dots (y_n, x_n), \quad x \in R^n, \quad y \in R \qquad n = 1,2, \dots, n$$

$$f(x) = w.x_i + b \tag{19}$$

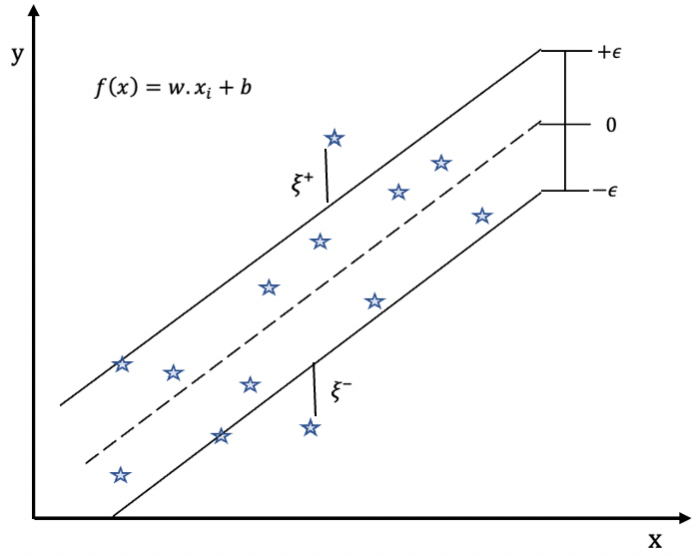
19'da yer alan denklemde x n boyutlu uzayda girdi vektörünü, y çıktı vektörünü, w ağırlık vektörünü, b sapmayı ifade etmektedir.



Şekil 22: Doğrusal DVR

DVR'nin DVM'den farkı sınıflandırma probleminde optimal aralığın belirlenerek bu aralıkta en iyi şekilde ayrılmış sınıflara ait veririn bulunmamasını sağlamaktır. Regresyon problemlerinde ise belirlenen hiper düzlem arasında maksimum veriyi alması istenmektedir. En uygun regresyon fonksiyonu için $f(x) = w.x_i + b$ denklemindeki w değerinin normunun minimize edilmesi gerekmektedir.

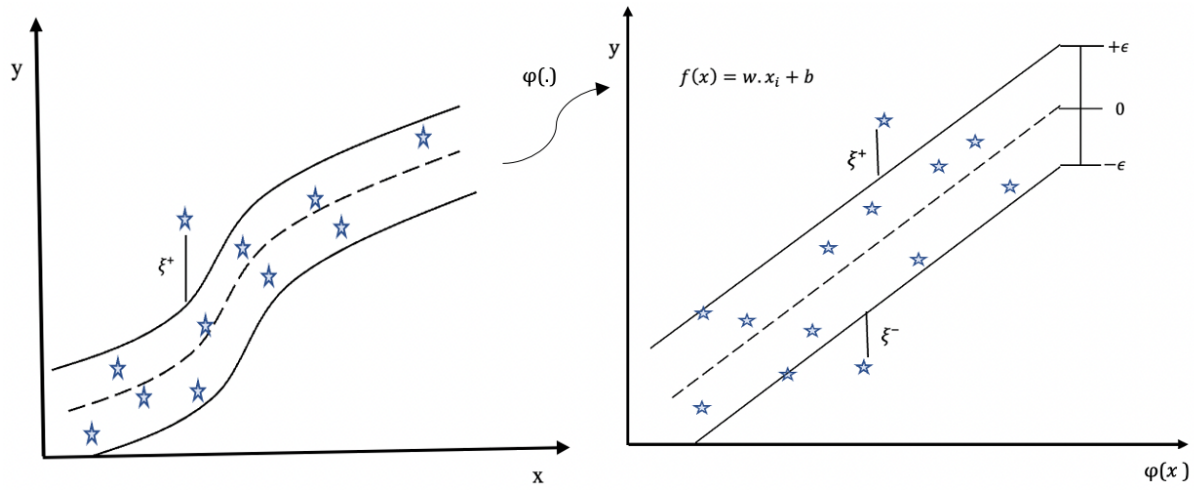
Tüm veri setleri için tüm gözlemler marjin aralığında olması mümkün olamamaktadır. Bu sebeple esneklik değişkeni (ξ) kullanılmaktadır.



Şekil 23: Doğrusal Doğrusal Fonksiyonu İçin **ϵ** Tüpü Ve Esneklik Değişkeni (DVR)

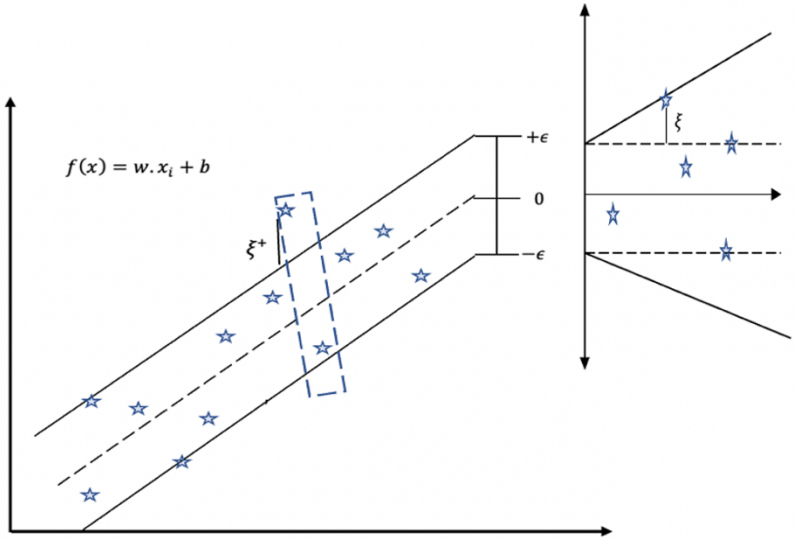
3.1.4 Doğrusal Olmayan Destek Vektör Regresyonu

Tüm veri setleri için doğrusal regresyon modeli uygun olmamaktadır. Bu sebeple doğrusal olmayan DVM'de olduğu gibi örnekler daha yüksek boyutlu uzaya taşınmaktadır.



Şekil 24: Çekirdek Yaklaşımı

Veriler yüksek boyutlu uzaya taşınırken çekirdek yaklaşımı kullanılmaktadır.



Şekil 25: DVR'de Kayıp Fonksiyonu

Şekil 25'de kayıp fonksiyon durumu gösterilmektedir. Çekirdek fonksiyonu $K(x_i, x_j) = \phi(x_i).\phi(x_j)$ tahmin fonksiyonuna eklendiğinde,

$$f(x) = \sum_{i=1}^{L} (\alpha_i^+ - \alpha_i^-)\, K(x_i, x_j) + b \qquad (20)$$

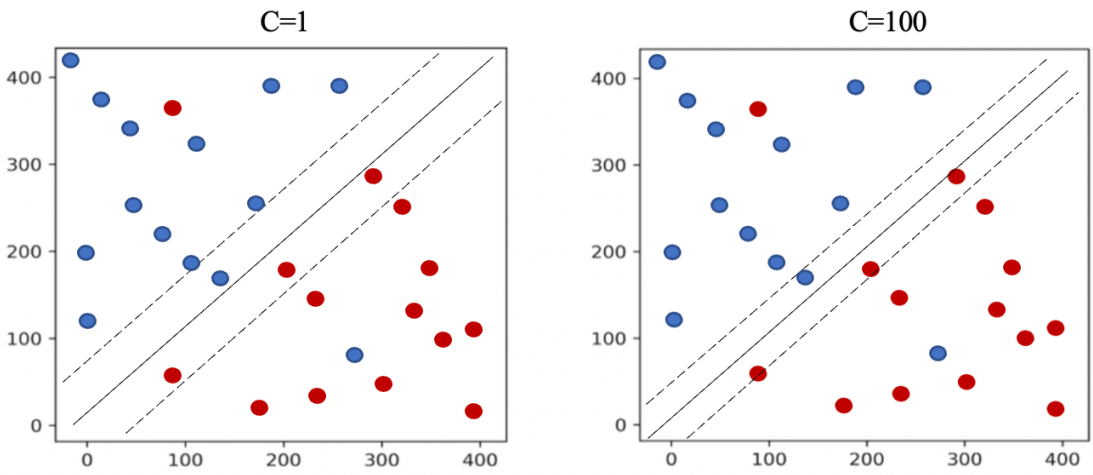
şeklinde gösterilmektedir.

DVR modellerinin tahmin doğruluğu için parametrelerinin belirlenmesi çok önemlidir. DVR algoritması parametrelerin seçimine karşı çok hassastır. Parametrelerin istenen tahmin değerini vermesi için cebirsel bir formülü bulunmamaktadır (Hsu ve ark., 2003).

Çizelge 1: DVR tabanlı tahmin modeli için kullanılan parametreler ve değer aralıkları

Regresyon Yöntemi	Parametre	Değer Aralığı
DVR	C	[0.1, 5000]
	ε	[0.0001, 100]
	γ	[0.0001, 50]

Sınıflandırma ve regresyon problemlerinde esneklik kazandıran çekirdek fonksiyonlarının kullanılması performansı arttırmaktadır. Doğrusal olarak ayrılamayan verilerde en uygun hiper düzlem haritalama yöntemi ile bulunmaktadır. Ancak doğru haritalama fonksiyonunu bulmak zordur (Güldoğan, 2017). Bu sebeple çekirdek fonksiyonları kullanılmaktadır. Veriler yüksek boyutlu uzaya taşınmadan iç çarpımı çekirdek fonksiyonları vermektedir. Bu çözümü sayesinde DVM, birçok alanda tercih edilmektedir.



Şekil 26: C' nin Belli Değeri İle Sahip Olduğu Marjin Aralığı

DVR'de C'nin aldığı değere göre marjin aralığı değişmektedir. Modelin yetersiz tahmininde C değeri arttırılmalı, aşırı tahmin durumunda ise bu değer azaltılmalıdır.

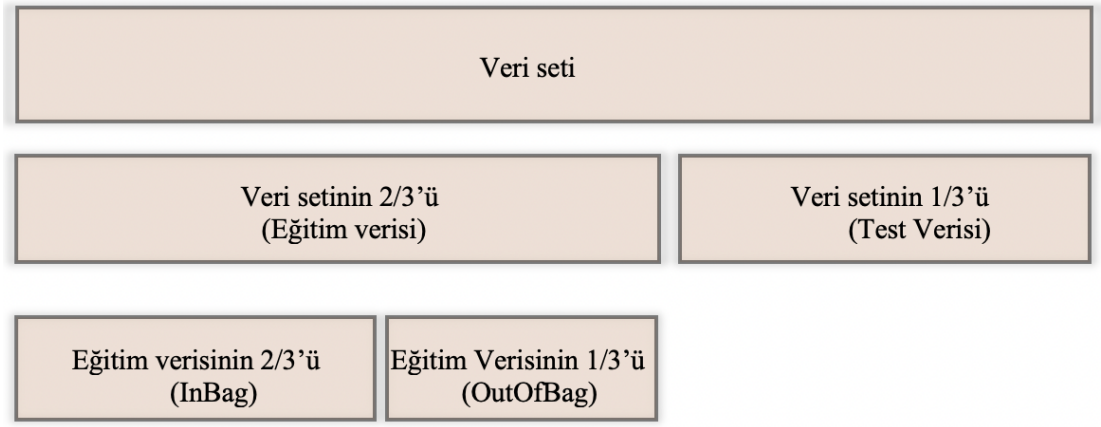
3.2. Rastgele Orman

Makine öğrenmesi algoritmalarından biri olan Karar Ağaçları (KA) en iyi bilinen ve sınıflandırma problemleri için kullanılan yöntemlerden bir tanesidir. KA yöntemi veri setini sınıflandırmada kullanılırken bazı problemlerle karşılaşarak ağacın doğru şekilde kurulamamasından dolayı yüksek performans sağlayamamaktadır. Bu sorunu ortan kaldırmak için Rastgele Orman (RO) algoritması geliştirilmiştir. RO algoritması birçok KA'nın bir araya gelerek oluşturduğu topluluktur.

RO, makine öğrenmesi yöntemlerinden biri olarak hem sınıflandırma hem de regresyon problemlerinde kullanılmaktadır. Leo Breimen tarafından geliştirilen RO yöntemi, Bagging ve alt grupların seçimi için rastgelelik yöntemi olan Random Subspace yöntemlerinin birleşmesi ile oluşmaktadır (Breiman, 2001).

RO algoritmasının Bagging yönteminden farkı değişkenin belirlenmesidir. İki önemli parametre olan J ve m sırasıyla; ormandaki ağaç sayısını ve rassal olarak seçilen tahmin edicilerini ifade etmektedir (Culter ve ark., 2012). m parametresinin belirlenmesinde p tane tahmin edicinin karekökü alınmaktadır (Dangeti, 2017). RO algoritmasında m ve p değişkenleri farklı sayıda olabilmektedir. Eğer aynı sayıdalarsa RO algoritmasının Bagging yönteminden farkı kalmamaktadır (Gareth ve ark., 2013).

RO yöntemi ile sınıflandırma yaparken mevcut verinin yaklaşık olarak 2/3'ü algoritmanın öğrenmesi için eğitim verisi olarak ayrılmaktadır. Eğitim verisinden "inbag" veri seti elde edilmektedir. 1/3'ü ise "out of bag (Oob)" veri setini oluşturmaktadır. Mevcut verinin 2/3'ü eğitim verisine ayrıldıktan sonra kalan 1/3'ü test verisini oluşturmaktadır (Çayır, 2014).



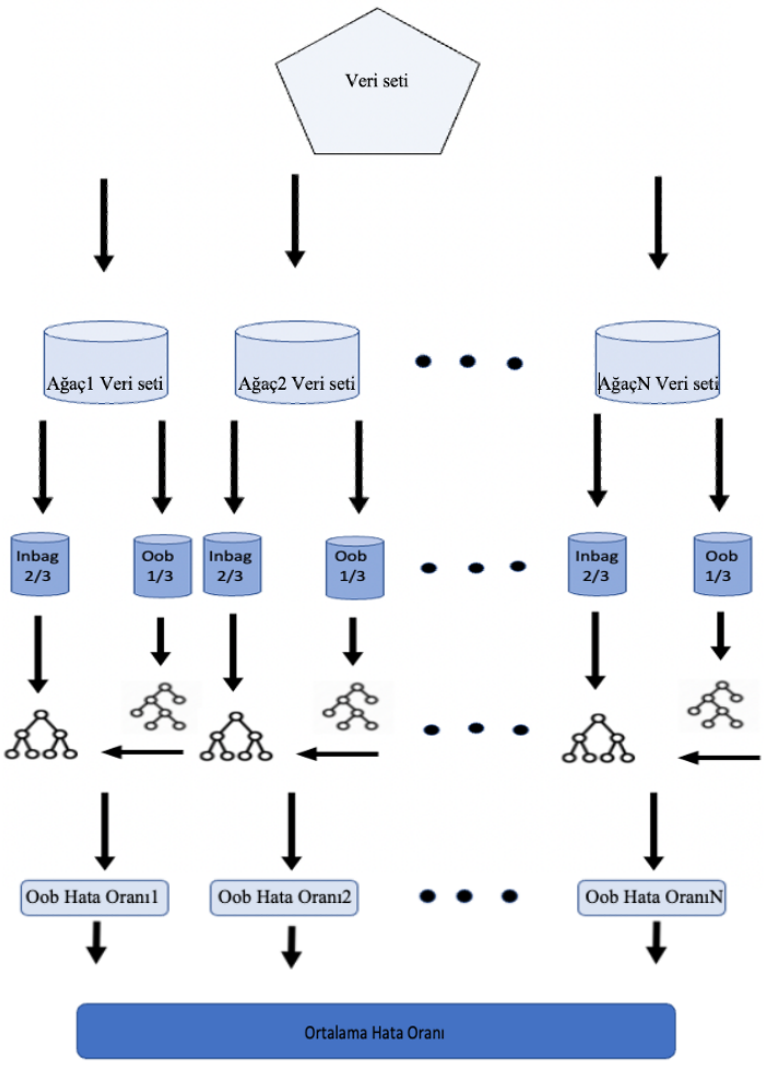
Şekil 27: Rastgele Orman Algoritması İle Veri Setinin Oluşturulması

Eğitim verisi ile karar ağaçları oluşturulmaktadır. Tüm tahmin değişkeleri içerisinde değişken seçmek için $p < m$ olacak şekilde rastgele seçilmektedir. Seçilen p tane tahminci içinden en yüksek bilgiye sahip değişkenle dallara ayırma işlemi gerçekleşecektir (Timofeev, 2004). Ormandaki ağaçların oluşturulmasında her ağaç diğerinden etkilenmeden oluşmaktadır. Ağaçlar oluşturulurken en iyi bölünmeyi sağmak için Gini indeksi kullanılmaktadır. Gini indeksi 2.1'deki eşitlikte ifade edilmektedir.

$$\sum\sum j \neq i \left(\frac{f(C_i, T)}{|T|}\right)\left(\frac{f(C_j, T)}{|T|}\right) \tag{21}$$

21'deki denklemde; T , eğitim veri kümesini, C_i , gözlemin ait olduğu sınıfı, $f(Ci, T)/|T|$ seçilen gözlemin C_i sınıfında olma olasılığını, $f(C_j\ , T)/|T|$ ise ele alınan gözlemin C_i dışındaki herhangi bir sınıfa ait oma olasılığını ifade etmektedir.

Oluşturulan ağaçlar ayrılan test verisi ile denenerek hata oranı belirlenmektedir. Oluşan tüm ağaçların ortalaması alınarak ortalama hata oranı belirlenmektedir. Hata oranı en yüksek olan ağaç en düşük, hata oranı en düşük olan ağaç ise en yüksek ağırlık ile ağırlıklandırılmaktadır (Atasever, 2011).



Şekil 28: Rastgele Orman Algoritması İle Oob İşleyiş Süreci

Rastgele Orman algoritması aşağıdaki verilen adımlar izlenerek işlemektedir (Abellan ve ark., 2017;Akman, 2010; Liaw ve Wiener, 2018);

Algoritma karar ağaçlarını oluşturmaya başlarken veri setindeki n tane gözlem sayısı Torbalama yöntemi kullanılarak N adet örneklem oluşturmaktadır. Mevcut veri setindeki örneklerin 2/3'ü Inbag verisi olarak kullanılmaktadır. Bu verilere eğitim verisi adı verilmektedir. Kalan 1/3'ü test veri olarak kullanılmaktadır ve hata oranını test etmektedir.

Inbag eğitim verisi ile oluşturulabilen en çok karar ağacı oluşturulmaktadır. Ağaçları oluşturulurken her düğümde m tane bağımsız değişkenden p tanesi rassal olarak seçilmektedir. Rastgele seçilmesinin sebebi ağacın aşırı uyum gösterme sorunuyla karşılaşması istenmemesindedir. Rassal olarak seçilen p tane tahminci değişkenden en iyi performansı sergileyen değişken ile dallara ayrılma işlemi gerçekleşmektedir. Orman oluşana kadar önceki adımlar tekrarlanmaktadır. Oluşan N tane her bir ağacın yapmış olduğu tahminlerden tekrardan yeni tahmin oluşturulmaktadır.

Ele alınan bir örneğin hangi sınıflarda kaç kez sınıflandırıldığına bakılmaktadır. İki sınıfı olan sınıflandırmada bir örnek tüm ağaçlardan en az %51 oy çokluğunu almış olduğu sınıfın adını almaktadır.

Orman topluğunu oluşturan her karar ağacının oob hata oranı yanlış sınıflandırma hatasını ifade etmektedir.

Rastgele Orman algoritmasına ait ağaç sınıflandırılması,

$$\{h(x, \theta_k), k = 1,2,..n\} \quad (22)$$

Denklem 22'de ifade edilmektedir. x girdi vektörünü, θ_k ise girdi vektörünü ifade etmektedir (Breiman,2001).

RO algoritması avantajlarından bir tanesi tüm veri setine ait kayıp verileri kullanabilme özelliğine sahiptir. Kayıp verilerin söz konusu olduğu değişken sürekli veriler içeriyorsa diğer gözlemlerin uç değeler olmasına karşın mevcut verinin ortalamasını almak yerine medyan değerleri alınarak kayıp gözlemlere atama yapılmaktadır. Kayıp verinin söz konusu olduğu değişken kategorik değişken ise eksik gözlemlere atama mevcut verilerin frekansları hesaplanarak yapılmaktadır (He, 2006).

RO yönteminde algoritmayı uygulamadan önce veri ön işlemesine gerek olmamaktadır, uygulama kendi içerisinde tüm veri setini işleyebilme özelliğine sahiptir. Herhangi bir şekilde veriyi dönüştürme, değiştirme veya yeniden ölçeklendirmesine gerek kalmadan tutarlı performans sergilemektedir.

Rastgele Orman algoritmasının sınıflandırma problemlerinde tercih edildiği gibi regresyon problemlerinde de tercih edilmektedir. RO regresyon yöntemi sınıflandırma yönteminden çok farkı bulunmamaktadır. Inbag veri setinden rassal olarak seçilen m adet değişkeni seçilmektedir. En iyi bölünmeyi sağlayan değişken seçilmektedir. Sınıflandırma yönteminde ağaç için $m = \sqrt{p}$ iken regresyon ağacı için $m = p/3$ olmaktadır.

Sınıflandırma yönteminde oy çoğunluğu dikkate alınırken regresyon yöntemi için ortalama dikkate alınarak yeni veri seti tahmin edilmektedir (Liaw ve Wiener, 2002). RO sınıflandırma yöntemi gibi regresyon yönteminde de dallanma kriterlerinde Gini indeksi kullanılmaktadır. Gini indeksinden faydalanırken iki parametre kullanılmaktadır. Bunlar, her düğümde belirlenecek örneklem sayısını ifade eden n parametresi ile oluşturulacak ağaç sayısını ifade eden N parametresidir (Çölkesen, 2015).

RO regresyon algoritması daha fazla ağaç ürettiğinde model performansı çok daha güçlü olmaktadır (Breiman, 2001). Bu sebeple farklı ağaç sayıları üretilerek model performans karşılaştırması önerilmektedir (Akman ve ark., 2011).

RO regresyon algoritmasında daha tutarlı tahminler yapılabilmektedir. Regresyonu sınıflandırmadan ayıran bir diğer özellik; sınıflandırmada bağımlı değişken kategorik iken regresyon uygulamalarında ise bağımlı değişken sürekli olmaktadır. Bu sebepten ötürü regresyon çalışmalarında üretilen her ağaç kendi tahminini üreterek karar aşamasında elde edilen tüm ağaçların tahminin ortalamaları alınmaktadır (Cutler ve ark., 2011).

3.3. K-En Yakın Komşular

K-en yakın komşular (KNN) yöntemi mevcut veriyi problemin farklılığına göre sınıflandırma ve regresyon tahmini yapabilen mesafe hesabına dayalı, kullanımı yaygın olarak tercih edilen makine öğrenmesi algoritmasıdır. k terimi komşu sayısını ifade etmektedir.

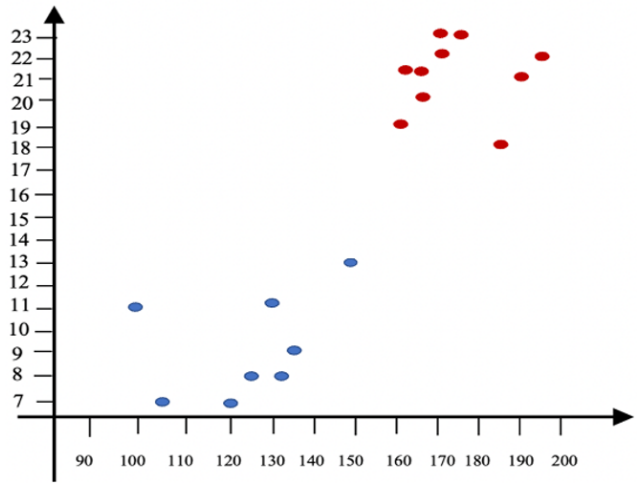
Algoritma yöntemi işlerken mevcut veriden ayrılan eğitim verisi gizlenmektedir. Yeni bir gözlemin sınıflandırılması ve tahmin edilmesi gerekene kadar model oluşturulmamaktadır. Yeni gözlemin sınıflandırılması ve tahmin edilmesi istendiğinde diğer algoritmalarda belirlenen parametrelerin aksine her seferinde eğitim veri seti kullanılarak sınıflandırma işlemi gerçekleşmektedir (Harrington, 2012). Sınıflandırma işlemi yapılırken her bir gözlemin diğer gözlemlere olan uzaklığı hesaplanmaktadır. Bir gözleme diğer gözlemlerden uzaklığı k adeti kadar komşu dikkate alınmaktadır. KNN algoritması gözlemin hangi sınıfa dahil olduğunu belirleyerek, gözleme en yakın olan k tane gözlemi bulmaktadır. Gözlem, k en yakın komşu içinden en çok benzerlik gösterdiği sınıfa atanmaktadır (Ayık ve ark., 2007). Elimizde boy, yaş ve gittiği okula göre gözlemleri oluşturan veri seti olduğunu varsayalım. Boy ve yaşı bilinmekte olan yeni öğrencinin gittiği okul tahmin edilsin. Bu veriler Çizelge 2'de yer almaktadır.

Çizelge 2. KNN örnek veri seti

Gözlem Sayısı	Boy(cm)	Yaş	Gittiği Okul
1	120	7	İlkokul
2	135	9	İlkokul
3	132	8	İlkokul
4	105	7	İlkokul
5	126	8	İlkokul
6	100	11	İlkokul
7	130	11	İlkokul
8	150	13	İlkokul
8	161	19	Üniversite

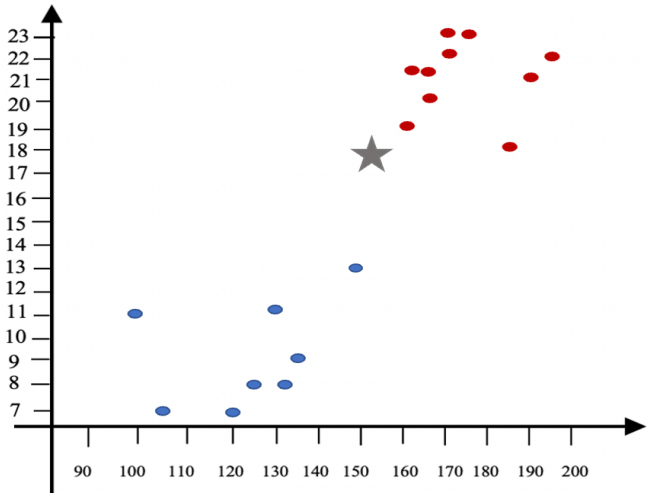
9	164	21	Üniversite
10	180	23	Üniversite
11	170	21	Üniversite
12	175	23	Üniversite
13	188	18	Üniversite
14	188	21	Üniversite
15	197	23	Üniversite
16	165	20	Üniversite
17	170	22	Üniversite

Çizelge 2'de verilen gözlemler, okuduğu okula göre Şekil 29'de yer alan iki boyutlu koordinat sisteminde gösterilmektedir. Gittiği okula göre, ilkokul mavi ve üniversite ise borda renk ile gösterilmektedir.



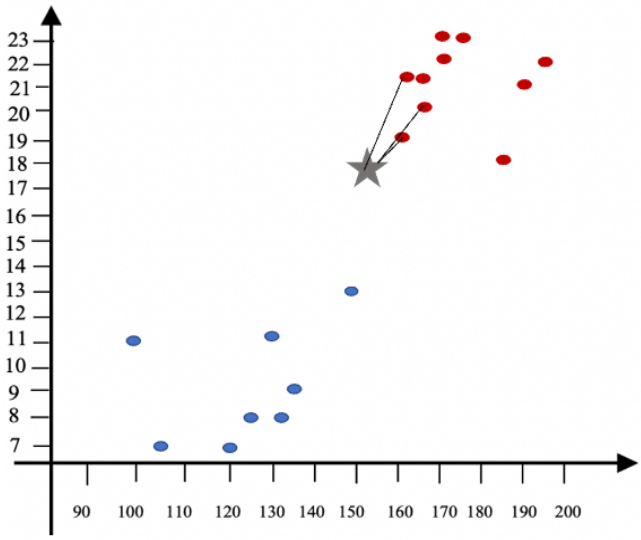
Şekil 29: KNN İki Boyutlu Düzlemde Örnek Çalışması

Yeni gözlem olarak 155 boyunda 18 yaşında belirlenen yeni bir öğrenci düşünelim. Şekil 30'de yeni gelen x öğrencisi gri renk ile gösterilmektedir.



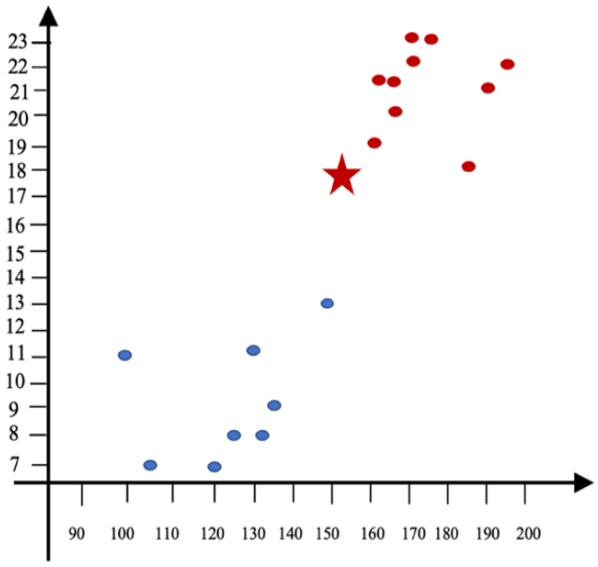
Şekil 30: İki Boyutlu Düzleme Yeni Bir Gözlemin Eklenmesi

KNN algoritması, gri renkte eklenen yeni gözlemin, algoritmada belirlenmekte olan k değerine göre en yakın gözlemler arasındaki mesafeye bakılmaktadır. Örneğin k değeri 10 belirlenmiş ise en yakın olan 10 komşu gözlem ile arasındaki mesafeye bakılarak bu gözler içerisinde en yakın mesafeye sahip olduğu sınıfa atanmaktadır. Bu örnekte k=3 olarak belirlendiğini varsayalım.



Şekil 31: k=3 Değerine Göre Komşu Belirleme İşlemi

Şekil 31 incelendiğinde tüm yakın komşuların üniversite öğrencisi olduğu görülmektedir. Bu sebeple yeni öğrenci için atama üniversite sınıfına olacak ve yeni öğrencide bordo rengine sahip olacaktır.



Şekil 32: Yeni Gelen Öğrencinin Sınıfa Atanması

KNN algoritmasında uygulama içerisinde belirlenen mesafe ölçütleri tüm veri setinin özniteliklerinin, numerik, kategorik ve ikili değerlere sahip olması gibi durumlara değişmekte olup genellikle Öklid uzaklık ölçütünün kullanımı tercih edilmektedir.

Öklid uzaklığı, kümeleme ve sınıflandırma analizlerinde en çok tercih edilen uzaklık ölçütüdür. Öklid uzaklığı, iki nokta arasındaki doğrusal uzaklık olmaktadır. P ve Q arasındaki Öklid uzaklığı $P = (x_1, x_2, x_3, \ldots, x_n)$ ve $Q = (y_1, y_2, y_3, \ldots, y_n)$ olmak üzere Eşitlik 3.1'e göre hesaplanmaktadır (Kreese ve Danko, 2012).

$$\left(\sqrt{\sum_{i=1}^{n}(x_i - y_i)}^{2}\right) \tag{23}$$

KNN algoritmasının kullanılan ağırlıklandırma yöntemi performansı arttırmada etkili olmaktadır. Uygulama içerisinde sürekli k değerinin değiştirilerek en etkili sonucu veren k değerinin belirlenmesi zaman kaybıdır. Uygulamada kullanılan söz konusu veri seti üzerinde k değerinin en uygun değerini belirlemek için uzaklıkların ağırlıklandırılması önemlidir. En uygun k değerini belirlemek için uygulanan

ağırlıklandırma işlemi test verisine en yakın olan k gözleminin mesafesine göre ağırlıklandırılarak başarılı bir performans sergiler (Özger ve Amasyalı, 2013).

k-En yakın komşu regresyonu parametrik olmayan bir makine öğrenmesi algoritmasıdır. Algoritmanın temeli örüntü tanıma yöntemine dayanmaktadır. En yakın komşu regresyonun amacı yeni eklenen noktaya en yakın mesafeyi önceden belirlenmiş olan sayıda gözlemleri bulmak ve tahmin gerçekleştirmektir. KNN regresyon yönteminde sürekli değişkenleri tahmin edilmektedir (Eyecioğlu, 2021).

KNN regresyon adımları aşağıdaki gibidir;

- Seçilmekte olan gözlem ile diğer gözlemler arasındaki mesafe hesaplanmaktadır.
- Mesafeler hesaplandıktan sonra bu hesaba göre sıralama yapılmaktadır.
- Kök ortalama kare hatası ve çapraz doğrulamadan faydalanılarak en yakın k komşu sayısı belirlenmektedir.

Belirlenen k en yakın komşularından faydalanarak ters mesafe ağırlıklı ortalama değeri hesaplanmaktadır.

4. MODEL PERFORMANS METRİKLERİ

Modellerin tahmin performansının (doğruluğunun) ölçülmesi için hata istatistik değerleri ve R^2 değeri dikkate alınmaktadır. R^2 modelin veri ile uyumunu göstermektedir. R^2 , 0 ile 1 arasında değer almaktadır. 1'e yakınlık modelin uygun olduğunu, başarılı tahmin gerçekleştirdiğini göstermektedir. R^2 pozitif (+) değer almaktadır ancak orijinden geçen regresyon modelinde negatif (-) değer alabilir. Negatif (-) değer almasının nedeni modelde sabit terim olduğu varsayımıdır.

$$R^2 = 1 - \frac{\sum \hat{u}^2}{\sum y^2} \tag{24}$$

Eşitlik 4.1'de yer alan denklemde $\sum \hat{u}^2$ tahmin eğrisine denk gelmeyen her noktayı ifade eden atıkların kareler toplamıdır. $\sum y^2$ ise her noktanın ortalamaya uzaklığının karesini ifade etmektedir.

Hata istatistikleri gerçek değerler ile öngörü değerleri arasındaki farkı göstermektedir. Performans değerlendirme için; Ortalama Kare Hata (MSE), Kök Ortalama Kare Hata(RMSE) metrikleri dikkate alınmaktadır.

MSE, gerçek veriyle tahmin edilen fonksiyonlar arasındaki mesafe ölçülerek karesi alınmaktadır. Tüm veriler için bu değerler toplanarak veri sayısına bölünmektedir.

$$MSE = \frac{1}{n}\sum_{j=1}^{n} e_j^2 \tag{25}$$

RMSE, MSE'nin kökü alınmış formülüdür. Verilerin, tahmin edilen regresyon modelinden ne kadar sapma olduğunu göstermektedir.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n} e_j^2} \tag{26}$$

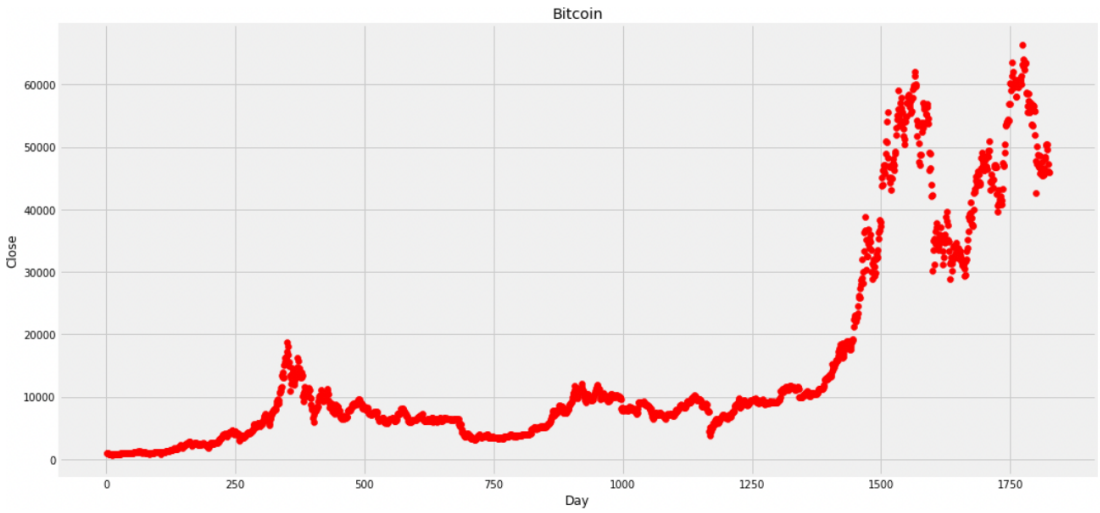
Eşitliklerde yer alan A_j gerçek değerleri, $e_j = A_j - P_j$ hataları oluşturmaktadır. P_j tahmin edilen değerleri ifade etmektedir.

5. UYGULAMA

Uygulamada 01/01/2017 yılı ile 31/12/2021 yılları arasında toplamda 1826 gözlemin oluşturan Bitcoin'e ait günlük kapanış verileri kullanılmaktadır. Mevcut veriler kullanılarak, Destek vektör regresyonu, Rastgele Orman regresyonu ve K-En Yakın Komşular regresyon modelleri ile bir sonraki gün kapanış değeri tahmin edilmektedir.

Çalışmanın amacı hangi makine öğrenmesi tahmin yönteminin zaman serisi tahmininde daha yüksek performans sergilediğini ve gerçek piyasa değerine en yakın tahmini yapabildiğine bakılmaktadır.

Uygulamada kullanılan 1826 gözlemin yer aldığı 5 yıllık verinin grafiği Şekil 33'de yer almaktadır.



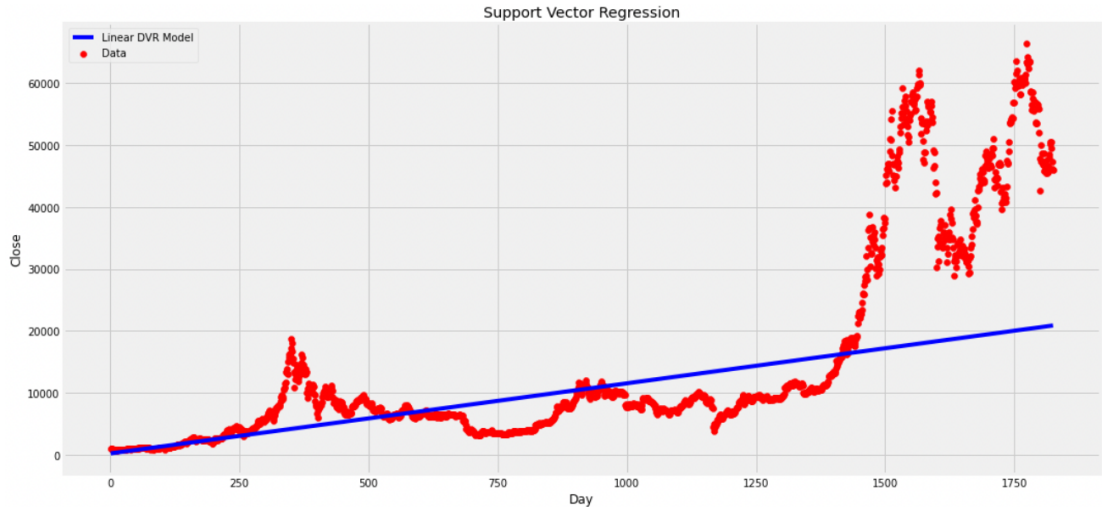
Şekil 33: Bitcoin Kapanış Verileri

Şekil 33'da gösterilen veriler, veri setimizi oluşturmaktadır. Gözlemler incelendiğinde zaman serilerin özelliği olan oynaklığın sürekli olduğu görülmektedir. Son yıllarda Bitcoin'e olan talebin çok fazla artarak devam eden zamanda tekrardan düşüş yaşadığı görülmektedir. Bu finansal piyasalara olan yatırımlarda belirsizliği göstermekte, ve yüksek kar kaybı riskini ifade etmektedir. 1826 gözlem kullanılarak 1827. gün yani 01/01/2022 günü kapanış fiyat tahmini yapılacaktır.

Çizelge 3. Bitcoin'e ait tanımlayıcı istatistikler

Değişken	Maks.	Min.	Ortalama	Standart Sap.
Bitcoin	66334.90	739.500	14959.47	16421.35

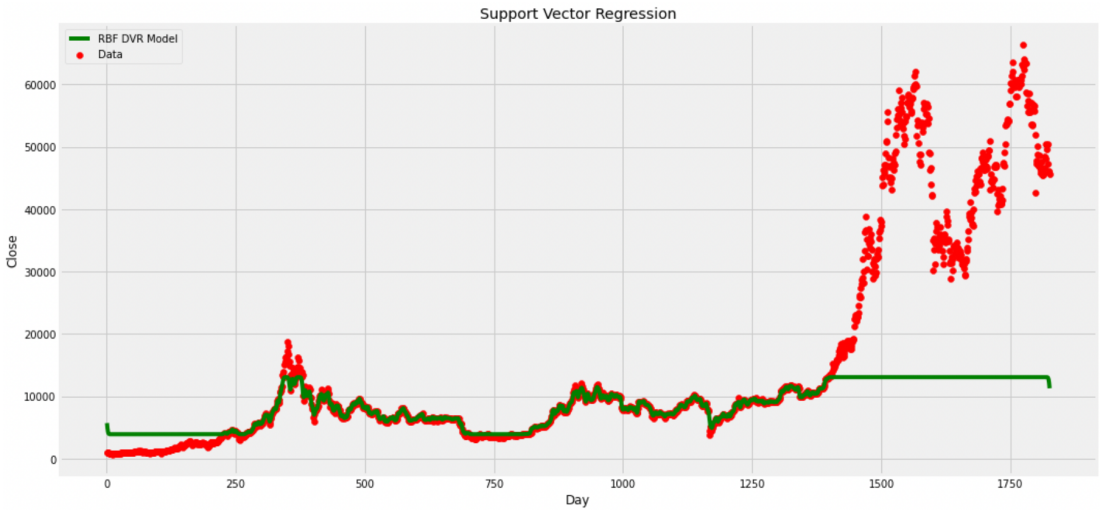
Şekil 34'de Doğrusal Destek Vektör regresyon tahminine ait grafik çıktısı yer almaktadır. Tahmin çıktıları üzerinden yorum yapılabilir fakat model doğruluğu ve performansı için hata istatistik değerleri ve R^2 değeri dikkate alınmaktadır.



Şekil 34: Doğrusal DVR

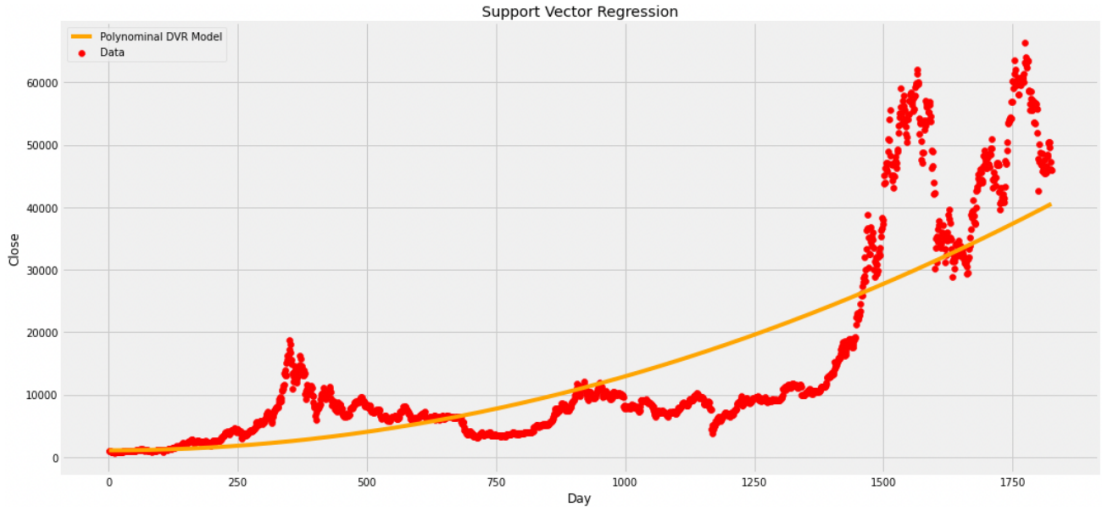
Şekil 34 incelendiğinde doğrusal DVR'nin zaman serisi üzerinde çalışmadığını ve veri setinin doğrusal olarak artmadığını göstermektedir. Bu durum zaman serilerinin özelliklerinden bir tanesidir. Veri setinin doğrusal olarak ayrılamadığını görülmektedir. Bu sebeple Doğrusal olmayan DVR denenmesi gerektiğini görmekteyiz.

Şekil 35'de Doğrusal olmayan model Radyal tabanlı DVR 'ye ait grafik yer almaktadır.



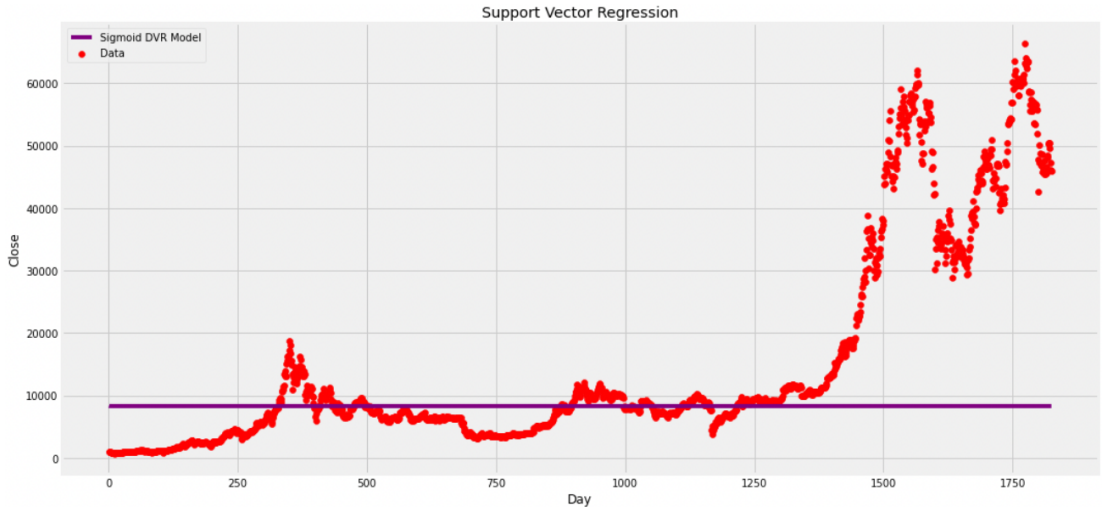
Şekil 35: RBF -DVR

Radyal tabanlı DVR incelendiğinde, zaman serisi üzerinde çalıştığı fakat etkili performans sergilemediği görülmektedir. Model tahmini ile gerçek veri karşılaştırıldığında, tahmin modelinin gerçek veriyi tamamen kapsamadığını göstermektedir. Bir diğer doğrusal olmayan DVR modeli olan Polinominal-DVR modeli şekil 36'da yer almaktadır.



Şekil 36: Polinominal -DVR

Polinominal DVR'nin yer aldığı Şekil 36 incelendiğinde verinin kısmen şekilde polinominal olarak arttığı görülmektedir. Bir diğer doğrusal olmayan DVR modeli olan sigmoid-DVR modeli Şekil 37'de yer almaktadır.

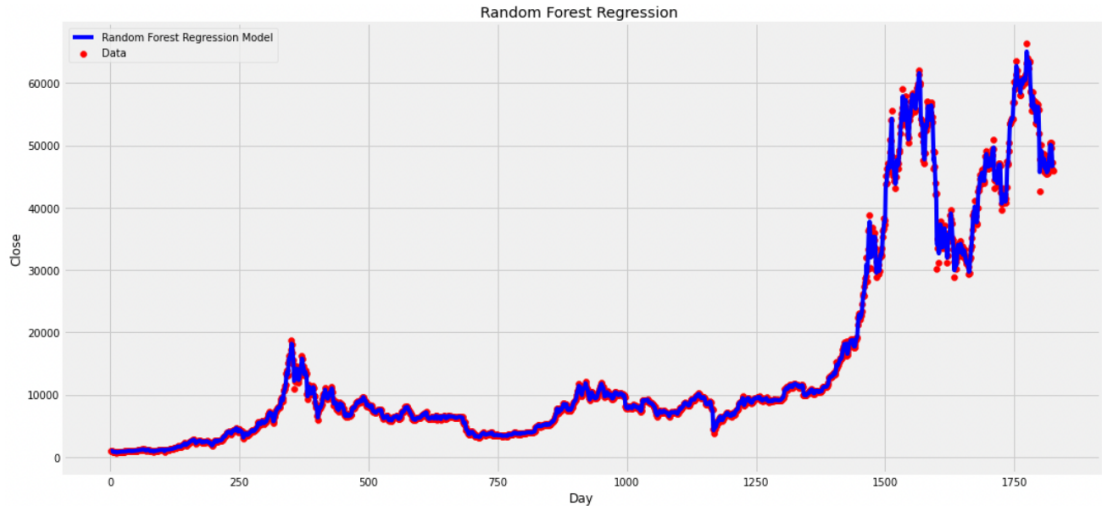


Şekil 37: Sigmoid-DVR

Şekil 37 incelendiğinde sigmoid-DVR'nin veri seti üzerinde çalışmadığı görülmektedir.

DVR'nin tahmin modelleri için; Doğrusal-DVR'nin çalışmadığını, doğrusal olmayan modeller incelendiğinde ise RBF-DVR'nin etkin olmayan bir tahmin performansı sergilediği, Polinominal-DVR'nin tahmin çıktısı gerçek veri artışı gibi olduğu ama tüm veriyi kapsamadığı, Sigmoid-DVR için ise hiç çalışmadığı görülmektedir.

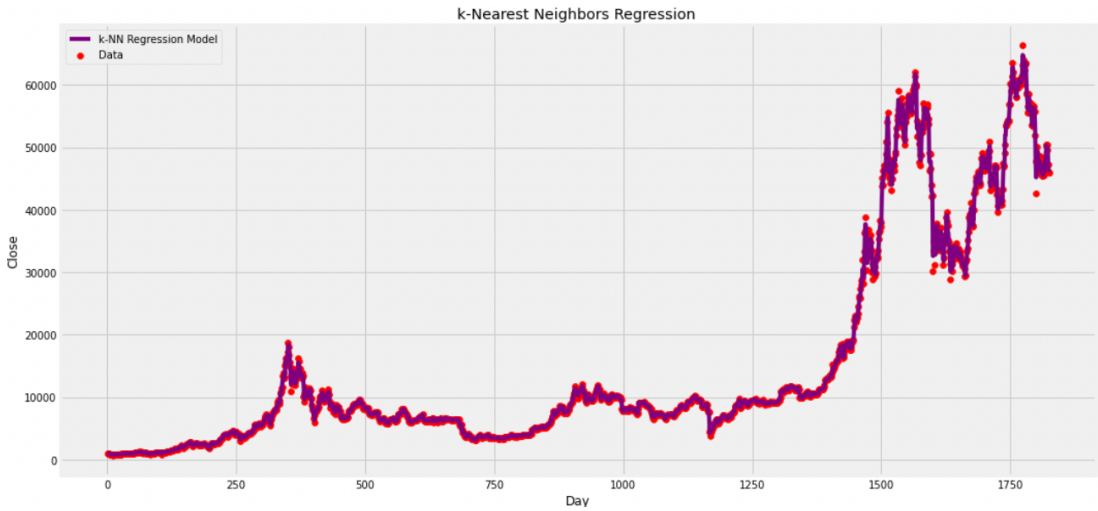
Rastgele orman regresyon tahmini Şekil 38'de yer almaktadır.



Şekil 38: Rastgele Orman Regresyonu

Rastgele orman regresyonuna ait grafik incelendiğinde tahmin algoritmasının gerçek veri ile model tahminin neredeyse hepsinin örtüştüğü, etkin tahmin yaptığı görülmektedir. Bu tahmin çıktısı dikkate alındığında zaman serilerinde Rastgele Orman tahmin algoritmasının kullanılmasının doğru olacağını ifade edebilmekteyiz.

KNN regresyonuna ait tahmin çıktısı şekil 39'de yer almaktadır.



Şekil 39: KNN Regresyonu

Şekil 39 incelendiğinde KNN regresyonunun etkin performans gösterdiği ve tahmin ile gerçek verilerin neredeyse hepsinin örtüştüğü görülmektedir.

Finansal zaman serileri için KNN regresyonunun tercih edilmesinin doğru olacağını grafik çıktıları göstermiş olsa da asıl etkinliği yorumlayabilmemiz için hata metriklerine ve R^2 değerine bakılması gerekmektedir.

Makine öğrenmesi tahmin algoritmalarının 1827. gün tahmin değerleri Çizelge 4'de yer almaktadır.

Çizelge 4. Makine öğrenmesi algoritmalarının tahmin değerleri

Regresyon Modelleri	**Tahmin Değerleri**
Doğrusal-DVR	21035.6350
RBF- DVR	10298.1122
Polinominal-DVR	40901.1741
Sigmoid-DVR	8318.2
Rastgele Orman	**45959.9884**
KNN	45848.3

Bitcoin'e ait bir sonraki günün gerçek kapanış değeri **46217.5** 'dir. Bu değere en yakın tahmin değerini veren algoritma Rastgele Orman regresyonu olmaktadır.

Tüm tahminlere ait hata istatistik değerleri Çizelge 5'de yer almaktadır. Hata değerinin en küçük değer alması istenmektedir ve en küçük değere sahip algoritma uygulamada en uygun model olacaktır.

Çizelge 5. Makine öğrenmesi tahmin algoritmalarına ait hata istatistikleri

Regresyon Modelleri	MSE	RMSE
Doğrusal-DVR	174790062.9245	13203.8425
RBF-DVR	234549256.0444	15327.8728
Polinominal-DVR	79452292.5628	8889.0798
Sigmoid-DVR	313140205.3927	17709.3134
RO	**819860.7498**	**905.2844**
KNN	1050977.3755	1025.8331

Model performans değerleri için hata istatistikleri incelendiğinde MSE ve RMSE hata değerlerinde en küçük istatistik değerini veren Rastgele Orman regresyon algoritması olmaktadır. İkinci en iyi hata değerine sahip olan model KNN regresyonudur, üçüncü en iyi değeri ise Polinominal-DVR vermektedir. Grafikler göz önüne alındığına bu sonuçlarla model tahmin grafiklerinin doğrusal olduğu görülmektedir. Çizelge 5'de yer alan DVR regresyon modellerinin hata istatistik değerinin çok fazla olduğu görülmektedir.

Bir diğer performans ölçütü R^2'dir. Tahmin algoritmalarına ait R^2 değerleri Çizelge 6'da yer almaktadır.

Çizelge 6. Makine öğrenmesi tahmin algoritmalarına ait $\boldsymbol{R^2}$ değerleri

Regresyon Modelleri	$\boldsymbol{R^2}$
Doğrusal-DVR	0.3509
RBF-DVR	0.1284
Polinominal-DVR	0.7050
Sigmoid-DVR	-0.1635
Rastgele Orman	**0.9997**
KNN	0.9993

Çizelge 6'da yer alan model performans ölçütü olan R^2 değerleri incelendiğinde, en yüksek değerin Rastgele Orman regresyonunun sahip olduğu görülmektedir. Bir sonraki en yüksek R^2 değerine sahip modelin KNN regresyon modelidir. DVR modelleri incelendiğinde en yüksek değerin Polinominal-DVR modeline ait olduğu görülmektedir.

6. SONUÇ

Bu çalışmada 01/01/2017 ile 31/12/2021 yılları arasını kapsayan Bitcoin'e ait günlük kapanış değerleri dikkate alınarak bir sonraki gün tahmini için belirlenen makine öğrenmesi algoritmaları uygulanmaktadır. Makine öğrenmesi tahmin yöntemlerinden bu çalışma için tercih edilen algoritmalar Destek Vektör Regresyonu, Rastgele Orman Regresyonu, K-en Yakın Komşular Regresyonu olmaktadır. Verinin özelliklerine göre Destek Vektör Regresyonu için doğrusal ve doğrusal olmayan yöntemler ile tahmin yapılmaktadır.

Zaman serisi tahminleri için belirlenen algoritmalar içinde grafik çıktıları dikkate alındığında DVR'nin etkili tahmin yapmadığı görülmektedir. Rastgele Orman regresyon tahmin grafiği dikkate alındığında neredeyse tüm verileri içine alacak şekilde etkili tahmin yaptığı görülmektedir. KNN regresyon tahmin grafiği incelendiğinde, Rastgele Orman algoritması kadar etkin performans sergilediği neredeyse tüm veriler ile örtüşecek şekilde tahmin çıktısı sergilediği görülmektedir. Sadece regresyon tahmininin performansı yorumlamak için tahmin grafiklerine bakmak yeterli olmamaktadır. Bu çalışma için dikkate alınan model performansını belirlemek için hata istatistik değeri olan MSE, RMSE ve diğer performans ölçütü olan R^2 değerlerine bakılmaktadır.

Tüm algoritmaların 1826 gözlemden yararlanarak 1827. gün tahmin değerleri karşılaştırıldığında gerçek değere en yakın tahmini yapan yöntemin Rastgele Orman regresyonunun olduğu ve en yakın ikinci tahmini yapan algoritmanın KNN regresyonunun sergilediği görülmektedir. DVR modelleri için hepsinin tahminlerinin gerçek değerden çok fazla sapma olduğu en yakın değeri Polinominal DVR'nin verdiği görülmektedir. Hata değerleri incelendiğinde MSE ve RMSE için en az hata istatistiğine sahip olan yöntemin Rastgele Orman olduğu ikinci olarak en az hataya sahip olan değişkenin KNN regresyonunun olduğu görülmektedir. R^2 değeri incelendiğinde en yüksek değere sahip algoritma Rastgele Orman regresyonuna ait iken ikinci en yüksek değere sahip yöntemin KNN regresyonu olduğu görülmektedir.

DVR modelleri incelendiğinde doğrusal ve doğrusal olmayan algoritmalar içinde en etkin performansı sergileyen Polinominal-DVR'den sonra Doğrusal-DVR ve en kötü tahmini gerçekleştiren ise Sigmoid DVR'nin olduğu görülmektedir.

Analizler dikkate alındığında zaman serisi olan ve Bitcoin'e ait kapanış değerlerinden oluşan veri setimizden yararlanarak bir sonraki gün tahmini için tercih edilen DVR, Rastgele Orman ve KNN regresyonları arasından en iyi performansı Rastgele Orman regresyonun ikinci olarak KNN regresyonunun ve üçüncü olarak iyi denilecek kadar performans sergileyen Polinominal-DVR olmaktadır. Bu sebeple zaman serileri için geleneksel yöntemlerin aksine makine öğrenmesi algoritmaları tercih edilebileceği görülmektedir. Dikkate alınan bu üç yöntemden en iyi performansa sahip Rastgele Orman regresyonunun ve KNN regresyonun finansal zaman serilerinde gelecek tahmini için kullanım tercihinin doğru olabileceği, gelecek zaman serisi çalışmalarına ışık tutabileceği görülmektedir.

KAYNAKÇA

Abellán, J., Mantas, C. J., & Castellano, J. G. (2017). A Random Forest Approach Using İmprecise Probabilities. *Knowledge-Based Systems*, *134*, 72-84.

Akca M. F. , (2020). "Nedir Bu Destek Vektör Makineleri? (Makine Öğrenmesi Serisi-2), *https://medium.com/deep-learning-turkiye/nedir-bu-destek-vektör-makineleri-makine-öğrenmesi-serisi-2-94e576e4223e,* (10.03.2022)

Akman, M. (2010). Veri Madenciliğine Genel Bakış Ve Random Forests Yönteminin İncelenmesi: Sağlık Alanında Bir Uygulama. *Unpublished Master's Thesis). Ankara University, Ankara*.

Akman, M., Genç, Y., & Ankaralı, H. (2011). Random Forests Methods And An Application İn Health Science. *Turkiye Klinikleri Journal Of Biostatistics*, *3*(1), 36-48.

Alpago, H. (2018). Bitcoin'den Selfcoin'e Kripto Para. *Uluslararası Bilimsel Araştırmalar Dergisi (Ibad)*, *3*(2), 411-428.

Alpaydin, E. (2020). *Introduction To Machine Learning*. Mıt Press.

Atasever, Ü. H. (2011). Uydu Görüntülerinin Sınıflandırılmasında Hızlandırma (Boosting), Destek Vektör Makineleri, Rastgele Orman (Random Forest) Ve Regresyon Ağaçları Yöntemlerinin Kullanılması. *Unpublished Master's Thesis). Erciyes University, Kayseri*.

Atış, D. (2014). Elektronik Para Birimleri. *Araştırma Raporu-Türkiye İş Bankası*.

Ayık, Y. Z., Özdemir, A., & Yavuz, U. (2007). Lise Türü Ve Lise Mezuniyet Başarısının, Kazanılan Fakülte İle İlişkisinin Veri Madenciliği Tekniği İle Analizi. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *10*(2), 441-454.

Baran, M. (2020). *Makine Öğrenmesi Yöntemleriyle Çoklu Etiketli Verilerin Sınıflandırılması* (Master's Thesis, Sosyal Bilimler Enstitüsü).

Blockchain, (T.Y). Hash Rate (Online). https://www.blockchain.com/tr/charts/n-transactions. (17.04.2022).

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32.

Breiman, L. (2001). "Random Forest ", Https://Www.Stat.Berkeley.Edu/~Breiman/Randomforest2001.Pdf. (11.04.2022)

Burges, C. J. (1998). A Tutorial On Support Vector Machines For Pattern Recognition. *Data Mining And Knowledge Discovery*, *2*(2), 121-167.

Chao, W. L. (2011). Machine Learning Tutorial. *Digital Image And Signal Processing*.

Chaum, D. (1982). Blind Signatures For Untraceable Payments (1983). In *Advances İn Cryptology*.

Chen, A. S., Leung, M. T., & Daouk, H. (2003). Application Of Neural Networks To An Emerging Financial Market: Forecasting And Trading The Taiwan Stock Index. *Computers & Operations Research*, *30*(6), 901-923.

Corsello, P. (2018). *What Is Cryptocurrency?. Numbers, Numbers And More Numbers.* Https://Medium.Com/Datadriveninvestor/What-Is-Cryptocurrency-20fb4009e718/. (13.04.2022).

Cutler, A., Cutler, D. R. &Stevens, J. R. (2011). Random Forest, *Machine Learning*, 45 (1): 157-176.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random Forests. In *Ensemble Machine Learning* (Pp. 157-175). Springer, Boston, Ma.

Çağır, A. (2014). Master Thesis. *Performance Comparısun of Locality Sensitive Hashing and Random Forests Algorithms for Handwritten Digits Recognition* . İstanbul, Türkiye: Kadir Has University.

Çölkesen, İ. (2015). *Yüksek Çözünürlüklü Uydu Görüntüleri Kullanarak Benzer Spektral Özelliklere Sahip Doğal Nesnelerin Ayırt Edilmesine Yönelik Bir Metodoloji Geliştirme* (Doctoral Dissertation, Fen Bilimleri Enstitüsü).

Dangeti, P. (2017). *Statistics For Machine Learning*. Packt Publishing Ltd.

Dıffıe, W., Hellman, E, M. (1976). *New Directions İn Cryptography*. Ieee Transactions On Informatıon Theory, Vol.22, N.6.

European Central Bank (Ecb), (2012). *Virtual Currency Schemes.* https://www.ecb.europa.eu/pub/pdf/other/virtualcurrencyschemes201210en.pdf /.(20.04.2022).

Eyecioglu, Ö., (2021). Bazalt/Panı Kompozitlerinin Dielektrik Özelliklerinin Tahmini İçin Makine Öğrenmesi Modellerinin Karşılaştırılması. *Avrupa Bilim Ve Teknoloji Dergisi*, (23), 817-826.

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). *An İntroduction To Statistical Learning: With Applications İn R*. Spinger.(Editors: Casella G., Fienberg, S., Olkin, I.). Springer, New York. 603.

Güldoğan, E., (2017). *Çeşitli Çekirdek Fonksiyonları Ile Oluşturulan Destek Vektör Makinesi Modellerinin Performanslarının İncelenmesi: Bilr Klinik Uygulam*a, İnönü Üniversitesi, Sağlık Bilimleri Enstitüsü, Biyoistatistik Ve Tıp Bilişimi Anabilim Dalı, Doktora Tezi

Halaburda, H., & Sarvary, M., (2016). Beyond Bitcoin. *The Economics Of Digital Currencies*.

Harrington, P. (2012). Machine Learning İn Action . 20 Balding Road. *Shelter Island: Manning Publications Co*.

He, Y. (2006). *Missing Data İmputation For Tree-Based Models*. University Of California, Los Angeles.

Hsu, C. W., Chang, C. C., & Lin, C.L. (2003). A Practical Guide To Support Vector Classification. Retrieved From Htpp://Www.Csie.Ntu.Edu.Tw/∼Cjlin.

İnci, S. , Alpen, İ. (2018). Bitcoin Devrimi.(1.Baskı). *Elma Yayınevi*.

Keskenler, M. F., & Keskenler, E. F. (2017). Geçmişten Günümüze Yapay Sinir Ağları Ve Tarihçesi. *Takvim-İ Vekayi*, *5*(2), 8-18.

Kresse, W. (2012). *Springer handbook of geographic information*(pp. 118-120). D. M. Danko (Ed.). Berlin: Springer.

Kızılkaya, Y. M., & Oğuzlar, A. (2018). Bazı Denetimli Öğrenme Algoritmalarının R Programlama Dili İle Kıyaslanması. *Karadeniz Uluslararası Bilimsel Dergi*, *37*(37), 90-98.

Liaw, A., & Wiener, M. (2002). Classification And Regression By Randomforest. *R News*, *2*(3), 18-22.

Liaw, A., Wiener, M., (2018). The R Project For Statistical Computing. *https://cran.r-project.org/web/packages/randomforest/randomforest.pdf.*/ (09.04.2022).

Mcleay, M., Radia, A., Ve Thomas, R. (2014). *Money In The Modern Economy: An Introduction*. www.youtube.com/Watch?V=Zite32hiwdk. (12.04.2022).

Öğücü, M. O. (2006). *Yapay sinir ağları ile sistem tanıma* (Doctoral dissertation, Fen Bilimleri Enstitüsü).

Özer, Z. B., & Amasyalı, M. F. (2013). Meta Öğrenme İle Knn Parametre Seçimi. *21. Sinyal İşleme Ve Uygulamaları Konferansı*.

Partners, S.(2015). What İs Virtual Money? [Online]. http://en.finance.sia-partners.com/what- virtual-money.(20.04.2022)

Prowmes, *Makine Öğrenmes*i, http://www.prowmes.com/blog/makine-ogrenmesi/ , (14.01.2022)

Şeker, E, Ş. (2009). Kör İmzalar (Blind Signatures). *https://bilgisayarkavramlari.com/2009/10/14/kor-imzalar-blind-signatures/* (12.04.2022).

Timofeev, R. (2004). Classification And Regression Trees (Cart) Theory And Applications. *Humboldt University, Berlin*, *54*.

Ünsal, E. M. (2017). Makro İktisat (11. Baskı). *Ankara: Murat Yayınları*.

Vapnik, V. (1999). *The Nature Of Statistical Learning Theory*. Springer Science & Business Media.

Yıldırım, F. (2015). Kripto Paralar, Blok Zinciri Teknolojisi Ve Uluslararası İlişkilere Muhtemel Etkileri. *Medeniyet Araştırmaları Dergisi*, *2*(4), 81-97.

Yüksel, B., & Ebru, A. (2015). Elektronik Para, Sanal Para, Bitcoin Ve Linden Dolari'na Hukuki Bir Bakiş. *İstanbul Üniversitesi Hukuk Fakültesi Mecmuası*, *73*(2), 173-220.

Wang, Z. Ve Xue, X. (2013). Multi-Class Support Vector Machine, *Springer, Support Vector Machines Applications,* Eddition 2014.

Wenzel, H., Smit, D., & Sardesai, S. (2019). A Literature Review On Machine Learning İn Supply Chain Management. In *Artificial Intelligence And Digital Transformation İn Supply Chain Management: Innovative Approaches For Supply Chains. Proceedings Of The Hamburg International Conference Of Logistics (Hıcl), Vol. 27* (Pp. 413-441). Berlin: Epubli Gmbh.

World Payments Report. (2021). Capgemini Research Institute.Https://Worldpaymentsreport.Com/WpContent/Uploads/Sites/5/2021/10/World-Payments-Report_2021_Web.Pdf/. (13.04.2022).

Zhang, J., Shan, R.,&Su, W. (2009). Applying Time Series Analysis Builds Stock Price Forecast Model. *Modern Applied Science*, *3*(5), 152-157.