MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE AĞ TRAFİĞİNİN SINIFLANDIRILMASI

2020

BİLGİSAYAR MÜHENDİSLİĞİ

BİTİRME PROJESİ TEZİ

BÜŞRA BAKKALCI

MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE AĞ TRAFİĞİNİN SINIFLANDIRILMASI

BÜŞRA BAKKALCI

Karabük Üniversitesi

Mühendislik Fakültesi

Bilgisayar Mühendisliği Bölümünde

Bitirme Projesi Tezi

Olarak Hazırlanmıştır.

KARABÜK

Mayıs 2020

Büşra BAKKALCI tarafından hazırlanan “MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE AĞ TRAFİĞİNİN SINIFLANDIRILMASI” başlıklı bu projenin Bitirme Projesi Tezi olarak uygun olduğunu onaylarım.

Dr. Öğr. Üyesi Emrullah SONUÇ .....

Bitirme Projesi Danışmanı, Bilgisayar Mühendisliği Anabilim Dalı

...../…../2020

Bilgisayar Mühendisliği bölümü, bu tez ile, Bitirme Projesi Tezini onamıştır.

Dr. Öğr. Üyesi Hakan KUTUCU .....

Bölüm Başkanı

"MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE AĞ TRAFİĞİNİN SINIFLANDIRILMASI" başlıklı projemde tüm bilgilerin akademik kurallara ve etik ilkelere uygun olarak elde edildiğini ve sunulduğunu; ayrıca bu kuralların ve ilkelerin gerektirdiği şekilde, bu çalışmadan kaynaklanmayan bütün atıfları yaptığımı beyan ederim."

Büşra BAKKALCI

ÖZET

Bitirme Projesi Tezi

MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE AĞ TRAFİĞİNİN SINIFLANDIRILMASI

Büşra BAKKALCI

Karabük Üniversitesi

Mühendislik Fakültesi

Bilgisayar Mühendisliği Bölümü

Tez Danışmanı:

Dr. Öğr. Üyesi Emrullah SONUÇ

Mayıs 2020, 35 sayfa

Teknolojinin ilerlemesi ve internetin gelişmesiyle birlikte insanların bilgiye ulaşmaları kolaylaşmış ve veri kavramı ön plana çıkmıştır. İnsanların verilerini paylaşmasıyla da veri yığınları halini almıştır. Bununla birlikte internet dünyasındaki bilgi kirliliği de giderek artmıştır. Bu kirlilik içerisinden anlamlı bilgiler ortaya çıkarabilmek ihtiyacı veri madenciliği ve makine öğrenmesi kavramlarının hayatımıza girmesini sağlamıştır.

Bu çalışmada makine öğrenmesi sınıflandırma algoritmaları kullanılmıştır. Ağ trafik verilerinde sınıflandırma yapılıp, makine öğrenmesi tekniklerinin veri setlerini sınıflandırmadaki başarıları analiz edilmiştir. 17 farklı veri setinde sınıflandırma yapılıp başarı sonuçları çıkarılmıştır.

Anahtar Sözcükler: Makine öğrenmesi, veri madenciliği, sınıflandırma, ağ trafiği

ABSTRACT

Senior Project Thesis

CLASSIFICATION OF NETWORK TRAFFIC WITH MACHINE LEARNING METHODS

Büşra BAKKALCI

Karabük University
Faculty of Engineering
Department of Computer Engineering

Project Supervisor:
Assis. Prof. Dr. Emrullah SONUÇ
May 2020, 35 pages

With the advancement of technology and the development of the internet, people's access to information has become easier and the concept of data has come to the fore. It has also become a pile of data when people share their data. However, information pollution in the internet world has increased steadily. The need to extract meaningful information from this pollution enabled the concepts of data mining and machine learning to enter our lives.

Machine learning classification algorithms are used in this study. Network traffic datas are classified and their success in classifying data sets of machine learning techniques is analyzed. The classification was made in 17 different data sets and success results were obtained.

Key Words : Machine Learning, Data Mining, Classification, Network Traffic

TEŞEKKÜR

Tez sürecim boyunca bana kıymetli zamanını ayırıp desteğini esirgemeyen, tecrübeleriyle bana yol gösteren ve beni her zaman çalışmaya teşvik eden değerli hocam Dr. Öğr. Üyesi Emrullah SONUÇ' a sonsuz teşekkürlerimi sunarım.

İÇİNDEKİLER

Sayfa

Sayfa

ŞEKİLLER DİZİNİ

Sayfa

TABLOLAR DİZİNİ

Sayfa

GRAFİKLER DİZİNİ

Sayfa

KISALTMALAR LİSTESİ

VPN : Virtual Private Network

P2P : Peer to Peer

VoIP : Voice Over Internet Protocol

FT **:** File Transfer

BÖLÜM 1

GİRİŞ

Günümüzde internetin yaygın olarak kullanılmasıyla birlikte insanlar verilere daha kolay ulaşabilmekte, daha çok miktarda veriyi bilgisayarlarında saklayabilmekte ve bu verileri işleyebilmektedir. Hatta dünyadaki bu verilerin %90'ının son iki yılda üretildiği söyleniyor. Bu kadar çok veri arasından anlamlı bilgiler ortaya çıkarmak gerekmektedir. Burada da karşımıza veri madenciliği kavramı çıkmaktadır. Boyut olarak çok büyük olan bu veri yığınlarının bir insan tarafından analiz edilmesi de oldukça zordur. Bunun için makine öğrenmesi algoritmalarından yararlanılmaktadır. Büyüyen verilerde güvenlik açıkları da meydana gelmektedir. Kablosuz ağ bağlantılarında tüm internet trafiğinin dışarıdan izlenme tehdidi vardır. Bu tehdidi en aza indirmek için birçok kişi tarafından kullanılan kablosuz ağları güvenli kullanabilmek için başvurulan yöntemlerden biri VPN' dür. Özel Sanal Ağlar (Virtual Private Network) teknolojisi uçtan uca şifreleme sağlar. Böylelikle verileri koruyarak güvenli iletişim sağlar. Bu çalışmada kullanılan veri seti Kanada Siber Güvenlik Enstitüsü tarafından oluşturulan VPN 2016 veri setidir. Çalışmada 14 farklı ağ trafik kategorisinden oluşan (VoIP, VPN-VoIP, P2P, VPN-P2P vb.) veri setlerinde (VPN-nonVPN dataset (ISCXVPN2016)) sınıflandırma yapılıp makine öğrenmesi algoritmalarının başarıları karşılaştırılmıştır. Yapılan çalışmada 14 ağ trafik kategorisinden oluşan 17 farklı veri setinde sınıflandırma yapılmıştır.

Ağ Trafik Kategorileri

BROWSING	VPN-BROWSING
CHAT	VPN-CHAT
STREAMING	VPN-STREAMING
MAIL	VPN-MAIL
VoIP	VPN-VoIP
P2P	VPN-P2P
FT	VPN-FT

1.1. LİTERATÜR ÖZETİ

Yapılan bu çalışmada ağ trafik kategorilerinin sınıflandırılması için kullanılan makine öğrenmesi yöntemleri araştırılmış, veri setleri ve elde edilen başarılar değerlendirilmiştir.

Serhat ÖZEKES ve Elif Nur KARAKOÇ tarafından yayınlanan "Makine Öğrenmesi Yöntemleriyle Anormal Ağ Trafiğinin Tespit Edilmesi" adlı araştırma makalesinde önerdikleri yöntemin karar ağacı ve rastgele orman algoritmalarıyla sınıflandırılması sonucu, gerçek zamanlı olarak saldırıları tespit etmede %100'e yakın bir başarım elde ettiğini görmüşlerdir.

"Şifreli İnternet Trafiğinin Gerçek Zamanlı Sınıflandırılması" adlı makalede Cihangir Beşiktaş ve Hacı Ali Mantar, makine öğrenmesi tabanlı bir yöntemi ele almaktadırlar. Bu yöntemde yeni bir akışın sınıflandırılması her bir uygulamanın karakteristiği ile yapılan ağırlıklı kosinüs benzerliği hesabına göre yapılmaktadır. Önerdikleri yöntemin yüksek başarı oranına sahip olduğunu ve trafik sınıflandırma performansını arttırdığını öngörmüşlerdir.

Fatih Ertam ve Engin Avcı yayınladıkları "Uç Öğrenme Makineleri Kullanılarak İnternet Trafik Bilgisinin Sınıflandırılması" adlı makalelerinde Destek Vektör Makineleri (SVM) ve Yapay Sinir Ağları (YSA) gibi önceki çalışmalarda oldukça fazla kullanılan klasik sınıflandırıcılar yerine Uç Öğrenme Makinesi (UÖM) algoritmasını kullanarak UÖM ile yapılan sınıflandırmada başarının daha yüksek olduğunu görmüşlerdir.

Fatih ERTAM “Kurumsal Bilgisayar Ağlarındaki Trafik Bilgisinin Akıllı Sistemler İle Sınıflandırılması” adlı çalışmasında internet trafiğinin sınıflandırılması için daha önce hiç kullanılmamış olan Uç Öğrenme Makineleri (UÖM) yöntemlerini kullanmıştır. UÖM’ün başarısını karşılaştırabilmek amacıyla da daha önce kullanılan makine öğrenmesi sınıflandırma algoritmalarından Destek Vektör Makineleri (SVM) ve Naive Bayes algoritmalarını veri setlerine uygulayarak karşılaştırmıştır.

UÖM algoritmalarıyla yapılan sınıflandırmanın diğer makine öğrenmesi algoritmalarına göre daha hızlı sonuçlar verdiğini gözlemlemiştir.

1.2. PROJENİN AMACI

Dünya üzerindeki birçok bilgisayar ağının aynı anda birbirlerine bağlanması sonucu veriler gün geçtikçe büyümektedir. Artan veri miktarının kontrolü ve verilere ulaşmak da zorlaşmaktadır. Bu büyük miktardaki verilerden anlamlı bilgiler ortaya çıkartmak, verileri kontrol etmek ve analiz etmek güçleşmektedir.

Ayrıca bazı ağlarda güvenlik açıkları meydana gelmekte ve bu güvenlik açıklarını kontrol etmek de zorlaşmaktadır. Bu yüzden ağlarda VPN hizmeti kullanarak, bağlantımızı güvenli hale getirir ve güvenli olmayan herhangi bir ağa bağlanırken bağlantımızı şifreleyerek kimliğimizin tespit edilememesini sağlarız.

Ağ trafiğinin güvenli bir şekilde sağlanabilmesi ve güvenli olmayan ağların tespit edilip sınıflandırılması amacıyla bu çalışma gerçekleştirilmiştir. Makine öğrenmesi yöntemleri ile ağ trafik kategorileri üzerinde sınıflandırma yapmak ve algoritmaların sınıflandırmadaki başarılarını değerlendirmek amaçlanmıştır.

BÖLÜM 2

MAKİNE ÖĞRENMESİ

Çok büyük miktardaki verilerin elle işlenmesi ve analiz edilmesi mümkün değildir. Bu yüzden makine öğrenmesi yöntemleri geliştirilmiştir. Makine öğrenmesi yöntemleri, geçmişteki veriyi kullanarak yeni veri için en uygun modeli bulmaya çalışır. Algoritmayı veri ile besleriz ve bu şekilde algoritma bu veriye dayanarak kendi mantığını oluşturur. Makine Öğrenmesi 3'e ayrılır;

1- Supervised Learning (Gözetimli Öğrenme)

Sınıflandırma (Classification)

- Logistic Regression
- Naive Bayes Algoritması
- K-NN Algoritması
- SVM (Support Vector Machines)
- Decision Tree
- Random Forest

Regresyon

- Linear Regression
- Logistic Regression

2- Unsupervised Learning (Gözetimsiz Öğrenme)

Kümeleme (Clustering)

- K-Means Algoritması
- Hierarchical Clustering

Birliktelik Kuralı (Association Rule)

- Apriori

Boyut Azaltma

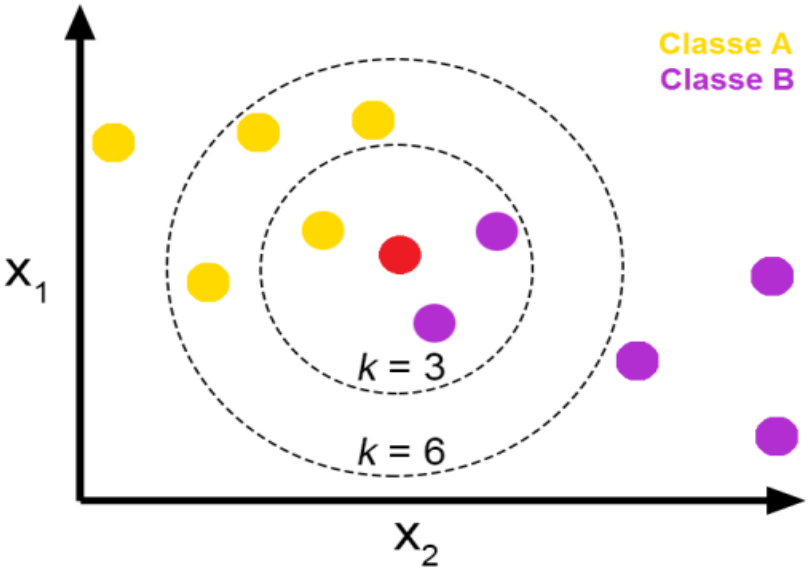
- PCA

3- Reinforcement Learning (Pekiştirmeli-Takviyeli Öğrenme)

Deep Learning (Derin Öğrenme) Algoritmaları

2.1. PROJEDE KULLANILAN YÖNTEMLER

2.1.1. K-NN (K Nearest Neighborhood) Algoritması



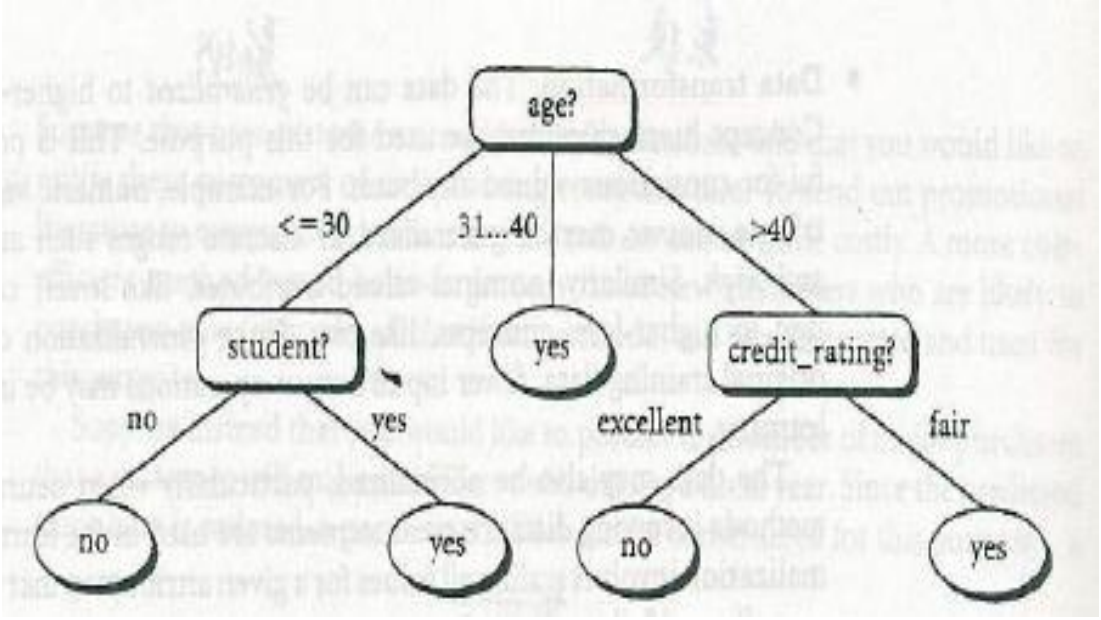
Şekil 1: K-NN Algoritması

İki boyutlu uzaydan belirlenmiş verilerimizin yeni bir veri geldiğinde nasıl sınıflandırılacağını belirleyen metottur. K değeri kadar elemanla sınıflandırma yapar. Örneğin k değeri 3 ise öklid uzaklığına göre kendine en yakın 3 noktayla sınıflandırma yapar.

Algoritma 4 adımdan oluşur;

1. Öncelikle K değeri belirlenir.
2. Diğer nesnelerden hedef nesneye olan öklid uzaklıkları hesaplanır.
3. Uzaklıklar sıralanır ve minimum uzaklığa bağlı olarak en yakın komşular bulunur.
4. En yakın komşular bulunup, sınıf belirlenir.

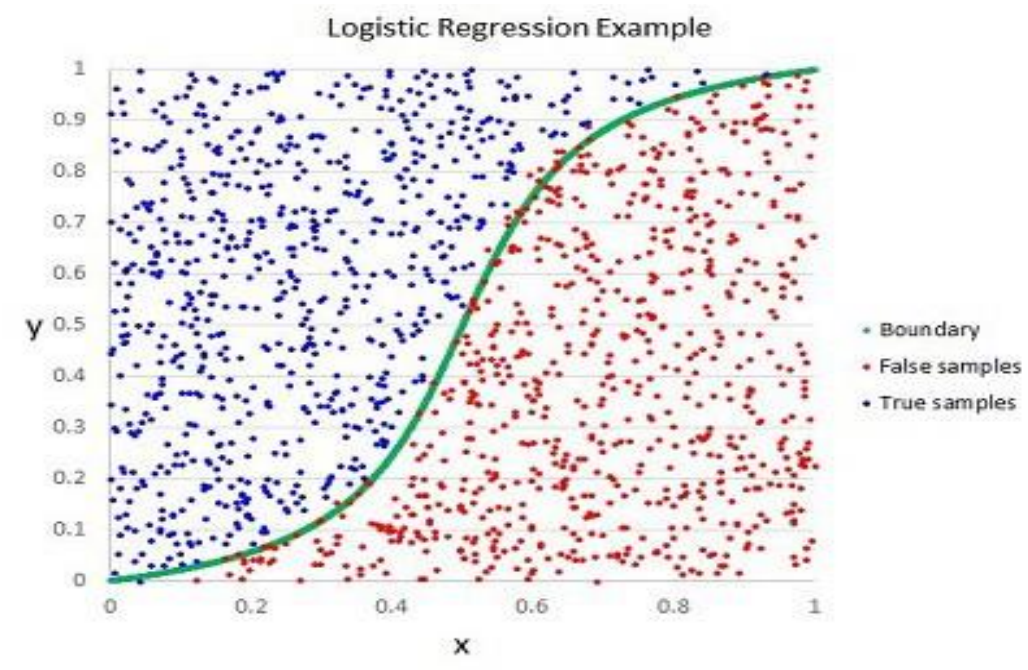
2.1.2. Karar Ağaçları (Decision Trees)



Şekil 2: Karar Ağaçları

Karar ağaçları; sınıflandırma, özellik ve hedefe göre karar düğümleri (decision nodes) ve yaprak düğümlerinden (leaf nodes) oluşan ağaç yapısı formunda bir model oluşturan sınıflandırma yöntemidir. Karar ağaçları, veri setini küçük parçalara bölerek geliştirir. Bir karar düğümü bir veya birden fazla dallanma içerebilir. İlk düğüme kök düğüm (root node) denir. Algoritmada amaç; karar vermeye en çok etki eden nitelikleri köklerde bulundurup, karmaşıklığı (entropy) en aza indirmektir.Ağaç tabanlı yöntemler; yüksek doğruluk, kararlılık ve yorumlanma kolaylığına sahiptir.

2.1.3. Lojistik Regresyon (Logistic Regression)



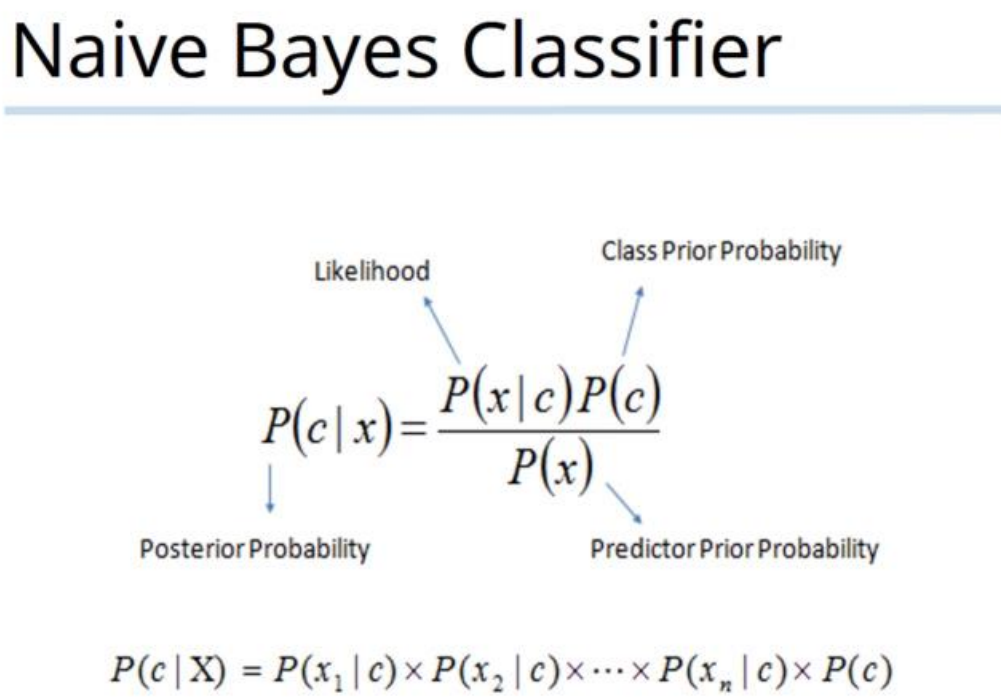
Şekil 3: Lojistik Regresyon

Lojistik regresyon, bir sonucu belirleyen bir veya daha fazla bağımsız değişken bulunan bir veri kümesini analiz etmek için kullanılan istatistiksel bir yöntemdir.

Lojistik regresyon, ikili sınıfları (doğru-yanlış, 0-1) öngörmek için kullanılır.

Lojistik regresyonun amacı; iki yönlü karakteristiği (bağımlı değişken = yanıt veya sonuç değişkeni) ile ilgili bir dizi bağımsız (açıklayıcı) değişken arasındaki ilişkiyi tanımlamak için en uygun logaritmik modeli (eğriyi) bulmaktır.

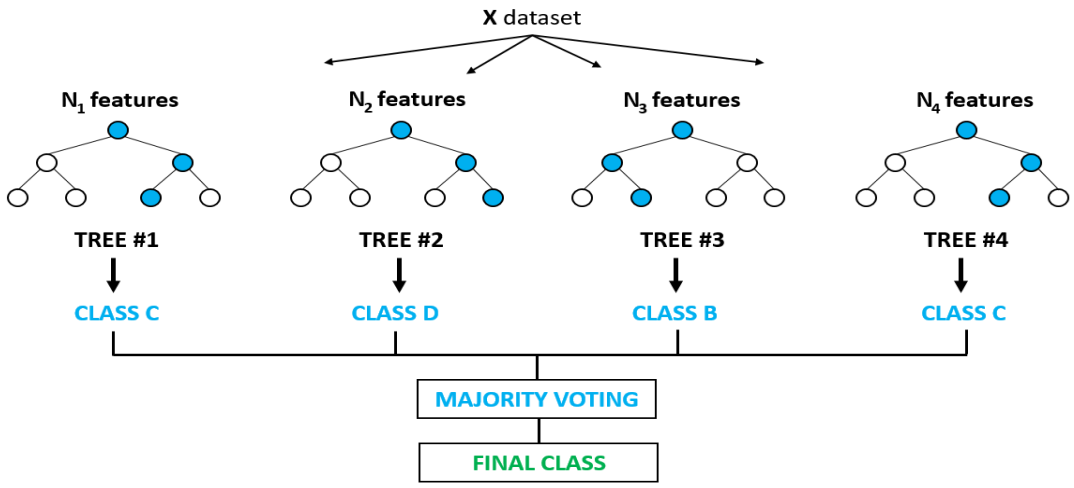
2.1.4. Naive Bayes Algoritması



Şekil 4: Bayes Teoremi

Naive Bayes sınıflandırıcısı, sınıflandırma görevi için kullanılan olasılıklı bir makine öğrenme modelidir. Sınıflandırıcının noktası, Bayes teoremine dayanır. Bayes teoremini kullanarak, x oluşumu göz önüne alındığında, c olma olasılığını bulabiliriz. Burada x kanıt, c ise hipotezdir. Algoritmanın çalışması; bir eleman için her durumun olasılığını hesaplar ve olasılık değeri en yüksek olana göre sınıflandırır.

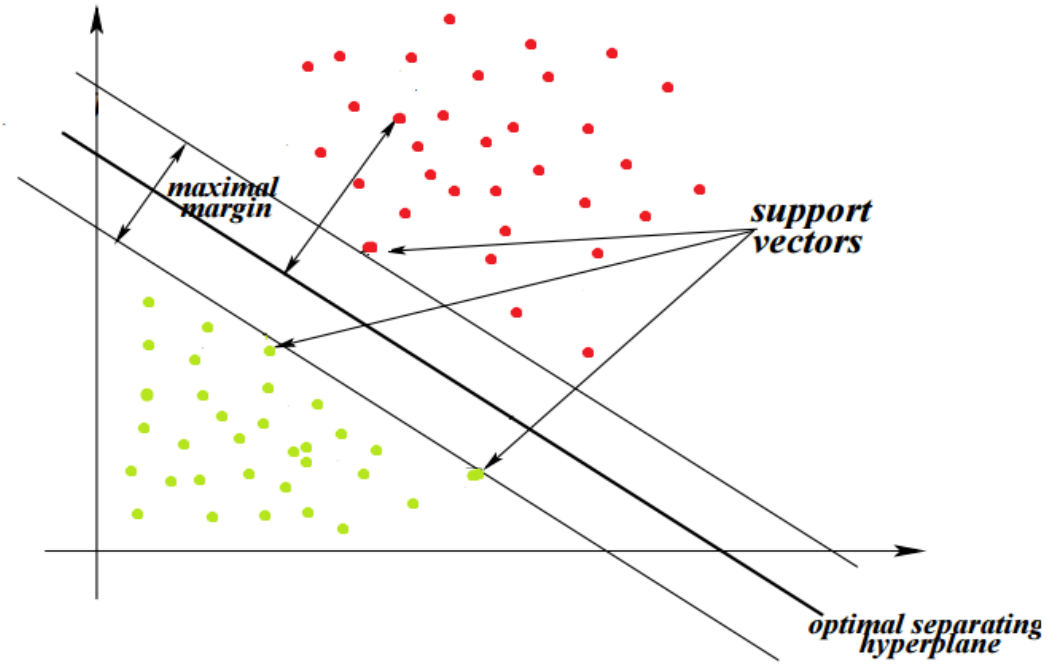
2.1.5. Random Forest (Rastgele Orman) Algoritması



Şekil 5: Random Forest Algoritması

Rastgele Orman denetimli bir öğrenme algoritmasıdır. Bir orman oluşturur ve bunu bir şekilde rastgele yapar. Kurduğu orman, çoğu zaman "bagging" yöntemiyle eğitilen karar ağaçları topluluğudur. Bagging yönteminin genel fikri, öğrenme modellerinin bir kombinasyonunun genel sonucu arttırmasıdır. Rastgele orman algoritması; birden fazla karar ağacı oluşturur ve daha doğru bir tahmin elde etmek için onları birleştirir.

2.1.6. Destek Vektör Makineleri (Support Vector Machines)

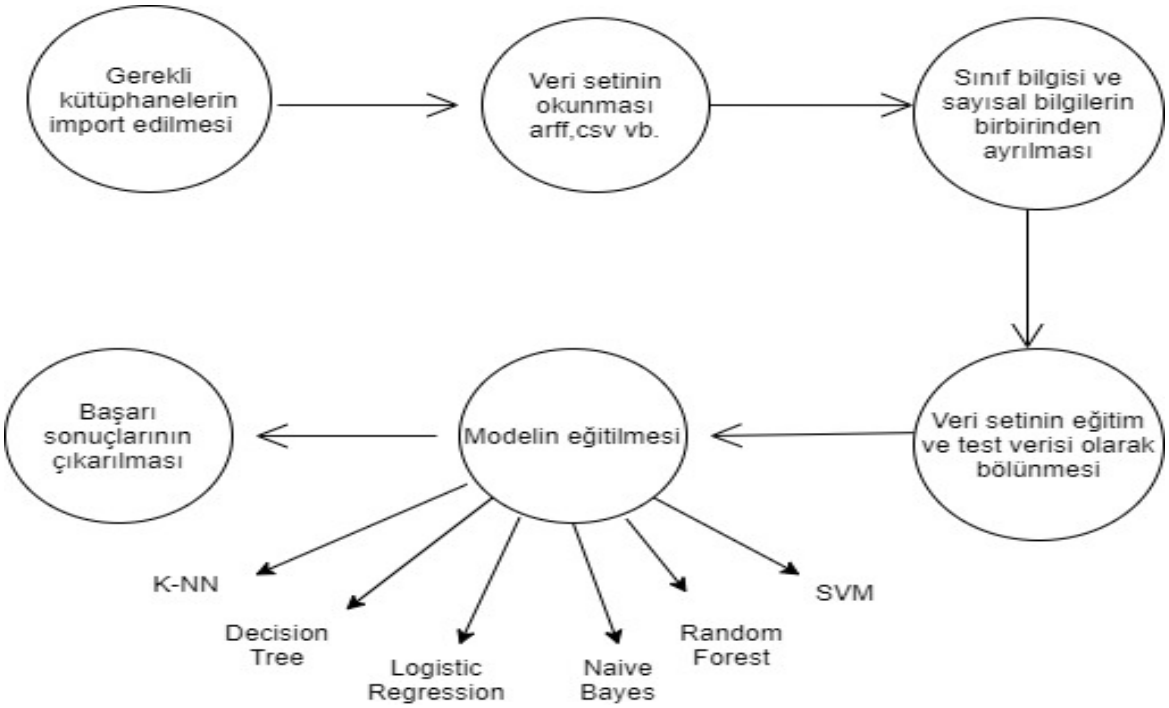


Şekil 6: Destek Vektör Makineleri ile Ayırma

Destek Vektör Makinesi (SVM), sınıflandırma veya regresyon problemleri için kullanılabilen denetimli bir makine öğrenmesi algoritmasıdır. Bununla birlikte, çoğunlukla sınıflandırma problemlerinde kullanılır. Bu algoritmada, her bir veri belirli bir koordinatın değeri olan her özelliğin değeri ile birlikte n-boyutlu boşluğa (n: sahip olduğunuz özelliklerin sayısı) bir nokta olarak çizilir. Ardından, iki sınıftan oldukça iyi ayrım yapan hiper düzlemi bularak sınıflandırma gerçekleştirilir. Destek vektörleri, sadece gözlemin koordinatlarıdır. Destek Vektör Makinesi, iki sınıfı (hiper düzlem / çizgi) en iyi ayıran bir sınırdır.

BÖLÜM 3

YAZILIMIN TASARIMI

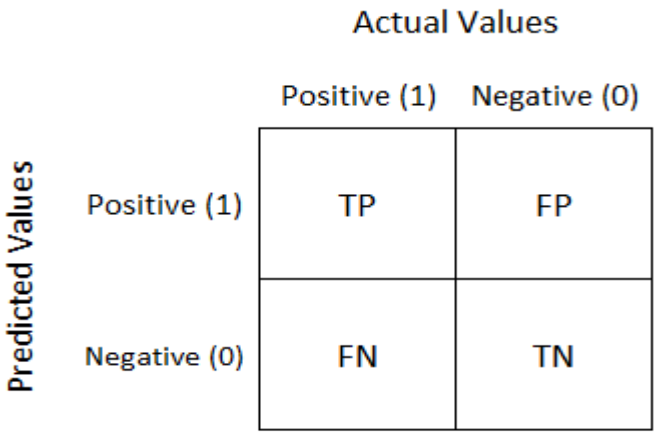


Şekil 7: Yazılımın Tasarımı

- Gerekli kütüphanelerin projeye eklenmesi
- Veri setinin okunması
- Sınıf bilgisi ve sayısal bilgilerin birbirinden ayrılması
- Veri setinin eğitim ve test verisi olarak bölünmesi
- Projede kullanılan modelin (algoritmanın) kodlanması ve eğitilmesi
- Başarı sonuçlarının çıkarılması

PERFORMANS DEĞERLENDİRME ÖLÇÜTLERİ

Confusion Matrix (Hata Matrisi)



Şekil 8: Hata Matrisi

1. Doğruya doğru demek (True Positive – TP) DOĞRU
2. Doğruya yanlış demek (True Negative – TN) YANLIŞ
3. Yanlışa doğru demek (False Positive – FP) YANLIŞ
4. Yanlışa yanlış demek (False Negative – FN) DOĞRU

Doğruluk Oranı (Accuracy Rate):

Doğru sınıflandırmanın toplama bölümüdür.

(TP+TN) / (TP+TN+FP+FN)

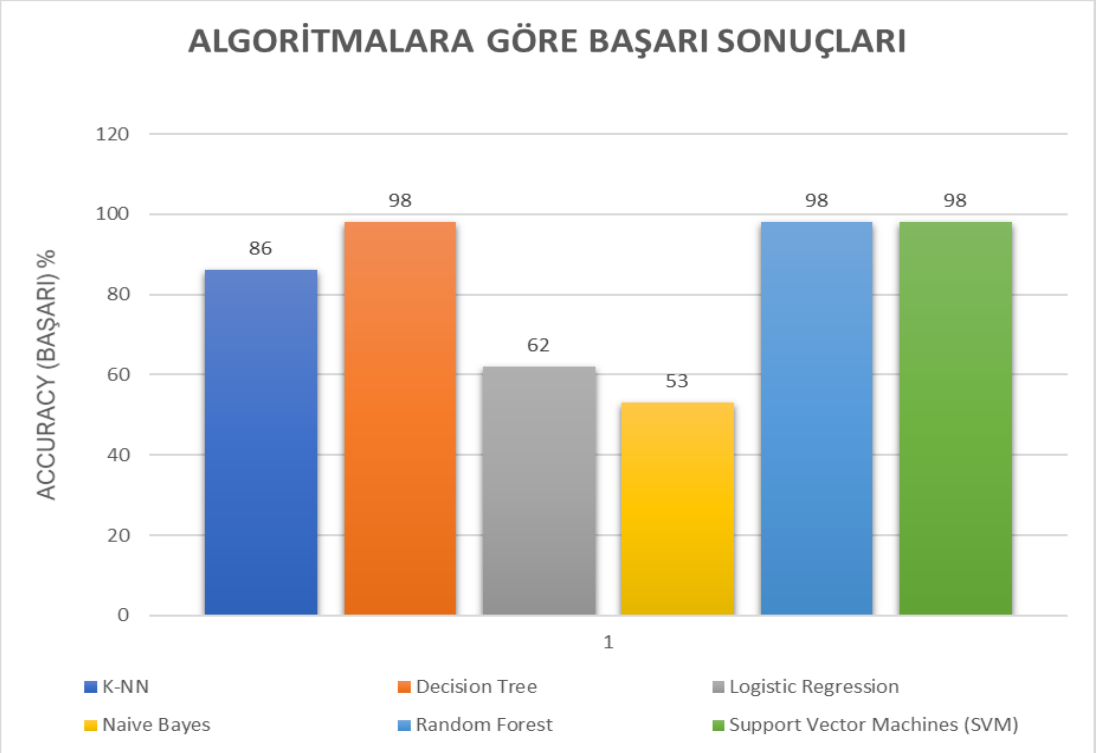
BÖLÜM 4

SONUÇLAR

Scenario A1

TimeBasedFeatures-Dataset-15s-VPN.arff

Toplam Örnek Sayısı = 18.758



Grafik 1: Algoritmalara Göre Başarı Sonuçları

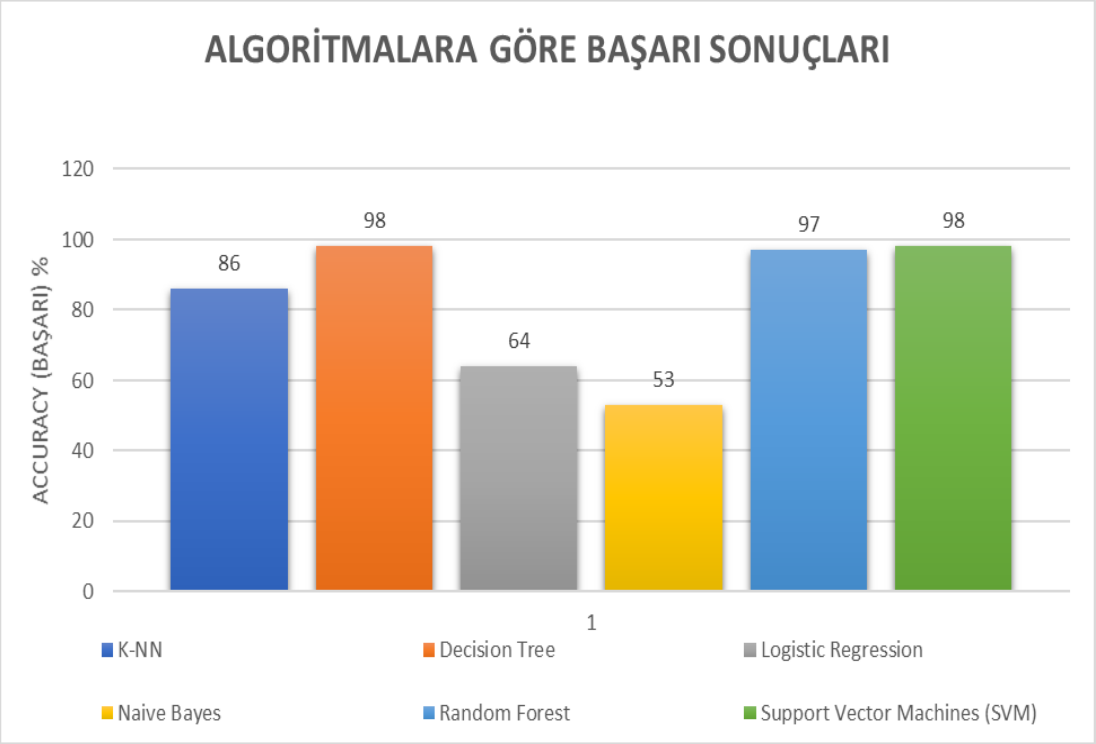
Algoritmalar	Hatalı Sınıflandırılan Örnek Sayısı
K-NN	2465
Decision Tree	231
Logistic Regression	6975
Naive Bayes	8768
Random Forest	307
SVM	231

Tablo 1: Algoritmalara Göre Hatalı Sınıflandırılan Örnek Sayıları

Scenario A1

TimeBasedFeatures-Dataset-30s-VPN.arff

Toplam Örnek Sayısı = 14.651



Grafik 2: Algoritmalara Göre Başarı Sonuçları

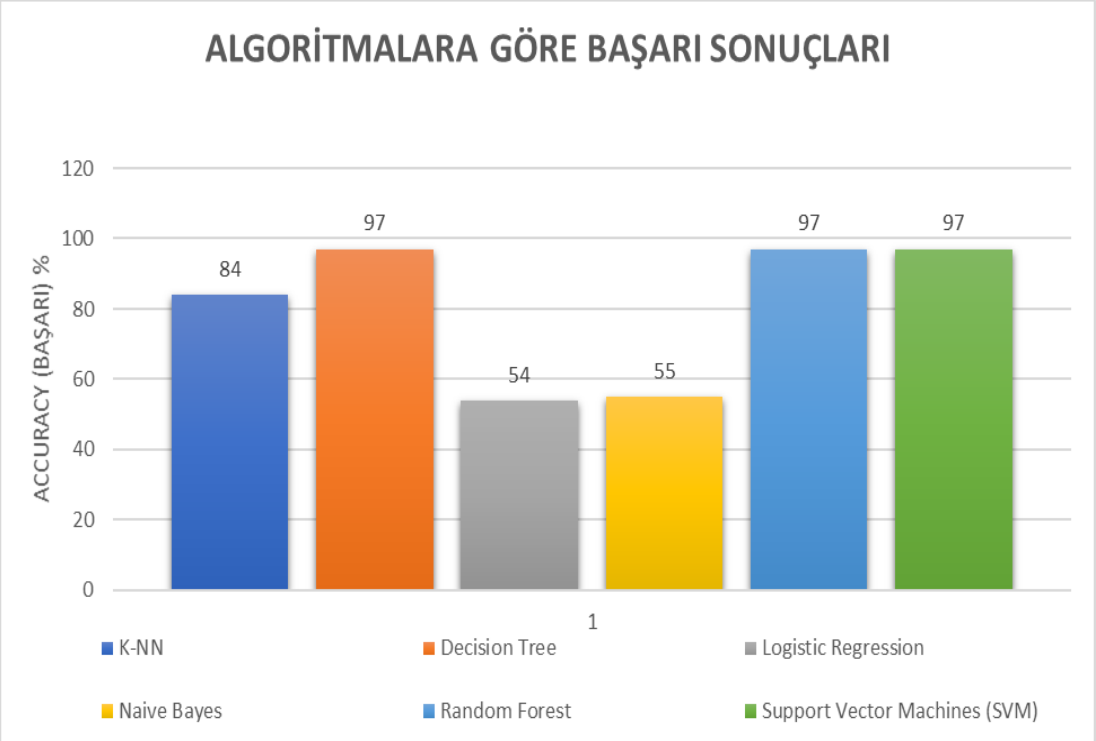
Algoritmalar	Hatalı Sınıflandırılan Örnek Sayısı
K-NN	2023
Decision Tree	229
Logistic Regression	5188
Naive Bayes	6820
Random Forest	296
SVM	229

Tablo 2: Algoritmalara Göre Hatalı Sınıflandırılan Örnek Sayıları

Scenario A1

TimeBasedFeatures-Dataset-120s-VPN.arff

Toplam Örnek Sayısı = 10.782



Grafik 3: Algoritmalara Göre Başarı Sonuçları

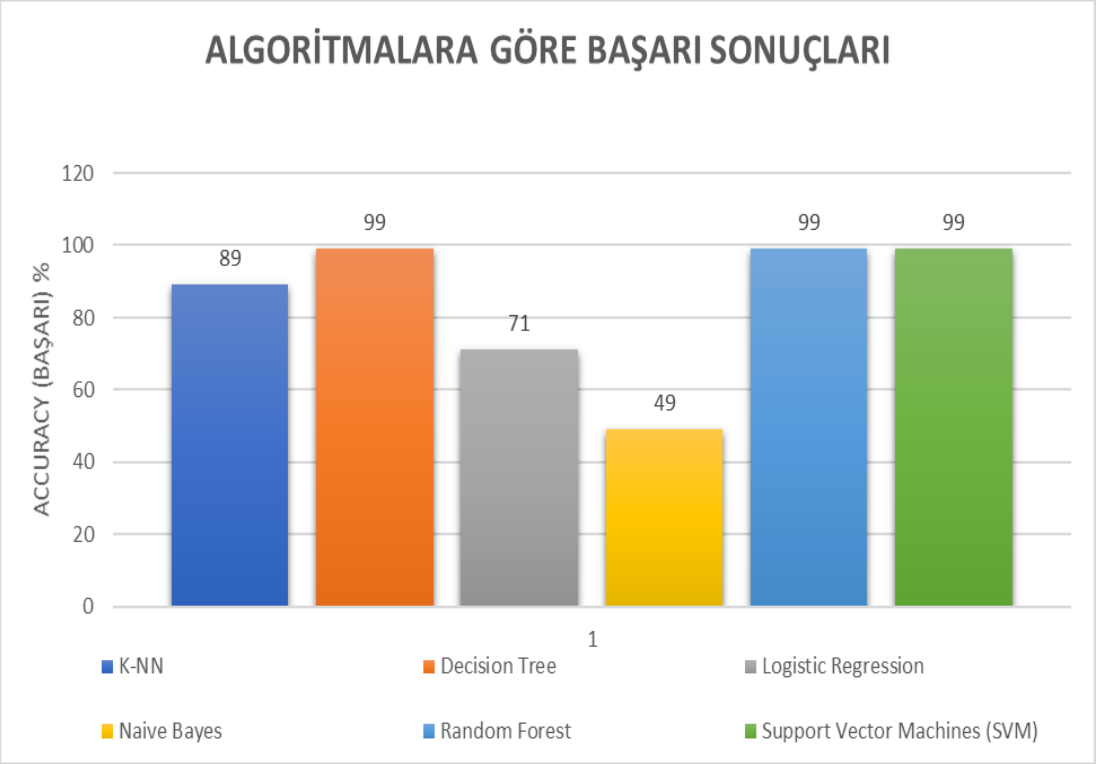
Algoritmalar	Hatalı Sınıflandırılan Örnek Sayıları
K-NN	1640
Decision Tree	232
Logistic Regression	4920
Naive Bayes	4796
Random Forest	269
SVM	232

Tablo 3: Algoritmalara Göre Hatalı Sınıflandırılan Örnek Sayıları

Scenario A2

TimeBasedFeatures-Dataset-15s-NO-VPN.arff

Toplam Örnek Sayısı = 8.965



Grafik 4: Algoritmalara Göre Başarı Sonuçları

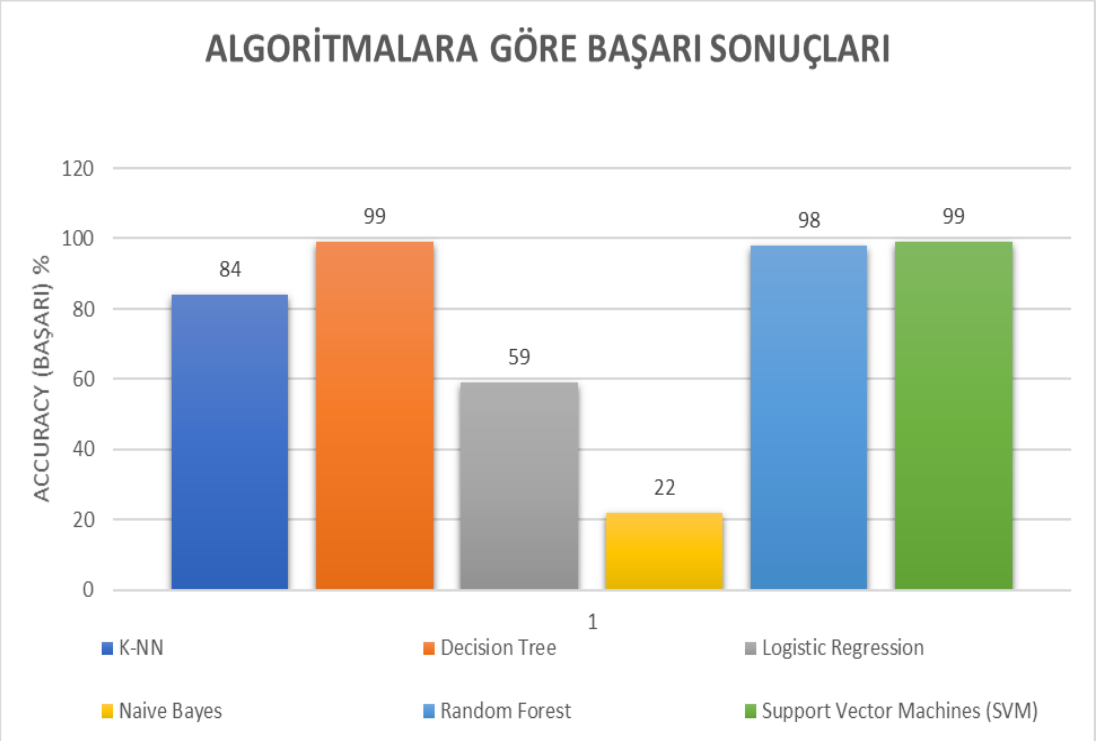
Algoritmalar	Hatalı Sınıflandırılan Örnek Sayısı
K-NN	947
Decision Tree	52
Logistic Regression	2564
Naive Bayes	4492
Random Forest	84
SVM	52

Tablo 4: Algoritmalara Göre Hatalı Sınıflandırılan Örnek Sayıları

Scenario A2

TimeBasedFeatures-Dataset-15s-VPN.arff

Toplam Örnek Sayısı = 9.793



Grafik 5: Algoritmalara Göre Başarı Sonuçları

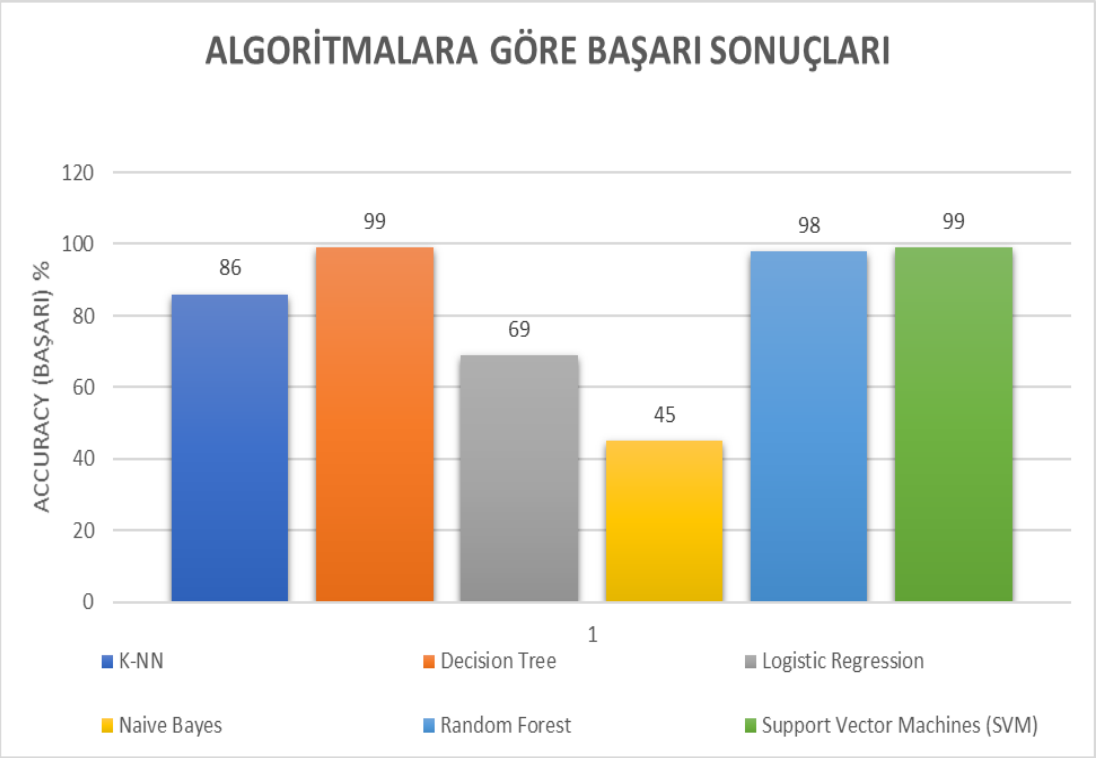
Algoritmalar	Hatalı Sınıflandırılan Örnek Sayısı
K-NN	1496
Decision Tree	89
Logistic Regression	3966
Naive Bayes	7581
Random Forest	130
SVM	89

Tablo 5: Algoritmalara Göre Hatalı Sınıflandırılan Örnek Sayıları

Scenario A2

TimeBasedFeatures-Dataset-30s-NO-VPN.arff

Toplam Örnek Sayısı = 6.917



Grafik 6: Algoritmalara Göre Başarı Sonuçları

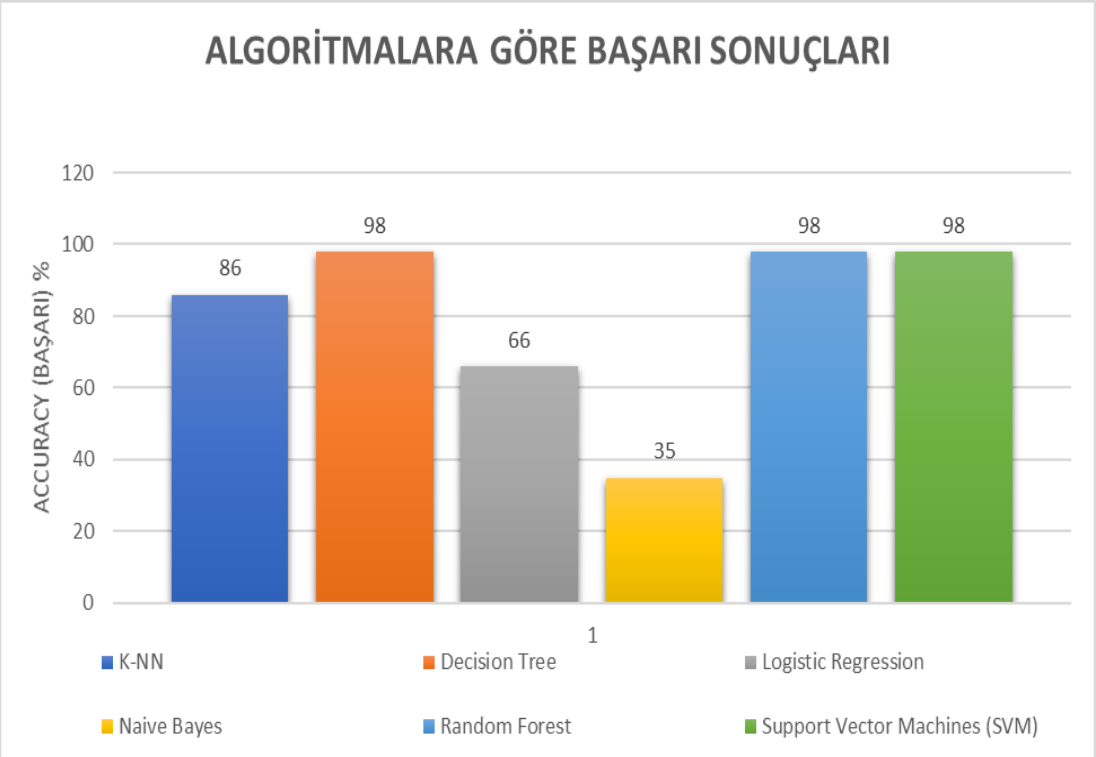
Algoritmalar	Hatalı Sınıflandırılan Örnek Sayısı
K-NN	963
Decision Tree	47
Logistic Regression	2087
Naive Bayes	3803
Random Forest	80
SVM	47

Tablo 6: Algoritmalara Göre Hatalı Sınıflandırılan Örnek Sayıları

Scenario A2

TimeBasedFeatures-Dataset-30s-VPN.arff

Toplam Örnek Sayısı = 7.734



Grafik 7: Algoritmalara Göre Başarı Sonuçları

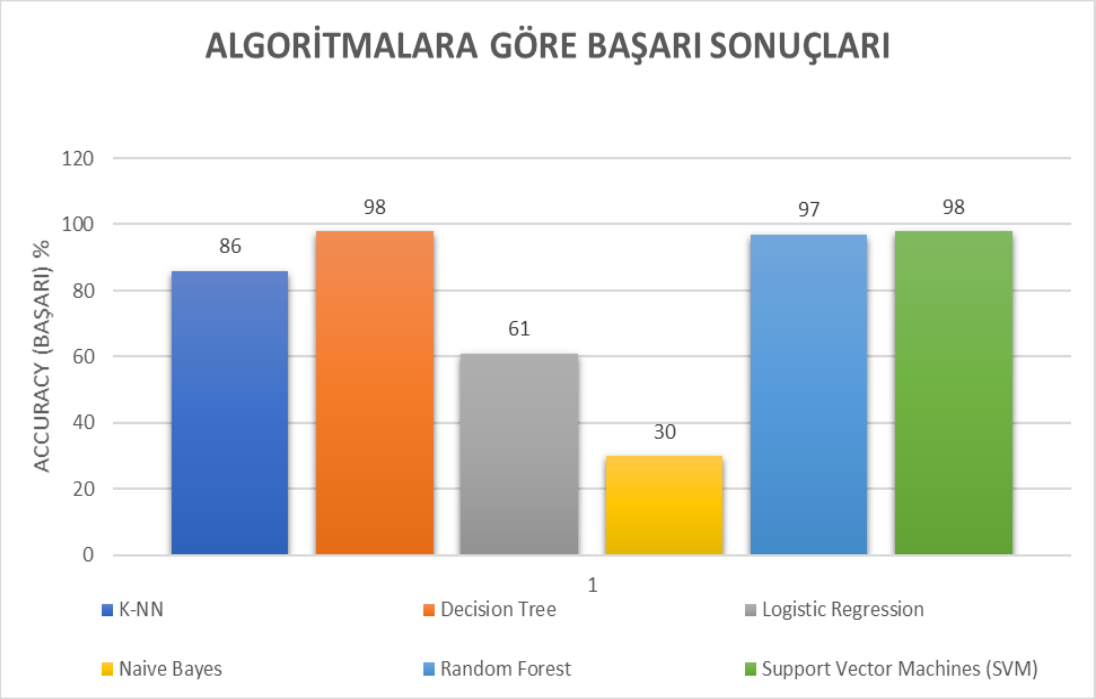
Algoritmalar	Hatalı Sınıflandırılan Örnek Sayısı
K-NN	1077
Decision Tree	80
Logistic Regression	2570
Naive Bayes	5003
Random Forest	125
SVM	80

Tablo 7: Algoritmalara Göre Hatalı Sınıflandırılan Örnek Sayıları

Scenario A2

TimeBasedFeatures-Dataset-60s-NO-VPN.arff

Toplam Örnek Sayısı = 8.580



Grafik 8: Algoritmalara Göre Başarı Sonuçları

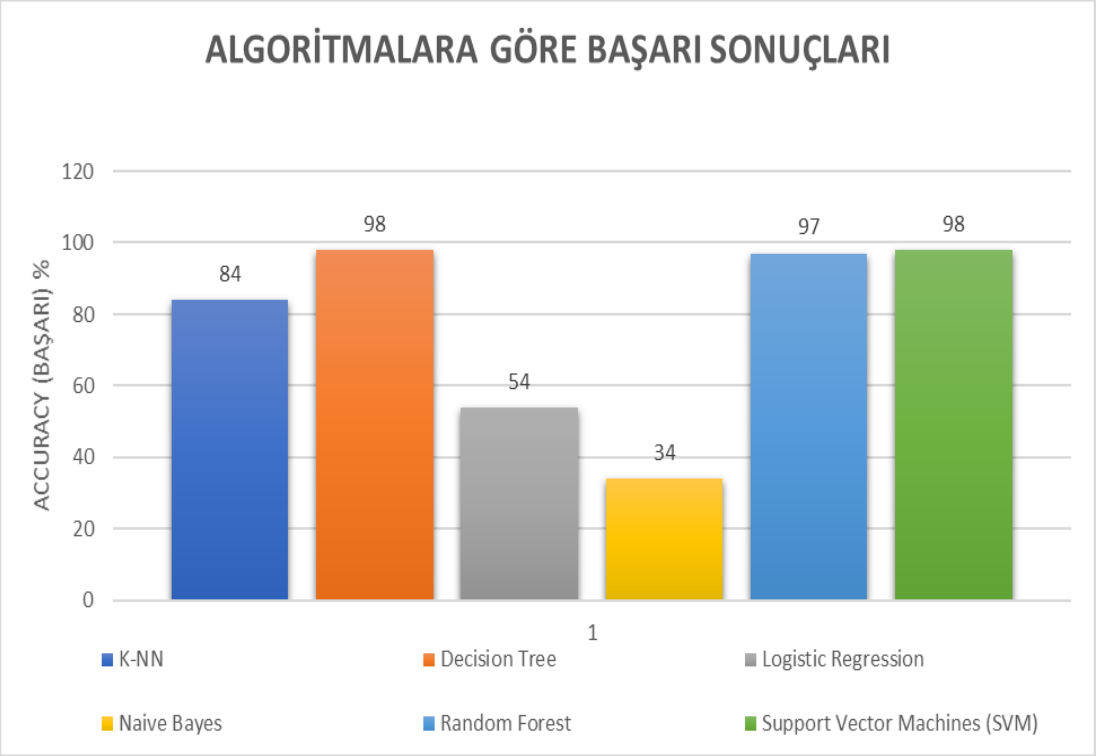
Algoritmalar	Hatalı Sınıflandırılan Örnek Sayısı
K-NN	1198
Decision Tree	155
Logistic Regression	3293
Naive Bayes	5944
Random Forest	185
SVM	155

Tablo 8: Algoritmalara Göre Hatalı Sınıflandırılan Örnek Sayıları

Scenario A2

TimeBasedFeatures-Dataset-60s-VPN.arff

Toplam Örnek Sayısı = 6.935



Grafik 9: Algoritmalara Göre Başarı Sonuçları

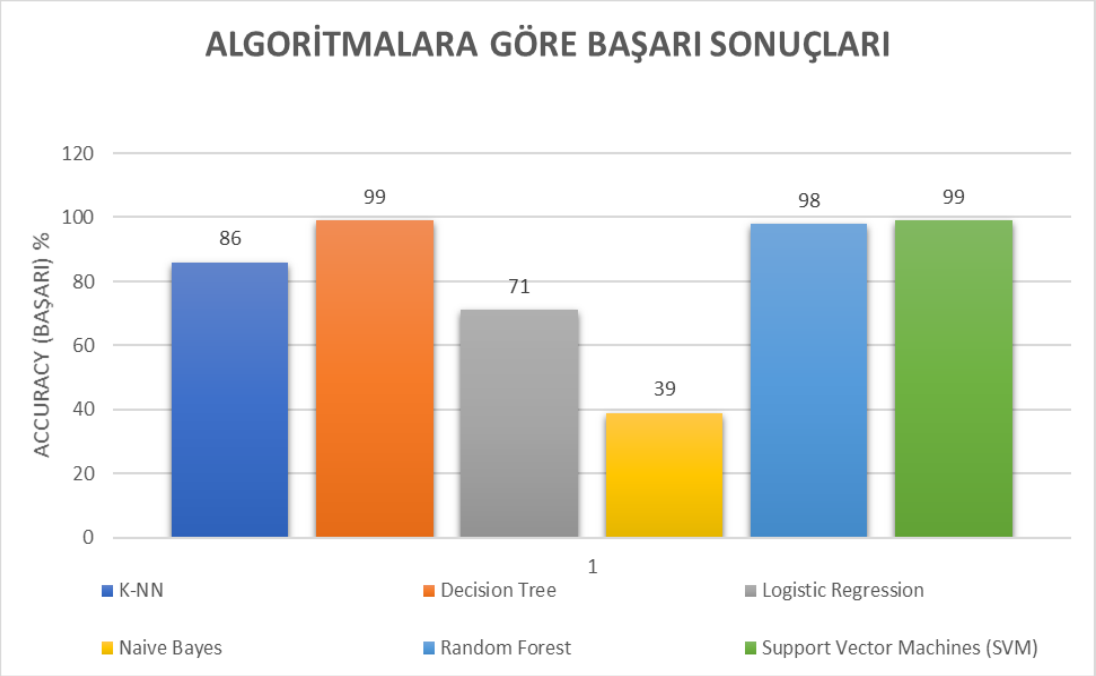
Algoritmalar	Hatalı Sınıflandırılan Örnek Sayısı
K-NN	1098
Decision Tree	103
Logistic Regression	3173
Naive Bayes	4520
Random Forest	147
SVM	104

Tablo 9: Algoritmalara Göre Hatalı Sınıflandırılan Örnek Sayıları

Scenario A2

TimeBasedFeatures-Dataset-120s-NO-VPN.arff

Toplam Örnek Sayısı = 5.151



Grafik 10: Algoritmalara Göre Başarı Sonuçları

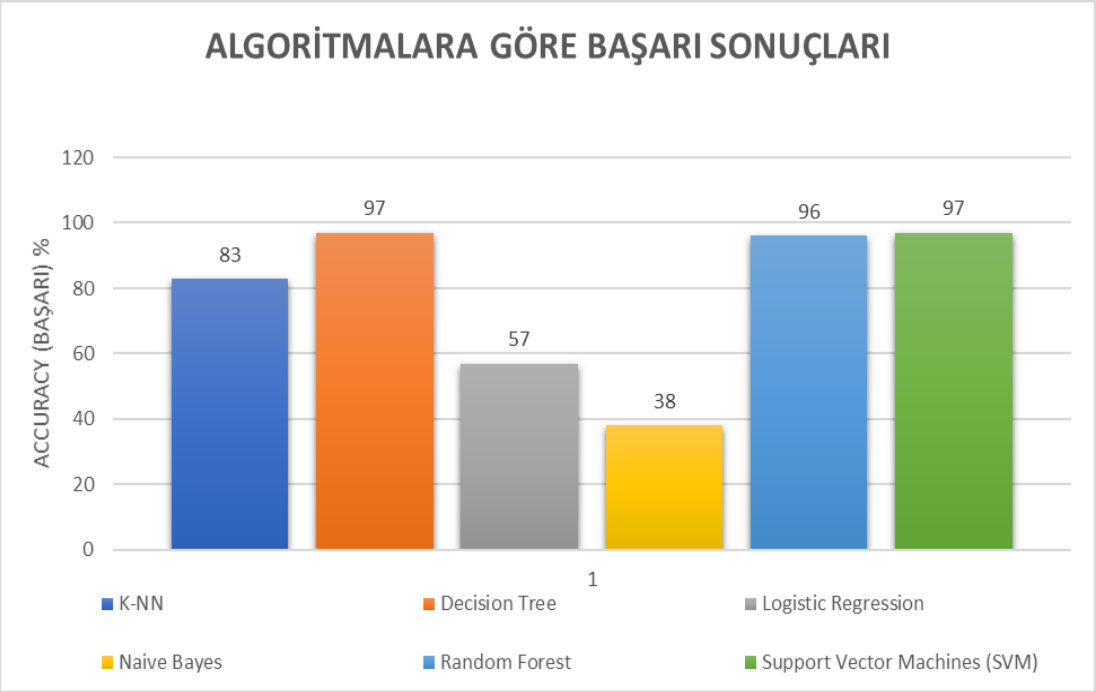
Algoritmalar	Hatalı Sınıflandırılan Örnek Sayısı
K-NN	721
Decision Tree	44
Logistic Regression	1449
Naive Bayes	3100
Random Forest	69
SVM	44

Tablo 10: Algoritmalara Göre Hatalı Sınıflandırılan Örnek Sayıları

Scenario A2

TimeBasedFeatures-Dataset-120s-VPN.arff

Toplam Örnek Sayısı = 5.631



Grafik 11: Algoritmalara Göre Başarı Sonuçları

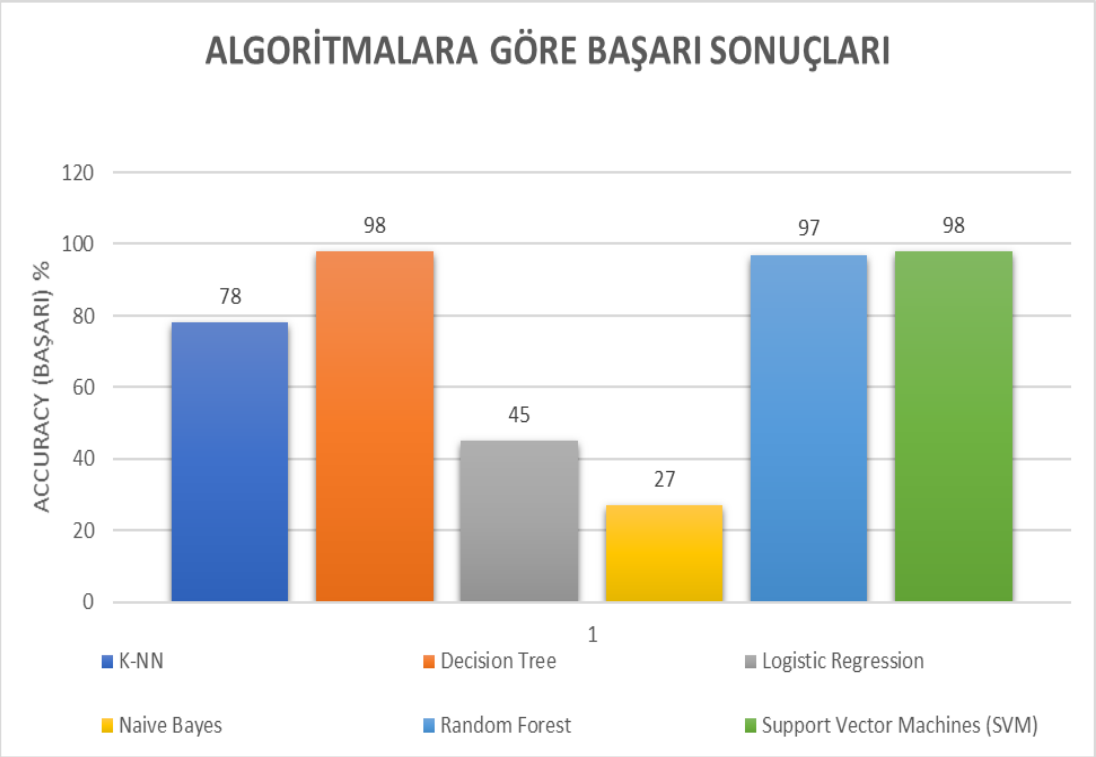
Algoritmalar	Hatalı Sınıflandırılan Örnek Sayısı
K-NN	949
Decision Tree	154
Logistic Regression	2384
Naive Bayes	3440
Random Forest	195
SVM	155

Tablo 11: Algoritmalara Göre Hatalı Sınıflandırılan Örnek Sayıları

Scenario B

TimeBasedFeatures-Dataset-15s.arff

Toplam Örnek Sayısı = 18.758



Grafik 12: Algoritmalara Göre Başarı Sonuçları

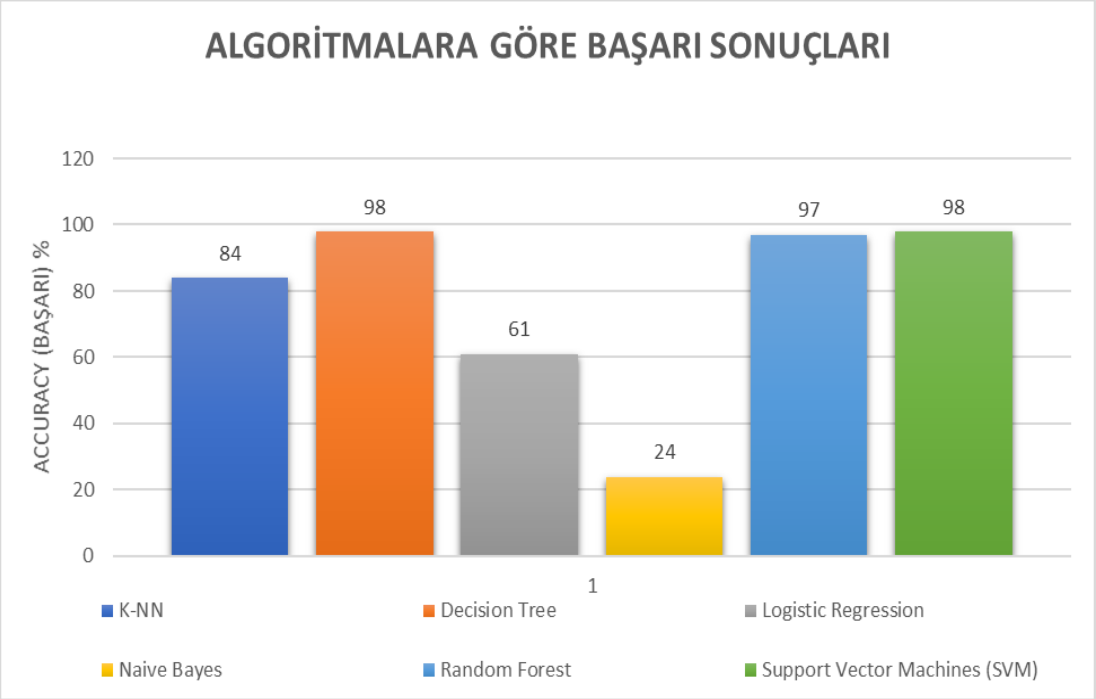
Algoritmalar	Hatalı Sınıflandırılan Örnek Sayısı
K-NN	3989
Decision Tree	321
Logistic Regression	10316
Naive Bayes	13616
Random Forest	424
SVM	323

Tablo 12: Algoritmalara Göre Hatalı Sınıflandırılan Örnek Sayıları

Scenario B

TimeBasedFeatures-Dataset-15s-AllinOne.arff

Toplam Örnek Sayısı = 18.758



Grafik 13: Algoritmalara Göre Başarı Sonuçları

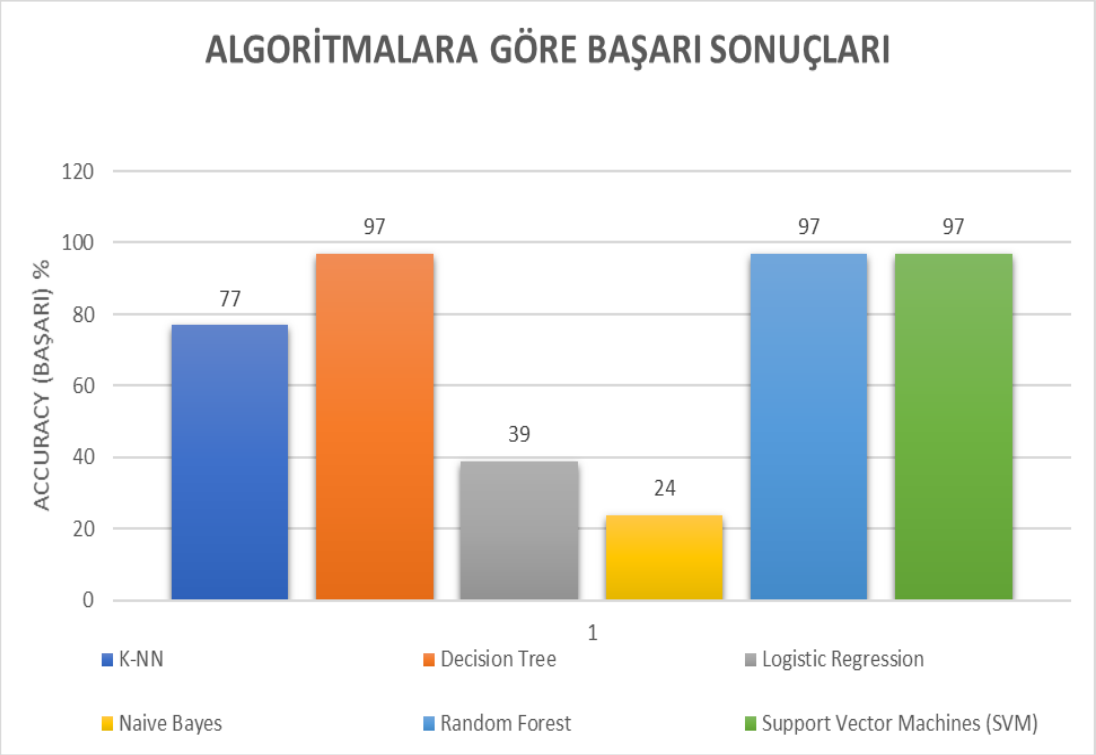
Algoritmalar	Hatalı Sınıflandırılan Örnek Sayısı
K-NN	2844
Decision Tree	296
Logistic Regression	7152
Naive Bayes	14179
Random Forest	406
SVM	296

Tablo 13: Algoritmalara Göre Hatalı Sınıflandırılan Örnek Sayıları

Scenario B

TimeBasedFeatures-Dataset-30s.arff

Toplam Örnek Sayısı = 14.651



Grafik 14: Algoritmalara Göre Başarı Sonuçları

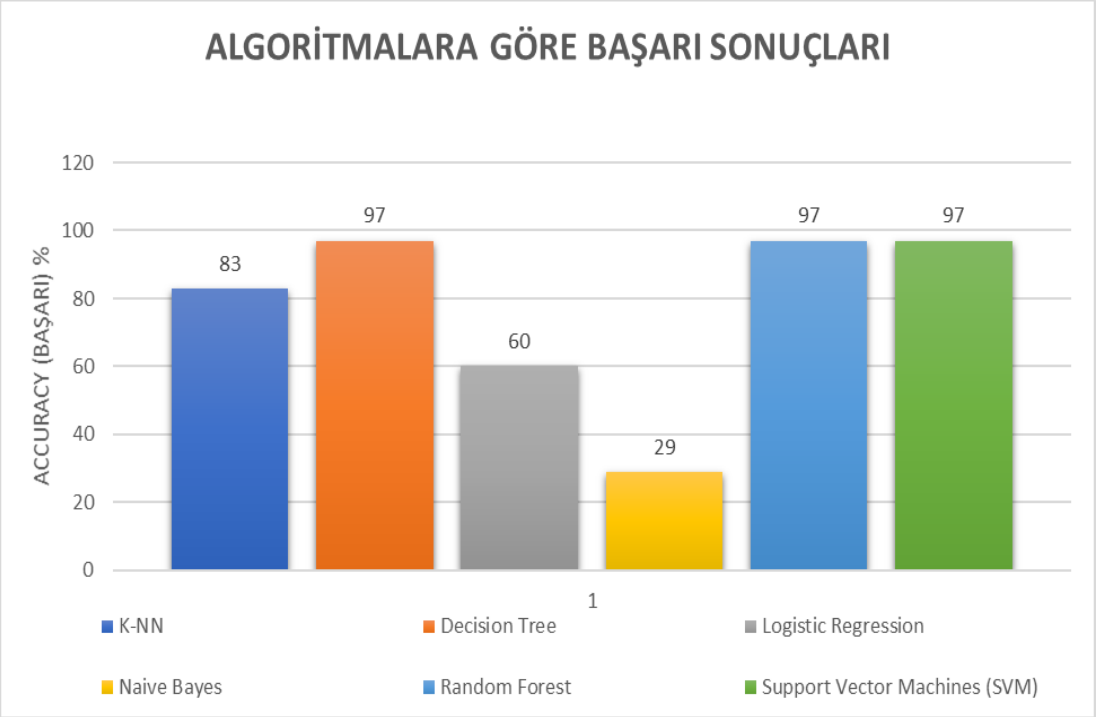
Algoritmalar	Hatalı Sınıflandırılan Örnek Sayısı
K-NN	3345
Decision Tree	311
Logistic Regression	8809
Naive Bayes	11029
Random Forest	401
SVM	311

Tablo 14: Algoritmalara Göre Hatalı Sınıflandırılan Örnek Sayıları

Scenario B

TimeBasedFeatures-Dataset-30s-AllinOne.arff

Toplam Örnek Sayısı = 14.651



Grafik 15: Algoritmalara Göre Başarı Sonuçları

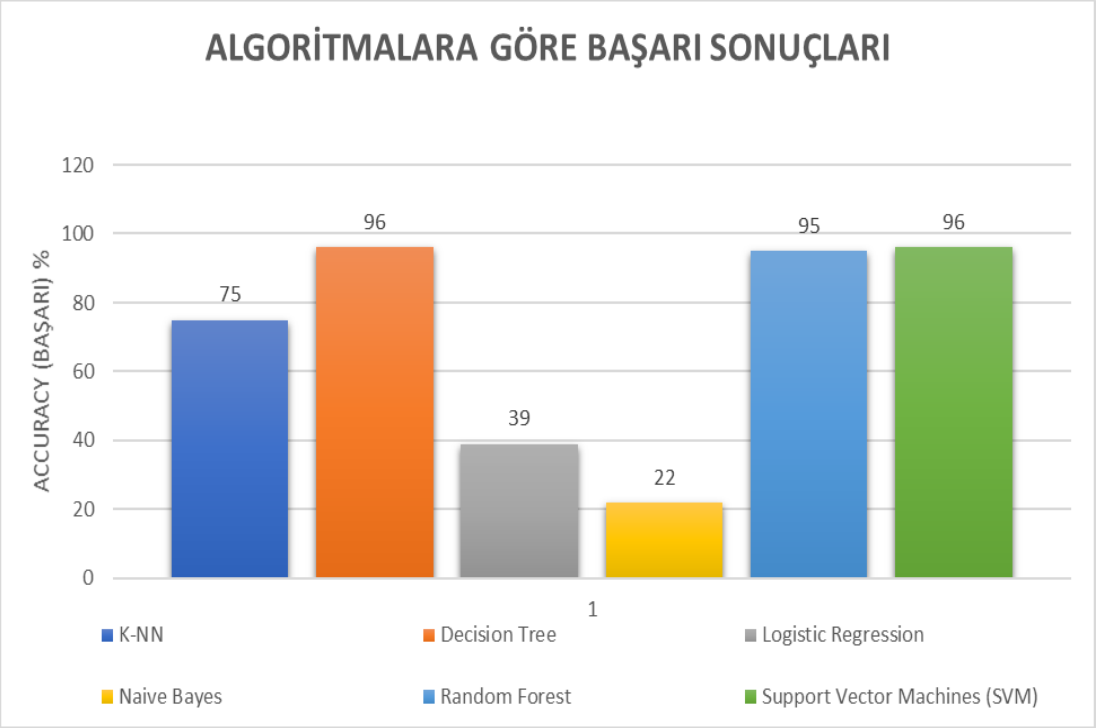
Algoritmalar	Hatalı Sınıflandırılan Örnek Sayısı
K-NN	2435
Decision Tree	301
Logistic Regression	5803
Naive Bayes	10367
Random Forest	375
SVM	301

Tablo 15: Algoritmalara Göre Hatalı Sınıflandırılan Örnek Sayıları

Scenario B

TimeBasedFeatures-Dataset-120s.arff

Toplam Örnek Sayısı = 10.782



Grafik 16: Algoritmalara Göre Başarı Sonuçları

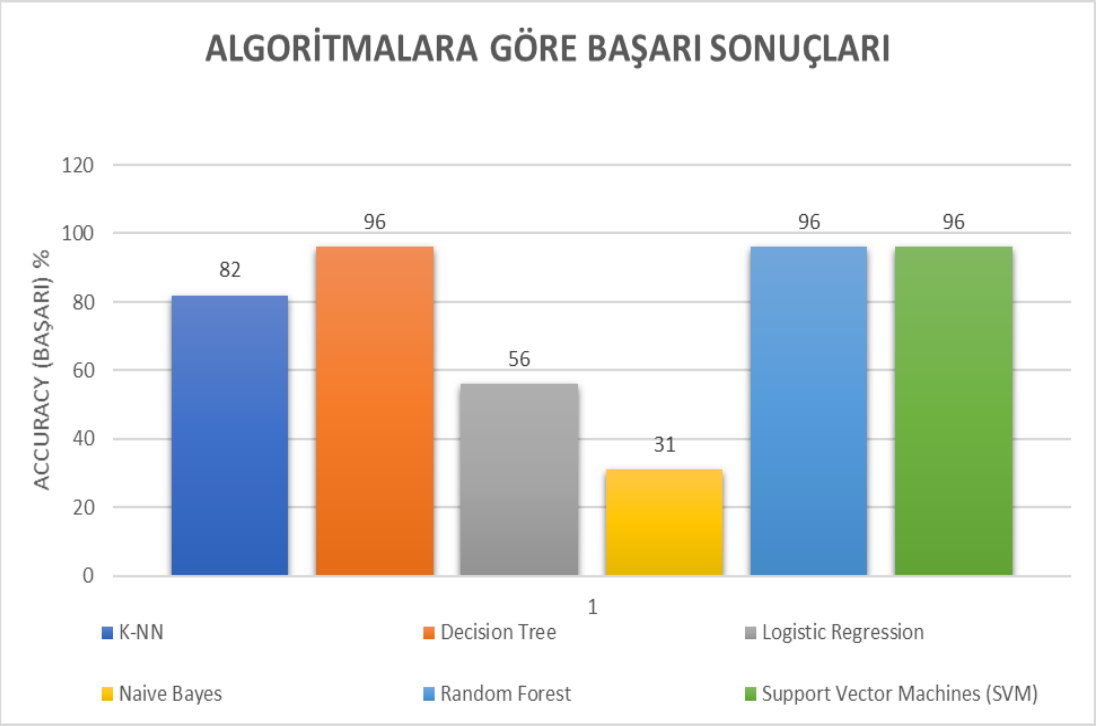
Algoritmalar	Hatalı Sınıflandırılan Örnek Sayısı
K-NN	2691
Decision Tree	388
Logistic Regression	6506
Naive Bayes	8353
Random Forest	472
SVM	389

Tablo 16: Algoritmalara Göre Hatalı Sınıflandırılan Örnek Sayıları

Scenario B

TimeBasedFeatures-Dataset-120s-AllinOne.arff

Toplam Örnek Sayısı = 10.782



Grafik 17: Algoritmalara Göre Başarı Sonuçları

Algoritmalar	Hatalı Sınıflandırılan Örnek Sayısı
K-NN	1936
Decision Tree	371
Logistic Regression	4710
Naive Bayes	7433
Random Forest	426
SVM	372

Tablo 17: Algoritmalara Göre Hatalı Sınıflandırılan Örnek Sayıları

TimeBasedFeatures-Dataset-	K-NN	Decision Tree	Logistic Regression	Naive Bayes	Random Forest	SVM
15s-VPN.arff	**%86**	**%98**	**%62**	**%53**	**%98**	**%98**
30s-VPN.arff	**%86**	**%98**	**%64**	**%53**	**%97**	**%98**
120s-VPN.arff	**%84**	**%97**	**%54**	**%55**	**%97**	**%97**
15s-NO-VPN.arff	**%89**	**%99**	**%71**	**%49**	**%99**	**%99**
15s-VPN.arff	**%84**	**%99**	**%59**	**%22**	**%98**	**%99**
30s-NO-VPN.arff	**%86**	**%99**	**%69**	**%45**	**%98**	**%99**
30s-VPN.arff	**%86**	**%98**	**%66**	**%35**	**%98**	**%98**
60s-NO-VPN.arff	**%86**	**%98**	**%61**	**%30**	**%97**	**%98**
60s-VPN.arff	**%84**	**%98**	**%54**	**%34**	**%97**	**%98**
120s-NO-VPN.arff	**%86**	**%99**	**%71**	**%39**	**%98**	**%99**
120s-VPN.arff	**%83**	**%97**	**%57**	**%38**	**%96**	**%97**
15s.arff	**%78**	**%98**	**%45**	**%27**	**%97**	**%98**
15s-AllinOne.arff	**%84**	**%98**	**%61**	**%24**	**%97**	**%98**
30s.arff	**%77**	**%97**	**%39**	**%24**	**%97**	**%97**
30s-AllinOne.arff	**%83**	**%97**	**%60**	**%29**	**%97**	**%97**
120s.arff	**%75**	**%96**	**%39**	**%22**	**%95**	**%96**
120s-AllinOne.arff	**%82**	**%96**	**%56**	**%31**	**%96**	**%96**

Tablo 18: 17 Farklı Veri Setinin 6 Farklı Algoritma İle Sınıflandırılması Sonucu Başarı Yüzdeleri

DEĞERLENDİRME

Bu projede ağ trafik kategorilerinden oluşan 17 farklı veri seti 6 farklı sınıflandırma algoritması ile sınıflandırılıp başarıları grafiklerde ve tabloda gösterilmiştir.

Tablodan ve grafiklerden de görüldüğü üzere, Decision Tree (Karar Ağaçları) ve SVM (Support Vector Machines) sınıflandırma algoritmalarının ortalama %97,76 başarı oranı ile diğer algoritmalara göre daha yüksek sonuçlar verdiği görülmektedir.

Daha sonra ortalama %97,17 başarı oranı ile Random Forest (Rastgele Orman) algoritması diğer algoritmalara göre doğru sınıflandırmada en iyi performansı göstermiştir.

KULLANILAN ARAÇLAR

Ağ trafiğinin sınıflandırılması ve makine öğrenmesi algoritmalarının kodlama aşaması için Anaconda yazılımının içerisindeki Jupyter Notebook programı kullanılmıştır. Programlama dili olarak da makine öğrenmesi projelerinde sıklıkla kullanılan Python tercih edilmiştir.

KAYNAKLAR

[1] “Machine Learning ve Python: A’dan Z’ye Makine Öğrenmesi (4) | Udemy,” . https://www.udemy.com/course/machine-learning-ve-python-adan-zye-makine-ogrenmesi-4/?start=0.

[2] Ş. E. Şeker, “Python ile Makine Öğrenmesi | Udemy,” 2018. https://www.udemy.com/course/makine-ogrenmesi/?start=0.

[3] “Destek Vektör Makineleri (Support Vector Machine) – Veri Bilimcisi.” https://veribilimcisi.com/2017/07/19/destek-vektor-makineleri-support-vector-machine/.

[4] “Naive Bayes Sınıflandırıcısı (Naive Bayes Classifier) – Veri Bilimcisi.” https://veribilimcisi.com/2017/07/20/naive-bayes-siniflandiricisi-naive-bayes/.

[5] “Lojistik Regresyon (Logistic Regression) – Veri Bilimcisi.” https://veribilimcisi.com/2017/07/18/lojistik-regresyon/.

[6] “K-En Yakın Komşu (K-Nearest Neighbors(KNN)) – Veri Bilimcisi.” https://veribilimcisi.com/2017/07/20/k-en-yakin-komsu-k-nearest-neighborsknn/.

[7] “Karar Ağaçları (Decision Trees) – Veri Bilimcisi.” https://veribilimcisi.com/2018/02/23/karar-agaclari-decision-trees/.

[8] “Sınıflandırma Metrikleri – Veri Bilimcisi.” https://veribilimcisi.com/2018/11/28/siniflandirma-metrikleri/.

[9] “Scikit-Learn.” https://scikit-learn.org/stable/

ÖZGEÇMİŞ

Büşra BAKKALCI 1995 yılında KADIKÖY' de doğdu. İlk, orta ve lise öğrenimini KOCAELİ' de tamamladı. 2014 yılında Karabük Üniversitesi Bilgisayar Mühendisliği Bölümü'nde öğrenime başlayıp 2020 (beklenen) yılında mezun oldu.

ADRES BİLGİLERİ

Adres : Fevzi Çakmak Mah. Dr. Zeki Acar Cad. Ferit Sok.
No:7 Kat:2 Darıca / KOCAELİ
Tel : (539) 2703167
E-posta : busrabakkalci.11@gmail.com