T.C.
VAN YÜZÜNCÜ YIL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

RASTGELE ORMANLAR YÖNTEMİ İLE ÖZELLİK SEÇİMİ KULLANILARAK VAN İLİNDE YAŞAYANLARIN TRAFİKTE ALGI VE TUTUMLARININ BELİRLENMESİ

YÜKSEK LİSANS TEZİ

HAZIRLAYAN: Vedat GÖRGÜLÜ
DANIŞMAN: Dr. Öğr. Üyesi Sanem ŞEHRİBANOĞLU

VAN-2021

T.C.
VAN YÜZÜNCÜ YIL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK ANABİLİM DALI

RASTGELE ORMANLAR YÖNTEMİ İLE ÖZELLİK SEÇİMİ KULLANILARAK VAN İLİNDE YAŞAYANLARIN TRAFİKTE ALGI VE TUTUMLARININ BELİRLENMESİ

YÜKSEK LİSANS TEZİ

HAZIRLAYAN: Vedat GÖRGÜLÜ

VAN-2021

KABUL VE ONAY SAYFASI

İstatistik Anabilim Dalı'nda Dr. Öğr. Üyesi Sanem ŞEHRİBANOĞLU danışmanlığında, Vedat GÖRGÜLÜ tarafından sunulan "Rastgele Ormanlar Yöntemi İle Özellik Seçimi Kullanılarak Van İlinde Yaşayanların Trafikte Algı ve Tutumlarının Belirlenmesi" isimli bu çalışma Lisansüstü Eğitim ve Öğretim Yönetmeliği'nin ilgili hükümleri gereğince 26/04/2021 tarihinde aşağıdaki jüri tarafından oy birliği / oy çokluğu ile başarılı bulunmuş ve Yüksek Lisans tezi olarak kabul edilmiştir.

Başkan: İmza: …….

Üye: İmza: …….

Üye: İmza: …….

Fen Bilimleri Enstitüsü Yönetim Kurulu'nun … / … / … tarih ve …… sayılı kararı ile onaylanmıştır.

İmza

……...

Enstitü Müdürü

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, ayrıca tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

(İmza)
Vedat GÖRGÜLÜ

ÖZET

RASTGELE ORMANLAR YÖNTEMİ İLE ÖZELLİK SEÇİMİ KULLANILARAK VAN İLİNDE YAŞAYANLARIN TRAFİKTE ALGI VE TUTUMLARININ BELİRLENMESİ

GÖRGÜLÜ, Vedat
Yüksek Lisans Tezi, İstatistik Anabilim Dalı
Tez Danışmanı: Dr. Öğr. Üyesi Sanem ŞEHRİBANOĞLU
Haziran 2021, 77 sayfa

Bu tez çalışmasında, Van ilinde yaşayanların trafik algı ve tutumlarını ölçmek amacıyla Rastgele Orman yöntemi uygulanmıştır. 2015 ile 2018 yıllarında Van ilinde yaşayanlara yönelik uygulanan anket çalışmasına katılan 773 adet katılımcıya ilişkin bilgiler kullanılarak Rastgele Ormanlar yöntemi ile öncelikle sınıflandırma işlemi yapılarak performans ölçütleri açıklanmıştır. Sonrasında bağımlı değişken ile kuramsal ilişkisi olduğu düşünülen 41 adet bağımsız değişkenin önemlerine göre özellik seçimi işlemi yapılmıştır.

Bu çalışmada, bağımlı değişken olarak anketin içerisinde yer alan "Trafik kurallarına harfi harfine uyarım", "Trafik kurallarının araba kullanma zevkini yok ettiğine inanıyorum" ve "Hız limitini 10-15 km/saat aşmakta sorun yoktur çünkü bunu herkes yapıyor" soruları birbirlerinden bağımsız olarak sırasıyla seçilerek katılımcıların verdikleri cevaplar üzerinden sınıflandırma işlemi uygulanmıştır. İlk olarak sınıflama işleminde baz alacağımız iki adet parametre olan Ağaç Sayısı ile Maksimum Değişken sayısı parametreleri önce tablo ve grafikler üzerinden ayrı ayrı, ardından hiperparametre optimizasyonu ile iki farklı şekilde hesaplanmıştır. Sonrasında sınıflandırma modeline ait Doğruluk yüzdesi, F-puanı, ROC eğrisi altında kalan alan (AUC değeri) hesaplanarak modelin başarı ve güvenilirlik ölçüsü hesaplanmıştır. Son olarak Kappa değerleri üzerinden sınıflar arasındaki uyumu ölçme işlemi uygulanmış olup bağımlı değişkenlerin her birinin ayrı ayrı olarak sınıflama performanslarını belirlemede önemi en yüksek olan bağımsız değişkenler arasından özellik seçimi işlemi yapılarak bağımlı değişken ile arasındaki ilişki yorumlanmıştır.

Anahtar kelimeler: Algı, AUC değeri, Değişken önemi, Doğruluk, F-puanı, Kappa, Özellik seçimi, Rastgele Ormanlar, ROC eğrisi, Trafik, Tutum, Van.

ABSTRACT

DETERMINING THE PERCEPTION AND ATTITUDE OF THE VAN LIVES IN THE TRAFFIC BY USING THE FEATURE SELECTION BY THE RANDOM FOREST METHOD

GÖRGÜLÜ, Vedat
M.Sc. Thesis Department of Statistics
Supervisor: Asst. Prof. Dr. Sanem ŞEHRİBANOĞLU
June 2021, 77 pages

In this thesis, the Random Forest method was applied in order to measures traffic perception and attitude of people living in Van province. Using the information about 773 workers who participated in the questionnaire participant conducted for those living in Van in 2015 and 2018, the Random Forest was firstly classified and performance criteria were explained. After that, features were selected according to the importance of 41 independent variables, which were thought to have a theoretical relationship with the dependent variable.

In this study, the questions of "I strictly obey traffic rules", "I believe that traffic rules destroy the pleasure of driving" and "It is okay to exceed the speed limit of 10-15 km/hour because everyone does it", which are included in the questionnaires independently. The classification process was applied based on the answers given by the participants by selecting them in order. Firstly, the two parameters that we will take as basis in the classification process, Number of Trees and Maximum Number of Variables, were calculated separately over tables and graphs, after that then in two different ways with hyperparameter optimization. Afterwards, the Accuracy percentage, F-score, area under the ROC curve (AUC value) of the classification model were calculated and the success and reliability measures of the model was calculated. Finally, the process of measuring the fit between the classes through Kappa values was applied and the relationship between the dependent variable was interpreted by feature selection the independent variables that have the highest importance in determining the classification performance of each of the dependent variables separately.

Keywords: Perception, AUC value, Variable importance, Accuracy, F-score, Kappa, Feature selection, Random Forests, ROC curve, Traffic, Attitude, Van.

ÖN SÖZ

Bu tez çalışmasında, her türlü ilgi ve yardımlarını esirgemeyen danışmanım Sayın Dr. Öğr. Üyesi Sanem ŞEHRİBANOĞLU' na teşekkürlerimi sunarım. Ayrıca tez jürimde yer alan ve desteklerini esirgemeyen Sayın Dr. Öğr. Üyesi Murat CANAYAZ' a ve Sayın Dr. Öğr. Üyesi Özlem BEZEK GÜRE' ye teşekkürlerimi sunarım. Türkiye İstatistik Kurumu'nda görev yapan ve desteklerini hiçbir zaman esirgemeyen çalışma arkadaşlarım ve dostlarım Sayın Fatih BEKTAŞ' a ve Sayın Seher KUŞÇULAR' a teşekkür ederim. Bugünlere gelmemde en büyük paya sahip olan ve eğitim hayatım boyunca desteklerini benden hiçbir zaman esirgemeyen babam İsmail GÖRGÜLÜ' ye, annem Raziye GÖRGÜLÜ' ye ve kardeşim Damla GÖRGÜLÜ' ye sonsuz teşekkürlerimi sunarım.

2021

Vedat GÖRGÜLÜ

İÇİNDEKİLER

Sayfa

Sayfa

Sayfa

ÇİZELGELER LİSTESİ

Çizelge **Sayfa**

Çizelge **Sayfa**

ŞEKİLLER LİSTESİ

Şekil **Sayfa**

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış bazı simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler	Açıklama
$\mathbf{A_i}$	Doğrulama metriği
b	Eğitim kümesine ait tekrarlı değerler içeren alt küme elemanları
$\mathbf{B_i}$	Bagging algoritmasına ait her bir önişlem sınıflamasını içeren küme
c	Eğitim veri setindeki mevcut doğru tahminlerin sayısı
$\mathbf{c_b}$	$C^*(x)$ sınıflayıcısına ait elemanların hata olasılıkları
$\mathbf{C_i}$	Tahmin değerlerinin ait olduğu sınıfların kümesi
$\mathbf{C_j}$	Tahmin değerlerinin dışında kalan sınıfların kümesi
$\mathbf{C^b(x)}$	b kümesine ait türetilen alt kümelere ait her bir sınıflayıcı
$\mathbf{C^*}$	Bagging algoritmasına ait en çok sınıfı içeren küme
$\mathbf{c^*}$	Eğitim veri seti içerisinde oluşturulan alt kümelere ilişkin doğru tahmin sayısı
d	Eğitim veri seti içerisinde mevcut tahminler ile yeni oluşturulan alt kümeye ilişkin doğru tahmin sayısı arasındaki fark

Simgeler	Açıklama
$\bar{d}$	Mevcut m. değişkenin ağaç düğümüne ait ortalama önem katsayısı
D	Eğitim veri setine ilişkin özellik sayısı
D'	Eğitim kümesi içerisinde rastgele eleman seçilerek oluşturulan alt kümeye ait eleman sayısı
err_b	$C^*(x)$ sınıflayıcısına ait her bir değere ilişkin hata olasılık değeri
h(x)	Sınıflama algoritmasına ait her bir sınıfı içeren küme
i	Sınıflayıcı içerisindeki her bir eleman
k	Eğitim verisine ilişkin sınıf sayısı
K	Kappa Skoru
m	Eğitim setinden iadeli şekilde çekilerek oluşturulan örneklem sayısı
n	Örneklem sayısı
n_j	İlgili j. sınıfın gözlem sayısı
N	Karar ağacı sayısı
$OOBError1_v$	v değişkeninin 1.tahmincisine ait hesaplanan OOB hata değeri
$OOBError2_v$	v değişkeninin 2.tahmincisine ait hesaplanan OOB hata değeri
p	Eğitim nesnelerini içeren alt kümelerin boyutu
p_k	k. düğüme ait pozitif gözlem sayısı
Pr(a)	İki değerlendiriciye yönelik uyumlu gözlemler toplamı
Pr(e)	Mevcut uyumun şansa bağlı çıkması olasılığı

Simgeler	Açıklama
SEd_i	Birbirinden bağımsız mevcut tüm d_i değişkenlerinin standart hatası
T	Eğitim veri setine ait örneklem sayısı
v	Önem derecesi hesaplanacak değişken
w_i^b	Boosting algoritması içerisindeki b sınıflayıcılarına ilişkin her bir ağırlık değeri
W_j	İlgili j. sınıfın ağırlığı
X	Örnek bir eğitim veri seti
X_i	Eğitim kümesine ait her bir eleman
X^b	X^* kümesinin alt kümelerine ilişkin her bir tahmin değeri
X^*	Her bir X_i değerine ait ağırlıklandırılmış değerler
$\widehat{X}^b$	b sınıfına ait eğitim nesnelerin kümesi
$\widehat{X}^b_i$	b sınıfına ait eğitim nesnelerini içeren her bir eleman
y	C^* sınıflayıcısına ait {-1, +1} aralığındaki her bir etiket
$\hat{y}$	Sınıflama algoritması sonucu oluşan tahminler kümesi
$\hat{y}_i$	i. sınıflama algoritmasına ait tahmin değeri
$\hat{y}_f$	Topluluk karar ağaçlarının verdiği sonuçlar arasında en çok tercih edilen tahmin değeri
$\beta(x)$	Basit çoğunluk oylaması ile belirlenen nihai C^* sınıfı
Θ_k	Eğitim verisine ilişkin rastgele vektör

Simgeler	Açıklama
δ	Basit çoğunluk oylamasıyla belirlenecek olan nihai $C^*(x)$ sınıflandırıcısına ait y Є {-1, +1} aralığındaki etiket değerleri

Kısaltmalar	Açıklama
AUC	Area Under the Curve
BAP	Bilimsel Araştırma Projesi
CART	Classification and Regression Tree
CV	Cross Validation
ÇŞB	Çevre ve Şehircilik Bakanlığı
EGM	Emniyet Genel Müdürlüğü
FI	Feature Importance
FN	False Negative
FP	False Positive
FPR	False Positive Rate
IB	InBag
KM	Kilometre
M.Ö.	Milattan Önce
OOB	Out Of Bag
ROC	Receiver Operating Characteristic
SE	Standart Error
TN	True Negative
TP	True Positive
TPR	True Positive Rate
TÜİK	Türkiye İstatistik Kurumu
YYÜ	Yüzüncü Yıl Üniversitesi

1. GİRİŞ

Veri kavramının gün geçtikçe daha çok önem kazanmasıyla beraber geçmişten süregelen birçok süreçte değişimler daha fazla hız kazanmıştır. Bu değişim en fazla verimlilik kavramının esas alındığı süreçte daha çok öne çıkmaktadır. Bu nedenle, değişimde bu derece yüksek etkisi bulunan veri kavramının çok iyi bir şekilde işlenerek analiz edilmesi gerekmektedir (Laudon, 2007).

Bilgi teknolojilerinde yaşanan gelişmelerin sonucu olarak 2000'li yılların başlarından itibaren birçok sektörde mevcut olan iş akış süreçlerinde verimlilik ve zaman önem kazanarak geçmişte mevcut birçok normun artık gelecekte etkili olmayacağını bizlere göstermiştir (Mansoori ve ark., 2019). Bu yeni normlar sayesinde ortaya çıkan her yeni teknolojinin fayda sağladığı belirli alanlar mevcut olmuştur. Teknolojinin gelişmesiyle birlikte bizlere fayda sağlayan alanların çalışma temelleri, eldeki çeşitli veri yapılarını işleyebilme ve yorumlayabilme kabiliyetiyle ilişkilidir. Makine öğrenmesi yöntemleri de eldeki mevcut veri setinin işlenebilmesi ve anlamlı sonuçlar çıkartabilmesi yönüyle literatürde yer bulmaktadır (Kılınç ve ark., 2020).

Veri madenciliğinin, mevcut veri tabanı içerisindeki veri yığınlarına ait bağlantıları bulmak amacıyla istatistiksel yöntemlerden ve makine öğrenmesi algoritmalarından faydalanan kompleks bir analiz yöntemi şeklinde tanımlaması yapılabilir. Bunların yanı sıra, veri madenciliği kavramını; istatistik ve yapay zekâ gibi disiplinlerle bağlantılı olarak bir arada fayda sağlayan geniş bir çerçeve olarak tanımlayabiliriz (Emel ve Taşkın, 2005).

Makine öğrenmesi, eldeki veriler üzerinde basit tanımlayıcı kavramları açıklayabilmek için insanlar tarafından üretilen istatistiksel teknikleri kullanır (Michie ve ark., 1994). Makine öğrenmesinin günümüzde bu kadar tercih edilmesindeki en önemli faktörler; bilgisayar teknolojileri dünyasında meydana gelen gelişmelerle birlikte ihtiyaç duyulan verilerin ne kadar büyük hacimli olursa olsun rahatlıkla depolanabilmesi, yer ve zaman fark etmeksizin istenildiği anda ulaşılabilmesi ve analiz için kullanılabilmesidir (Kızılkaya ve Oğuzlar, 2018).

Günümüzde daha iyi analizler yaparak daha güvenilir ve başarılı kararlar verilebilmesini sağlayan çeşitli makine öğrenmesi yöntemleri geliştirilmiş olup zamanla

daha da geliştirilmeye devam etmektedir. Bu yöntemlerden başlıcaları; sınıflama yöntemleri, kümeleme yöntemleri, karar ağaçları ve yapay sinir ağları gibi birçok yöntem sayesinde eldeki verilerden anlamlı sonuçlar çıkarılarak etkili tahminler yapılabilmektedir (Kolay ve Erdoğmuş, 2016).

Trafik kazalarının meydana gelmesinde birçok etken mevcut olup bu kadar fazla etkenin bir arada yer alması trafik kazalarının oluşum sebebini belirleme sürecini daha karmaşık bir hale getirmektedir. Bu nedenle trafik güvenliğinin tam anlamıyla sağlanarak bu durumun daha da sürdürülebilir hale gelmesi oldukça zor bir hale gelmektedir (Özden ve Acı, 2018). Trafik kazalarına ilişkin başlıca sebepler; insan, yol ve çevre faktörü olarak sıralanabilir. Trafik kazaları sonuçlarının bazıları telafi edilebilir olacağı gibi bazıları da telafi edilemez sonuçlar doğurmaktadır. O sebeple trafik kazalarının sayılarını en aza indirebilmek ve sonuçlarını telafi edilebilir kılabilmek için ülkeler bu konularda çeşitli politikalar ve uygulamalar geliştirmektedir (Yavuz ve ark., 2021).

Bu tez çalışmasında, Van ilinde yaşayan bireylerin trafik kurallarına yönelik algı ve tutumlarını belirlemeye yarayan tarafımızca seçilen 3 adet sorunun bir topluluk öğrenme yöntemi olan Rastgele Ormanlar yöntemi kullanılarak sınıflandırma işlemi ve özellik seçimi yapılmıştır. Bu sayede trafikte can ve mal güvenliğinin sağlanabilmesi için yetkili mercilerin sisteminde yeni veriler oldukça trafik denetlemelerine yönelik yeni politikaların oluşturulmasında ve bu konu ile ilgili daha etkili uygulamalarla trafik kazalarının önüne geçilmesinde fayda sağlanabilmesi amaçlanmaktadır.

Bu amaca yönelik ilgili tez çalışmasında; konu ile ilgili yapılan literatür çalışmaları, veri madenciliği ile yöntemlerinin genel özellikleri, veri madenciliğinde makine öğrenmesi ile ilgili özellikler ve tekniklerinin açıklamaları, ağaç tabanlı topluluk öğrenme yöntemlerinin tanım ve özellikleri, rastgele ormanlar yönteminin tanımı ve özellikleri, Trafik kurallarına yönelik tutum ve davranışlar hakkında belirli tanımlar ve özellikler, Van iline ait trafik kurallarına yönelik algı ve tutumlarına yönelik uygulama verileri üzerinde rastgele ormanlar algoritması ile yapılan uygulama sonuçları ve son olarak yapılan uygulama sonucunda elde edilen bulgulara dayanarak sonuçların değerlendirmeleri ve yorumlamaları yapılarak bizlere gelecekte sağlayacağı faydaları hakkında bilgiler verilmiştir.

1.1. Veri Madenciliği

1.1.1. Veri madenciliği tanımı ve özellikleri

Teknoloji'nin her geçen gün ilerlemesi ve gelişmesi ile birlikte telefonlardan, tabletlerden ve bilgisayarlar üzerinden elde ettiğimiz büyük oranlı verileri dijital ortamlarda üretebilmek, saklayabilmek ve bunlara istediğimiz an erişebilmek çok kolay bir hal almıştır. Fakat kolaylaşan bu durum büyük hacimli veriler içerisinden anlamlı sonuçlar çıkarabilmek doğrultusunda bizlere birtakım zorluklar getirmektedir. Büyük hacimli ve birden fazla değişkene sahip birbirlerinden bağımsız bu veri yığınları içerisinden anlamlı bir bütün oluşturma işlemine ise genel anlamda veri madenciliği adı verilmiştir (Kalikov, 2006).

Veri madenciliği yöntemi ayrıca, veri tabanı içerisindeki veri yığınlarına ait bağlantıları bulmak amacıyla istatistiksel yöntemlerden ve makine öğrenmesi algoritmalarından faydalanan karmaşık bir analiz yöntemi olarak tanımlanabilir. Bunların yanı sıra, veri madenciliği kavramını; istatistik ve yapay zekâ gibi disiplinlerle birlikte fayda sağlayan geniş bir çerçeve içerisinde tanımlayabiliriz (Emel ve Taşkın, 2005).

1.1.2. Veri madenciliğinin gelişimi

Veri madenciliği uygulamalarının ilk dönemlerinde eldeki büyük hacimli verilerin, içinde bulunduğu dönemin teknik imkanlarının kısıtlı olması dolayısıyla işlenebilmesi ve verimli sonuçlar alınabilmesi önemli bir endişe kaynağıydı. İlerleyen dönemlerde daha büyük hacimli verilerin de işleme kapasitelerinin artması ve daha kısa sürede bu hesaplamaların yapılabilmesi ile bu endişeler yerini başka kaygılara bırakmıştır (Coenen, 2004). Bu kaygılar ise, çok büyük hacimli veri yığınları içerisinde bilgiye ulaşma sürecinde birtakım yeni bilgiler elde edebilmektir (Giudici, 2003).

Veri madenciliği sürecinin gelişimi ile bilgi çağında elde ettiğimiz büyük hacimli verileri analiz etme aşamasında elde edilen sonuçların doğruluğunu arttırmak için istatistik, makine öğrenmesi, veri tabanları gibi yöntemlere dayalı çeşitli teknikler geliştirilmeye başlanmıştır (Öğüt, 2002).

1.1.3. Veri madenciliğinin kullanım amaçları

Veri madenciliği süreçleri genel olarak yapılacak çalışmaya ait iki adet temel soruya cevap aramaktadır. Bunların ilki model kurmak, bir diğeri ise betimlemeler yapmaktır (Moshkovich ve ark., 2002). Tahmin modelleri eldeki verilere ilişkin eğilim ve bağlantıları bulmaya yararken tanımlayıcı bilgi ise araştırmacı için ilgi çekici örüntüleri ve ilişkileri görselleştirmesi bakımından araştırmacıya fayda sağlar (Melody ve Kumar, 2001).

1.1.4. Veri madenciliği yöntemleri

Veri madenciliğine ait yöntemlerin dayandığı temellere göre ilki klasik yöntemler diyebileceğimiz istatistiksel yöntemler olup ikincisi de genellikle istatistik temelli olan fakat daha çok yapay zekâ ve makine öğrenmesine dayalı modern yöntemlerdir. İşlevlerine göre ise sınıflandırma ve regresyon yöntemleri tahmine, kümeleme ve birliktelik kuralları yöntemleri tanımlama yapmayı amaçlamaktadır (Özekes, 2003).

Veri madenciliği yöntemleri, işlevlerine göre 3 grupta toplanır. Bunlar;

1. Sınıflandırma (Classification) Yöntemleri,
2. Regresyon (Regression) Yöntemleri,
3. Kümeleme (Clustering) Yöntemleri,

1.2. Makine Öğrenmesi

1.2.1. Makine öğrenmesi tanımı

Modern çağa ait bilgi toplumunda, veriye yönelik karar destek sistemlerine ilişkin son yıllara ait yazılım ve donanım alanlarında meydana gelen gelişmelerin yanı sıra büyük hacimdeki verilerin de işlenerek uygulanan yöntemlerin artmasıyla birlikte bilgiye ulaşmak daha pratik hale gelmiştir (Çakır, 2005).

Makine öğrenmesi, büyük hacimli veri setlerinin anlaşılmasına ilişkin yeni seviyelere ulaşması dolayısıyla özellikle kâr amaçlı işletmelere büyük oranda fayda sağlamaktadır. Süreç içerisinde uygun makine öğrenmesi yöntemlerini kullanarak

işletmelerdeki değişimleri ve uygulanacak yeni politikaları belirlemede kendilerine oldukça fayda sağlamaktadırlar (Hurwitz ve Kirsch, 2018).

Veri madenciliğinde elektronik ortamda saklanan verilere ilişkin önemli bağlantılar ortaya çıkarılarak tahminler ve belirli sınıflamalar oluşturmak için belirli kuralların öğrenilmesine ihtiyaç vardır. Araştırmacının bu süreçte en önemli görevi ise makine üzerinden öğrenilen bilgileri incelemek, gürültülü verileri temizlemek ve veriyi daha düzenli hale getirmektir (Weiss ve Indurkhya, 1998).

Makine öğrenmesi yönteminin temel amacı, en doğru modelin bulunmasının yanı sıra bağımsız değişkenlerin hangilerinin en önemli olduğunu ortaya çıkarmaktır (Louppe, 2014). Öğrenme aşamasının en önemli tarafı, daha sonradan test verisi olarak sisteme dahil edilecek yeni verilere, daha önceden kurulan model dahilinde en doğru tahminleri üretmektir (Turgut, 2012).

Karar ağaçları temeline dayalı makine öğrenmesi yöntemlerinin temel amaçları şunlardır (Van Diepen ve Franses, 2006):

a) Daha önceden eğitim verisi aracılığıyla kurulan modele ait sınıflara ilişkin olası elemanları belirlemek.
b) Veri seti içerisinde yer alan birbirine benzer kategorileri birleştirmek ve belli bir gruba dahil etmek istediğimiz sürekli değişkenleri kesikli değişkenler haline getirmek.
c) Belirli vakalara ilişkin elemanları düşük, orta ve yüksek risk gruplarına ayırmak.
d) Parametrik modellerde kullanmak için eldeki veriler içerisinde önemlilik derecesi en yüksek öz değerleri belirlemek.
e) Eldeki geçmiş veriler üzerinden gelecekte gerçekleşmesi beklenen olayları ve durumları tahmin etmek.
f) Belli öz değerlerin önem değerlerini belirlemek.

Veri madenciliği ile makine öğrenmesi yöntemlerinde ortak olan özellikler aşağıdaki gibidir (Çakır, 2005);

1. İki yöntem de klasik istatistik yöntemlerinde karşımıza çıkan kuramsal problemlerin yanı sıra sınıflandırma veya ampirik öğrenme problemleriyle ilgilenir.
2. İki yöntem de eldeki test verileri üzerinden belli kurallar bütününü ortaya çıkararak kurulan bir model dahilinde yeni veriler üzerinde sınıflandırma veya tahminler oluşturur.

3. Veriden elde edilen bilginin somutlaştırılması iki yöntemin de ortak amaçlarının başında gelmektedir.
4. İki yöntemde de makinenin görevi veriyi öğrenerek içerisindeki kurallar dahilinde yeni bilgiler üretmek ve eldeki bilgiyi kullanmak, araştırmacının görevi ise yönteme dair eksiklikleri gidermektir.

1.2.2. Makine öğrenmesi teknikleri

1.2.2.1. Denetimli öğrenme

Denetimli öğrenme; Bir araştırmacının, eldeki veri setini eğitim ve test olarak iki gruba ayırarak eğitim veri seti üzerinden öğrenme sağlanarak ortaya yeni bir model çıkarılması, daha sonra da diğer grup olan test veri setini modele dahil ederek girdi verileri üzerinden söz konusu modelin sınıflandırma ve regresyon gibi uygulamalarının performansını sınama işlemidir. Eğitim işlemini uygulamanın asıl amacı, algoritmanın doğru sınıflama yapabilmesi için eğitim veri setine ait sınıflandırma önemliliği yüksek olan bağımsız değişkenleri ortaya çıkarmaktır. Denetimli öğrenme tekniğine ait en iyi örnek sınıflandırma yöntemidir (Learned-Miller, 2014).

1.2.2.2. Denetimsiz öğrenme

Denetimsiz öğrenme; veri kümesindeki yapıyı veya veri modelleri arasındaki ilişkiyi araştıran ve bu ilişkilere bakarak modeller düzenleyen bir yöntemdir. Denetimsiz öğrenme tekniğine ait en iyi örnek kümeleme yöntemidir (Dayan, 2008).

1.2.2.3. Yarı denetimli öğrenme

Yarı denetimli öğrenme, denetimli öğrenme ile denetimsiz öğrenme arasında bir yere sahip olup bağımlı değişken içeren veriler ile içermeyen verilerin bir arada kullanıldığı bir yöntemdir. Uygulama aşamasında eğitim veri setine ait bağımlı değişkene sahip verilerin yanı sıra bağımlı değişkeni olmayan daha fazla veriyi içerisinde barındırır (Chapelle ve ark., 2006).

1.2.2.4. Takviyeli öğrenme

Takviyeli öğrenme; ardışık karar verme sistemleri ile ilişkili bir makine öğrenmesi yöntemidir. Makine öğrenmesi yönteminin nihai hedefi, eğitim verisinden çıkarılan bilgi ve kuralları test verisine de uygulayarak ileriye yönelik tahminler oluşturmaktır. Takviyeli öğrenme yönteminde öğrenme işlemi bir ajan aracılığıyla gerçekleşir. Bu ajan çevreyle iletişim kurarak belli sonuçları toplar ve bu sonuçlar aracılığıyla bilgi toplayarak karar verme işlemini uygular. Söz konusu işlemler karşısında aldığı kararlar doğrultusunda ödül alır. Sonrasında bu ödülü daha da arttırmak için denemeler yaparak aynı uygulamayı aşamalı bir şekilde sürdürmeye devam eder (Sutton ve Barto, 2014).

1.2.2.5. Ağaç tabanlı topluluk öğrenme yöntemleri

Ağaç Tabanlı Topluluk Öğrenme Yöntemleri; aynı probleme ilişkin ortak çözümleri içinde barındıran, güvenilirlik bakımından ise en yüksek tahminleri üreten modeli ortaya çıkarmayı amaçlayan, bir diğer deyişle nihai kararlarda birden farklı görüşü ele alarak onlar arasındaki güvenilirliği en yüksek olan sonuçları bizlere sunmayı hedefleyen yöntemler olarak tanımlanmaktadır (Polikar, 2006).

Topluluk öğrenme yöntemlerinin ortaya çıktığı ilk yıllarda makine öğrenmesi yönteminde genel bir sorun olan olan yüksek varyans problemini azaltmaya ve bu sayede doğruluğu arttırmaya yönelik kullanılmasına karşın kısa süre içerisinde makine öğrenmesi yöntemleri içerisinde en sık karşımıza çıkan öz değer seçimi, kayıp özellikleri saptama ve veri ön işlemeye yönelik başarılı çözümler kazandırmıştır (Polikar, 2012).

İstatistik biliminde, 1977 yılında Tukey'in veri setine ikili lineer regresyon modelini önermesi ile topluluk öğrenme yöntemine ilk adım atılmıştır (Rokach, 2010). Bugün gelinen topluluk öğrenme mantığına pek de yakın sayılmayacak çalışmalardan sonra 1990'lı yıllarda Hansen ve Salomon, kendi çalışmalarında tahmin performansını yükseltmek amacıyla sinir ağlarına benzer topluluk öğrenme yaklaşımını önermişlerdir (Hansen ve Salomon, 1990). Bugün ise hala yaygın olarak tercih edilen Ağaç Tabanlı Topluluk Öğrenme yöntemlerinden Adaboost algoritmasının temeli Schapire tarafından ortaya atılmış, Freund ve Schapire de sınıflama oranı düşük olan sınıflayıcıları birleştirerek sınıflama oranı yüksek bir yöntem geliştirmişlerdir (Freund ve ark., 1996).

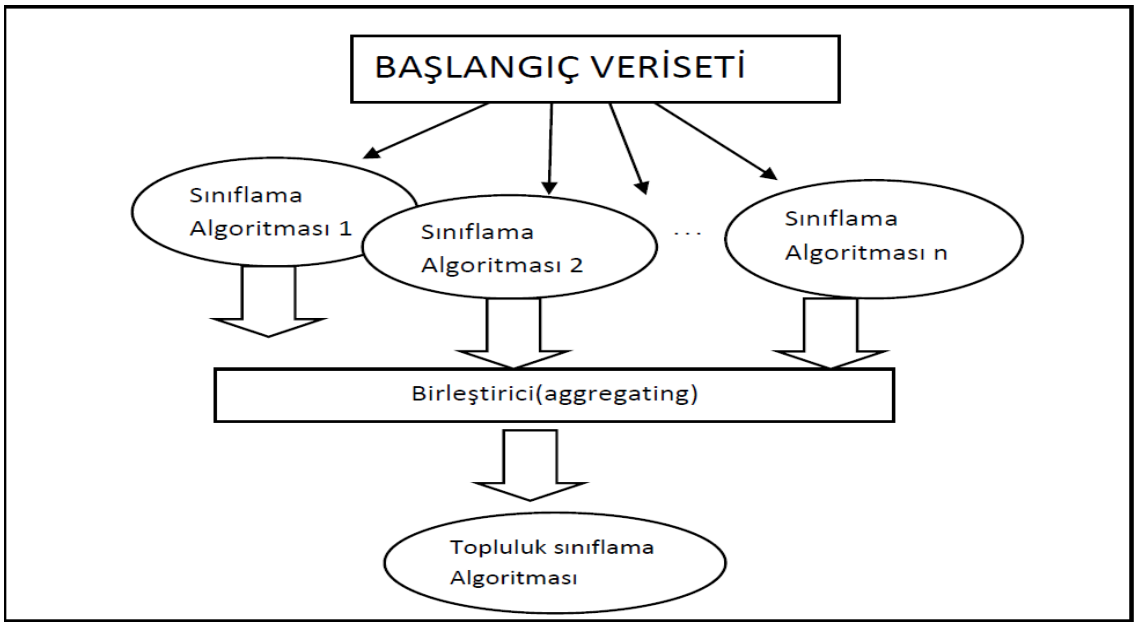
Topluluk öğrenme yöntemi, birden fazla sınıflayıcının bir araya getirilmesinden ziyade sınıflayıcılar tarafından elde edilen tahminlerin birleştirilmesi yöntemini kullanarak ortak bir tahminin üretilmesi esasına dayanır (Zhou, 2012).

Topluluk öğrenme yöntemlerinin, yüksek varyans problemini azaltmasının yanında, yanlılığa ilişkin problemi de gidermesi ile model performansını yükseltmesine de faydası olmaktadır (Bartlett ve Shawe-Taylor, 1999).

Topluluk öğrenme yöntemleri, temel sınıflayıcılara göre model performansını arttırmak amacıyla veri setine en uygun birleştirme tekniğinin kullanımı ve en uygun sınıflayıcının modele uygulanması konusunda fayda sağlar. Bunun yanı sıra modelin veri setini ezberleme problemi olan aşırı uyum (overfitting) probleminin azaltılmasında da fayda sağlamaktadır (Zhang ve ark., 2017).

Birleştirme tekniklerine ilişkin, eğitime ayrılan veri setlerine yönelik örneklem seçiminde ve işlem adımlarında farklı topluluk öğrenme yöntemlerinin kullanımları mümkündür. Topluluk öğrenme yöntemleri arasında en çok tercih edilen Bagging (Torbalama) ve Boosting (Arttırma) yöntemleridir (Rokach, 2010).

Topluluk öğrenme algoritmalarının çalışma prensibi, Şekil 1.1'de gösterilmektedir.



Şekil 1.1. Topluluk öğrenme stratejileri (Akman, 2010).

Topluluk öğrenme yöntemlerine matematiksel olarak basit bir şekilde açıklamak gerekirse mod işleminden örnek verebiliriz. Örneğin sınıflama algoritmasına ait oluşan her bir sınıflamayı h = {h_1, h_2, h_3, …, h_n} olarak, bu sınıflama algoritması sonucunda oluşan tahminleri de $\hat{y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$ olarak ifade edersek;

$h_i(x)=\hat{y}_i$: i. sınıflama algoritmasına ilişkin tahmin değeri

$\hat{y}_f$: Topluluk karar ağaçlarının verdiği sonuçlar arasında en çok tercih edilen tahmin değeri

$\hat{y}_f$ = mod{h_1, h_2, …, h_n} olarak tanımlanır (Raschka, 2018).

1.2.2.5.1. Bagging (Torbalama) tekniği

Bagging yöntemi, hata varyansını düşürmek amacıyla 1996 yılında Breiman tarafından geliştirilmiştir. Her bir sınıflandırıcıya ait farklı eğitim setleri elde etmek için yeniden örnekleme tekniğini kullanır. Diğer bir deyişle her bir örneklemin yerine koymak şartıyla (iadeli) çekilerek/seçilerek, söz konusu örneklemlerden model kurma işlemi olarak tanımlanır (Breiman, 2001). Yöntem, öğrenilmiş sınıflayıcıların veriden elde ettiği çeşitli çıktılarını tek bir tahmin halinde birleştirmek için geliştirilmiştir. Bagging yöntemi, sınıflama ve tahmin doğruluğunu yükseltmeyi amaçlamaktadır (Rokach ve Maimon, 2014).

Bagging yöntemini basit bir şekilde açıklarsak; çalışma konusuna ait eğitim setinden iadeli bir şekilde eşit oranda m adet örnek içeren yeni örneklemler üretilir. T adet Bootstrap örneklemleri olan B_1, B_2, …, B_T üretilir ve her bir Bootstrap içerisindeki her B_i için bir C_i sınıflayıcısı oluşur. Son sınıflayıcı olan C^*, C_1, C_2, …, C_T sınıflayıcılarının en fazla tahmin ettiği sınıfı baz alarak elde edilmektedir (Bauer ve Kohavi, 1999).

Bagging yönteminin uygulama aşamasında In Bag (IB) eğitim verisi içerisinden yaklaşık %63.2 kadar orijinal veri rassal olarak seçilir. Geriye kalan verilerden bazıları ise veri türetimi işlemi yapılarak %100'e tamamlanmaktadır. Bu sayede belli bir miktarda eğitim seti de elde edilmiş olur. Eğitim setlerinin her biri aynı değişken üzerinde uygulanarak alınan kararların her biri ağırlıklı oylama işlemi üzerinden birleştirmesi yapılır (Zhou, 2009).

IB veri setinden seçilen örneklemlerin %63.2 olmasının sebebini ise "Eş. 1.1" de gösterilen matematiksel tanım ile gösterebiliriz (Bulut, 2017);

$$1-(1/n)^n \approx e^{-1} = 0.368 \quad (1.1)$$

İlgili değişkene ait eleman sayısı olan n değerini sonsuza doğru yakınsattığımızda doğal logaritma tabanı olan e değerinin tersini elde ederiz. Bu sebeple n sayısının sonsuza doğru ilerletirken mevcut veri setinin orijinalinde %36.8'ini yüksek ihtimalle işleme almamış olacağız. Bagging yönteminde seçilmeyen değerlerin de seçilmesini sağlayabilmemiz amacıyla IB veri setinden %63.2 kadar (1-0.368=0.632) örneği rassal olarak seçerek yeni bir IB veri seti oluşturmayı amaçlarız (Bulut, 2017).

Bagging yönteminin uygulamasında takip edilen yollara ilişkin matematiksel gösterim "Eş. 1.2" ve "Eş. 1.3" de gösterilmiştir (Skurichina ve Duin, 2002);

1. Tekrarlı bir b = 1,2,… B oluşturulur
   a) X eğitim setine ait rastgele olarak xb alt kümesi seçilir.
   b) Eğitim verisi içerisinden seçilen xb elemanları kullanılarak bir C*(x) sınıflayıcısına ait elemanlar oluşturulur (C*(x)=0 dahil olmak üzere).
2. Oluşturulan elemanlar arasında seçilecek olan nihai C*(x) sınıflayıcısı ise, b=1, 2,…,B elemanları içerisinden basit çoğunluk oylaması yöntemiyle seçilir;

$$\beta(x)= \underset{y \in \{-1,+1\}}{argmax} \sum_b \delta_{C^b(x)),y}, \quad (1.2)$$

$$\delta_{i,j} = f(x) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \quad (1.3)$$

Burada; $\delta_{i,j}$ basit çoğunluk oylamasıyla belirlenecek olan nihai $C^*(x)$ sınıflandırıcısına ait y Є {-1, +1} aralığındaki etiket değerleri olup β(x) ise basit çoğunluk oylaması ile belirlenen nihai C^* sınıfıdır.

1.2.2.5.2. Boosting (Arttırma) tekniği

Boosting tekniği, makine öğrenmesi yöntemleri içerisinde yakın zamanda en fazla gelişme gösteren tekniklerden biri olarak gösterilir. Buradaki temel prensip, veri setine ilişkin her bir örneğin belli bir standart doğrultusunda ağırlıklandırılarak bir alt kümesinin oluşturulduğu ve bu alt kümeden faydalanan bir takım öğrenici oluşturmaktır (Kilimci ve

Ganiz, 2016). Bu öğreniciler arasından zayıf olanlar eğitim verileri üzerinde belirli dağılımları uygulayarak tekrarlı olarak çalışmaya devam ederler. Başlangıç aşamasında tüm örneklerin ağırlıkları birbirlerine eşittir. Yapılacak her yineleme işleminde, önceki sınıflamalara ilişkin yapılan eğitim hatalarına göre yanlış sınıflandırılan örneklerin ağırlıklarında güncellemeler yapılır. Her bir adımda, bir öncesinde yanlış olarak tahmin edilen örnekler doğru olarak tahmini yapılanlardan daha sık olarak seçilerek bu işlemler devam eder. Nihai karar ise bireysel sınıflayıcının tahmin ettiği sınıfların ağırlıklı olarak çoğunluk oyuna yönelik oluşturulmaktadır (Ren ve ark., 2016).

Bu tekniğin uygulamasındaki asıl amaç yanlış kararlar üzerinde yoğunlaşarak sınıflandırmanın genel başarısını yükseltmektir. Bunu yaparken de ilgili sınıflama yöntemlerini iyileştirerek seçilme olasılıklarını arttırmayı hedefler (Breiman, 1996).

Boosting yönteminin uygulamasında takip edilen yollara ilişkin matematiksel gösterim "Eş. 1.4" ve "Eş. 1.5" de gösterilmiştir (Skurichina ve Duin, 2002);

1. Tekrarlı bir b = 1,2, …, B oluşturulur
   a) X eğitim kümesine öncelikle ağırlıklandırma işlemi uygulanarak X* = $w_1^b X_1$, $w_2^b X_2$, …, $w_n^b X_n$ dönüşümü uygulanır. Burada w_i^b (i=1, 2, …, n), her bir X_i (i=1, 2, …, n) değerine ait ağırlık katsayıları olup oluşturulan bu X^* kümesinin alt kümesine ait her bir x^b alt kümesine ait C*(x) sınıflayıcısına ait elemanlar oluşturulur.
   b) Sınıflayıcıya ait elemanların hata olasılıkları hesaplanır.

$$C_b = \frac{1}{2}\log\frac{1-err_b}{err_b} \tag{1.4}$$

   c) Her bir değere ilişkin hata olasılık değeri olan errb katsayıları 0,5 ve üzeri olanlar seçilerek tekrardan ağırlıklandırılır ve hata olasılık değerleri minimize edilene kadar bu işlem devam eder.
2. Ağırlıklandırılmış değerler ile oluşturulan her bir değere ait sınıflayıcılar içerisinde, nihai C*(x) sınıflayıcısı, basit çoğunluk yöntemiyle seçilir.

$$\beta(x) = \underset{y \in \{-1,+1\}}{argmax} \sum_b \delta_{C^b(x)),y} \tag{1.5}$$

1.2.2.5.3. Random subspace (Rastgele altuzay) tekniği

Random Subspace yöntemi, Ho (1998) tarafından önerilen karmaşık kararların genellemesindeki doğruluğu arttırmak amacıyla belli örnek veri setindeki belirlenen özellikler arasından daha küçük bir alt küme seçen ve bu işlemi yineleyerek doğru modelleyiciye ulaşmayı hedefleyen bir toplu öğrenme yöntemidir (Çimen, 2020).

Bu yöntem, Bagging işlemi ile benzerlik gösterir fakat yönteme tüm örnekler dahil edilmez. Bunun yerine ilgili veri kümesine ait rastgele olarak seçilen bir özellik kümesinden faydalanılır (Breiman, 2001). D özellikleri olan bir veri kümesinde, Rastgele Altuzay D' özelliklerini D'<D olması şartıyla rassal olarak seçer. Orijinal veri setindeki özellikler içerisinde büyük bir kısmı kapsaması koşuluyla S farklı özelliğin alt kümesini almak için yine S kadar tekrarlama işlemi uygulanır. Sonrasında oluşturulan S adet temel sınıflandırıcının nihai kararını oylama işlemi üzerinden gerçekleştirir (Ho, 1998).

Yürütülen bazı çalışmalar içerisinde özellik sayısı D'nin eğitim nesnelerinin sayısı olan n katsayısından çok daha fazla olduğu durumlarda Rastgele Altuzay tekniğinin başarılı bir performans gösterdiği ortaya konulmuştur (Panov ve Dzerovski, 2007).

Rastgele alt uzay yönteminin matematiksel gösterimi aşağıdaki şekildedir (Skurichina ve Duin, 2002);

X: örnek bir eğitim veri seti olmak üzere

X_i (i=1, 2, 3, …, n) olmak üzere X = X_1, X_2, …, X_n: örnek veri setine ait n-boyutlu vektörlere ayrıştırılır

X_i = x_{i1}, x_{i2}, …, x_{ip}: rastgele olarak p-boyutlu bileşenlerine ayrılır (p<n). Bu sayede farklılaştırılmış bir eğitim veri seti elde edilir.

$\hat{X}^b$ = $\hat{X}^b_1$, $\hat{X}^b_2$, …, $\hat{X}^b_n$ olmak üzere n boyutlu yeni bir eğitim nesneleri elde edilir. Oluşturulan bu n boyutlu eğitim nesnesine ait her bir eleman;

$\hat{X}^b = \hat{X}^b_1, \hat{X}^b_2, \ldots, \hat{X}^b_n$ olmak üzere p boyutlu vektörlere ayrılarak her bir vektör elemanını "Eş. 1.6" da gösterildiği şekilde r-boyutlu vektörlere ayrıştırır;

$$\hat{X}^b = \hat{X}^b_{i1}, \hat{X}^b_{i2}, \ldots, \hat{X}^b_{ir} \ (i=1,2,..,p) \tag{1.6}$$

x^b_{ij}: (j=1,2,…,p) vektörü, rastgele olarak n-boyutlu x_{ij} (j=1,2,…,n) (p<n) boyutlu özelliklere sahip X eğitim veri setinden rastgele alt uzay yöntemiyle tekrarlı şekilde

üretilerek yeni sınıflandırıcılar oluşturulur. Oluşturulan sınıflandırıcılar ise nihai kararda basit çoğunluk oyuyla birleştirilir. Rastgele alt uzay yöntemi "Eş. 1.7" deki şekilde birleştirilir;

1) B = 1, 2, …, B olmak üzere;

(a) n boyutlu $\hat{X}^b$ eğitim verisi orijinal X veri setinden seçilerek türetilir.

(b) Türetilen her veri setinden $C^b(x)$ sınıflandırıcıları oluşturulur.

2) Sınıflandırıcılar birleştirilir $C^b(x)$, b= 1, 2, …, B (Nihai kararda basit çoğunluğun oyuyla birleştirme işlemi sağlanır). Nihai birleştirme formülü;

$$\beta(x) = \underset{y \in \{-1,+1\}}{argmax} \sum_b \delta_{C^b(x)),y} \quad : y \in \{-1, +1\} \tag{1.7}$$

1.3. Trafik Kurallarına Yönelik Tutum ve Davranışlar

1.3.1. Trafik, yol güvenliği ve kaza tanımı

İnsanların, ihtiyacı olan şeylerin hepsinin sabit bir yerde olması diye bir şeyin imkânı olmadığı için insanların ve eşyaların zaman zaman yer değiştirmesi önemli bir gerekliliktir. İnsanlar ve eşyalar için gerekli yer değişimini sağlamak ise ulaşım hizmetleri ile gerçekleştirilmektedir. Bu hizmetlerin pratik, ekonomik ve güvenli bir şekilde sağlanması için bulunduğu topluma belirli kazanımlar sağlamak önemli bir rol oynar (İyinam ve ark., 1999).

Trafik kavramını; insanların, hayvanların ve vasıtaların yollarda gidiş ve geliş hareketlerinin genel bir tanımı olarak alabiliriz. Trafik ve yol güvenliği kavramını da yine bu yolları kullanan insanların, hayvanların ve vasıtaların gidiş ve geliş hareketlerinin emniyet kuralları dahilinde gerçekleşmesi olarak tanımlayabiliriz (Yıldız ve Karaca, 2005).

Yaşama dair bütün alanlarda yer alan güvenlik hissiyatı kişilerin içinde bulunduğu sosyal ve üretim ilişkilerinde başta olmak üzere birçok noktada etkisini göstermektedir. Fiziki çevreden gelecek tehlikelerin en başında ise trafik güvenliği ihlalleri gelmektedir. İnsan, vasıta ve hayvan etkileşiminden ortaya çıkacak sorunların belirli kurallar çerçevesinde çözüme kavuşturmak trafik güvenliğinin temel amacıdır (Onay, 2013).

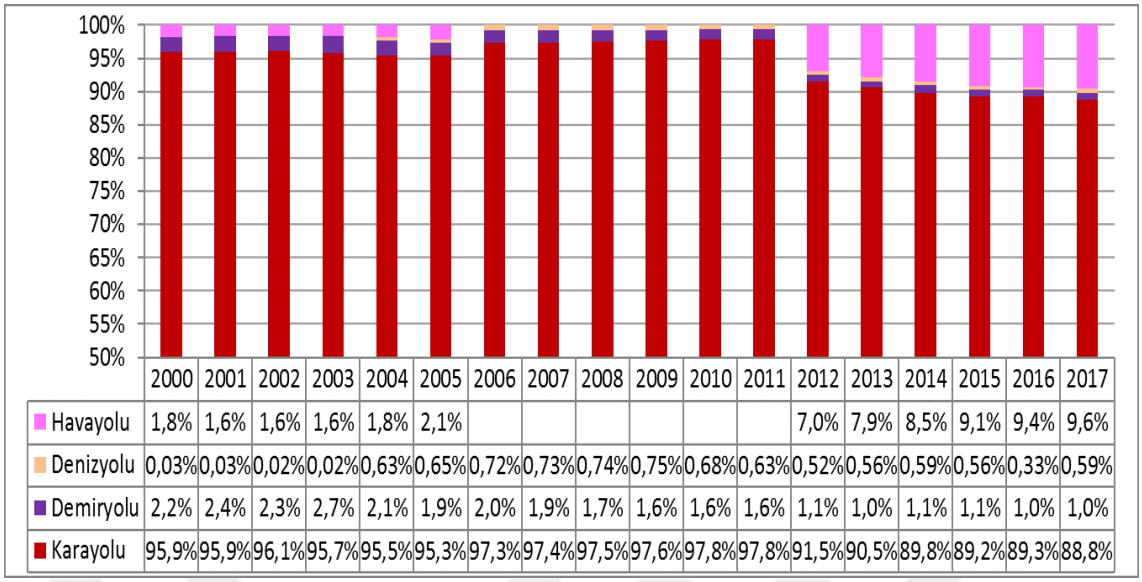
1.3.2. Karayolları trafiğinin tarihsel gelişimi

Karayolları trafiğinin başlangıcı tekerleğin icadına kadar uzanmaktadır. İnsanoğlu ilk çağlardan beri belirli ihtiyaçlarını karşılamak amacıyla bir yerlere ulaşmak için sürekli bir hareketlilik içerisindeydi. Bu hareketlilik ilk zamanlarda yürümeyle daha sonrasında ise hayvanları kullanarak devam etmiştir. Milattan önce (M.Ö.) 2000'li yıllarda tekerleğin keşfi insanoğlu için karayolu trafik hareketlerinin başlangıcı olmuştur. Zaman ilerledikçe taşıtlar icat edilmiş ve ilk M.Ö. 200'lü yıllarda ilk araçlar karayollarında görülmeye başlamıştır (İnce, 2009).

19'uncu yüzyılın sonlarından itibaren insanların iradesine sunulan motorlu araçlar zamanla sosyal, bilimsel ve ekonomik gelişmelerle aynı çizgide gelişim göstererek hem sayıları artmış hem de donanımsal olarak gelişme göstermişlerdir. Bu gelişmeler karayolu ulaşımına dair avantajların yanı sıra kaza ve tıkanıklıklar gibi belirli yapısal sorunları da beraberinde getirmekle birlikte insanoğluna yönelik yeni bir ilgi ve uğraş olan trafik kavramını da beraberinde getirmiştir (Tosun, 2004).

1.3.3. Türkiye'de karayollarının durumu

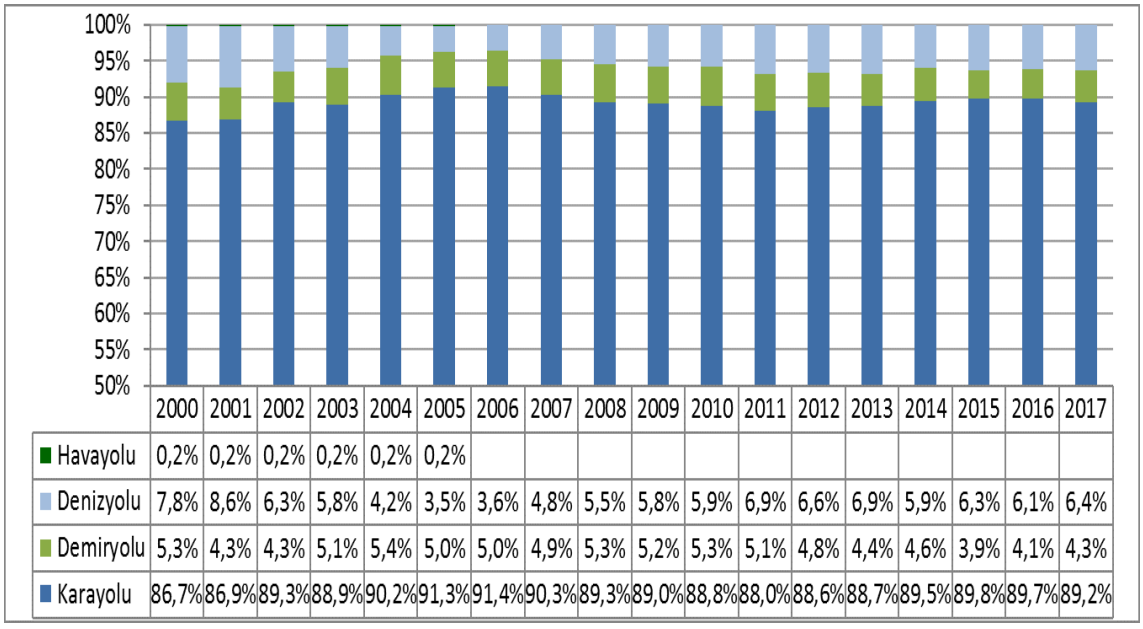
Çevre ve Şehircilik Bakanlığının 2018 yılı resmi verilerine göre kamuoyuna paylaştığı Yurt içi yolcu taşıma oranları dağılımı Şekil 1.2'de, Yurt içi yük taşıma oranları dağılımı Şekil 1.3'de, Ulaşım yollarına göre yurt içi yolcu ve yük taşımacılığı dağılımı Şekil 1.4'de yüzdelik rakamlarla bizlere sunulmuştur (ÇŞB, 2018).



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Havayolu	1,8%	1,6%	1,6%	1,6%	1,8%	2,1%							7,0%	7,9%	8,5%	9,1%	9,4%	9,6%
Denizyolu	0,03%	0,03%	0,02%	0,02%	0,63%	0,65%	0,72%	0,73%	0,74%	0,75%	0,68%	0,63%	0,52%	0,56%	0,59%	0,56%	0,33%	0,59%
Demiryolu	2,2%	2,4%	2,3%	2,7%	2,1%	1,9%	2,0%	1,9%	1,7%	1,6%	1,6%	1,6%	1,1%	1,0%	1,1%	1,1%	1,0%	1,0%
Karayolu	95,9%	95,9%	96,1%	95,7%	95,5%	95,3%	97,3%	97,4%	97,5%	97,6%	97,8%	97,8%	91,5%	90,5%	89,8%	89,2%	89,3%	88,8%

Şekil 1.2. Yurt içi yolcu taşıma oranları (yolcu-km üzerinden % oran) (ÇŞB, 2018).

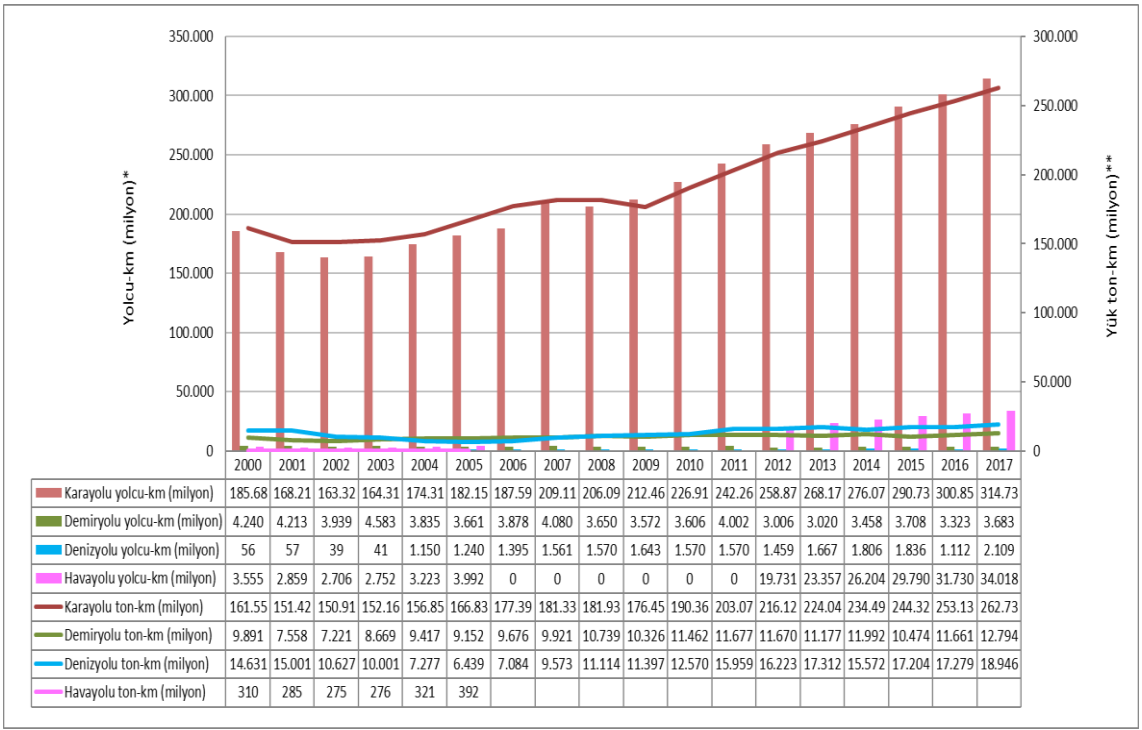
Şekil 1.2'de yolcu-km üzerinden yüzdelik dilimlerle sunulan yurt içi yolcu taşıma oranlarının 2000 yılında %95.9'luk bir paya sahip olduğu, 2013 yılından itibaren yüzdelik olarak düşüşe geçse bile 2000-2017 yılları arasında her zaman çok tercih edilen ulaşım şekli olduğunu bizlere göstermektedir.



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Havayolu	0,2%	0,2%	0,2%	0,2%	0,2%	0,2%												
Denizyolu	7,8%	8,6%	6,3%	5,8%	4,2%	3,5%	3,6%	4,8%	5,5%	5,8%	5,9%	6,9%	6,6%	6,9%	5,9%	6,3%	6,1%	6,4%
Demiryolu	5,3%	4,3%	4,3%	5,1%	5,4%	5,0%	5,0%	4,9%	5,3%	5,2%	5,3%	5,1%	4,8%	4,4%	4,6%	3,9%	4,1%	4,3%
Karayolu	86,7%	86,9%	89,3%	88,9%	90,2%	91,3%	91,4%	90,3%	89,3%	89,0%	88,8%	88,0%	88,6%	88,7%	89,5%	89,8%	89,7%	89,2%

Şekil 1.3. Yurt içi yük taşıma oranları (ton-km üzerinden % oran) (ÇŞB, 2018).

Şekil 1.3'de ton-km üzerinden yüzdelik dilimlerle sunulan yurt içi yük taşıma oranlarının 2000 yılında %86.7'lik bir paya sahip olduğu ve 2017 yılında %89.2'lik bir paya sahip olarak 2000-2017 yılları arasında diğer ulaşım yollarına göre en çok tercih edilen yük taşımacılığı olduğunu bizlere göstermektedir.



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Karayolu yolcu-km (milyon)	185.68	168.21	163.32	164.31	174.31	182.15	187.59	209.11	206.09	212.46	226.91	242.26	258.87	268.17	276.07	290.73	300.85	314.73
Demiryolu yolcu-km (milyon)	4.240	4.213	3.939	4.583	3.835	3.661	3.878	4.080	3.650	3.572	3.606	4.002	3.006	3.020	3.458	3.708	3.323	3.683
Denizyolu yolcu-km (milyon)	56	57	39	41	1.150	1.240	1.395	1.561	1.570	1.643	1.570	1.570	1.459	1.667	1.806	1.836	1.112	2.109
Havayolu yolcu-km (milyon)	3.555	2.859	2.706	2.752	3.223	3.992	0	0	0	0	0	0	19.731	23.357	26.204	29.790	31.730	34.018
Karayolu ton-km (milyon)	161.55	151.42	150.91	152.16	156.85	166.83	177.39	181.33	181.93	176.45	190.36	203.07	216.12	224.04	234.49	244.32	253.13	262.73
Demiryolu ton-km (milyon)	9.891	7.558	7.221	8.669	9.417	9.152	9.676	9.921	10.739	10.326	11.462	11.677	11.670	11.177	11.992	10.474	11.661	12.794
Denizyolu ton-km (milyon)	14.631	15.001	10.627	10.001	7.277	6.439	7.084	9.573	11.114	11.397	12.570	15.959	16.223	17.312	15.572	17.204	17.279	18.946
Havayolu ton-km (milyon)	310	285	275	276	321	392												

Şekil 1.4. Ulaşım yollarına göre yurt içi yolcu ve yük taşımacılığı (ÇŞB, 2018).

Şekil 1.4'de sunulan ulaşım yollarına göre yurt içi yolcu ve yük taşımacılığı grafiğine göre 2017 yılında karayolu ile km bazında 300 milyondan fazla yolcu taşımacılığı yapıldığı ve yine 2017 yılında km bazında 250 milyon tondan fazla yük taşımacılığı yapıldığını bizlere göstermektedir.

1.3.4. Türkiye'de motorlu kara taşıtları

Türkiye'de son yıllarda otomotiv kredilerinin hacimlerinin de artmasıyla ülke bazında otomotiv ürünlerinin satışı da artış göstermiştir. Bu durum aynı bazında son yıllarda refah düzeyini, kişi başına düşen geliri ve gayrisafi milli hasılanın da artış gösterdiğini bizlere göstermektedir (Eken ve Çiçek, 2009).

Son yıllarda meydana gelen ekonomik ve teknolojik gelişmelerin neticesinde motorlu taşıt sayısında artış gerçekleşmiştir. Türkiye İstatistik Kurumu (TÜİK) verilerine göre 2020 yılının aralık ayı sonu itibariyle Emniyet Genel Müdürlüğünün (EGM) idari kayıtlarından elde edilen Trafiğe kayıtlı motorlu kara taşıtı sayıları tablosuna ait 2015 ve 2020 yılları arasındaki rakamlar (yol ve iş makinelerine ait rakamların boş olmasından dolayı hariç tutularak) Çizelge 1.1'de gösterilmiştir (TÜİK, 2021).

Çizelge 1.1. Trafiğe kayıtlı motorlu kara taşıtı sayıları (TÜİK, 2021).

YIL	Otomobil	Minibüs	Otobüs	Kamyonet	Kamyon	Motorsiklet	Traktör	Özel Amaçlı Taşıtlar	Toplam
2015	10.589.337	449.213	217.056	3.255.299	804.319	2.938.364	1.695.152	45.732	19.994.472
2016	11.317.998	463.933	220.361	3.442.483	825.334	3.003.733	1.765.764	50.818	21.090.424
2017	12.035.978	478.618	221.885	3.642.625	838.718	3.102.800	1.838.222	60.099	22.218.945
2018	12.398.190	487.527	218.523	3.755.580	845.462	3.211.328	1.885.952	63.359	22.865.921
2019	12.503.049	493.373	213.358	3.796.919	844.481	3.331.326	1.908.999	65.470	23.156.975
2020	13.099.041	493.395	212.407	3.938.732	859.670	3.512.576	1.958.727	70.309	24.144.857

Çizelge 1.1'de görüleceği üzere 2015 yılında toplam 19.994.472 olan trafiğe kayıtlı motorlu taşıt sayısı her geçen yıl artış göstererek 2020 yılında 24.144.857 adete ulaşmıştır. 2020 yılında trafiğe kayıtlı araçlar arasında en yüksek payı 13.099.041 adet ile otomobiller elinde tutmaktadır.

Araştırma çerçevesinin Van ili olması dolayısıyla TÜİK tarafından 26 Ocak 2021 tarihinde kamuoyuna paylaşılan 2020 yılı Motorlu kara taşıtları istatistikleri kapsamında yer alan 2020 yılı il bazında motorlu kara taşıtları dağılımı Çizelge 1.2'de gösterilmiştir (TÜİK, 2021).

Çizelge 1.2. 2020 yılı Van iline kayıtlı motorlu kara taşıtları sayısı (TÜİK, 2021).

VAN İLİNE KAYITLI TAŞIT TİPLERİ	SAYI
Otomobil	27.065
Minibüs	7.209
Otobüs	693
Kamyonet	22.136
Kamyon	5.947
Motorsiklet	3.412
Traktör	9.321
Özel Amaçlı Taşıtlar	801
Toplam	76.584

Çizelge 1.2'de görüleceği üzere 2020 yılı itibariyle Van iline kayıtlı toplamda 76.584 taşıt bulunmaktadır. Bu taşıtlar arasında en büyük payı 27.065 adet ile otomobiller, ikinci sırayı da 22.136 adet ile kamyonetler oluşturmaktadır.

1.3.5. Türkiye'de trafik kazaları

Toplumlar açısından ulaştırma faaliyetlerinin önemliliği her geçen gün artmakla birlikte bu durum birtakım problemleri de doğurmaktadır. Bu problemlerin en başında trafik kazaları, çevre kirliliği, enerji ihtiyaçları, yaşanan trafik tıkanıklığından kaynaklı zaman kaybı gibi problemler baş göstermektedir. Trafik kazalarından meydana gelen maddi ve manevi kayıpların kişi bazında ciddi boyutlarda olmasının yanı sıra ülke ekonomisinde de önemli sorunlara yol açtığı inkâr edilemez bir gerçektir (Tuncuk, 2004).

Trafik kazalarının yaşanmasında bir veya birden fazla etken söz konusu olmakla birlikte bu faktörlerin dayandığı temelleri insan davranışları (Sürücü, yolcu, yaya), taşıt özellikleri (Aracın modeli, üretim yılı vb.), yol faktörü ile çevre faktörü olarak örnekleyebiliriz (Tuncuk, 2004).

TÜİK tarafından 2020 yılında Emniyet Genel Müdürlüğü'nün idari kayıtlarından toplanarak kamuoyuna paylaşılan yıllara göre trafik kazalarına neden olan kusurlara ait veriler içerisinde 2019 yılına ait Türkiye'de yaşanan trafik kazalarına neden olan kusurların sayıları ve yüzdesel dağılımları Çizelge 1.3'de gösterilmiştir (TÜİK, 2020).

Çizelge 1.3. 2019 yılına ait Türkiye'de yaşanan trafik kazalarına neden olan kusurların sayıları ve yüzdesel dağılımları (TÜİK, 2020).

Kusurlar	Sayıları	Toplam kusura oranı (%)
Sürücü Kusuru	180.042	88.02
Yolcu Kusuru	2.572	1.26
Yaya Kusuru	16.726	8.18
Yol Kusuru	1.045	0.51
Araç Kusuru	4.153	2.03
Toplam	204.538	100

Çizelge 1.3'de görüleceği üzere 2019 yılında Türkiye'de meydana gelen kazaların %88.02'si sürücü kusurlarından kaynaklı olup bu durum büyük bir çoğunluğu kapsamaktadır. Sürücü kusurlarından sonra kazaya sebebiyet veren faktörler arasında ikinci sırayı %8.18 ile yayalardan kaynaklı kusurlar almaktadır.

TÜİK tarafından 2020 yılında Emniyet Genel Müdürlüğü'nün idari kayıtlarından toplanarak kamuoyuna paylaşılan yıllara göre Türkiye'de taşıt cinlerine göre trafiğe kayıtlı ve trafik kazasına karışan taşıtlara ait veriler içerisinde 2019 yılına ait Türkiye'de taşıt cinslerine göre trafiğe kayıtlı ve trafik kazasına karışan taşıtların sayıları Çizelge 1.4'de gösterilmiştir (TÜİK, 2020).

Çizelge 1.4. 2019 yılına ait Türkiye'de taşıt cinslerine göre trafiğe kayıtlı araçlar ile trafik kazasına karışan taşıtların sayısı (TÜİK, 2020).

Taşıt Cinsi	Trafiğe Kayıtlı Taşıt Sayısı	Trafik Kazasına Karışan Taşıt Sayısı	Oran (%)
Otomobil	12.503.049	149.111	1.2
Otobüs	213.358	6.036	2.8
Minibüs	493.373	8.395	1.7
Kamyon	844.481	12.181	1.4
Kamyonet	3.796.919	41.849	1.1
Motorsiklet	3.331.326	45.711	1.4
Diğer	1.974.469	17.417	0.9
Toplam	23.156.975	280.700	1.2

Çizelge 1.4'de görüleceği üzere 2019 yılında Türkiye'de taşıt cinsine göre trafik kazalarına karışan araçların dağılımına göre rakamsal bazda en fazla kazaya karışan araç tipi 149.111 adet ile otomobiller olup trafiğe kayıtlı araç sayıları bakımından oransal bazda da en fazla kazaya karışan taşıt cinsi %2.8 ile otobüslerdir.

1.3.6. Türkiye'de trafik güvenliği

Trafik güvenliği tanımı en temel haliyle meskûn ve meskûn olmayan mahallerde trafik kazaları sonucunda yaşanan can kaybı, maddi hasar ve yaralanma gibi istenmeyen durumları minimuma indirgemeyi amaçlayan bir kavram olarak ortaya çıkmaktadır (Şimşek ve ark., 2009).

Gelişim sürecinde olan ülkemizde, geçtiğimiz her gün trafiğe çıkan araç ve sürücü sayılarında artış gözlenmektedir. Bu durumun getirdiği bir sonuç olarak trafik güvenliğine olan ihtiyaç aynı düzeyde artmaktadır (Topuz, 2015). Trafik güvenliğinin sağlanması ile elde edilen kazanımlar hem trafik kazalarını hem de kazalar sonucunda ortaya çıkan can ve mal güvenliğini azaltmaya yöneliktir (Şimşek ve ark., 2009).

Türkiye'de uzun yıllardan beri önemini kaybetmeden süregelen trafik güvenliği kavramı, kişilerin karşı karşıya oldukları önem düzeyi yüksek toplumsal ve güvenlik

sorundur. İnsan, taşıt ve çevre arasındaki etkileşim aşamasında yaşanabilecek problemlerin belirlenmesi, çözüm önerilerinin bulunması ve söz konusu konularda belirli yöntemlerin geliştirilmesi trafik güvenliği kavramının temel amacını oluşturmaktadır (Çelik, 2014).

Trafikte yaşanan ölüm vakalarının en önemli sebebi kural ihlalleri olmakla beraber bu durumu engellemek için alınması gereken en etkili yolun ülkelerde trafik denetimlerinin sistematik bir şekilde uygulanması olacaktır. Trafikte uygulanacak olan denetim faaliyetleri sürücü ve yayaları yaşanabilecek kural ihlallerine karşın caydırıcı bir güç olmak için psikolojide öğrenme kuramı ilkeleri dahilinde trafik kurallarına uyma oranını arttırmaya ve trafik güvenliğine yönelik normların zamanla oluşmasını amaçlamaktadır (Sümer ve Kaygısız, 2015).

Sayıştay Başkanlığı tarafından 2008 yılında yayınlanan Trafik Kazalarını Önleme Faaliyetleri raporuna göre trafik güvenliğini sağlamaya yönelik uygulanan denetimlerin birkaçı aşağıdaki şekildedir (Anonim, 2008);

- Radarla Hız Kontrolü
- Alkol Kontrolü
- Yaya Denetimi
- Trafik Bilgi Sistemi
- Trafik Para Cezaları
- Ceza Puanı Uygulaması
- Psikoteknik Değerlendirme
- Trafik İşaretlemeleri.

2. KAYNAK BİLDİRİŞLERİ

Dietterich (1998), tarafından yapılan çalışmada topluluk öğrenme yöntemlerinden Bagging, Boosting ve Randomization yöntemlerini karşılaştırılmış, Bagging yönteminin deneysel sınıflandırmada en fazla başarıya ulaşan yöntem olduğuna değinilmiştir.

Ho (1998), tarafından yapılan bir çalışmada ise karar ormanlarının oluşturulmasında Random Subspace yöntemi aracılığıyla rastgele alt uzaylarının seçimine değinilmiştir.

Brieman (2001), tarafından yapılan çalışmada Rastgele Ormanlar yönteminin çalışma prensibi, gini kriterine bağlı değişken önem seçimi, algoritmanın özellikleri, gürültülü veriler karşısında ne kadar doğru tahminler ürettiği, algoritmanın arkasında çalışan karar ağaçları ve bağımsız değişkenler ile tahmin etmek istenilen bağımlı değişken arasındaki korelasyona değinilmiştir.

Skurichina ve Duin (2002), tarafından yapılan çalışmada Bagging, Boosting ve Random Subspace yöntemlerinin özellikleri detaylı bir şekilde anlatılmış olup çeşitli simülasyon verilerinin yanı sıra gerçek zamanlı veriler üzerinde uygulaması yapılarak model performansları kıyaslanmıştır.

Timofeev (2004), tarafından yapılan çalışmada ise Sınıflandırma ve Regresyon Ağaçları (CART) algoritmasının teorisine ve yapılan uygulama örneklerine değinilmiştir.

Polikar (2006), tarafından yapılan çalışmada karar verme süreçlerinde topluluk öğrenme yöntemlerine değinilmiş olup sınıflama aşamasında Bagging, Boosting ve Adaboost yöntemlerinden faydalanarak özellik seçimi işlemleri ve ağaç oluşturmaya yönelik örnekler ile bunların farklarına değinilmiştir.

Zhou (2009), tarafından yürütülen bir çalışmada Topluluk Öğrenimi (Ensemble Learning) yöntemlerinin özellikleri, avantajları ve tutarlılığına değinilmiştir.

Akman (2010), tarafından yürütülen çalışmada sağlık alanında veri seti kullanılarak Rastgele Ormanlar yöntemiyle özellik seçimi yöntemine değinilmiş olup sonrasında Bagging, CART ve Rastgele Orman yöntemlerinin performansı karşılaştırılmıştır.

Besnah ve ark. (2011), tarafından yürütülen bir çalışmada Etiyopya'nın başkenti Addis Ababa'da meydana gelen trafik kazalarının sonucunda olası yaralanmalara ilişkin

sebeplere ilişkin verilere CART ve Rastgele Orman algoritmaları üzerinden sınıflandırma modelleri kurularak kazalara ilişkin sebepler yorumlanmış ve ilerleyen dönemlerdeki trafik tedbirleri tartışılmıştır.

Krishnaveni ve Hemalalatha (2011), tarafından yürütülen çalışmada trafik kazaları ve sonucunda yaşanan ölüm ve yaralanmalara ilişkin derlenen veriler üzerinden Naive Bayes, J48, AdaBoostM1, PART ve Rastgele Orman algoritmalarını uygulayarak oluşturulan sınıflandırma modellerinin performansları karşılaştırılmıştır.

Rokach ve Maimon (2014), tarafından yürütülen çalışmada veri madenciliği süreçlerinde sınıflama ve kümeleme yöntemlerinde sıklıkla kullanılan popüler karar ağaçları ve topluluk öğrenme yöntemlerinin teorisine ve Rastgele Ormanlar yönteminde özellik seçimine değinerek bölüm sonlarında uygulama örneklerine yer verilmiştir.

Lessmann ve ark. (2015), tarafından yürütülen bir çalışmada kredi puanlama türlerine ait örnek bir veri seti üzerinde bireysel ve topluluk öğrenme yöntemlerine ait olan farklı sınıflama algoritmaları uygulanarak elde edilen sonuçlara göre model performansları kıyaslanmıştır.

Liu ve ark. (2015), tarafından yürütülen bir çalışmada finansal sahtekarlıkların belirlenmesinde Rastgele Orman algoritması ile özellik seçimi işlemine değinilerek kısmi korelasyon analiziyle çok boyutlu ölçekleme işlemi yapılmıştır.

Hou ve ark. (2017), tarafından yürütülen bir çalışmada sıkışık ve sıkışık olmayan trafik durumuna sahip ağlar üzerindeki mevcut yolculuk sürelerini ve trafiğin akış sürelerini zamansal ve mekânsal olarak derleyerek Rastgele Orman algoritması üzerinden tahmin etmeye yönelik bir model kurmuşlardır.

Ekelik (2019), tarafından yürütülen çalışmada inşaat sektöründe faaliyet gösteren bir firmanın dijital reklamlar üzerinden iletişim kurduğu müşterilerine ait başvuru sürecine ve özelliklerine ait toplanan veriler üzerinden kullanıcıların satış ofisine gelip gelmemesi durumları Rastgele Orman algoritması ile sınıflandırılmış olup değişken önemine göre ortaya çıkan etkenler yorumlanmıştır.

Bezek Güre (2019), tarafından yapılan çalışmada PISA 2015 verilerine göre Türkiye'deki öğrencilerin matematik başarılarını etkileyen faktörlerin belirlenmesi amacıyla Rastgele Ormanlar yöntemiyle özellik seçimine değinilmiş olup çalışmanın sonunda Rastgele Ormanlar yöntemi ile Çok Katmanlı Algılayıcı Yapay Sinir Ağları ve Radyal Tabanlı Fonksiyon Yapay Sinir Ağları yöntemlerinin sınıflama performansları

karşılaştırılmıştır. Rastgele Ormanlar yönteminin, diğer iki yönteme göre daha düşük hatalarla modeli tahmin ettiği gözlemlenmiştir.

Kruber ve ark. (2019), tarafından yapılan çalışmada otonom araçların sürüş güvenliklerini geliştirmek amacıyla trafik senaryoları üzerine olası risklerin ve hataların önlenmesine yönelik toplanan veriler üzerinden Rastgele Ormanlar sınıflandırıcısı uygulayarak olası kaza risklerini sınıflamış ve aynı zamanda yaşanacak olan kaza senaryolarını kümeleme analizi uygulayarak riskleri gruplamıştır.

Schonlau ve Zou (2020), tarafından yapılan çalışmada Rastgele Orman yönteminin özelliklerine değinilerek bölüm sonunda 2 farklı veri seti ile uygulamalara yer verilmiştir.

3. MATERYAL VE YÖNTEM

3.1. Rastgele Ormanlar (Random Forest) Yöntemi

3.1.1. Tanım ve algoritma

Rastgele Ormanlar yöntemini ilk olarak 2001 yılında Leo Breiman geliştirmiştir. Breiman tarafından 1996 yılında geliştirilen Bagging yöntemiyle 1998 yılında Ho tarafından önerilmiş olan ve rastgelelik yöntemiyle alt grupların seçimi amaçlanan The Random Subspace yönteminin birleştirilmesiyle oluşan yeni bir yöntemdir. Breiman tarafından bu yöntemin geliştirilmesi aşamasında, Amit ve Geman tarafından 1997 yılında ortaya çıkarılan, her bir düğümde en iyi ayırımın rastgelelik yöntemiyle yapılan seçimler sonucunda belirlenen bir çalışmadan etkilenilmiştir (Breiman, 2001).

Rastgele Ormanlar, topluluk öğrenme yöntemi olup birbirinden bağımsız CART karar orman topluluğunu oluşturur. Oluşan karar ormanlarından ortaya çıkan sonuçlar bir araya getirilerek nihai tahmin işlemi yapılır. Rastgele Ormanlar yönteminde oluşturulan ağaçlar, seçilen Bootstrap örneklemleri ve her bir düğümün kırılımında rastgele olarak seçilen m adet tahminci ile ortaya çıkar. Burada dikkat edilecek en önemli husus söz konusu m adet tahmincinin toplam tahminci sayısından daha küçük bir sayıda olmasıdır. Oluşturulan karar ağacına herhangi bir budama işlemi uygulanmaz ve bu ağaçlar en geniş halinde bırakılır (Cutler ve ark., 2013).

Rastgele Ormanlar yöntemini Bagging yönteminden ayıran en önemli fark temelde kaç değişkenin seçimidir. Örnek vermek gerekirse toplam m adet örnek içeren bağımsız değişken içerisinden p adet örnek seçtiğimizde eğer p ve m değişkenleri aynı sayıda ise rastgele orman yönteminin Bagging yönteminden hiçbir farkı olmaz, fakat m adet örneğin içerisinden tarafımızca seçilen p adet örnek mevcut m kadar örnekten daha az bir sayıda olursa kullanılan Rastgele Ormanlar yönteminin, hem sınıflama hatası hem de Out Of Bag (OOB) hatası minimum seviyelerde tutulabilir (James ve ark., 2013).

Rastgele Ormanlar yöntemi, CART oluşturulmasına göre değişim göstermektedir. Eğer sınıflandırma amacıyla kullanılıyorsa her yaprak düğümünü sadece bir sınıfa ait üyeleri içerecek şekilde oluşturulur. Eğer regresyon için kullanılacaksa, nihai olan yaprak

düğümünde minimum birim kalana kadar ağaçların bölünme işlemi devam eder (Cutler ve ark., 2013).

Rastgele Ormanlar yönteminin sıklıkla tercih edilmesinin nedeni, bilinen makine öğrenme yöntemleri içerisinde tahmin geçerliliğinin çok yüksek olması, kurulan modeli çok daha kolay ve başarılı bir şekilde yorumlayabilmemizi sağlaması ve belirli optimizasyon tekniklerini bizlere sunmasıdır (Qi, 2012).

Rastgele Ormanlar yönteminde eğitim verileri ile kurulacak modele sonradan veri ekleyerek güvenilirliği yüksek tahminler yapma ve ilgili veri setindeki değişkenlerin önem derecelerini hesaplama işlemleri de mümkündür. Bu işlem sayesinde değişken sayısı yüksek olan veri setlerinde bağımlı değişkenin seyrini değiştiren asıl değişkenleri ele alarak modeli indirgemek de mümkün olmaktadır. Bu sayede modele yönelik daha etkin tahminler yapılmasına olanak sağlamaktadır (Akman, 2010).

Rastgele Orman yöntemi için önem arz eden diğer kavramlar ise IB ve OOB kavramlarıdır. Rastgele Ormanlar yönteminde her bir ağacı oluştururken, orijinal veri setindeki gözlem sayısı n ile aynı olacak şekilde Bagging yöntemi ile N adet örneklem oluşturulmaktadır (Dietterich, 1998). Orijinal veri setindeki gözlemlerin: 2/3'ü IB verisi olarak örneklemde kullanırken, 1/3'ü de kurulan modelin iç hata oranını test etmek amacıyla OOB verisi olarak örneklem dışında kalmaktadır (Ho, 1998). IB verisi ile en büyük genişlikte CART karar ağacı oluşturulmaktadır. Bu ağaç oluşturulurken her düğümde mevcut tüm tahmin değişkenleri içerisinden en iyi değişkenleri seçmek yerine, her düğüm bölünmesinde toplam m tane bağımsız değişkenden p tanesi $p<m$ olacak şekilde rastgele seçilmektedir. Çünkü ağacın aşırı büyümesi ve aşırı uyum göstermesi bizler için istenilen bir durum değildir. Seçilen bu p tane tahminci değişkenden bilgi kazancı en yüksek olan değişken ile dallara ayırma işlemi gerçekleşmektedir (Timofeev, 2004). Orman oluşturulacak olan N ağaç sayısı elde edilinceye kadar önceki adımlar tekrar edilmektedir. Ayrıca N tane ağacın ayrı ayrı yapmış olduğu sınıf tahminleri bir araya getirilerek yeni bir tahminde bulunulmaktadır. İncelenen bir gözlemin hangi kategorilerde kaç defa sınıflandırdığı sayılarak belirli bir oy çokluğuna göre sınıflandırma işlemi uygulanmaktadır. Örneğin iki kategori içeren bir yapıya sahip sınıflandırma modelinde, bir gözlem, tüm ağaçlar üzerinden en az %51 oy çokluğunu aldığı sınıfın niteliklerini içermektedir (Breiman, 2001).

Rastgele Ormanlar algoritması genel olarak sınıflandırma, özellik seçimi ve regresyon konu başlıkları altında kullanımlarıyla karşımıza çıkmaktadır (Lahouar ve Slama, 2015).

Rastgele Ormanlar yönteminin uygulaması sade ve anlaşılması gayet kolaydır. Uygulamada çalışmayı yürütecek olan analistin belirlemesi gerekli olan iki adet parametre vardır, bu parametreler oluşturulacak ağaçların sayısı ve seçilecek değişkenlerin sayısıdır. (Akman, 2010).

Rastgele Ormanlar yönteminin algoritması aşağıda verilen adımlar izlenerek kurulmaktadır (Abellan ve ark., 2017; Liaw ve Wiener, 2018);

1) Her bir karar ağacı oluşturulurken, orijinal veri seti içerisindeki n adet gözlem sayısı ile aynı olacak şekilde Torbalama yöntemiyle N adet örneklem oluşturulmaktadır.

2) Orijinal veri setindeki gözlemlerin: 2/3'ü IB verisi olarak eğitim amaçlı örneklemde bulunurken, 1/3'ü de kurulan modelin diğerlerinden bağımsız olarak iç hata oranını test etmek için OOB verisi olarak örneklemin dışında kalmaktadır.

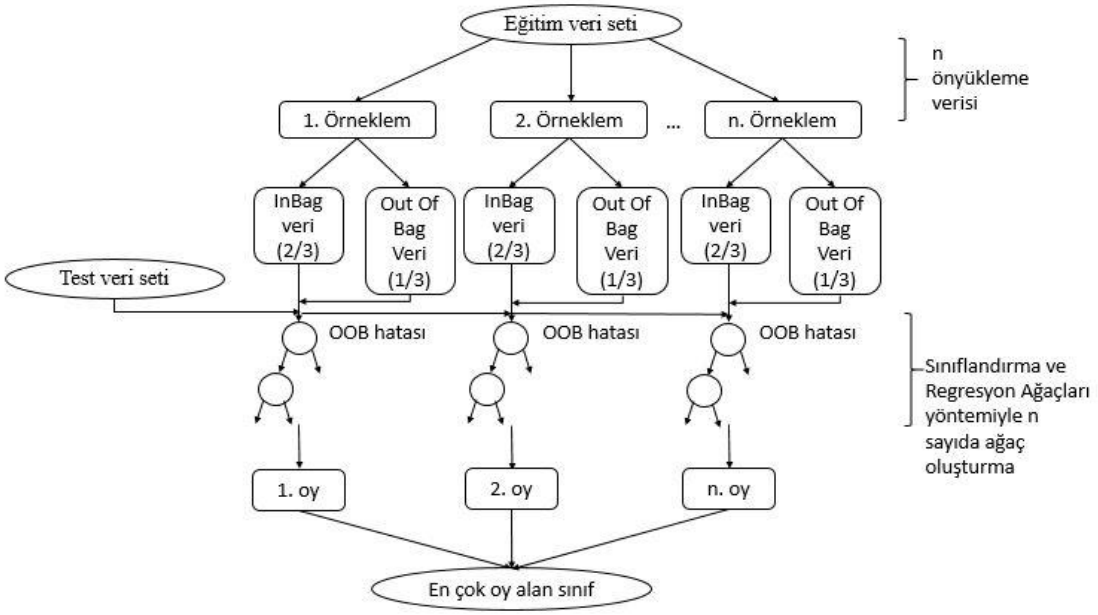
3) IB verisi ile en büyük genişlikte CART karar ağacı oluşturulmaktadır. Bu ağaç oluşturulurken her düğümde mevcut olan tüm tahmin değişkenlerinin içerisinden en iyi değişkenleri seçmek yerine, her düğüm bölünmesinde toplam m tane bağımsız değişkenden p tanesi rastgele seçilmektedir ($p<m$). Çünkü söz konusu ağacın aşırı uyum problemi göstermesi istenmemektedir. Seçilen bu p adet tahminci değişkenden bilgi kazancı en yüksek olan değişken ile dallara ayrılma gerçekleşmektedir. Bu değişkenin hangi değerine göre ayrımın olacağına Gini indeksi ile karar verilmektedir. Hesaplanan değere göre veri seti her düğümde iki alt dala ayrılmaktadır. Bu işlem her düğüm için yeni oluşturulacak dal kalmayıncaya kadar tekrar edilmektedir ve elde edilen bu karar ağaçları budanmamaktadır.

4) Ormanı oluşturacak olan N adet ağaç sayısı elde edinceye kadar önceki adımlar tekrar edilmektedir. Ardından N adet ağacın ayrı ayrı yapmış olduğu sınıf tahminleri bir araya getirilerek yeni bir tahminde bulunulmaktadır. İncelenen bir gözlemin hangi kategorilerde kaç kez sınıflandırıldığı sayılmaktadır. Her gözleme, ağaç setleri üzerinden belirlenen bir oy çoğunluğu ile sınıf ataması yapılmaktadır. Örneğin iki kategoriye sahip bir sınıflandırma modelinde, bir gözlem, tüm ağaçlar üzerinden en az %51 oy çoğunluğunu aldığı sınıfın etiketini taşımaktadır ve bu sınıf, onun tahmin edilmiş sınıf değeri olmaktadır. Sınıflandırma ağaçları için en çok oyu alan sınıf, en son tahmin olarak

seçilirken, regresyon ağaçları için yapılan oylamanın ortalaması alınarak nihai tahmin yapılmaktadır.

5) Bireysel ağaçlarda kullanılmayan OOB gözlemleri ile yapılan tahminler ise ormanın iç hata oranının kestirimini yapmak için kullanılmaktadır. Ormanı oluşturan her karar ağacının OOB hata oranı hesaplanmaktadır. Yanlış sınıflandırmanın yüzdesi rastgele ormanlar sınıflandırma hatası oranı olarak belirlenmektedir.

Rastgele Ormanlar yönteminin genel şeması Şekil 3.1'de tanımlanmaktadır (Ayas, 2014);



Şekil 3.1. Rastgele Ormanlar yönteminin genel şeması (Ayas, 2014).

Rastgele Ormanlar yöntemlerine ilişkin ağaç tipi sınıflandırıcısı aşağıdaki şekilde uygulanır (Breiman, 2001);

$\{h(x, \theta_k), k = 1,\ldots\}$

x: Girdi verisi.

θ_k: Rastgele Vektör.

Rastgele Ormanlar yönteminde mevcut olan dalların arasındaki en iyi sınıflandırıcıyı belirlemek amacıyla "Eş. 3.1" de gösterildiği şekilde gini indeksinden faydalanılır (Pal, 2005);

$$\sum\sum j \neq i(\int(C_i, T)/|T|)(\int(C_j, T)/|T|) \quad (3.1)$$

Burada;

T: Eğitim veri kümesi

C_i: Verinin ait olduğu sınıf

$\int(C_i, T)/|T|$: Seçilen verinin C_i sınıfına ait olma olasılığı

$\int(C_j, T)/|T|$: Seçilen verinin C_i sınıfı dışındaki sınıflardan birine ait olma olasılığını göstermektedir.

3.1.1.1. Rastgele ormanlar regresyonu

Rastgele Ormanlar yönteminin regresyon alanında uygulamasının temel prensibi, rastgele olarak seçilen birden fazla karar ağacının her birinin farklı eğitim kümeleri ile eğitilmesiyle ortaya çıkan kararların ortalamaları alınarak nihai bir tahmin sınıfının oluşması işlemidir (Breiman, 2001). Bu uygulamaya ilişkin dallanma kriterlerinin belirlenmesinde Gini indeksi yönteminden faydalanılır. Bu yöntemi kullanırken de iki adet parametreden faydalanılmaktadır. Bu parametreler ise ilki her bir düğümde seçilecek örneklem sayısı olan n ile oluşturulacak ağaç sayısı olan N parametreleridir (Çölkesen, 2015).

Rastgele Ormanlar yönteminin regresyon için kullanıldığı çalışmalarda da model indirgemesi aynı şekilde uygulanmakta ve daha etkin tahminler yapmak mümkün olmaktadır. Aradaki tek fark ise bağımlı değişken sınıflandırma için kategorik iken regresyon çalışmalarımızda bağımlı değişkenimiz sürekli değişkendir. Dolayısı ile regresyon uygulamasında her bir ağaç kendi tahminlerini üretir ve nihai karar aşamasında yaprak düğümünde minimum birim kalana kadar devam ederek elde edilen bireysel tahminlerin ortalamaları alınmaktadır (Cutler ve ark., 2011).

3.1.1.2. Rastgele ormanlar sınıflandırıcısı

Günümüzde çeşitli sınıflama örneklerinde sıklıkla tercih edilen Rastgele Ormanlar algoritmasının sınıflama problemlerinde etkili bir performans gösterdiği görülmektedir (Akar ve Güngör, 2005).

Rastgele Ormanlar yönteminde, {h(x, θ_k) k= 1,..,} şeklinde ağaç tipi sınıflandırıcılardan fayda sağlanır. Buradaki;

x: girdi verileri

θ_k: rastgele vektör olarak tanımlanmaktadır (Breiman, 2001).

Rastgele Ormanlar yöntemi uygulamasında gerçek veri setinden yer değiştirmeli olarak yeni bir veri seti oluşturulur ve söz konusu yeni oluşturulan veri setine rastgele özellik seçimi yöntemi uygulanarak yeni bir ağaç geliştirilir. Bu yeni geliştirilen ağaca hiçbir budama işlemi yapılmaz (Archer, 2008).

Rastgele Ormanlar sınıflandırıcıları ile bir ağacın üretilmesi için kullanıcı tarafından kullanılan 2 adet parametreden faydalanılır. Bu parametreler ise en optimal bölünmeyi belirlemek amaçlı her bir düğümde kullanılacak değişken sayısı ile geliştirilecek ağaç sayılarıdır (Pal, 2005). Başlangıçta kullanılacak değişken sayısı kullanıcı tarafından rastgele seçilerek bu değişkenler genelleştirilmiş hataları minimize edecek şekilde OOB arttırılır ya da azaltılır. Değişken sayısının optimum değerden az olması korelasyonu ve gücü azaltırken, fazla olması da bu değerleri arttırır. İkinci parametre olan maksimum ağaç sayısı da belirlenerek bu ağaçlar üzerinden bölünme işlemi sağlanır (Özkan, 2013). Bölünme işlemlerinde ise homojen sınıf dağılımlarına ait düğümlerin seçilmesiyle her bir örneğin ilgili sınıfta olma olasılığı işlemini uygular (Pal, 2005). Bu işlemin uygulanmasında Gini indeksinden faydalanır. Sınıf homojenliğini sağlamak için Gini indeksini minimum değerde tutar. Bu indeks sıfıra ulaşınca bu durum her bir yaprağın belirli bir sınıfa ait olduğunu gösterir ve sınıflandırma işlemi tamamlanır (Watts ve ark., 2011). Kaç adet ağaç üretmek isteniyorsa ilgili her bir düğüm için en iyi dalın belirlenmesi amaçlanarak o kadar ağaç üretilir (Liaw ve Wiener, 2002).

3.1.2. Özellik seçimi (Feature selection)

Değişken önem derecesi, söz konusu değişkenin tahmin etme gücünü ölçmektedir. Tahminci değişkenlerin önem derecesi değişkenlerin seçiminde ve kurulacak olan karar ormanlarını etkin bir şekilde yorumlamamıza fayda sağlamaktadır (Breiman ve Cutler, 2005). Bazı istatistiksel analizleri uygulama aşamasından öncesinde, yüksek boyuta sahip öz değerlerin indirgenmesi işleminde temel bileşenler analizi işlemi uygulansa dahi bu yöntemin tahmin için önemli bilgileri yakalayamama gibi dezavantajları mevcuttur. Bu

sebeple temel bileşenler analizinin aksine önemlilik derecesi yüksek olan değişkenleri ele alarak model kurma işlemi daha sık tercih edilir (Cutler ve ark., 2011).

Rastgele Orman yönteminde tahminler oluşturulurken doğrudan özellik seçimi işlemi gerçekleştirilir. Değişken önem derecesinin hesaplanmasının amacı modelin gösterdiği performansı güçlendirerek aşırı uyum problemini ortadan kaldırmak ve veriyi üreten süreçteki etkenleri daha iyi bir şekilde ortaya çıkarmaktır (Qi, 2012).

3.1.2.1. Standart yöntem

Rastgele Ormanlar yönteminde, söz konusu m. değişkenin önem derecesi bulunurken genel olarak uygulanan yöntem ise karar ağacının oluşturulması işleminden sonra OOB test verisinin içerisinde mevcut doğru sınıfa ait tahmin sayısı (c_i) baz alınır. Sonrasında ise OOB verisinde mevcut olan m. değişkenin mevcut değerlerinin kendi içerisindeki yerleri değiştirilerek elde edilen yeni OOB verisi, içerisindeki doğru sınıfa ait tahmin sayısı (c_i^*) kaydedilir. Elimizde mevcut olan doğru tahminlerin sayısı ile elde edilen yeni doğru sınıf tahmin sayıları arasındaki fark alınır ($d_i=c_i-c_i^*$). Mevcut m. değişken için ilgili fark alma işlemi karar ormanı içerisindeki tüm ağaçlar için tekrar edilir ve elde edilen mevcut d_i değerlerinin ortalamaları alınarak söz konusu m. değişken için ağaç düğümüne ait ortalama önem katsayısı ortaya çıkar. Mevcut veri setindeki tüm değişkenler içerisinde de aynı işlemler uygulanır ve söz konusu karar ağaçlarının birbirinden bağımsız ve d_i'lerin normal dağıldığı varsayımı altında standart hatası hesaplanır. İlk bulduğumuz m. değişkenin ortalama önem derecesinin hesaplanan standart hata değerine bölünme işlemi yapılarak m. değişkenin önem derecesi değeri "Eş. 3.2" deki gibi kabaca hesaplanır (Akman, 2010);

$$\text{Önem Derecesi Skoru} = \frac{\bar{d}}{SEd_i} \tag{3.2}$$

$\bar{d}$ = Mevcut m. değişkenin ağaç düğümüne ait ortalama önem katsayısı
SEd_i: Birbirinden bağımsız mevcut tüm d_i değişkenlerinin standart hatası.

Değişken önem derecesini birbirine benzerlik gösteren iki yöntem ile hesaplayabiliriz. Bunlardan ilki Gini önemliliğine dayalı yöntem, diğeri de Permutasyona dayalı değişken önemliliğidir (Breiman ve Cutler, 2004).

3.1.2.2. Gini önemliliğine dayalı yöntem

Gini önemliliğine dayalı yöntem, Rastgele Ormanlar yöntemine yönelik karar ormanlarının oluşması aşamasında gini indeksinden faydalanarak elde edilen bir yöntemdir. Bildiğimiz üzere gini indeksi bir mevcut veri seti içerisindeki homojenlik seviyesini belirlemeye yarayan sayısal bir ölçektir. Örnek vermek gerekirse iki sınıftan oluşan bir sınıflandırma probleminde belirli bir k düğümüne ait pozitif gözlemlerin oranı $p(j|k)$, negatif gözlem oranı da 1- $p(j|k)$'ye eşittir. Bu durumda Gini indeksi "Eş. 3.3" de gösterilen şekilde hesaplanmaktadır (Cutler ve ark., 2012);

$$\text{Önem Derecesi Skoru } (\text{Gini}_k) = 2*\sum_j p(j|k)) * (1 - p(j|k)) \quad (3.3)$$

Burada;
$p(j|k)$: k. düğüme ait pozitif gözlem oranı.
1- $p(j|k)$: k. düğüme ait negatif gözlem oranı.

Bir düğüm ne kadar homojen olursa, Gini indeksi değeri de aynı doğrultuda azalacaktır. Örnek verecek olursak önem derecesi hesaplanacak olan v değişkeni üzerinde bölünmenin gerçekleşmesi durumunda elde ettiğimiz yeni düğümün Gini indeksi değeri de düşecektir. Her bir tekil karar ağacına ait v değişkenine ait gini değeri, bu iki değer arasındaki fark alınarak hesaplanmaktadır. Karar ormanına ait tüm ağaçlar oluştuktan sonra kullanıcı tarafından seçilen v değişkeninin bulunduğu ağaçlara ait Gini önemliliği değerleri toplanarak söz konusu v değişkenine ait Gini yöntemine dayalı önemlilik derecesi hesaplanmış olur (Cutler ve ark., 2013).

3.1.2.3. Permutasyona dayalı değişken önemlilik

Rastgele Ormanlar yöntemine göre permutasyona dayalı önem derecesi aşağıdaki şekilde hesaplanmaktadır (Akman, 2010; Cutler ve ark., 2012, Cutler ve ark., 2013);

1. İlk olarak OOB gözlemleri karar ağaçlarından geriye bırakılır ve tahmin sonucu ortaya çıkan değişkenler belirlenir.
2. OOB değişkenindeki diğer tahminci olarak belirlenen değişkenleri sabit tutarak v değişkenine ait gözlem değişkenler rassal olarak karıştırılır. Bu işlemin sonucunda birbirinden bağımsız iki adet tahmin değeri ortaya çıkar.
3. Bu iki adet tahmincinin OOB hata değerleri hesaplanarak aralarındaki fark alınır.
4. Ortaya çıkan birbirinden bağımsız fark değerleri karar ağacı sayısına bölünerek ortalamaları alınır ve buradan önemlilik skorları elde edilir.
5. Ortaya çıkan bu skor değerlerine göre her bir değişkenin önemlilikleri hesaplanarak o önemliliklerin sıralaması oluşturulur.

Permutasyona dayalı değişken önemliliğine dayalı skorların hesaplanmasında matematiksel gösterim "Eş. 3.4" de gösterilmiştir (Tong ve ark., 2021);

$$\text{Önem Derecesi Skoru (P)} = \frac{1}{N} * \sum_{v=1}^{N} OOBError2_v - OOBError1_v \qquad (3.4)$$

Burada;

v: önem derecesi hesaplanacak değişken

$OOBError1_v$: v değişkeninin 1.tahmincisine ait hesaplanan OOB hata değeri

$OOBError2_v$: v değişkeninin 2.tahmincisine ait hesaplanan OOB hata değeri

N: Karar ağacı sayısı

3.1.3. Farklı sınıf büyüklükleri

Rastgele Orman uygulamaları esnasında birbirinden farklı gözlem büyüklüklerine sahip veri setleri tahmin işlemlerinde en büyük sorunların başında gelmektedir (Akman, 2010). Klasik bir tahmin işleminde gözlem sayısı en yüksek olanına odaklanacağı için doğruluk performansına ilişkin hata oranının yüksek olmasına sebebiyet vermektedir. Rastgele ormanlar yöntemi ise değişkenler arasındaki dengesiz veri sayılarına rağmen

bizlere hata oranı en düşük ve en tutarlı sonuçları sunmak için etkili bir yöntem kullanarak veriler ağırlıklandırılır. Bu yöntem, gözlem sayısı küçük olan değişkenlerin etkilerine daha fazla odaklanır ve etkin tahminci değişkenleri gözden kaçırmamaya çalışmaktadır (Cutler ve ark., 2013).

Farklı sınıf büyüklüklerine yönelik ağırlıklandırma kriterine ait matematiksel gösterim "Eş. 3.5" de gösterilen şekildedir (Zhang ve ark., 2016);

$$W_j = \frac{N}{k * n_j} \tag{3.5}$$

Burada;
W_j = İlgili j. sınıfın ağırlığı
N = Toplam gözlem sayısı
k = Sınıf sayısı
n_j = İlgili j. sınıfın gözlem sayısı olarak tanımlanmaktadır.

3.1.4. Rastgele ormanlar ve kayıp veri

Rastgele Ormanlar yönteminde, veri setine ilişkin kayıp verileri kullanabilme yeteneği bulunmaktadır. Bu işlemin gerçekleşmesi aşamasında kayıp verilerin bulunduğu değişkenimiz sürekli değerler içeriyorsa bu değişkene ait eldeki mevcut değerlerin uç değerler barındırması riskine karşın ortalama yerine daha güvenilir olan medyan değeri alınarak eksik kısımlara atama işlemi gerçekleştirilir. Eğer kayıp verilerin bulunduğu değişkenimiz kategorik ise mevcut ilgili değişkenimize ait eldeki mevcut verilerimizin frekansları hesaplanarak maksimum frekans değerine sahip olan kategori eksik verilerin bulunduğu kısımlara atanır (He, 2006).

Eksik veri probleminin giderildiği veri setine Rastgele Ormanlar yöntemi uygulanır ve ortaya çıkarılan modelden örnekler arası yakınlık matrisleri oluşturulur. Matriste elde edilen uzaklık değerleri üzerinden ağırlıklandırma ölçülerinin elde edilmesinde faydalanılır. Elde ettiğimiz değerin eksik olan veriye atama işlemi gerçekleştirilir. Eğer eksik değer içeren değişkenimiz kategorik değerlerden oluşuyor ise uzaklık katsayısı maksimum olanın kategorisi eksik değere atanır. Kayıp değer atama

işlemi tamamlandıktan sonra tekrardan Rastgele Ormanlar yöntemi uygulanarak yeni bir uzaklık matrisi oluşturulur ve aynı şekilde kayıp değer atama işlemleri devam eder. Mevcut süreç 5 kez tekrarlanır. Mevcut yöntemin uzaklığa dayalı en yakın komşuluk olması dolayısıyla rastgelelik çerçevesinde geçerliliği yüksek olmaktadır (Yılmaz, 2014).

3.1.5. Tahmin ve performans ölçütleri

3.1.5.1. Temel ölçütler

Breiman'a (2001) göre Rastgele Ormanlar yönteminde kurulan karar ormanının hata oranı, karar ormanındaki karar ağaçlarının sayısı arttıkça belirli bir limite yakınsar. Ağaç tabanlı sınıflayıcılarda oluşan karar ormanlarının hata oran değeri, kendi içerisinde bireysel olarak oluşan karar ağaçlarının tahmin performanslarına ve birbirleri arasındaki korelasyonun düşüklüğüne doğrudan bağlıdır. Bunun yanı sıra her düğümde seçilen rastgele değişkenlerin her birinin hata oranı değerlerinin Boosting algoritmasına göre daha düşük olmasını sağlamaktadır (Breiman, 2001).

Rastgele Ormanlar yönteminin performans ölçütlerinin belirlenmesinde kullanılan temel ölçütler ve bu ölçütlerin en iyi şekilde anlaşılabilmesi için detaylı açıklamaları aşağıdaki gibidir (Gutierrez, 2015);

TP: Gerçek Pozitif (True Positive): Bu değer, gerçekte pozitif grupta olup model tarafında da doğru bir şekilde pozitif olarak sınıflandırılan verilerin sayısıdır.

TN: Gerçek Negatif (True Negative): Bu değer, gerçekte negatif gruptan olup model tarafından da doğru bir şekilde negatif olarak sınıflandırılan verilerin sayısıdır.

FP: Yanlış Pozitif (False Positive): Bu değer, gerçekte pozitif grupta olup model tarafından yanlış bir şekilde negatif olarak sınıflandırılan verilerin sayısıdır.

FN: Yanlış Negatif (False Negative): Bu değer, gerçekte negatif grupta olup model tarafından pozitif olarak sınıflandırılan verilerin sayısıdır.

Toplam veri sayısı "Eş. 3.6" deki gibi hesaplanmaktadır.

$$\text{Toplam Veri Sayısı} = TP+TN+FP+FN \tag{3.6}$$

Doğruluk (Accuracy): Doğru sınıflandırılan veri sayısının, tüm verilerin sayısına oranıdır. Doğruluk değeri "Eş. 3.7" da gösterilen şekilde hesaplanmaktadır.

$$\text{Doğruluk} = (TP+TN)/(TP+TN+FP+FN) \quad (3.7)$$

Duyarlılık (Sensitivity-Recall): Doğru olarak sınıflandırılan pozitif verilerin sayısının, tüm pozitif verilerin sayısına oranıdır. Duyarlılık katsayısı "Eş. 3.8" da gösterilen şekilde hesaplanmaktadır.

$$\text{Duyarlılık} = TP/(TP+FN) \quad (3.8)$$

Seçicilik (Specificity): Doğru olarak sınıflandırılan negatif verilerin sayısının, tüm negatif verilerin sayısına oranıdır. Seçicilik katsayısı "Eş. 3.9" de gösterilen şekilde hesaplanmaktadır.

$$\text{Seçicilik} = TN/(TN+FP) \quad (3.9)$$

Hassasiyet (Precision): Doğru olarak sınıflandırılan pozitif verilerin sayısının, pozitif olarak sınıflandırılan tüm verilerin sayısına oranıdır. Hassasiyet katsayısı "Eş. 3.10" de gösterilen şekilde hesaplanmaktadır.

$$\text{Hassasiyet} = TP/(TP+FP) \quad (3.10)$$

F Puanı (F Score): Duyarlılık ve hassasiyet metriklerinin harmonik ortalaması alınarak hesaplanmaktadır. F puanı değerinin amacı, birden fazla sınıflandırma modelinin performanslarını kıyaslamada kullanılır. F skorlarının hesaplanması "Eş. 3.11" de gösterilen şekildedir.

$$\text{F Puanı} = 2*\text{Hassasiyet}*\text{Duyarlılık}/(\text{Hassasiyet}+\text{Duyarlılık}) \quad (3.11)$$

Kappa Katsayısı: Kappa katsayısı, verilerin kategorik değerlendirilmesinde birden fazla sınıfın birbiriyle uyumunu belirlemede yol gösterir. Eğer iki sınıf arasında karşılaştırma

yapılacak ise Cohen's Kappa katsayısı kullanılırken ikiden fazla sınıfın birbirleriyle karşılaştırılmasında Fleiss's Kappa katsayısı kullanılır.

Cohen's Kappa katsayısının hesaplanması "Eş. 3.12" de gösterilen şekildedir.

$$\text{Cohen's Kappa: } K = \frac{\Pr(a)-\Pr(e)}{1-\Pr(e)} \tag{3.12}$$

Burada;
Pr(a): İki değerlendiriciye yönelik uyumlu gözlemler toplamı.
Pr(e): Mevcut uyumum şansa bağlı çıkması olasılığı.

Cohen's Kappa değeri -1 ile +1 arasında değerler alır. Eğer;
K: -1 ise mevcut iki sınıf bütünüyle birbirlerinin tersini değerlendiriyor.
K: 0 ise mevcut iki sınıfın uyumu şansa bağlı olarak değerlendiriliyor.
K: +1 ise mevcut iki sınıfın birbiri ile tamamen uyumlu olarak değerlendiriliyor.

Fleiss's Kappa değeri ise "Eş. 3.13" deki şekilde hesaplanmaktadır;

$$\text{Fleiss's Kappa: } K = \frac{\bar{P}-\bar{P}(e)}{1-\bar{P}(e)} \tag{3.13}$$

Burada;
$\bar{P} - \bar{P}(e)$: Rastgeleliğin dışında sınıflar arasında uyuşma olması olasılığı.
$\bar{P} - \bar{P}(e)$: Mevcut sınıflar arasında şansa bağlı gözlemlerin olması olasılığı.

Fleiss's Kappa değeri 0 ile 1 arasında değerler alır. Eğer;
K: 0 ise tüm sınıflar birbirleriyle uyumludur.
K: +1 ise mevcut tüm sınıfların arasındaki uyum tamamıyla şansa bağlıdır ve onun dışında herhangi bir uyuşma durumu bulunmamaktadır.

3.1.5.2. Yakınlık matrisi (Proximity matrix)

İki sınıflı modele ait yakınlık matrisi Çizelge 3.1'de gösterilen şekildedir (Breiman, 2001).

Çizelge 3.1. Örnekler arası yakınlık matrisi.

		Gerçek Sınıf Değeri	
		Pozitif	Negatif
Modelin Sınıf Tahmini	Pozitif	Doğru Pozitif Sayısı (TP)	Yanlış Pozitif Sayısı (FP)
	Negatif	Yanlış Negatif Sayısı (FN)	Doğru Negatif Sayısı (TN)

3.1.5.3. ROC eğrisi altında kalan alan

Bu eğri çizilirken iki metrikten faydalanılır ve bu metriklerin hesaplaması "Eş. 3.14" ve "Eş. 3.15" de gösterilen şekilde yapılır (Hossin ve Sulaiman, 2015);

$$\text{Gerçek Pozitif Oranı (True Positive Rate-TPR)} = TP/(TP+FN) \quad (3.14)$$

$$\text{Yanlış Pozitif Oranı (False Positive Rate-FPR)} = FP/(FP+TN) \quad (3.15)$$

Bu değerler hesaplandıktan sonra X ve Y eksenleri oluşturulur ve bu değerler eksenlerin koordinatlarına atanırlar. Bu iki metriklerin sonucunda elde edilen oranların hesaplanması ile elde edilen X ve Y noktaları üzerinden eğriler oluşturulur. Eğrinin altında kalan alanın büyük olması, modelin de başarısının yüksekliği ile doğru orantılıdır. Eğrinin altında kalan alan ile F-puanı aynı mantık ve doğrultuda bizlere model başarısı hakkında bilgi sağlar. (Hossin ve Sulaiman, 2015).

3.1.6. Hata oranı tahmini

Kurulan modele ilişkin performans kriterleri, mevcut öğrenme algoritmalarının yanı sıra eğitim ile test verilerinin hacmi, sınıfların dağılımları veya hatalı

sınıflandırmalara bağlı olarak farklılıklar gösterebilir. Bunun gibi durumları ortadan kaldırmak amacıyla geliştirilen çeşitli yöntemler mevcuttur (Zhang ve ark., 2017).

3.1.6.1. Holdout yöntemi

Bu yöntemde kullanılacak olan veri seti, eğitim ve test veri olarak iki kısma ayrılır. Eğitim aşamasında modeli oluşturabilmek için eğitim verilerinden faydalanılır. Bu veriler yardımıyla aynı zamanda modele ilişkin temsil yeteneği en fazla olan değişkenler ve en uygun performans ölçütleri eldeki veri setimizin %80'lik kısmını ayırdığımız eğitim verileri üzerinden gerçekleştirilir. Geriye kalan %20'lik test veri seti ise kurulan modelin genel anlamda performansını değerlendirmek amacıyla kullanılır. Holdout yöntemine ilişkin 2 temel dezavantaj mevcut olup bunlardan birincisi veri setinin düşük hacimli olması durumunda test verilerinin hacminin düşük olması, diğeri ise eğitim ve test veri setlerinin bir defaya mahsus birbirinden ayrılması dolayısıyla modelin hata oranında bu ayrımdan kaynaklı temel sorunlar yaşanabileceği yönündedir (Page, 2015).

3.1.6.2. Tekrarlı holdout yöntemi

Birbirinden farklı alt veri setlerini oluşturmak amacı ile her aşamada eldeki veri setine ait belirli bir oranda eğitime ayrılan veri setleri işlemi Holdout yöntemiyle birkaç kez tekrarlanır. Ancak bu modelin en büyük dezavantajı birbirinden farklı veri setlerinin üst üste gelerek hata oranında yükseltici etkiye sebep olması ve yanıltıcı sonuçlara yer açmasıdır (Dua ve Chowriappa, 2013).

3.1.6.3. Üçlü ayırma yöntemi

Üçlü ayırma yöntemini, Holdout yönteminden ayıran en büyük fark, eldeki veri setini eğitim ve test verisinin yanında doğrulama kısmını da ekleyerek üç kısıma ayırmasıdır. Bu yöntemin uygulama safhasında eğitim veri seti üzerinden kurulan modele ait parametrelerin doğrulama verisine göre detaylı olarak optimizasyonu yapılır ve model tekrardan revize edilir. Son olarak test veri seti ise son hali verilmiş modelin performans

çıktılarını değerlendirmek için kullanılır. Bu sayede de kullanılacak modelin seçimi ile model performansının tahmini aynı anda elde edilmektedir (Nordman, 2011).

3.1.6.4. Çapraz doğrulama yöntemi

Çapraz doğrulama yöntemine ait mevcut olan 2 temel hedef bulunmaktadır. Bu temel hedefler aşağıdaki gibidir (Refaeilzadeh ve ark., 2009);

- Bir algoritma aracılığıyla eldeki mevcut veriler üzerinden elde edilen model performansını ölçmek,
- İki ya da daha fazla algoritmanın performansını ölçmek ve mevcut veriye yönelik en iyi algoritmayı seçmek veya bir modele ait iki ya da daha fazla değişken performanslarını birbirleriyle kıyaslamaktır.

Çapraz doğrulama yöntemi, k-katlı çapraz doğrulama ve birini dışarıda bırakarak çapraz geçerleme yöntemleri olarak iki şekilde uygulanmaktadır. İlki olan k-kat çapraz doğrulamada eldeki mevcut veri setini k adet eşit parçalara ayrılır. Eldeki k adet parçadan her defasında bir adet test verisi, geriye kalan k-1 adet veri de eğitim verisi olmak üzere ayrılır. Bu işlemlerin sonucunda elimizde k adet hata oranı mevcut olup bu hata katsayılarının ortalaması alınarak modeldeki genel tahmin ortalaması hesaplanır (Saharidis ve ark., 2011). Elimizde kalan hata sayısı olan k değerinin düşük olması, elde edilen tahminciye ait olan sapmanın düşük ve modele ait gerçek tahminciye ait olan sapmanın da yüksek olması anlamına gelmektedir (Remesan ve Mathew, 2014). Sapmadan kaynaklı olarak çıkan hatalar gerçek değer ile kurulan model tahmini arasındaki farkı vermektedir. Bu sebeple bu farkın az olması için seçilecek k katsayı değeri ele alınırken çok düşük değerlerden kaçınmamız gerekmektedir (Scott, 2012).

Çapraz doğrulama yönteminin uygulamasına matematiksel gösterim "Eş. 3.16" da gösterilen şekildedir (Akman, 2010);

$$\text{Cross validation (CV)} = \frac{1}{k}\sum_{i=1}^{k} A_i \qquad (3.16)$$

Burada;

A_i: Doğrulama metriği

k: Doğrulama sayısı olarak tanımlanmaktadır.

3.1.7. Rastgele ormanlar yönteminin avantaj ve dezavantajları

Rastgele Ormanlar yönteminin algoritmasında modeli oluşturma aşamasında sunulan avantajları aşağıda verilmiştir (Breiman, 2001; Akman ve ark., 2011; Yao ve ark., 2013; Yılmaz, 2014; Abellán ve ark., 2017; Liaw ve Wiener, 2018; Minitab, 2018a);

1) Belli sayıda tahmincilerin arasından en uygun olanlarını seçme işleminde başarılı bir performans göstermektedir. Şu ana kadar literatürde yer alan yöntemler arasında en iyi tahmincileri veren algoritmaların başında yer alır.
2) Veri ön işlemesine gerek olmaksızın genel haliyle çalışma içerisindeki tüm verileri işleyebilme yeteneğine sahip bir algoritmadır. Diğer algoritmalardan farklı olarak ekstra bir şekilde eldeki verileri dönüştürme işlemi, değiştirme işlemi veya yeniden ölçeklendirme işlemi yapmaya gerek bırakmadan tutarlı sonuçlar verir.
3) Veri setindeki belirli sütunlarda mevcut olan uç değerlerden etkilenme problemi diğer sınıflama yöntemlerine göre en düşüktür.
4) Eldeki veri yapısının çeşitli özelliklerine göre esnek bir çalışma prensibine sahiptir. 2'den fazla kategorik veriler de dahil olmak üzere eldeki veri setine ait bağımlı değişkenlerin sürekli ve kategorik değişkenlerden oluşması yine de bu yöntemin etkili çalışma performansında sıkıntı yaratmamaktadır.
5) Kayıp veri problemlerine ilişkin değer atama yöntemlerinden faydalanılarak bu yöntem ile doğru ve güçlü tahminler elde edilebilir.
6) Rastgele Ormanlar yöntemi, diğer sınıflama yöntemleri içerisinde özellikle tekil ağaç tabanlı öğrenme yöntemlerine göre çok daha doğru ve tutarlı sonuçlar elde edilerek en başarılı alternatif modeller ile de rekabet etmesi dolayısıyla iddialı bir yöntemdir.
7) Birbirinden bağımsız karar ağacının rastgele seçilerek birleştirilmesi sebebiyle eğitim verisindeki aşırı uyum problemine karşı diğer sınıflama yöntemlerine göre çok daha dayanıklı bir yöntemdir.
8) OOB verileri üzerinden modelin kendi kendisini test etmek gibi bir özelliği mevcuttur. Bu özelliği Rastgele Ormanlar yöntemimizin modelden elde ettiği tahminlerde ekstra olarak bir güvenilirlik sağlar.
9) Rastgele Ormanlar yönteminin içerisinde mevcut olan algoritma oldukça hızlı bir çalışma prensibine sahip olması dolayısıyla çalışma anında çıktı ve sonuçları hızlı bir

şekilde verir. Bunun sebebi ise değişken önem dereceleri yüksek olan az sayıda değişken ile model kurma işlemini sağlamasıdır.

Rastgele Ormanlar yönteminin algoritmasında modeli oluşturma aşamasında sunulan dezavantajları ise aşağıda belirtilen şekilde verilmiştir (Breiman, 2001; Akman ve ark., 2011; Yao ve ark., 2013; Yılmaz, 2014; Abellán ve ark., 2017; Liaw ve Wiener, 2018; Minitab, 2018a):

1) CART gibi klasik tekil öğrenimli karar ağaçlarında olduğu gibi ortaya çıkan karar ağacı yapısı, rastgele ormanlar yönteminde maalesef elde edilememektedir. Çünkü topluluk öğrenme yöntemi olan rastgele ormanlar yöntemi birden fazla karar ağacının birleşmesiyle ortaya çıktığı için bu işlemleri aşama aşama göreceğimiz bir karar ağacı grafiği yerine alternatif olarak ısı haritası benzeri alternatif grafiklerden fayda sağlanır.
2) Yapay Sinir Ağları veya Lojistik Regresyon gibi yöntemlerde elde ettiğimiz güven aralıklarını Rastgele Ormanlar modelinde çalışma prensibinden dolayı maalesef elde edememekteyiz.
3) Rastgele Ormanlar yöntemi uygulamalarını düşük bellekli bilgisayarlarda yürütürken, oluşturulan birden fazla karar ağacını bellekte tutulması, yakınlık matrisi ve süreç sonrası vb. işlemler sırasında çok fazla belleğe ihtiyaç duyması sebebiyle ciddi sıklıkta sorunlar yaşanmaktadır.

4. BULGULAR

4.1. Uygulamanın Amacı

Ekonomik gelişmelerin de etkisiyle ülkemizde her geçen gün trafiğe çıkan araç ve sürücü sayılarının da artmasıyla trafikte yaşanabilecek maddi ve can kayıpları risklerinin de artışı trafik kurallarına verilmesi gereken önemin her geçen gün arttığını göstermektedir. Nüfus artışıyla birlikte yollardaki yaya sayısında artışı bu durumun daha da ciddiyet kazanmasını bu duruma yönelik bilimsel çalışmalarla desteklemesi gerekliliğini göstermektedir.

Çalışmamızda sıkı denetime tabi olan ve verilen eğitimlerle de desteklenen trafik kurallarına yönelik insanların algı ve tutumlarını ölçmek amaçlanmış olup 2015 ve 2018 yıllarında Van iline ait yaya ve sürücülere yönelik anket yoluyla alınan bilgiler sonucunda insanların trafik algı ve tutumlarına yönelik hangi etkenlerin daha çok öne çıktığı belirlenmeye çalışılmıştır.

4.2. Uygulama Kapsamı ve Veri Yapısı

Uygulamanın örneklemi için Şehribanoğlu (2019) tarafından Van Yüzüncü Yıl Üniversitesi Bilimsel Araştırmalar Projesi (BAP) desteği ile yürütülen çalışma dahilinde Van İlinde Yaşayanların Trafik İşaretleri Bilgisi ve Trafik Kurallarına Bakış Açıları Üzerine Bir Araştırma adlı çalışmasında kullanılan verilerden yararlanılmıştır. Bu veri seti 2015 ve 2018 yıllarına ait Van il merkezlerinde rastgele örneklem yöntemiyle belirlenen cevaplayıcılara ilişkin farklı zaman aralıklarında ve bölgelerde uygulanan anketlerden oluşmaktadır. Eksik ve hatalı veriler çalışmadan çıkarıldıktan sonra toplamda 738 kişiye uygulan anket kullanılmıştır.

Yapılan bu çalışmada katılımcıların demografik bilgileri dışında trafiğe yönelik algı tutumları ölçülmeye çalışılmış olup 42 adet soru içeren anket uygulaması üzerinden her bir soruda genel anlamda tek bir yargıya yönelik bilgi ve tutumlara cevap aranmıştır. Cevaplayıcıların tercihleri 1-kesinlikle katılmıyorum, 2-katılmıyorum, 3-kararsızım, 4-katılıyorum, 5-kesinlikle katılmıyorum seçeneklerinden oluşmaktadır.

Çalışmanın ana prensibi, bu anket içerisinde trafik kurallarına ilişkin algı ve tutumlara yönelik dikkat çeken 3 adet soruyu etkileyen en önemli faktörleri belirlemektir. Seçilen bu değişkenlerin hem sınıflama performanslarını incelemek hem de kurallara uyma algısını içeren bağımlı değişkenler üzerinde hangi bağımsız değişkenlerin etkili olduğunu özellik seçimi yöntemiyle belirlemek amaçlanmıştır.

Uygulanan anket çalışması sonucunda 2015 ile 2018 yıllarında elde edilen kullanıcı verileri rastgele ormanlar yönteminden faydalanılarak analiz edilmiştir. Analiz aşamasında Python 3 programında RandomForestClassifier kütüphanesinden faydalanılmıştır. Yapılan çalışmada iki sınıf üzerinden sınıflandırma performanslarının değerlendirilmesi amaçlanmış olup orijinalinde verilen likert cevaplara göre 5 sınıflı olan veriler içerisinde bağımlı değişken olarak baz alacağımız değişkenlerin Çakır ve Doğan (2017), tarafından kendi çalışmalarında yapıldığı gibi trafik algı ve tutumların belirlenen kategorilere göre anlamlı düzeyde sınıflandırmalar oluşturabilmek için yapılan ankete ait 5 dereceli cevaplar 2 dereceye indirgenmiştir. Çalışma aşamasında belirgin bir tanımlama içermediği ve sınıflama performansını düşürdüğü için araştırmanın bulgularını olumsuz etkileyebileceği düşüncesiyle kararsız olan 3-kararsızım değişkeni çalışma kapsamının dışında bırakılmıştır. Geriye kalan 1-kesinlikle katılmıyorum ve 2-katılmıyorum değişkenleri 0-katılmıyorum olarak tek bir sınıf içerisine alınırken, 4-katılıyorum ve 5-kesinlikle katılıyorum değişkenleri ise 1-katılıyorum olarak tek bir sınıf içerisinde değerlendirilmiştir. Analiz aşamasında uygulanan Rastgele Ormanlar yönteminin çalışma prensibi gereği elimizdeki veri setinin 2/3'ü (Eğitim verileri) ile model eğitilecek olup geriye kalan 1/3'ü (Test verileri) ile eğitim sonucu kurulan modelin başarısı test edilmiştir.

4.3. Uygulamanın Gerçekleştirilmesi

Trafik kurallarına yönelik algı ve tutuma yönelik 3 adet düşünceyi ele almamız dolayısıyla çalışmanın her birinde farklı bağımlı değişken seçilerek 3 aşamada incelenmiş olup elde edilen bulgular son kısımda tartışılmıştır. Bağımlı değişkenimizi seçtikten sonra geriye kalan tüm değişkenler bizim bağımsız değişkenimiz olacaktır. Uygulama aşamasında kullanılacak ölçüm değişkenleri Çizelge 4.1'de verilmiştir.

Çizelge 4.1. Uygulamaya ilişkin ölçüm değişkenleri.

Ankete İlişkin Tanımlayıcı Özellikler	
Soru 1-) Kazaların, engelleyici ve önleyici tedbirlere daha çok önem vererek engellenebileceğine inanıyorum.	
Ortalama	4.052
Std. Sapma	0.564
Medyan	4.00
Soru 2-) Trafik kazalarının kötü yol standartları yüzünden olduğuma inanıyorum.	
Ortalama	3.352
Std. Sapma	0.832
Medyan	3.00
Soru 3-) Trafik kazalarını insanların yeterince eğitilmemesine bağlıyorum.	
Ortalama	4.066
Std. Sapma	0.606
Medyan	4.00
Soru 4-) Koşullar uygun olduğunda sürücülerin hız limitlerini aşmasında sorun olmadığını düşünüyorum.	
Ortalama	2.605
Std. Sapma	0.942
Medyan	3.00
Soru 5-) Birçok sürücünün yeteneklerini göstermek için hız yaptığına inanıyorum.	
Ortalama	3.884
Std. Sapma	0.814
Medyan	4.00
Soru 6-) Araba sürme zevkine ancak sürat yapıldığında varılacağına inanıyorum.	
Ortalama	1.745
Std. Sapma	0.936
Medyan	2.00
Soru 7-) Araba kullanmanın heyecan verici olması benim için önemlidir.	
Ortalama	2.511
Std. Sapma	0.869
Medyan	2.00
Soru 8-) Güçlü bir arabaya sahip olmanın tek anlamı onu tam kapasite kullanmaktır.	
Ortalama	2.950
Std. Sapma	0.831
Medyan	3.00
Soru 9-) Riske girmek hiçbir zaman eğlenceli değildir.	
Ortalama	4.010
Std. Sapma	0.759
Medyan	4.00
Soru 10-) Trafik kurallarının araba kullanma zevkini yok ettiğine inanıyorum.	
Ortalama	2.061
Std. Sapma	0.795
Medyan	2.00
Soru 11-) Sürücülerin tamamıyla hız kurallarına uyması gerektiğini düşünüyorum.	
Ortalama	4.647
Std. Sapma	0.768
Medyan	4.00
Soru 12-) Araba kullanırken hız yapmanın ve heyecanın ayrılmaz bir ikili olduğuma inanıyorum.	
Ortalama	2.044
Std. Sapma	0.691
Medyan	2.00
Soru 13-) Yetişkinlerin, trafikte eğlence ve heyecana gereksinim duyduğuna inanıyorum.	
Ortalama	2.128
Std. Sapma	0.789
Medyan	2.00
Soru 14-) Hiçbir zaman hızlı araba kullanmayı istemem.	
Ortalama	3.793
Std. Sapma	0.812
Medyan	4.00
Soru 15-) Alkol almış bir kişinin arabasına binebilirim.	
Ortalama	2.255
Std. Sapma	0.932
Medyan	2.00

Çizelge 4.1. Uygulamaya ilişkin ölçüm değişkenleri (devam).

Ankete İlişkin Tanımlayıcı Özellikler	
Soru 16-) Sarhoş birisiyle arabaya kesinlikle binmem	
Ortalama	3.652
Std. Sapma	0.968
Medyan	4.00
Soru 17-) Tanıdığım ve güvendiğim birisinin arabasına alkol almış olsa da binebilirim.	
Ortalama	2.361
Std. Sapma	0.888
Medyan	2.00
Soru 18-) Alkollü araba kullanmak sanıldığı kadar riskli değildir.	
Ortalama	1.514
Std. Sapma	0.764
Medyan	1.00
Soru 19-) Trafikte bir kişiyi yaralarsam hayatıma hiçbir şey olmamış gibi devam edemem.	
Ortalama	3.888
Std. Sapma	0.952
Medyan	4.00
Soru 20-) Bir kazaya neden olacaksam kimsenin yaralanmamış olmasını umarım.	
Ortalama	4.448
Std. Sapma	0.803
Medyan	5.00
Soru 21-) Trafik kazalarının bir kader olduğuna inanıyorum.	
Ortalama	2.060
Std. Sapma	0.780
Medyan	2.00
Soru 22-) Araba kullanmanın beni huzursuz edeceğine inanıyorum.	
Ortalama	1.783
Std. Sapma	0.968
Medyan	2.00
Soru 23-) Araba kullanmanın beni zorlamayacağını düşünüyorum.	
Ortalama	3.840
Std. Sapma	0.830
Medyan	4.00
Soru 24-) Sinirli olduğumda trafikte daha fazla hata yapacağıma inanıyorum.	
Ortalama	3.895
Std. Sapma	0.797
Medyan	4.00
Soru 25-) Sinirli olduğumda trafik kurallarını ihlal edebileceğimi düşünüyorum.	
Ortalama	3.454
Std. Sapma	0.893
Medyan	3.00
Soru 26-) Trafikte bazen risk almak gerekir.	
Ortalama	2.402
Std. Sapma	0.884
Medyan	2.00
Soru 27-) Trafik kurallarına harfi harfine uymak gerekir.	
Ortalama	4.013
Std. Sapma	0.702
Medyan	4.00
Soru 28-) Bir sürücünün risk alması ve bazı trafik kurallarını çiğnemesi onun daha az güvenli bir sürücü olduğu anlamına gelmez.	
Ortalama	2.969
Std. Sapma	0.860
Medyan	3.00
Soru 29-) Bir yaya olarak tüm trafik kurallarına harfi harfine uyarım.	
Ortalama	4.160
Std. Sapma	0.915
Medyan	4.00
Soru 30-) Trafiğin aksamamasını sağlamak için bazen kuralları esnetmek gerekir	
Ortalama	2.925
Std. Sapma	0.795
Medyan	3.00

Çizelge 4.1. Uygulamaya ilişkin ölçüm değişkenleri (devam).

Ankete İlişkin Tanımlayıcı Özellikler	
Soru 31-) Koşullar uygun olduğunda bence hız limitlerini aşmakta sorun yoktur.	
Ortalama	2.605
Std. Sapma	0.910
Medyan	3.00
Soru 32-) Hız limitini 10-15 km/saat aşmakta sorun yoktur çünkü bunu herkes yapıyor.	
Ortalama	2.498
Std. Sapma	0.869
Medyan	2.00
Soru 33-) Trafiğin aksamamasını sağlamak için bazen kuralları esnetmek gerekir	
Ortalama	2.909
Std. Sapma	0.822
Medyan	3.00
Soru 34-) Bazen trafik kurallarının çiğnenmesini göz ardı etmek gerekir.	
Ortalama	2.069
Std. Sapma	0.771
Medyan	2.00
Soru 35-) Trafik akışını sağlamak, kurallara her zaman uymaktan daha önemlidir.	
Ortalama	2.250
Std. Sapma	0.165
Medyan	3.00
Soru 36-) Trafik akışını sağlamak adına uyulması imkânsız birçok kural olduğuna inanıyorum.	
Ortalama	2.867
Std. Sapma	0.751
Medyan	3.00
Soru 37-) Zamanında varmak için bazen trafik kurallarını esnetmenin hiçbir sakıncası yoktur.	
Ortalama	2.079
Std. Sapma	0.729
Medyan	2.00
Soru 38-) Her zaman trafik kurallarına uymaktansa akıcı bir şekilde araba kullanmak daha iyidir.	
Ortalama	2.122
Std. Sapma	0.818
Medyan	2.00
Soru 39-) Şehir içinde emniyet kemerini takmanın gereksiz olduğunu düşünüyorum.	
Ortalama	1.734
Std. Sapma	0.938
Medyan	2.00
Soru 40-) Emniyet kemerinin hayati bir önem taşıdığına inanıyorum.	
Ortalama	4.785
Std. Sapma	0,.691
Medyan	5.00
Soru 41-) Trafik işaret ve levhalarının gereksiz yere fazla olduğunu düşünüyorum.	
Ortalama	2.301
Std. Sapma	0.879
Medyan	2.00
Soru 42-) Yollardaki trafik işaret ve levhalarını okuma ve anlama becerisine sahibim.	
Ortalama	3.992
Std. Sapma	0.638
Medyan	4.00

Çizelge 4.1'de söz konusu 42 adet değişkenin ham veri halindeki durumlarına ait ortalamaları, standart sapmaları ve medyanları verilmektedir.

4.3.1. Trafik kurallarına harfi harfine uyarım değişkenine yönelik tutum ve sınıflandırma ölçeği

Trafik kurallarına karşı algı ve tutumların sınıflandırılmasına yönelik uygulama aşamasında bağımlı değişken olarak ele alacağımız ilk ölçeğimiz anket uygulamamızın 27. sorusu olan "Trafik kurallarına harfi harfine uyarım" değişkenidir. Çalışmamızda veri setinden yola çıkarak Rastgele Ormanlar algoritması ile yeni bir sınıflama modeli oluşturduk. Söz konusu modelin uygulanmasındaki temel amaç veri setindeki bağımsız değişkenlerin hangilerinin bağımlı değişkenimiz olan trafik kurallarına harfi harfine uymaya yönelik tutumların sınıflanmasına katkı yaptığını belirlemek ve yapılan sınıflamanın performansını ölçmektir.

Veri setimizdeki bağımlı değişkenimiz olan "Soru 27-) Trafik kurallarına harfi harfine uyarım" sorusuna ilişkin cevaplara ilişkin sınıf değerleri Çizelge 4.2'deki gösterilmiştir.

Çizelge 4.2. Veri seti içerisinde bağımlı değişkene ilişkin toplam sınıf bilgileri.

Soru 27-) Trafik kurallarına harfi harfine uyarım	
Sınıf Değeri	Denek Sayısı
0	38
1	700

Çizelge 4.2'de görüleceği üzere sınıf değeri 0-katılmıyorum olan sınıfın denek sayısı 38 kişi, sınıf değeri 1-katılıyorum olan sınıfın denek sayısı da 700 kişi olduğundan Python 3 programında class_weight fonksiyonundan faydalanılarak sınıf değerleri ağırlıklandırılmıştır. Bu işlem sonucunda denek sayısının düşük olduğu sınıfa yüksek ağırlık, denek sayısı fazla olan sınıfa da düşük ağırlık verilerek sınıfların dağılımları dengelenmiştir. Sınıf değerlerine ilişkin uygulanacak sınıf ağırlıkları Çizelge 4.3'de gösterilmiştir.

Çizelge 4.3. Sınıf değişkenlerini dengelemek için kullanılacak ağırlıklar.

Soru 27-) Trafik kurallarına harfi harfine uyarım	
Sınıf Değeri	Sınıf Ağırlığı
0	1.8
1	0.1

Bağımlı değişkene göre eğitim ve test veri setleri Çizelge 4.4 ve Çizelge 4.5'de gösterilmiştir.

Çizelge 4.4. Bağımlı değişkene göre eğitim veri seti.

Soru 27-) Trafik kurallarına harfi harfine uyarım	
Sınıf Değeri	Denek Sayısı
0	23
1	471

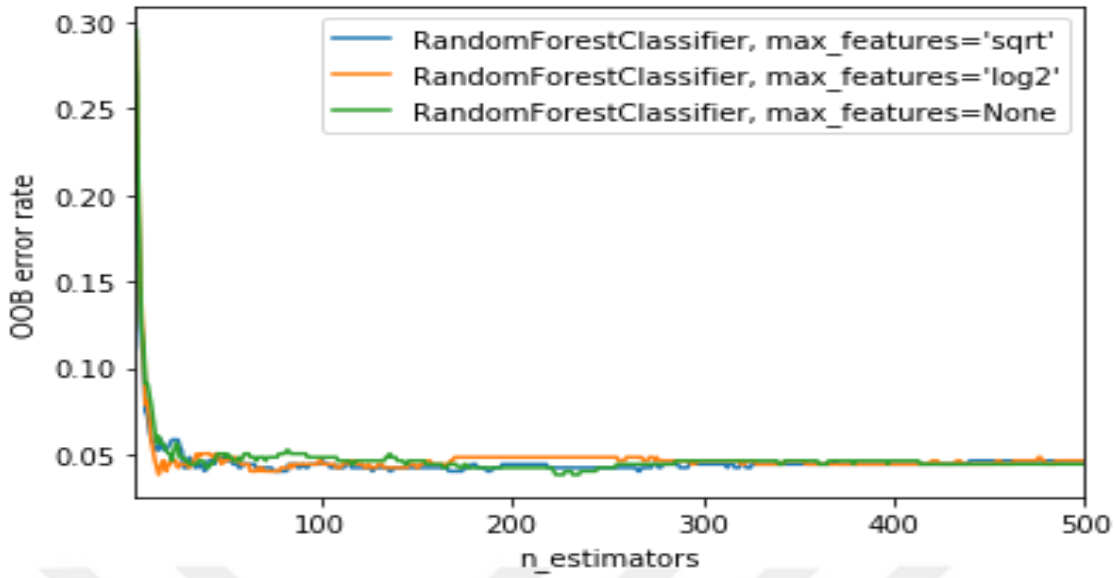
Çizelge 4.4'de görüleceği üzere bağımlı değişken baz alınarak toplam veri setinin 2/3'ü ile oluşturulan eğitim veri setine ilişkin sınıf değeri 0-katılmıyorum olan değişkene ilişkin denek sayısı 23 kişi, sınıf değeri 1-katılıyorum olan değişkene ilişkin denek sayısı 471 kişidir.

Çizelge 4.5. Bağımlı değişkene göre test veri seti.

Soru 27-) Trafik kurallarına harfi harfine uyarım	
Sınıf Değeri	Denek Sayısı
0	15
1	229

Çizelge 4.5'de görüleceği üzere bağımlı değişken baz alınarak toplam veri setinin 1/3'ü ile oluşturulan test veri setine ilişkin sınıf değeri 0-katılmıyorum olan değişkene ilişkin sınıf sayısı 15 kişi, sınıf değeri 1-katılıyorum olan değişkene ilişkin sınıf sayısı ise 229 kişidir.

Modelimizin kurulumuna ilişkin öncelikle belirlenmesi gereken kriter ağaç sayısı kriteridir. Modelin ilk kurulumunda ağaç sayısı keyfi olarak belirlenerek bu grafik yardımıyla optimum ağaç sayısı tekrardan belirlenir. Seçilecek olan ilk 500 adet ağaç sayılarına ilişkin kurulan rastgele ormanlar modelinin hata oranlarına ilişkin çıktılar Şekil 4.1'de verilmiştir.



Şekil 4.1. Ağaç sayılarına göre OOB hata grafiği.

Şekil 4.1'de görüleceği üzere kullanılan ağaç sayısı arttıkça OOB hata oranlarının da sabitleştiği görülmektedir. Yeni kurulacak olan parametrede kullanılacak ağaç sayısı değeri n_estimators kırılımlarının değerlerinin üçünün de sabit olduğu minimum değer olan 400 değeri olarak belirlenebilir.

Kurulacak modele ilişkin birinci kriterimiz olan kullanılacak ağaç sayısı değerini belirledikten sonra ikinci kriterimiz olan modelde kullanılacak maksimum değişken sayısı max_features kriteridir. Veri setimizde toplamda 42 değişken olmasından dolayı $\sqrt{42} \cong 7$ olarak belirlenerek 400 ağaca ilişkin Çizelge 4.6'da gösterileceği üzere optimal değer hesaplanmıştır.

Çizelge 4.6. Modelde kullanılacak maksimum değişken sayısı.

Ağaç Sayısı	Gini yöntemiyle en önemli bulunan maksimum değişken sayısı	Gini yöntemiyle en önemli bulunan maksimum değişken sayısı modele girdiğinde hata oranı
400	3	0.615
400	**5**	**0.614**
400	6	0.615
400	7	0.615
400	15	0.615
400	21	0.615
400	35	0.615
400	41	0.615

Çizelge 4.6'da görüleceği üzere modelimizde kullanılacak maksimum değişken sayısı gini kriteri ile belirleyeceğimiz değişken önem düzeylerinde mevcut ilk 5 değişken olmuştur.

Modelimizin kurulumuna ilişkin baz alacağımız ağaç sayısı ve modelde kullanacağımız maksimum değişken sayısını Python 3 programının GridSearchCV kütüphanesini kullanarak hiperparametre optimizasyonu yardımıyla kısa yoldan çözümü yapılmaktadır. Bu işlemin uygulama ve kullanımı basit olup optimal değerlerini elde etmek istediğimiz 2 adet kritere ilişkin (Ağaç sayısı ve modelde kullanılacak maksimum değişken sayısı) denemek istediğimiz değerlerin her birini sözlüğe aktararak makine tarafından gerekli iterasyonların yapılmasıyla her bir kritere ilişkin en etkin parametre değerleri program tarafından ekrana yazdırılır. Çizelge 4.7'de yapılan hiperparametre optimizasyonu sonucunda modelin kurulması için kriterlerin en etkili parametre değerleri verilmektedir.

Çizelge 4.7. Hiperparametre optimizasyonu sonucunda en optimal parametre değerleri.

Ağaç Sayısı (n_estimators)	Maksimum Değişken (max_features)
400	5

Çizelge 4.7'de görüleceği üzere yukarıda uyguladığımız uzun işlemler sonucunda elde ettiğimiz parametre değerlerinin tek bir kütüphane yardımıyla kısa yoldan çözümü mevcuttur. Bu işlem sonucunda modelimizde uygulayacağımız ağaç sayısı kriterimiz 400, maksimum değişken sayımız ise 5 olarak belirlenmiştir.

Modelin kurulması için gerekli parametrelerin optimal değerlerinin belirlenme işlemi tamamlandıktan sonra bu parametreler ile algoritma yeniden çalıştırılarak Çizelge 4.8'deki sınıflandırma çizelgesi elde edilmektedir.

Çizelge 4.8. OOB test verisi sonucu sınıflandırma çizelgesi.

	Gerçek Sınıflar	
Tahmin Sınıfı	1-Katılıyorum	0-Katılmıyorum
1-Katılıyorum	219 (TP)	10 (FP)
0-Katılmıyorum	7 (FN)	8 (TN)

Çizelge 4.8'de görüleceği üzere optimal parametreler ile kurulan sınıflama modelinde gerçek sınıf değerlerine ilişkin yapılan tahminler verilmiştir. Yukarıda elde

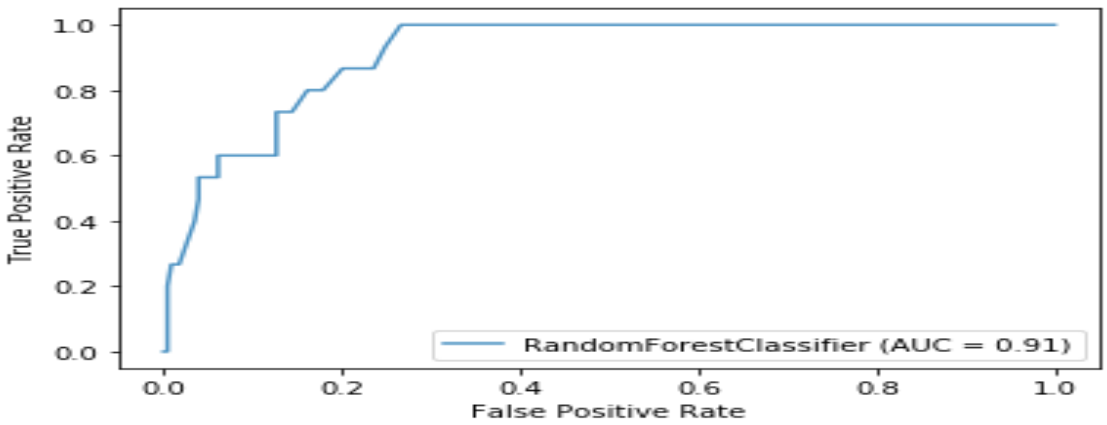
ettiğimiz çizelgenin bir diğer adı da yakınlık matrisidir. Test verisi için sınıflandırma çizelgesi sonucunda elde edilen performans ölçütleri Çizelge 4.9'da verilmiştir.

Çizelge 4.9. Test verisi için sınıflandırma çizelgesine ait performans ölçütleri.

Doğruluk	%93.03
Yanlış Sınıflandırma Oranı	%6.97
Hassasiyet	%95.33
Doğru Pozitif Oranı	%89.75
F-puanı	%92.45
Kappa	%44.78

Çizelge 4.9'da görüleceği üzere test verisine ilişkin doğruluk değerinin %93.03 gibi yüksek bir rakam çıkması kurulan sınıflama modelinin iyi bir performans gösterdiği anlamına gelmektedir. Aynı zamanda bu modelin başarısını F-puanının %92.45 olması performansın güvenilirliğini kanıtlamaktadır. Kappa istatistiğinin de %44.78 olması da mevcut sınıfların birbirleri ile kabul edilebilir düzeyde uyum gösterdiğini göstermektedir. Kappa skorunun sıfırdan büyük olması modele ilişkin sınıfların aralarındaki uyum için istenilen bir durumdur.

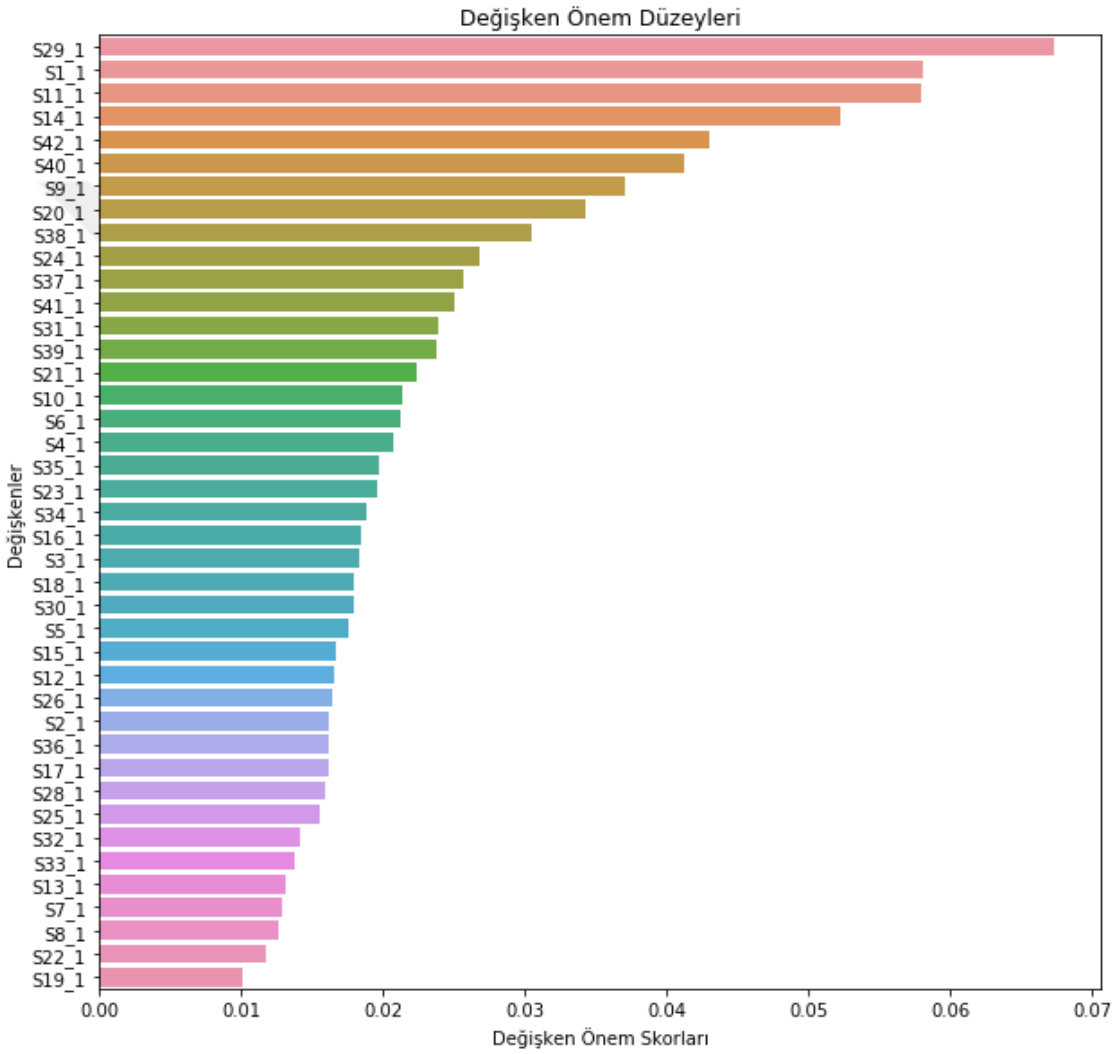
Rastgele Ormanlar algoritmasını kullanarak elde ettiğimiz sınıflandırma modelinin değerlendirilmesinde en önemli ölçütlerden biri de ROC eğrisidir. ROC eğrisi farklı sınıflar için olasılık değerlerini kullanarak eğri altında kalan alana göre modelin performansının iyi olup olmadığını göstermektedir. AUC değerinin en ideal değeri 1 olmakla birlikte bu değerin 0.5'den büyük olması model performansının ne kadar başarılı olduğunu göstermektedir. Kurulan modele ilişkin ROC eğrisi ve AUC değeri Şekil 4.2'deki gösterilmiştir.



Şekil 4.2. Kurulan modele ilişkin ROC eğrisi ve AUC değeri.

Şekil 4.2'de kurulan modele ilişkin eğri altında kalan AUC değerinin 0.91 olması modelin gayet başarılı bir performans sağladığını göstermiştir.

Sınıflandırma modelinin performans kriterlerini de tamamladıktan sonra bağımlı değişkenimiz olan "Soru 27-) Trafik kurallarına harfi harfine uyarım" bağımlı değişkeninin sınıflamasında en etkili değişkenlere ait önem düzeyleri Şekil 4.3'deki grafikte verilmiştir.



Şekil 4.3. Kurulan modele ilişkin değişken önem düzeyleri grafiği.

Şekil 4.3'te modelin içerisinde mevcut olan tüm değişkenlerin önem düzeyleri gösterilmektedir. Modelin kurulumunda baz aldığımız değişken sayısı olan, aynı zamanda maksimum değişken parametre değeri olan ilk 5 değere ait değişken önem düzeyleri sırasıyla Çizelge 4.10'da verilmektedir.

Çizelge 4.10. Kurulan modele ilişkin ilk 5 değişkenin önem düzeyleri.

Değişkenler	Önem Düzeyleri
Soru 29-) Bir yaya olarak tüm trafik kurallarına harfi harfine uyarım.	0.067
Soru 1-) Kazaların çoğu, engelleyici ve önleyici tedbirlere daha çok önem vererek engellenebileceğine inanıyorum.	0.058
Soru 11-) Sürücülerin tamamıyla hız kurallarına uyması gerektiğini düşünüyorum.	0.058
Soru 14-) Hiçbir zaman hızlı araba kullanmayı istemem.	0.052
Soru 42-) Yollardaki trafik işaret ve levhalarını okuma ve anlama becerisine sahibim.	0.043

4.3.2. Trafik kurallarının araba kullanma zevkini yok ettiğine inanıyorum değişkenine yönelik tutum ve sınıflandırma ölçeği

Trafik kurallarına karşı algı ve tutumların sınıflandırılmasına yönelik uygulama aşamasında bağımlı değişken olarak ele alacağımız ikinci ölçeğimiz ise anket uygulamamızın 10. sorusu olan "Trafik kurallarının araba kullanma zevkini yok ettiğine inanıyorum" değişkenidir.

Veri setimizdeki bağımlı değişkenimiz olan "Soru 10-) Trafik kurallarının araba kullanma zevkini yok ettiğine inanıyorum" sorusunun cevaplarına ilişkin sınıf değerleri Çizelge 4.11'deki gösterilmiştir.

Çizelge 4.11. Veri seti içerisinde bağımlı değişkene ilişkin toplam sınıf bilgileri.

Soru 10-) Trafik kurallarının araba kullanma zevkini yok ettiğine inanıyorum	
Sınıf Değeri	Denek Sayısı
0	679
1	53

Çizelge 4.11'de görüleceği üzere sınıf değeri 0-katılmıyorum olan sınıfın denek sayısı 679 kişi, sınıf değeri 1-katılıyorum olan sınıfın denek sayısı da 53 kişi olduğundan sınıf değerlerine ilişkin uygulanacak ağırlıklar Çizelge 4.12'de gösterilmiştir.

Çizelge 4.12. Sınıf değerlerini dengelemek için kullanılacak ağırlıklar.

Soru 10-) Trafik kurallarının araba kullanma zevkini yok ettiğine inanıyorum	
Sınıf Değeri	Sınıf Ağırlığı
0	0.1
1	1.3

Bağımlı değişkene göre eğitim ve test veri setlerine ilişkin sınıf değerlerine ait denek sayıları Çizelge 4.13 ve Çizelge 4.14'de gösterilmiştir.

Çizelge 4.13. Bağımlı değişkene göre eğitim veri seti.

Soru 10-) Trafik kurallarının araba kullanma zevkini yok ettiğine inanıyorum	
Sınıf Değeri	Denek Sayısı
0	451
1	39

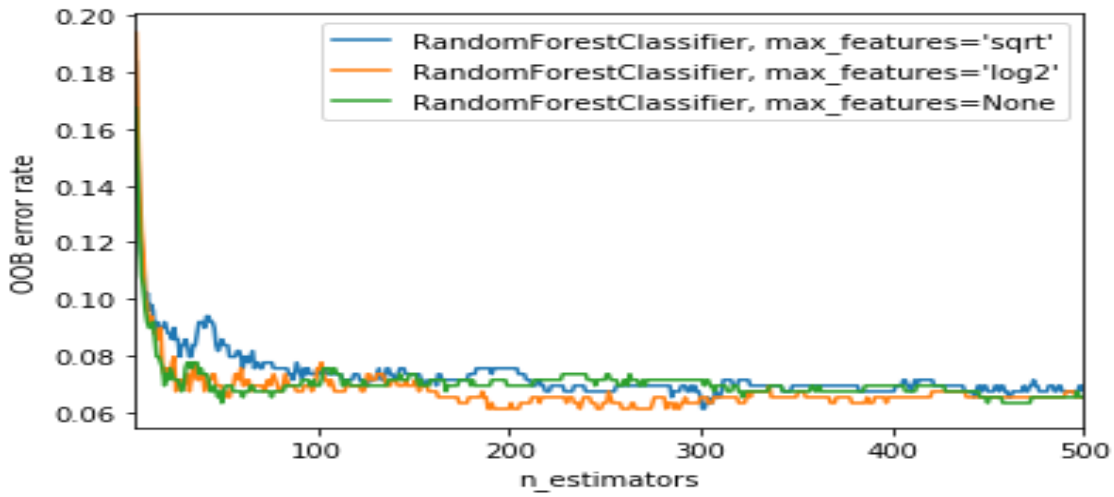
Çizelge 4.13'de görüleceği üzere eğitim veri setine ilişkin sınıf değeri 0-katılmıyorum olan denek sayısı 451 kişi, 1-katılıyorum olan değişkene ait denek sayısı 39 kişidir.

Çizelge 4.14. Bağımlı değişkene göre test veri seti.

Soru 10-) Trafik kurallarının araba kullanma zevkini yok ettiğine inanıyorum	
Sınıf Değeri	Denek Sayısı
0	228
1	14

Çizelge 4.14'te görüleceği üzere test veri setine ilişkin sınıf değeri 0-katılmıyorum olan denek sayısı 228 kişi, 1-katılıyorum olan değişkene ait denek sayısı 14 kişidir.

Modelimizin kurulumuna ilişkin seçilecek olan ilk 500 adet ağaç sayılarına ilişkin kurulan rastgele ormanlar modeline ait hata oranlarına ilişkin çıktılar Şekil 4.4'de verilmiştir.



Şekil 4.4. Ağaç sayılarına göre OOB hata grafiği.

Şekil 4.4'de görüleceği üzere kullanılacak ağaç sayısının 100 olduğu durumda diğer durumlara göre OOB hatasının daha kararlı olduğu görülmektedir. Yeni kurulacak olan parametrede kullanılacak ağaç sayısı değeri 100 olarak belirlenmiştir.

Çizelge 4.15. Modelde kullanılacak maksimum değişken sayısı.

Ağaç Sayısı	Gini yöntemiyle en önemli bulunan maksimum değişken sayısı	Gini yöntemiyle en önemli bulunan maksimum değişken sayısı modele girdiğinde hata oranı
100	3	0.662
100	**5**	**0.661**
100	6	0.662
100	7	0.662
100	15	0.663
100	21	0.663
100	35	0.663
100	41	0.663

Çizelge 4.15'de görüleceği üzere modelimizde kullanılacak maksimum değişken sayısı gini kriteri ile belirleyeceğimiz değişken önem düzeylerinde mevcut ilk 5 değişken olmuştur.

Modelimizin kurulumuna ilişkin baz alacağımız ağaç sayısı ve modelde kullanacağımız maksimum değer sayısını belirlemek için yapılan hiperparametre optimizasyonu sonucunda söz konusu kriterlere ilişkin en etkili parametre değerleri Çizelge 4.16'da verilmektedir.

Çizelge 4.16. Hiperparametre optimizasyonu sonucunda en optimal parametre değerleri.

Ağaç Sayısı (n_estimators)	Maksimum Değişken (max_features)
100	5

Çizelge 4.16'da görüleceği üzere bu işlem sonucunda modelimizde uygulayacağımız ağaç sayısı kriterimiz 400, maksimum değişken sayımız ise 5 olarak belirlenmiştir.

Modelin kurulması için gerekli parametrelerin optimal değerlerinin belirlenme işlemi tamamlandıktan sonra bu parametreler ile algoritma yeniden çalıştırılarak Çizelge 4.17'deki sınıflandırma çizelgesi elde edilmiştir.

Çizelge 4.17. OOB test verisi sonucu sınıflandırma çizelgesi.

	Gerçek Sınıflar	
Tahmin Sınıfı	0-Katılmıyorum	1-Katılıyorum
0-Katılmıyorum	219 (TP)	9 (FP)
1-Katılıyorum	7 (FN)	7 (TN)

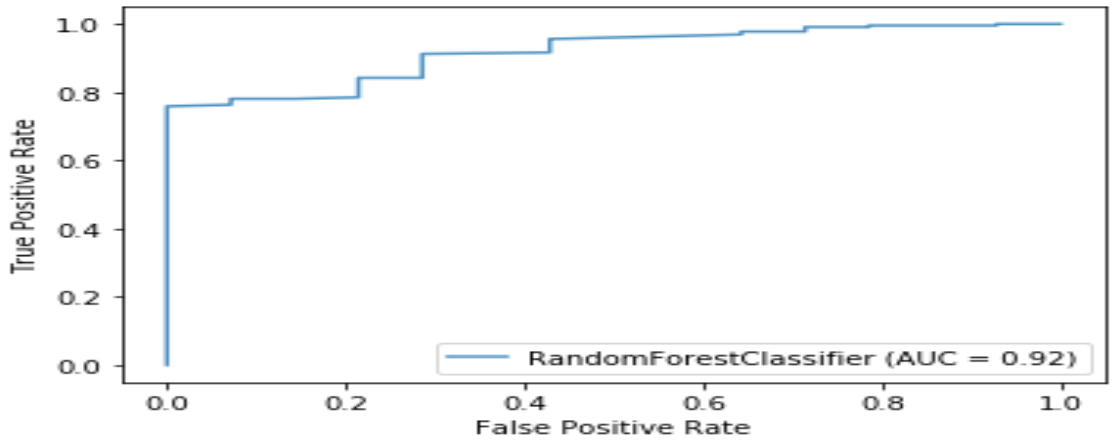
Çizelge 4.17'de tahmin sınıfı ve gerçek sınıflar bizlere verilmiş olup sınıflandırma çizelgesine ait performans ölçütleri Çizelge 4.18'de verilmiştir.

Çizelge 4.18. Test verisi için sınıflandırma çizelgesine ait performans ölçütleri.

Doğruluk (Accuracy)	%93.39
Yanlış Sınıflandırma Oranı	%6.61
Hassasisyet (Precision)	%99.05
Doğru Pozitif Oran (Duyarlılık/Pozitif Öngörü)	%93.49
F-puanı	%96.19
Kappa	%43.16

Çizelge 4.18'de görüleceği üzere test verisine ilişkin Doğruluk değerinin %93.39 gibi yüksek bir rakam çıkması iyi bir performans gösterdiğinin kanıtıdır. Bu durumu bizlere F-puanının %96.19 olması da göstermektedir. Kappa istatistiğinin %43.16 olması ise iki sınıfın arasındaki uyumun olduğunu göstermektedir.

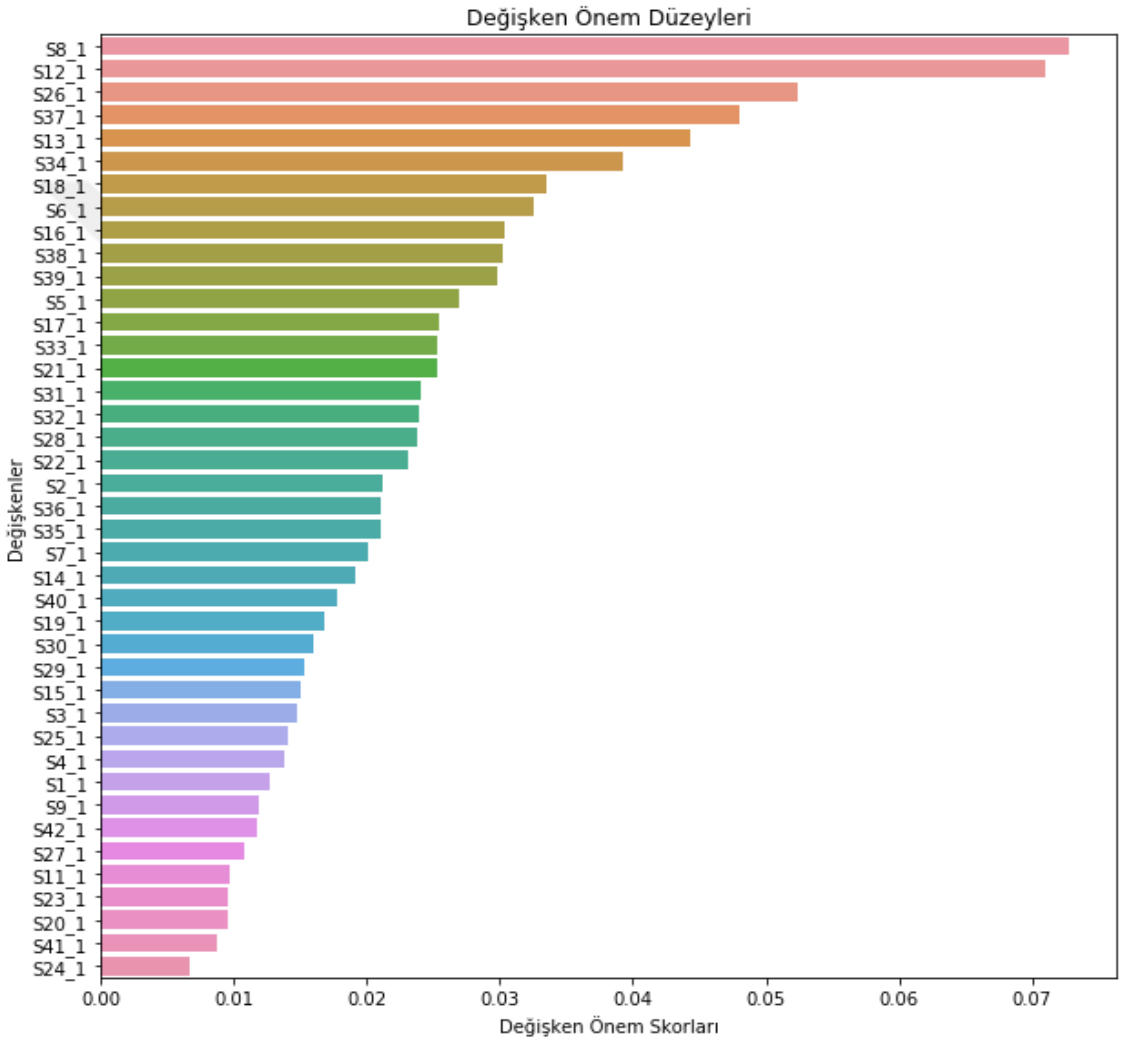
Kurulan modele ilişkin ROC eğrisi ve AUC değeri Şekil 4.5'de bizlere gösterilmiştir.



Şekil 4.5. Kurulan modele ilişkin ROC eğrisi ve AUC değeri.

Şekil 4.5'de kurulan modele ilişkin eğri altında kalan AUC değerinin 0.92 olması modelin gayet başarılı bir performans gösterdiğini bizlere sunmaktadır.

Sınıflandırma modelinin performans kriterlerini de tamamladıktan sonra bağımlı değişkenimiz olan "Soru 10-) Trafik kurallarının araba kullanma zevkini yok ettiğine inanıyorum" bağımlı değişkeninin sınıflamasında en etkili değişkenlere ait önem düzeyleri Şekil 4.6'daki grafikte verilmiştir.



Şekil 4.6. Kurulan modele ilişkin değişken önem düzeyleri grafiği.

Şekil 4.6'da modelin içerisinde mevcut olan tüm değişkenlerin önem düzeyleri bizlere sunulmuştur. Fakat bizim asıl olarak dikkate alacağımız değişkenler modelin kurulumunda baz aldığımız aynı zamanda maksimum değişken parametre değeri olan ilk

5 değerdir. Bu ilk 5 değişkenin açıklamaları ile değişken önem düzeyleri sırasıyla Çizelge 4.19'da bizlere sunulmuştur.

Çizelge 4.19. Kurulan modele ilişkin ilk 5 değişkenin önem düzeyleri.

Değişkenler	Önem Düzeyleri
Soru 8-) Güçlü bir arabaya sahip olmanın tek anlamı onu tam kapasite kullanmaktır.	0.073
Soru 12-) Araba kullanırken hız yapmanın ve heyecanın ayrılmaz bir ikili olduğuna inanıyorum.	0.071
Soru 26-) Trafikte bazen risk almak gerekir.	0.052
Soru 37-) Zamanında varmak için bazen trafik kurallarını esnetmenin hiçbir sakıncası yoktur.	0.048
Soru 13-) Yetişkinlerin, trafikte eğlence ve heyecana gereksinim duyduğuna inanıyorum.	0.044

4.3.3. Hız limitini 10-15 km/saat aşmakta sorun yoktur çünkü bunu herkes yapıyor değişkenine yönelik tutum ve sınıflandırma ölçeği

Trafik kurallarına karşı algı ve tutumların sınıflandırılmasına yönelik uygulama aşamasında bağımlı değişken olarak ele alacağımız üçüncü ve son ölçeğimiz ise anket uygulamamızın 32. sorusu olan "Hız limitini 10-15 km/saat aşmakta sorun yoktur çünkü bunu herkes yapıyor" değişkenidir. Çalışmamızda veri setinden yola çıkarak Rastgele Ormanlar algoritması ile yeni bir sınıflama modeli oluşturacağız.

Veri setimizdeki bağımlı değişkenimiz olan "Soru 32-) Hız limitini 10-15 km/saat aşmakta sorun yoktur çünkü bunu herkes yapıyor" sorusuna ilişkin cevaplar 0-katılmıyorum ve 1-katılıyorum olarak tarafımızca yeniden kategorize edilerek model tarafından tahmin edilmeye çalışılmıştır. Cevaplara ilişkin sınıf değerleri Çizelge 4.20'de gösterilmiştir.

Çizelge 4.20. Veri seti içerisinde bağımlı değişkene ilişkin toplam sınıf bilgileri.

Soru 32-) Hız limitini 10-15 km/saat aşmakta sorun yoktur çünkü herkes yapıyor	
Sınıf Değeri	Denek Sayısı
0	411
1	67

Çizelge 4.20'de görüleceği üzere sınıf değeri 0-katılmıyorum olan sınıfın denek sayısı 411 kişi, 1-katılıyorum olan sınıfın denek sayısı 67 kişidir. Sınıf değerlerine ilişkin uygulanacak sınıf ağırlıkları Çizelge 4.21'de gösterilmiştir.

Çizelge 4.21. Sınıf değerlerini dengelemek için kullanılacak ağırlıklar.

Soru 32-) Hız limitini 10-15 km/saat aşmakta sorun yoktur çünkü bunu herkes yapıyor	
Sınıf Değeri	Sınıf Ağırlığı
0	0.1
1	0.6

Bağımlı değişkene göre eğitim ve test veri setlerine ilişkin sınıf değerlerine ait denek sayıları Çizelge 4.22 ve Çizelge 4.23'de gösterilmiştir.

Çizelge 4.22. Bağımlı değişkene göre eğitim veri seti.

Soru 32-) Hız limitini 10-15 km/saat aşmakta sorun yoktur çünkü bunu herkes yapıyor	
Sınıf Değeri	Denek Sayısı
0	270
1	50

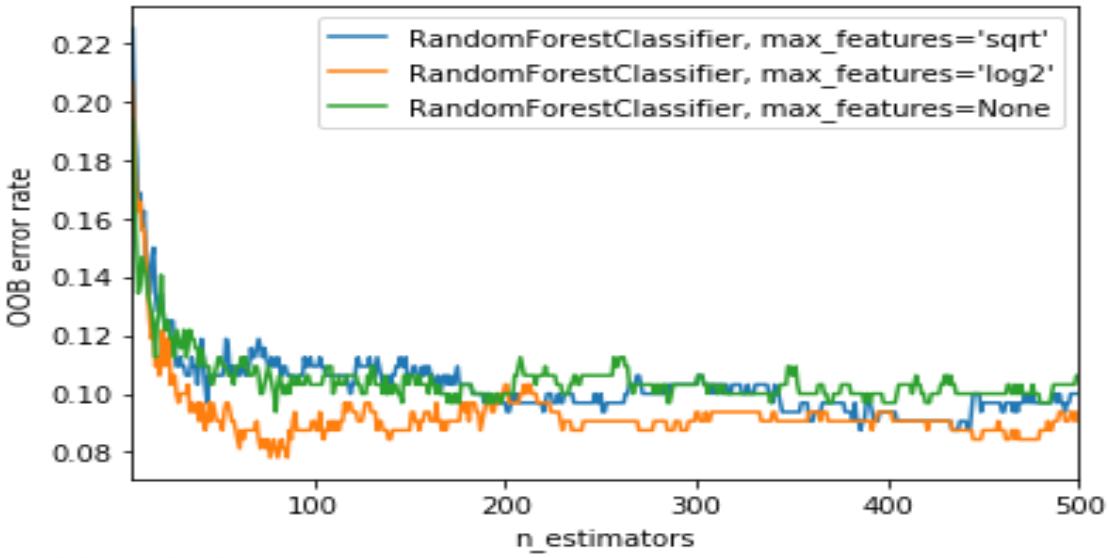
Çizelge 4.22'de görüleceği üzere eğitim veri setine ilişkin sınıf değeri 0-katılmıyorum olan değişkene ait denek sayısı 270 kişi, sınıf değeri 1-katılıyorum olan değişkene ait denek sayısı 50 kişidir.

Çizelge 4.23. Bağımlı değişkene göre test veri seti.

Soru 32-) Hız limitini 10-15 km/saat aşmakta sorun yoktur çünkü bunu herkes yapıyor	
Sınıf Değeri	Denek Sayısı
0	141
1	17

Çizelge 4.23'de görüleceği üzere test veri setine ilişkin sınıf değeri 0-katılmıyorum olan değişkene ait sınıf sayısı 141 kişi, sınıf değeri 1-katılıyorum olan değişkene ait sınıf sayısı ise 17 kişidir.

Modelimizin kurulumuna ilişkin seçilecek olan ilk 500 adet ağaç sayılarına ilişkin kurulan rastgele ormanlar modelinin hata oranlarına ilişkin çıktılar Şekil 4.7'de verilmiştir.



Şekil 4.7. Ağaç sayılarına göre OOB hata grafiği.

Şekil 4.7'de görüleceği üzere kullanılacak ağaç sayısının 400 olduğu durumda diğer durumlara göre OOB hatasının daha kararlı olduğu görülmektedir. Yeni kurulacak olan parametrede kullanılacak ağaç sayısı değeri 400 olarak belirlenmiştir.

Çizelge 4.24. Modelde kullanılacak maksimum değişken sayısı.

Ağaç Sayısı	Gini yöntemiyle en önemli bulunan maksimum değişken sayısı	Gini yöntemiyle en önemli bulunan maksimum değişken sayısı modele girdiğinde hata oranı
400	3	0.102
400	5	0.102
400	**6**	**0.102**
400	7	0.103
400	15	0.103
400	21	0.103
400	35	0.103
400	41	0.103

Çizelge 4.24'de görüleceği üzere dikkate aldığımız durum ilk 6 değişkene kadar OOB hataları sabit olması dolayısıyla baz alacağımız değişken sayısı ilk 6 değişken olmuştur.

Modelimizin kurulumuna ilişkin hiperparametre optimizasyonu sonucunda modelin kurulması aşamasındaki en etkili kriterler Çizelge 4.25'de verilmektedir.

Çizelge 4.25. Hiperparametre optimizasyonu sonucunda en optimal parametre değerleri.

Ağaç Sayısı (n_estimators)	Maksimum Değişken (max_features)
400	6

Çizelge 4.25'de görüleceği üzere bu işlem sonucunda modelimizde uygulayacağımız ağaç sayısı kriterimiz 400, maksimum değişken sayımız ise ilk 6 değişken olmuştur.

Modelin kurulması için gerekli parametrelerin optimal değerlerinin belirlenme işlemi tamamlandıktan sonra bu parametreler ile algoritma yeniden çalıştırılarak Çizelge 4.26'daki sınıflandırma çizelgesi elde edilmiştir.

Çizelge 4.26. OOB test verisi sonucu sınıflandırma çizelgesi.

	Gerçek Sınıflar	
Tahmin Sınıfı	0-Katılmıyorum	1-Katılıyorum
0-Katılmıyorum	131 (TP)	10 (FP)
1-Katılıyorum	6 (FN)	11 (TN)

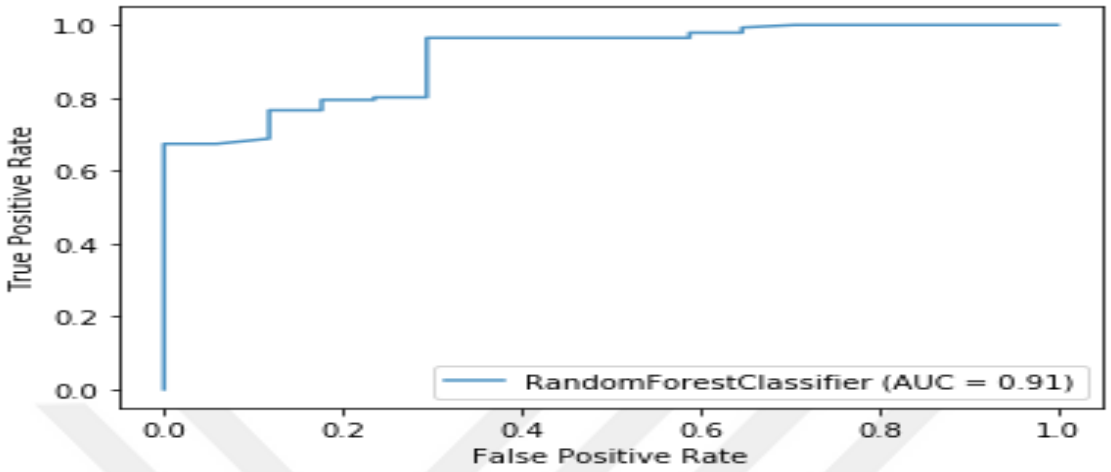
Çizelge 4.26'da görüleceği üzere optimal parametreler ile kurulan sınıflama modelinde gerçek sınıf değerleri ile tahmin sınıflarına ilişkin değerler bizlere sunulmuştur. Test verisi için sınıflandırma çizelgesi sonucunda elde edilen performans ölçütleri Çizelge 4.27'de verilmiştir.

Çizelge 4.27. Test verisi için sınıflandırma çizelgesine ait performans ölçütleri.

Doğruluk (Accuracy)	%89.87
Yanlış Sınıflandırma Oranı	%10.13
Hassasisyet (Precision)	%92.91
Doğru Pozitif Oran (Duyarlılık/Pozitif Öngörü)	%95.62
F-puanı	%94.25
Kappa	%52.22

Çizelge 4.27'de görüleceği üzere test verisine ilişkin Doğruluk değerinin %89.87 gibi yüksek bir rakam çıkması iyi bir performans gösterdiğinin kanıtıdır. Bu durumu bizlere F-puanının %94.25 olması da göstermektedir. Kappa istatistiğinin %52.22 olması ise iki sınıfın arasındaki uyumun olduğunu göstermektedir ve bu değerin sıfırdan büyük olması sınıflandırma problemlerinde istenilen bir durumdur.

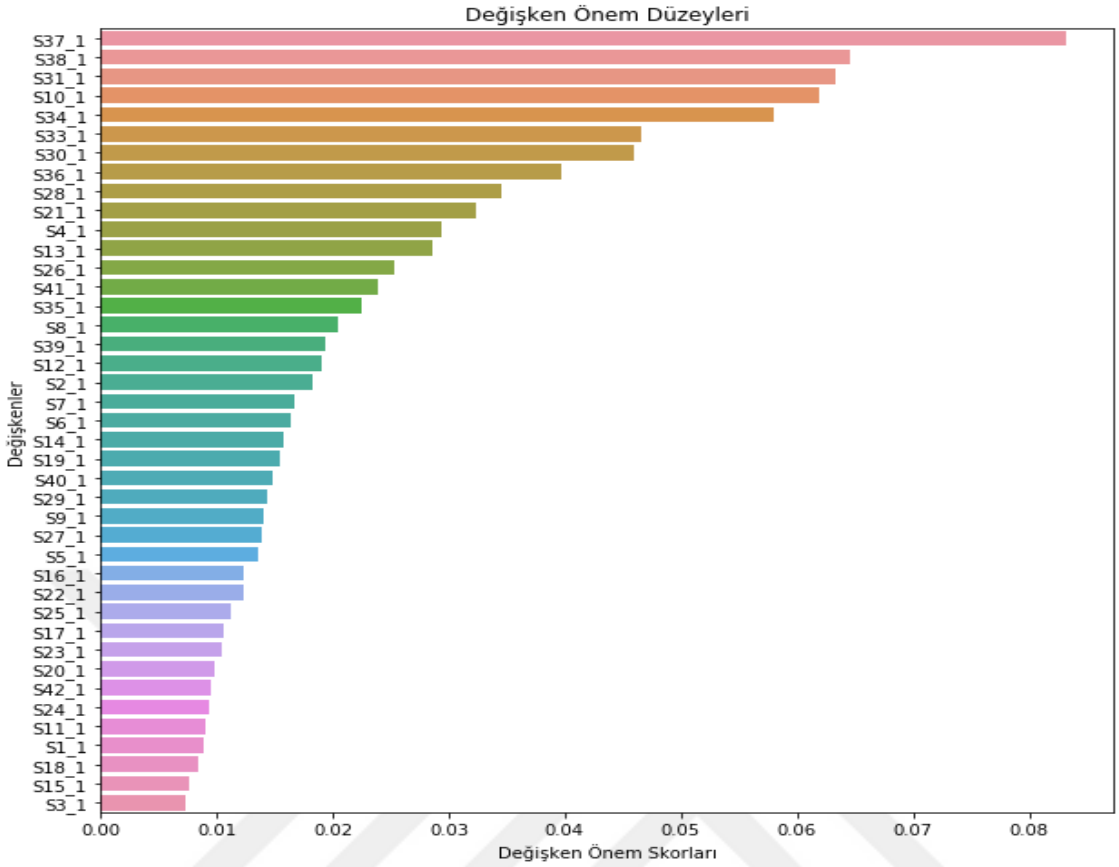
Kurulan modele ilişkin ROC eğrisi ve AUC değeri Şekil 4.8'de bizlere gösterilmiştir.



Şekil 4.8. Kurulan modele ilişkin ROC eğrisi ve AUC değeri.

Şekil 4.8'de kurulan modele ilişkin eğri altında kalan AUC değerinin 0.91 olması modelin gayet başarılı bir performans gösterdiğini bizlere sunmaktadır.

Sınıflandırma modelinin performans kriterlerini de tamamladıktan sonra bağımlı değişkenimiz olan "Soru 32-) Hız limitini 10-15 km/saat aşmakta sorun yoktur çünkü bunu herkes yapıyor" bağımlı değişkeninin sınıflamasında en etkili değişkenlere ait önem düzeyleri Şekil 4.9'daki grafikte verilmiştir.



Şekil 4.9. Kurulan modele ilişkin değişken önem düzeyleri grafiği.

Şekil 4.9'da modelin içerisinde mevcut olan tüm değişkenlerin önem düzeyleri bizlere sunulmuştur. Fakat bizim asıl olarak dikkate alacağımız değişkenler modelin kurulumunda baz aldığımız aynı zamanda maksimum değişken parametre değeri olan ilk 6 değişkendir. Bu ilk 6 değişkenin açıklamaları ile değişken önem düzeyleri sırasıyla Çizelge 4.28'de bizlere sunulmuştur.

Çizelge 4.28. Kurulan modele ilişkin ilk 6 değişkenin önem düzeyleri.

Değişkenler	Önem Düzeyleri
Soru 37-) Zamanında varmak için bazen trafik kurallarını esnetmenin hiçbir sakıncası yoktur.	0.083
Soru 38-) Her zaman trafik kurallarına uymaktansa akıcı bir şekilde araba kullanmak daha iyidir.	0.065
Soru 31-) Koşullar uygun olduğunda bence hız limitlerini aşmakta sorun yoktur.	0.063
Soru 10-) Trafik kurallarının araba kullanma zevkini yok ettiğine inanıyorum.	0.062
Soru 34-) Bazen trafik kurallarının çiğnenmesini göz ardı etmek gerekir.	0.058
Soru 33-) Trafiğin aksamamasını sağlamak için bazen kuralları esnetmek gerekir.	0.047

5. TARTIŞMA VE SONUÇ

Çalışmada Van iline ait trafiğe olan bakış açılarına yönelik uygulama verileri içerisinde cevaplayıcıların trafiğe karşı algı ve tutumlarını ölçen 3 adet değişken Rastgele Ormanlar yöntemiyle analiz edilmiştir.

"Trafik kurallarına harfi harfine uyarım" değişkenine yönelik AUC değeri %91, "Trafik kurallarının araba kullanma zevkini yok ettiğine inanıyorum" değişkenine yönelik değişkenine yönelik AUC değeri %92, "Hız limitini 10-15 km/saat aşmakta sorun yoktur çünkü bunu herkes yapıyor" değişkenine yönelik AUC değeri ise %91 seviyesidir. Bu değerlere bakarak söz konusu 3 adet bağımlı değişken için de yapılan sınıflandırmanın gayet iyi bir seviyede olduğu söylenebilir.

Bağımlı değişkenlerimizden ilki olan "Soru 27-) Trafik kurallarına harfi harfine uyarım" değişkenine yönelik yapılan sınıflandırma çalışmasında %93.44 sınıflandırma başarısı elde edilmiştir. Katılımcıların trafik kurallarına harfi harfine uymasına yönelik yapılan sınıflandırma başarısının artışında gini yöntemiyle belirlenen 5 adet değişkenin bu sınıflandırmada önem düzeylerinin yüksek olarak belirleyici olması öne çıkmaktadır. Bu değişkenler önem düzeyleri bakımından sırasıyla "Soru 29-) Bir yaya olarak tüm trafik kurallarına harfi harfine uyarım", "Soru 1-) Kazaların çoğu, engelleyici ve önleyici tedbirlere daha çok önem vererek engellenebileceğine inanıyorum", "Soru 11-) Sürücülerin tamamıyla hız kurallarına uyması gerektiğine inanıyorum", "Soru 14-) Hiçbir zaman hızlı araba kullanmayı istemem" ve "Soru 42-) Yollardaki trafik işaret ve levhalarını okuma ve anlama becerisine sahibim" değişkenleridir. Bu durumu yorumlayacak olursak trafik kurallarına harfi harfine uyan kişilerin trafik algı ve tutumlarında genel olarak yaya olarak da tüm trafik kurallarına uyduğu, kazaların çoğunun engelleyici ve önleyici tedbirlere daha çok önem vererek engellenebileceğine inandığı, sürücülerin tamamıyla hız kurallarına uyması gerektiğini düşündüğü, hiçbir zaman hızlı araba kullanmayı istemediği ve yollardaki trafik işaret ve levhalarını okuma ve anlama becerisine sahip olduğu söylenebilir. Bu sonuçlar aynı zamanda Selimoğlu (2014), tarafından trafik kazalarının nedenleri, sonuçları ve kazaların önlenmesine ilişkin öneriler adlı çalışmasındaki sonuçlar ile de benzerlik göstermektedir.

Bağımlı değişkenlerimizden ikincisi olan “Soru 10-) Trafik kurallarının araba kullanma zevkini yok ettiğine inanıyorum” değişkenine yönelik yapılan sınıflandırma çalışmasında %93.39 sınıflandırma başarısı elde edilmiştir. Katılımcılar içerisinde trafik kurallarının araba kullanma zevkini yok ettiği tutumuna ilişkin 5 adet değişken önem düzeyleri bakımından belirleyici olarak öne çıkmaktadır. Bu değişkenler önem düzeyleri bakımından sırasıyla “Soru 8-) Güçlü bir arabaya sahip olmanın tek anlamı onu tam kapasite kullanmaktır”, “Soru 12-) Araba kullanırken hız yapmanın ve heyecanın ayrılmaz bir ikili olduğuna inanıyorum”, “Soru 26-) Trafikte bazen risk almak gerekir”, “Soru 37-) Zamanında varmak için bazen trafik kurallarını esnetmenin hiçbir sakıncası yoktur”, “Soru 13-) Yetişkinlerin, trafikte eğlence ve heyecana gereksinim duyduğuna inanıyorum” değişkenleridir. Bu durumu yorumlayacak olursak; trafik kurallarının araba kullanma zevkini yok ettiğine inanan kişilerin algı ve tutumlarında ekseriyetle güçlü bir arabaya sahip olmanın tek anlamının onu tam kapasite kullanmak olduğu, Araba kullanırken hız yapmanın ve heyecanın ayrılmaz bir ikili olduğu, trafikte bazen risk almanın gerekli bir durum olduğu, zamanında varmak için bazen trafik kurallarının esnetilmesinde hiçbir sakınca görmediği ve yetişkinlerin trafikte eğlence ve heyecana gereksinim duyduğu düşünceler olduğu söylenebilir. Bu sonuçlar aynı zamanda Çavdar ve ark. (2008), tarafından trafik kazalarına sebep olan yüksek hız kusurlarının denetimi ve aktif güvenlik sistemleri ile kontrolü çalışmasındaki sonuçlar ile de benzerlik göstermektedir.

Bağımlı değişkenlerimizden sonuncusu olan “Soru 32-) Hız limitini 10-15 km/saat aşmakta sorun yoktur çünkü bunu herkes yapıyor” değişkenine yönelik yapılan sınıflandırma çalışmasında %89.87 sınıflandırma başarısı elde edilmiştir. Katılımcıların hız limitini 10-15 km/saat aşılmasının herkesin yaptığını örnek göstererek sorun görmediği algısına yönelik inancına olan çalışmada 6 adet değişkenin bu sınıflandırmada önem düzeylerinin yüksek olarak belirleyici olması öne çıkmaktadır. Bu değişkenler önem düzeyi bakımından sırasıyla “Soru 37-) Zamanında varmak için bazen trafik kurallarını esnetmenin hiçbir sakıncası yoktur”, “Soru 38-) Her zaman trafik kurallarına uymaktansa akıcı bir şekilde araba kullanmak daha iyidir”, “Soru 31-) Koşullar uygun olduğunda bence hız limitlerini aşmakta sorun yoktur”, “Soru 10-) Trafik kurallarının araba kullanma zevkini yok ettiğine inanıyorum”, “Soru 34-) Bazen trafik kurallarının çiğnenmesini göz ardı etmek gerekir” ve “Soru 33-) Trafiğin aksamamasını sağlamak için

bazen kuralların esnetilmesi gerekir" değişkenleridir. Bu durumu yorumlayacak olursak; hız limitini 10-15 km/saat aşılmasının herkesin yaptığını örnek göstererek sorun görmediği düşüncesine sahip kişilerin algı ve tutumlarında ekseriyetle zamanında varmak için bazen trafik kurallarını esnetmenin hiçbir sakıncası olmadığı, her zaman trafik kurallarına uymaktansa akıcı bir şekilde araba kullanmanın daha iyi olduğu, koşulların uygun olması halinde hız limitlerini aşmanın sorun olmadığı, trafik kurallarının araba kullanma zevkini yok ettiği, bazen trafik kurallarının çiğnenmesinin göz ardı edilmesi gerektiği ve trafiğin aksamamasını sağlamak için bazen kuralların esnetilmesi gerektiği gibi düşünceler olduğu söylenebilir. Bu sonuçlar aynı zamanda Kaçaroğlu ve ark. (2004), tarafından hız ihlali yapan sürücülerin ihlale ilişkin nedensel atıflarının ve kişisel özelliklerinin incelenmesi çalışmasındaki sonuçlar ile de benzerlik göstermektedir.

Böylece Rastgele Ormanlar yönteminin avantaj ve dezavantajlarında (Bölüm 3.1.7.) bahsedildiği üzere birbirinden bağımsız karar ağaçlarının rastgele seçilerek birleştirilmesi sonucunda hem eğitim verisinde aşırı uyum problemine karşı dayanıklılık göstermiş olup hem de OOB verileri üzerinden modelin kendi kendini test etmesiyle güvenilir sonuçlar veren bir model elde etmiştir. Bunları yaparken de değişken önem derecesi yüksek az sayıda değişken üzerinden sade ve anlaşılır sonuçlar vermektedir.

Diğer taraftan veriyi ağırlıklandırma işlemi uygulamamıza rağmen algoritma azınlık sınıfı tahmin etme konusunda bizlere elle tutulur bir kazanım sağlamadığı görülmektedir. Bu durum Breiman ve ark. (2003) yılında yapılan çalışma sonucunda Rastgele Ormanlar algoritmasının dengeli verilerle çalıştığında dengeli olmayan verilere ekstra bir kazanım oluşturmadığı sonucunu bizlere kanıtlamıştır.

Rastgele Ormanlar yönteminden faydalanılarak oluşturulan sınıflandırma modelinde ilgili veri setine sonradan dahil edilecek yeni kullanıcıların trafik kurallarına yönelik algı ve tutumları belirli bir hata oranı dahilinde sınıflandırılabilir ve söz konusu kişinin trafik kurallarına yönelik algı ve tutumlarında bu model sayesinde ön bilgiye sahip olunabilir. Bu sayede trafikte can ve mal güvenliğini sağlayabilmek için ilgili mercilerin trafik denetlemelerine yönelik yeni politikalar almasında ve bu konu üzerine daha etkili uygulamalarla trafik kazalarının önüne geçilmesinde fayda sağlayabilir.

KAYNAKLAR

Abellán, J., Mantas, C. J., Castellano, J. G., 2017. A random forest approach using imprecise probabilities, ***Knowledge-Based Systems,* 134**: 72-84.

Akar, Ö., Güngör, O., 2015. Integrating multiple texture methods and ndvi to the random forest classification algorithm to detect tea and hazelnut plantation areas in northeast Turkey. ***International Journal of Remote Sensing,* 36** (2): 442-464.

Akman, M., 2010. ***Veri Madenciliğine Genel Bakış ve Random Forest Yönteminin İncelenmesi: Sağlık Alanında Bir Uygulama*** (Yüksek Lisans Tezi). Ankara Üniversitesi, Sağlık Bilimleri Enstitüsü Biyoistatistik Anabilim Dalı, Ankara.

Akman, M., Genç, Y., Ankaralı, H., 2011. Random forest yöntemi ve sağlık alanında bir uygulama. ***Türkiye Klinikleri Biyoistatistik Dergisi,* 3** (1): 36-48.

Anonim, 2008. Trafik Kazalarını Önleme Faaliyetleri. Performans Denetim Raporu, ***https://www.sayistay.gov.tr/tr/Upload/62643830/files/raporlar/diger/Trafik%20Kazalar%C4%B1n%C4%B1%20%C3%96nleme%20Faaliyetleri%20Performans%20Denetimi%20Raporu.pdf*,** Sayıştay Başkanlığı, Ankara. Erişim tarihi: 09.02.2021.

Archer, K. J., 2008. Empirical characterization of random forest variable importance measure. ***Computational Statistics & Data Analysis,* 52** (4): 2249-2260.

Ayas, S., 2014. ***Mikroskobik İmgelerde Tüberküloz Bakterisinin Rastgele Ormanlar Yöntemiyle Sınıflandırılması*** (Yüksek Lisans Tezi). Karadeniz Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Trabzon.

Bauer, E., Kohavi, R., 1999. An empirical comparison of voting classification algorithms: Bagging, boosting and variants. ***Machine Learning,* 36:** 105–139.

Bartlett, P., Shawe-Taylor, J., 1999. Generalization performance of support vector machines and other pattern classifiers. ***Advances in Kernel Methods—Support Vector Learning,* 141:** 43-54.

Besnah, T., Ejigu, D., Abraham, A., Snasel, V., Kromer, P., 2011. Pattern recognition and knowledge discovery from road traffic accident data in Ethiopia: Implications for improving road safety. ***World Congress on Information and Communication Technologies (WICT).*** 11-14 Aralık 2011, Mumbai. 1241-1246.

Bezek Güre, Ö., 2019. ***Öğrencilerin Matematik Başarılarını Etkileyen Faktörlerin Rastgele Orman, Çok Katmanlı Algılayıcı ve Radyal Tabanlı Fonksiyon Yapay Sinir Ağları Yöntemleri İle Tahminleme Yeteneği Açısından Karşılaştırılması: Türkiye Örneği*** (Doktora Tezi). Van Yüzüncü Yıl Üniversitesi, Fen Bilimleri Enstitüsü, Van.

Breiman, L., 1996. Bias, variance and arcing classifiers. ***Technical Report***, 1-25.

Breiman, L., 2001. Random forests. ***Machine Learning***, **45** (1): 5–32.

Breiman, L., Chen, C., Liaw, A., 2003. Using random forest to learn imbalanced data. https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf. Berkeler University of California Department of Statistics. Erişim tarihi: 16.03.2021.

Breiman, L., Cutler, A., 2005. Random Forest. ***https://www.stat.berkeley.edu/users/breiman/RandomForests/cc_papers.htm***. Salford Systems for the Commercial Release of the Software. Erişim tarihi: 12.06.2020.

Bulut, F., 2017. Örnek tabanlı sınıflandırıcı topluluklarıyla yeni bir klinik karar destek sistemi, ***Journal of the Faculty of Engineering and Architecture of Gazi University,*** **32** (1): 65-76.

Chapelle, O., Schölkopf, B., Zien, A., 2006. IEEE Transactions on Neural Networks, 20. ***Semi-Supervised Learning*** (Editors: Chapelle, O., Schölkopf, B., Zien, A.). The MIT Press, 3, London, 481.

Coenen, F., 2004., Data mining: Past, present and future. ***The Knowledge Engineering Review,*** **00** (0): 1-24.

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, Section in Machine Learning. ***Ensemble Machine Learning,*** Springer, New York, 329.

Cutler, A., Cutler, D. R., Stevens, J. R., 2013. Tree-based Methods, Chap. 5. ***High-Dimensional Data Analysis in Cancer Research***. Springer, New York, 400.

Çakır, M., 2005. ***Firma Başarısızlığının Dinamiklerinin Belirlenmesinde Makine Öğrenmesi Teknikleri: Ampirik Uygulamalar ve Karşılaştırmalı Analizi*** (Türkiye Cumhuriyet Merkez Bankası İstatistik Genel Müdürlüğü, Uzmanlık Yeterlilik Tezi). TCMB, İstatistik Genel Müdürlüğü, Ankara.

Çakır, O., Doğan, M. C., 2017. İlkokul öğretmenlerinin postmodern eğitim anlayışına ilişkin görüşleri. ***Eğitim Kuram ve Uygulama Araştırmaları Dergisi,*** **3** (1): 38-52.

Çavdar, A., Uçar, M., Kılıçaslan, İ., 2008. Trafik kazalarına sebep olan yüksek hız kusurlarının denetimi ve aktif güvenlik sistemleri ile kontrolü. ***Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi,*** **23** (1): 187-198.

Çelik, İ. H., 2014. ***Sürücü Davranışı ve Sürücü Kişiliği Arasındaki İlişki Analizi*** (Yüksek Lisans Tezi). Gümüşhane Üniversitesi, Sosyal Bilimler Enstitüsü, Afet Yönetimi Anabilim Dalı, Gümüşhane.

Çevre ve Şehircilik Bakanlığı (ÇŞB), 2018. Ulaştırma Türlerine Göre Taşınan Yolcu ve Yük Miktarı, 2018. ***https://cevreselgostergeler.csb.gov.tr/ulastirma-turlerine-gore-tasinan-yolcu-ve-yuk-miktari-i-85789***. Çevre ve Şehircilik Bakanlığı Çevresel Göstergeler. Erişim Tarihi: 09.02.2021.

Çimen, E., 2020. A random subspace based conic functions enseble classifier. ***Turkish Journal of Electrical Engineering & Computer Sciences, TÜBİTAK***, 28: 2165-2182.

Çölkesen, İ., 2015. ***Yüksek Çözünürlüklü Uydu Görüntüleri Kullanarak Benzer Spektral Özelliklere Sahip Doğal Nesnelerin Ayırt Edilmesine Yönelik Bir Metodoloji Geliştirme*** (Doktora Tezi). İstanbul Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Geomatik Mühendisliği Anabilim Dalı, İstanbul.

Dayan, P., 2008. Library of Congress Cataloging-in-Publication Data. Section Unsupervised Learning. (Editor: R. A.Wilson, & F. C. Keil), ***The MIT Encyclopedia of the Cognitive Sciences***. The MIT Press. London. 857-859.

Dietterich, T., 1998. An experimental "Comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization"**,** ***Machine Learning,*** **40**: 139-157.

Dua, S., Chowriappa, P., 2013. Validation and Benchmarking, Chap. 9. ***Data Mining for Bioinformatics,*** CRC Press, Florida. 348.

Ekelik, H., 2019. ***Dijital Reklam Verilerinden Yararlanarak Potansiyel Konut Alıcılarının Rastgele Orman Yöntemiyle Sınıflandırılması*** (Yüksek Lisans Tezi). Marmara Üniversitesi**,** İstanbul.

Emel, G. G., Taşkın, Ç., 2005. Genetik algoritmalar ve uygulama alanları. ***Uludağ Üniversitesi İ.İ.B.F. Dergisi,*** **21** (1): 129–152.

Eken, M. H., Çiçek, M., 2009. Türkiye'de otomotiv sektöründeki ürünlerin kredilerle finansmanının satışlara etkisi. ***Maliye Finans Yazıları,*** **23** (84): 61-77.

Freund, Y., Schapire, R.E., Schapire, R., 1996. Experiments with a new boosting algorithm. ***Thirteenth International Conference on ML***, 22th January 1996, 1-16.

Giudici, P., 2003. Statistical Data Mining, Chap. 5. ***Applied Data Mining, Statistical Methods for Business and Industry,*** Wiley, Italy. 376.

Gutierrez, D. D., 2015. ***Machine Learning and Data Science: An Introduction to Statistical Learning Methods with R***. Technics Publications, San Francisco 498.

Hansen L. K., Salamon, P., 1990. Neural network ensembles. ***IEEE Trans Pattern Anal Mach Intell,*** **10**: 993-1001.

He, Y., 2006, ***Missing Data İmputation For Tree-Based Models*** (Master Thesis). California University, USA.

Hou, Y., Edara, P., Chang, Y., 2017. Road network state estimation using random forest ensemble learning. ***International Conference on Intelligent Transportation.*** 16-19 Ekim 2017, Yokohama, 1-6.

Ho, T. K., 1998. The random subspace method for constructing decision forests. ***IEEE Transaction on Pattern Analysis and Machine Intelligence***, **20** (8): 832-844.

Hossin, M., Sulaiman, M. N., 2015. A review on evaluation metrics for data classification evaluations. ***International Journal of Data Mining & Knowledge Management Process,*** **5** (2): 1-12.

Hurwitz, J., Kirsch, D., 2018. Understanding Machine Learning, Chap. 1. ***Machine Learning for Dummies*** (C.A. BURCHFİELD). John Wiley & Sons, Inc, Hoboken, New Jersey. 432.

İnce, M., 2009. ***Motorlu Taşıt ve Sürücü Kusurlarından Kaynaklanan Trafik Kazalarının İstatistiksel Olarak Analiz Edilmesi*** (Yüksek Lisans Tezi), Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Konya.

İyinam, F., İyinam, Ş., Ergün, M., 1999. Kullanıcı olarak insan faktörünün karayolu güvenliği üzerindeki etkileri. ***II. Ulaşım ve Trafik Kongresi.*** 29-30 Eylül 1999, Ankara. 251-262.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. ***An Introduction to Statistical Learning with Applications in R,*** (Editors: Casella G., Fienberg, S., Olkin, I.). Springer, New York. 603.

Kaçaroğlu, G., Amado, S., Akün, E., 2004. Hız ihlali yapan sürücülerin ihlale ilişkin nedensel atıflarının ve kişilik özelliklerinin incelenmesi. ***Türk Psikoloji Yazıları,*** **7** (13): 1-23.

Kalikov, A., (2006). ***Veri Madenciliği ve Bir E-Ticaret Uygulaması*** (Yüksek Lisans Tezi). Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Kılınç, M., Tarhan, Ç., Aydın, C., 2020. Kitle fonlaması projelerinin karar ağacı ve rastgele orman algoritmalarıyla sınıflandırılması. ***Journal of Information Systems and Management Research,*** **2** (2): 16-25.

Kızılkaya, Y. M., Oğuzlar, A., 2018. Bazı denetimli öğrenme algoritmalarının R programlama dili ile kıyaslanması. ***Karadeniz Uluslararası Bilimsel Dergi,*** **37** (37): 90-98.

Kilimci Z.H., Ganiz M.C., 2015, Evaluation of classification models for language processing. ***International Symposium on Innovations in Intelligent Systems and Applications.*** 2-4th September 2015, Madrid, Spain. 1-8.

Kolay, N., Erdoğmuş, P., 2016. The classification of breast cancer with machine learning techniques. ***In Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT).*** 26-27 Nisan 2016, İstanbul. 1-4.

Krishnaveni, S., Hemalatha, M., 2011. A perspective analysis of traffic accident using data mining techniques. ***International Journal of Computer Applications,*** **23** (7): 40-48.

Kruber, F., Wurst, J., Morales, E. S., Chakraborty, S., Botsch, M., 2019. Unsupervised and supervised learning with the random forest algorithm for traffic scenario clustering and classification. ***IEEE Symposium on Intelligent Vehicle.*** 9-12 Haziran 2019, Paris. 1-8.

Lahouar, A., Slama, J. B. H., 2015. Day-ahead load forecast using random forest and expert input selection. ***Energy Conversion and Management,*** **103:** 1040-1051.

Laudon, K. C., 2007. Ethical and Social Issues in Information Systems, Chap. 4. ***Management Information Systems: Managing the Digital Firm*** (Editor: Bob Horan). Pearson Education, India. 120.

Learned-Miller, E. G., 2014. Introduction to Supervised Learning, ***https://people.cs.umass.edu/~elm/Teaching/Docs/supervised2014a.pdf***. University of Massachusetts, USA. Erişim tarihi: 23.10.2020.

Lessmann, S., Baesens, B., Seow, H. V., Thomas, L. C., 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. ***European Journal of Operational Research***, **247** (1): 124–136.

Liaw A., Wiener, M., 2002. Classification and Regression By Random Forest, ***https://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf***. The Newsletter of the R Project, USA. Erişim Tarihi: 13.02.2021.

Liaw, A., Wiener, M., 2018. The R Project for Statistical Computing. ***https://cran.r-project.org/web/packages/randomForest/randomForest.pdf***. Erişim tarihi: 06.07.2020.

Liu, C., Chan, Y., Alam Kazmi, S. H., Fu, H., 2015. Financial fraud detection model: based on random forest. ***International Journal of Economics and Finance***, **7** (7): 178–188.

Louppe, G., 2014. ***Understanding Random Forest From Theory to Practice*** (PhD Thessis). University of Liege Faculty of Applied Sciences Department of Electrical Engineering & Computer Science, Liege.

Mansoori, Y., Karlsson, T., Lundqvist, M., 2019. The influence of the learn startup methodology on entrepreneurcoach relationships in the context of a startup accelerator. ***Technovation,*** **84:** 37-47.

Melody, Y., Kumar, A., 2001. An evaluation of self-organizing map networks as a robust alternative to factor analysis in data mining applications, ***Information Systems Research,*** **12** (2): 177-194.

Michie, D., Spiegelhalter, D. J., Taylor, C. C., 1994. Machine Learning, Chap. 1. ***Machine Learning, Neural and Statistical Classification.*** Overseas Press, USA. 3.

Minitab, 2018. Introduction to Random Forests. ***https://www.minitab.com/content/dam/www/en/uploadedfiles/content/products/spm/IntroRF.pdf***. Erişim tarihi: 13.09.2020.

Moshkovich, H.M., Mechitov, A.I., Olson, D.L., 2002. Rule induction in data mining: effect of ordinal scales, ***Expert Systems with Applications,*** **22** (4): 303-311.

Nordman, A., 2011. Data Mining Lecture 6: Evaluating the Performance of a Model. ***http://staffwww.itn.liu.se/~aidvi/courses/06/dm/lectures/lec6.pdf***. Erişim tarihi: 20.09.2020.

Onay, H., 2013. ***Trafikte Sosyal Sorumluluk Projelerinin Trafik Güvenliğine Etkilerinin İrdelenmesi*** (Yüksek Lisans Tezi). Gazi Üniversitesi, Fen Bilimleri Enstitüsü, Ankara.

Öğüt, S., 2002. Veri Madenciliği Kavramı ve Gelişim Süreci. ***http://www.sertacogut.com/blog/wp-content/uploads/2009/03/sertac_ogut_-_veri_madenciligi_kavrami_ve_gelisim_sureci.pdf***. Yeditepe Üniversitesi, İletişim Fakültesi Görsel İletişim Tasarımı Bölümü, İstanbul.

Özekes, S., 2003, Veri madenciliği modelleri ve uygulama alanları. ***İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi,*** **2** (3): 65-82.

Özden, C., Acı, Ç., 2018. Makine öğrenmesi yöntemleri ile yaralanmalı trafik kazalarının analizi: Adana örneği. ***Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi,*** **24** (2): 266-275.

Özkan, Y., 2013. Sınıflandırma ve Regresyon Ağaçları, Bölüm 4. ***Veri Madenciliği Yöntemleri*** (Editör: Dr. Rıfat ÇÖLKESEN). Papatya Yayınları, İstanbul. 240.

Page, D., 2015. Evaluating Machine-Learning Methods. ***http://pages.cs.wisc.edu/~dpage/cs760/evaluating.pdf***. Erişim tarihi: 15.11.2020.

Panov, P., Džeroski, S., 2007. Combining bagging and random subspaces to create better ensembles. ***International Conference on Intelligent Data Analysis***, 3-7th December 2007, Heidelberg, Berlin. 118-129.

Pal, M., 2005. Random forest classifier for remote sensing classification. ***International Journal of Remote Sensing***, **26** (1): 217-222.

Polikar, R., 2006. Ensemble based systems in decision making. ***IEEE Circuits and Systems Magazine***, **6** (3): 21–44.

Polikar, R., 2012 Ensemble Learning, Chap. 1. ***Ensemble Machine Learning*** (Editor: C. ZHANG ve Y. MA). Springer Science, Business Media, Boston. 1-34.

Qi, Y., Random Forest for Bioinformatics, 2012. ***https://www.cs.cmu.edu/~qyj/papersA08/11-rfbook.pdf***. Ensemble Machine Learning. Erişim tarihi: 12.10.2020.

Raschka, S., 2018. Machine Learning Lecture Notes, ***http://pages.stat.wisc.edu/~sraschka/teaching/stat479-fs2018/***. University of Wisconsin-Madison, USA. Erişim tarihi: 18.10.2020.

Refaeilzadeh, P., Tang, L., Liu, H., 2009. Cross-validation, Chap. 100. ***Encyclopedia of Database Systems,*** Springer**,** New York. 4355.

Remesan, R., Mathew, J., 2014. Machine Learning and Artificial Intelligence-Based Approaches, Chap. 3. ***Hydrological Data Driven Modelling: A Case Study Approach***, Springer, London. 264.

Ren, D., Qu, F., Ke, L., Zhang, Z., Xu, H., Wang, X., 2016. A gradient descent boosting spectrum modeling method based on back interval partial least squares, ***Neurocomputing,*** **171** (1): 1038-1046.

Rokach, L., 2010. Introduction to Ensemble Learning, Chap. 2. ***Pattern Classification Using Ensemble Methods***. World Scientific, London. 242.

Rokach, L., Maimon, O., 2014. Decision Forests, Chap. 9. ***Data Mining With Decision Tree Theory and Applications*** (Editors: H. BUNKE & P. S. P. WANG) World Scientific Publishing, New Jersey. 328.

Saharidis, G. K. D., Androulakis, I.P., Ierapetritou, M.G., 2011. Model building using bi-level optimization. ***Journal of Global Optimization,* 49** (1): 49–67.

Selimoğlu, E., 2014. Trafik kazalarının nedenleri, sonuçları ve kazaların önlenmesine ilişkin öneriler. ***Ziraat Mühendisliği Dergisi,* 361:** 51-54.

Schonlau, M., Zou, R.Y., 2020. The random forest algorithm for statistical learning. ***The Stata Journal: Promoting Communications on Statistics and Stata,* 20** (1): 3-29.

Scott, F. R., 2012. Understanding the Bias-Variance Tradeoff. ***http://scott.fortmann-roe.com/docs/BiasVariance.html***. Erişim tarihi: 12.11.2020.

Skurichina, M., Duin, R. P. V., 2002. Bagging, boosting and the random subspace method for linear classifiers, ***Pattern Analysis & Applications,* 5:** 121-135.

Sutton, R. S., Barto, A. G., 2014. Multi-Armed-Bandits, Chap. 2. ***Reinforcement Learning An Introduction*** (F. Bach). The MIT Press, London. 525.

Sümer, N., Kaygısız, Ö., 2015. Türkiye'de denetleme, cezalar ve trafik güvenliği göstergeleri arasındaki ilişkiler: 2008-2012 yılları analizi. ***Türkiye Halk Sağlığı Dergisi,* 13** (3): 193-205.

Şehribanoğlu, S., 2019. Van ilinde yaşayanların trafik işaretleri bilgisi ve trafik kurallarına bakış açıları üzerine bir araştırma. ***Trafik ve Ulaşım Araştırmaları Dergisi,* 2** (1): 1-15.

Şimşek Çetin, Ö., Gültekin Akduman, G., Alisinanoğlu, F., 2009. Çocuklarda trafik güvenliği eğitiminin önemi. ***Ulaşım ve Trafik Güvenliği Dergisi,* 3** (1): 5.

Timofeev, R., 2004. ***Classification and Regression Trees (CART) Theory and Applications*** (Master Thesis). Humboldt University, Berlin.

Tong, j., Zhang, J., Dong, E., Du, S., 2021. Severity classification of parkinson's disease based on permutation-variable importance and persistent entropy. ***Applied Sciences,* 11** (4): 18-34.

Topuz, E., 2015. ***Trafikte Hız Sorununun Trafik Güvenliğine Etkisi, Denetimi ve Çözüm Önerileri*** (Yüksek Lisans Tezi). T.C. Polis Akademisi Güvenlik Bilimleri Enstitüsü, Ankara.

Tosun, M., 2004. Geçmişten günümüze trafik. ***Polis Dergisi,* 39**: 171.

Tuncuk, M., 2004. ***Coğrafi Bilgi Sistemleri Yardımıyla Trafik Kaza Analizi: Isparta Örneği*** (Yüksek Lisans Tezi). Süleyman Demirel Üniversitesi, Fen Bilimleri Enstitüsü, Isparta.

Turgut, H., 2012. ***Veri Madenciliği Süreci Kullanılarak Alzheimer Hastalığı Teşhisine Yönelik Bir Uygulama*** (Yüksek Lisans Tezi). Süleyman Demirel Üniversitesi, Isparta.

Türkiye İstatistik Kurumu (TÜİK), 2020. Ölümlü Yaralanmalı Trafik Kazasına Neden Olan Kusur Sayısı, 2019. ***https://data.tuik.gov.tr/Bulten/Index?p=Karayolu-Trafik-Kaza-Istatistikleri-2019-33628***. TÜİK Veri Portalı. Erişim Tarihi: 10.02.2021.

Türkiye İstatistik Kurumu (TÜİK), 2020. Taşıt Cinslerine Göre Kayıtlı Taşıt, Ölümlü Yaralanmalı Kazaya Karışan Taşıt, Ölen ve Yaralanan Sürücü Sayısı, 2019. ***https://data.tuik.gov.tr/Bulten/Index?p=Karayolu-Trafik-Kaza-Istatistikleri-2019-33628***. TÜİK Veri Portalı. Erişim Tarihi: 10.02.2021.

Türkiye İstatistik Kurumu (TÜİK), 2021. Trafiğe Kayıtlı Motorlu Kara Taşıtları Sayısı, 2015-2020. ***https://tuikweb.tuik.gov.tr/PreTablo.do?alt_id=1051***. Ulaştırma İstatistikleri. Erişim Tarihi: 09.02.2021.

Türkiye İstatistik Kurumu (TÜİK), 2021. İllere Göre Motorlu Kara Taşıtları Sayısı, 2020. ***https://data.tuik.gov.tr/Bulten/Index?p=Motorlu-Kara-Tasitlari-Eylul-2020-33657***. TÜİK Veri Portalı. Erişim Tarihi: 10.02.2021.

Van Diepen, M., Franses, P. H., 2006. Evaluating chi-squared automatic interaction detection. ***Informatıon Systems,* 31** (8): 814-831.

Watts, J. D., Powell, S. L., Lawrence, R. L., Hilker, T., 2011. Improved classification of conservation tillage adoption using high temporal and synthetic satellite imagery, ***Remote Sensing of Environment,* 115** (2011): 66-75.

Weiss, S. M., Indurkhya, N., 1998. Machine Learning, Chap. 1. ***Predictive Data Mining: A Practical Guide (The Morgan Kaufmann Series in Data Management Systems)***, Morgan Kaufman Publishers Inc., San Fransisco. 228.

Yao, D., Yang, J., Zhan, X., 2013. A novel method for disease prediction: Hybrid of random forest and multivariate adaptive regression splines. ***Journal of Computers,* 8** (1): 170-177.

Yavuz, A.A., Ergül, B., Aşık, E., 2021. Trafik kazalarının makine öğrenmesi yöntemleri kullanılarak değerlendirilmesi. ***Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi,* 13** (1): 66-73.

Yıldız, M.C., Karaca, M., 2005. Otomobil sürücülerinin trafik ve yol güvenliği konusundaki görüşlerine ilişkin sosyolojik bakış, ***Dumlupınar Üniversitesi Sosyal Bilimler Dergisi,*** (12): 1-13.

Yılmaz, H., 2014. ***Random Forest Yönteminde Kayıp Veri Probleminin İncelenmesi ve Sağlık Alanında Bir Uygulama*** (Yüksek Lisans Tezi). Osmangazi Üniversitesi, Sağlık Bilimleri Enstitüsü Biyoistatistik Anabilim Dalı, Eskişehir.

Zhang, C., Li, Y., Yu., Z., Tian., F., 2016. A weighted random forest approach to improve predictive performance for power system transient stability assesment. ***IEEE PES Asia-Pasific Power and Energy Conference,*** 25- 28 October 2016, China. 1259-1263.

Zhang, Y., Liu, B., Cai, J., Zhang, S., 2017. Ensemble weighted extreme learning machine for imbalanced data classification based on differential evolution. ***Neural Comput Applications,* 28:** 259-67.

Zhou, Z., 2009. Ensemble Learning, Section E. ***Encyclopedia of Biometrics*** (Editor: S.Z. Lİ). Springer Reference, Boston. 1664.

Zhou, Z. H., 2012. Ensemble Methods, Chap. 1. ***Ensemble Methods: Foundations and Algorithms*** (Editor: Z.H. Zhou). Chapman and Hall/CRC, Florida. 236.

ÖZ GEÇMİŞ

Vedat Görgülü, ilk, orta ve lise eğitimini Aydın'ın Nazilli ilçesinde tamamladı. 2008 yılında Afyon Kocatepe Üniversitesi Fen-Edebiyat Fakültesi İstatistik Bölümü'ne başladı ve 2012 yılında mezun oldu. 2016 Ağustos ayından itibaren Türkiye İstatistik Kurumu Van Bölge Müdürlüğü'nde İstatistikçi olarak görev yapmaktadır.

T.C
VAN YÜZÜNCÜ YIL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
LİSANSÜSTÜ TEZ ORİJİNALLİK RAPORU

Tarih: 27/05/2021

Tez Başlığı / Konusu:

RASTGELE ORMANLAR YÖNTEMİ İLE ÖZELLİK SEÇİMİ KULLANILARAK VAN İLİNDE YAŞAYANLARIN TRAFİKTE ALGI VE TUTUMLARININ BELİRLENMESİ

Yukarıda başlığı/konusu belirlenen tez çalışmamın Kapak sayfası, Giriş, Ana bölümler ve Sonuç bölümlerinden oluşan toplam 66 sayfalık kısmına ilişkin, 27/ 05 / 2021 tarihinde şahsım/tez danışmanım tarafından Turnitin intihal tespit programından aşağıda belirtilen filtreleme uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı % 4 (Dört.) dır.

Uygulanan filtreler aşağıda verilmiştir:

- Kabul ve onay sayfası hariç,
- Teşekkür hariç,
- İçindekiler hariç,
- Simge ve kısaltmalar hariç,
- Gereç ve yöntemler hariç,
- Kaynakça hariç,
- Alıntılar hariç,
- Tezden çıkan yayınlar hariç,
- 7 kelimeden daha az örtüşme içeren metin kısımları hariç (Limit inatch size to 7 words)

Van Yüzüncü Yıl Üniversitesi Lisansüstü Tez Orijinallik Raporu Alınması ve Kullanılmasına İlişkin Yönergeyi inceledim ve bu yönergede belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini bilgilerinize arz ederim.

Tarih ve İmza

Adı Soyadı: Vedat GÖRGÜLÜ

Öğrenci No: 18910001315

Anabilim Dalı: İstatistik

Programı:

Statüsü: Y. Lisans X Doktora ☐

DANIŞMAN ONAYI
UYGUNDUR

Dr. Öğr. Üyesi Sanem ŞEHRİBANOĞLU

(Unvan, Ad Soyad, İmza)

ENSTİTÜ ONAYI
UYGUNDUR

(Unvan, Ad Soyad, İmza)