

**T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ**

**MEKANSAL ZEKA YÖNTEMLERİYLE YENİ NESİL
HAREKETLİLİK KAPSAMINDA ULAŞIM HİZMETLERİNİN
SÜRDÜRÜLEBİLİR YÖNETİMİ**

ALİ DOĞAN GÜMÜŞSOY

**YÜKSEK LİSANS TEZİ
ULAŞIM TEKNOLOJİLERİ ANABİLİM DALI
AKILLI ŞEHİR VE ULAŞIM TEKNOLOJİLERİ PROGRAMI**

DANIŞMAN: PROF. DR. ARİF ÇAĞDAŞ AYDINOĞLU

TEMMUZ 2025

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

MEKANSAL ZEKA YÖNTEMLERİYLE YENİ NESİL
HAREKETLİLİK KAPSAMINDA ULAŞIM
HİZMETLERİNİN SÜRDÜRÜLEBİLİR YÖNETİMİ

ALİ DOĞAN GÜMÜŞSOY

YÜKSEK LİSANS TEZİ
ULAŞIM TEKNOLOJİLERİ ANABİLİM DALI
AKILLI ŞEHİR VE ULAŞIM TEKNOLOJİLERİ
PROGRAMI

DANIŞMAN: PROF. DR. ARİF ÇAĞDAŞ AYDINOĞLU

TEMMUZ 2025

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL

**SUSTAINABLE MANAGEMENT OF
TRANSPORTATION SERVICES WITHIN THE SCOPE
OF NEXT GENERATION MOBILITY USING
GEOSPATIAL INTELLIGENCE**

ALİ DOĞAN GÜMÜŞSOY

**A THESIS OF MASTER OF SCIENCE
DEPARTMENT OF TRANSPORTATION
TECHNOLOGIES
SMART CITY AND TRANSPORTATION
TECHNOLOGIES PROGRAM**

ADVISOR: PROF. DR. ARİF ÇAĞDAŞ AYDINOĞLU

JULY 2025

YÜKSEK LİSANS JÜRİ ONAY FORMU

GTÜ Lisansüstü Eğitim Enstitüsü Yönetim Kurulunun 30/06/2025 tarih ve 2025/34 sayılı kararıyla oluşturulan jüri tarafından 21/07/2025 tarihinde tez savunma sınavı yapılan Ali Doğan GÜMÜŞSOY'un tez çalışması Ulaşım Teknolojileri Anabilim Dalı Akıllı Şehir ve Ulaşım Teknolojileri Programında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

JÜRİ

ÜYE

(TEZ DANIŞMANI) : Prof. Dr. Arif Çağdaş AYDINOĞLU

ÜYE

: Doç. Dr. Göker AKSOY

ÜYE

: Prof. Dr. Nursu TUNALIOĞLU ÖCALAN

ONAY

Gebze Teknik Üniversitesi Lisansüstü Eğitim Enstitüsü Yönetim Kurulunun

...../...../..... tarih ve/..... sayılı kararı.

İMZA/MÜHÜR

ÖZET

Günümüzde yaşanan hızlı nüfus artışı ve yüksek kentleşme eğilimi sonucunda kentsel hareketliliğin verimli bir şekilde karşılanması oldukça önemli hale gelmiştir. Büyüyen ve kalabalıklaşan kentlerde ulaşım talebinin artması sonucunda bireysel araç sahipliliği artarak kentlerde ulaşım altyapısı üzerinde yük oluşturmaktadır. Kent ulaşımında yaşanan bu sorunların yanında Bilgi ve İletişim Teknolojileri alanında yaşanan gelişmeler sonucunda yeni nesil hareketlilik kavramı ortaya çıkmaktadır. Yeni nesil hareketlilik kavramı içerisindeki paylaşımlı hareketlilik, kentsel ulaşım sorunlarına sunmuş olduğu çözümler ile ön plana çıkmaktadır. Bir paylaşımlı hareketlilik çözümü olan ve kent sakinlerinin ulaşım taleplerini toplu bir şekilde karşılanmasını sağlayan toplu taşıma sistemlerinin sürdürülebilir yönetimi kentsel hareketlilik açısından oldukça önemli bir konudur. Toplu taşıma sistemlerinin sürdürülebilir yönetimi büyük ölçüde yolcu talebinin yüksek doğrulukta tahmin edilebilmesine bağlı olmaktadır. Bu kapsamda yolcu talebinin tahmin edilmesine yönelik gerçekleştirilen çalışmalarda makine öğrenmesi tekniklerinin kullanımı yaygınlaşmıştır. Gerçekleştirilen tez çalışmasında, mekânsal zekâ yaklaşımı doğrultusunda Coğrafi Bilgi Sistemleri (CBS) ve yapay zekâ tekniklerinin kullanımıyla toplu taşıma sistemlerine yönelik ölçeklenebilir yolcu talep tahmin modelleri geliştirilmiştir. Çeşitli kentsel ilgi noktaları ve farklı demografik yapıları içermesi sebebiyle İstanbul ili Tuzla ve Pendik ilçeleri ile Kocaeli ili Çayırova, Darıca ve Gebze ilçeleri çalışma bölgesi olarak seçilmiştir. Yolcu talep tahmin modeli, çalışma bölgesi sınırları içerisinde bulunan M4 Metro ve Marmaray istasyonlarına yönelik geçmiş akıllı kart verilerinin kullanımı ile üretilmiştir. Çalışma kapsamında yolcu talebindeki dalgalanmaların tahmini için zamansal ve temel yolcu talebinin tahmini için mekânsal faktörler tahmin modeline dahil edilmiştir. Zengin veri setinin makine öğrenmesi tekniklerinde kullanımından önce Filtre ve Sarmalayıcı özellik seçimi yöntemleri uygulanmıştır. Oluşturulan veri setleri ile Rastgele Orman ve eXtreme Gradient Boosting (XGBoost) algoritmalarına dayalı modeller kurulmuştur. Algoritmalara has hiper parametrelerin oluşturulan veri setlerine uygun şekilde manuel ve Izgara Arama yöntemi ile ayarlanmasının ardından K-Katlı Çapraz Doğrulama tekniği ile modeller kurulmuştur. Model performansları performans karşılaştırma metrikleri ve hesaplama maliyeti açısından değerlendirilmiştir. En yüksek performanslı tahmin modeli ile çalışmada kullanılan faktörlerin önem düzeyi sıralamaları oluşturularak yolcu talebi ile ilişkisi yorumlanmış ve mekânsal faktörlerin model doğruluğunu artırıcı etkisi ortaya çıkmıştır. Seçilen tahmin modeli daha sonrasında mevcut ve proje aşamasındaki istasyonlara yönelik gelecek yolcu talep tahminleri oluşturulmuştur.

Anahtar Kelimeler: Coğrafi Bilgi Sistemleri, Makine Öğrenmesi, Rastgele Orman, Yeni Nesil Hareketlilik, Yolcu Talep Tahmini, XGBoost

ABSTRACT

As a result of the rapid population growth and high urbanization trend experienced today, it has become very important to meet urban mobility efficiently. As a result of the increase in transportation demand in growing and crowded cities, individual vehicle ownership increases and creates a burden on the transportation infrastructure in cities. In addition to these problems experienced in urban transportation, the concept of new generation mobility emerges as a result of the developments in the field of Information and Communication Technologies. Shared mobility within the concept of new generation mobility stands out with the solutions it offers to urban transportation problems. Sustainable management of public transportation systems, which is a shared mobility solution and enables the transportation demands of city residents to be met collectively, is a very important issue in terms of urban mobility. Sustainable management of public transportation systems largely depends on the ability to predict passenger demand with high accuracy. In this context, the use of machine learning techniques has become widespread in studies carried out to predict passenger demand. In the thesis study, scalable passenger demand prediction models for public transportation systems were developed using Geographic Information Systems and artificial intelligence techniques in line with the spatial intelligence approach. Due to their various urban points of interest and different demographic structures, Tuzla and Pendik districts of Istanbul and Çayirova, Darıca and Gebze districts of Kocaeli province were selected as the study area. The passenger demand forecasting model was produced using historical smart card data for M4 Metro and Marmaray stations located within the boundaries of the study area. Within the scope of the study, temporal factors were included in the forecasting model for the estimation of fluctuations in passenger demand and spatial factors were included in the forecasting model for the estimation of base passenger demand. Before using the big data set in machine learning techniques, Filter and Wrapper feature selection methods were applied. Models based on Random Forest and XGBoost algorithms were established with the created data sets. After the hyperparameters specific to the algorithms were adjusted manually and with the Grid Search method in accordance with the created data sets, models were established with the K-Fold Cross Validation technique. Model performances were evaluated in terms of performance comparison metrics and computational cost. By creating the importance level rankings of the factors used in the study with the highest performance forecasting model, their relationship with passenger demand was interpreted and the effect of spatial factors on increasing model accuracy was revealed. The selected forecast model was then used to create future passenger demand forecasts for existing and projected stations.

Keywords: Geographic Information Systems, Machine Learning, Random Forest, New Generation Mobility, Passenger Demand Forecast, XGBoost

TEŞEKKÜR

Gerçekleştirdiğim tez çalışması sürecinde kıymetli bilgi ve tecrübeleriyle desteğini esirgemeyen değerli hocam ve akademik danışmanım Prof. Dr. Arif Çağdaş AYDINOĞLU'na,

Yüksek lisans eğitimim boyunca her zaman yanımda olan, desteği ve güveninin hiçbir zaman esirgemeyen değerli hocam Prof. Dr. Şükran YALPIR'a,

Tez çalışmam boyunca yol gösterici katkılar sunan değerli iş arkadaşım Arş. Gör. Süleyman ŞİŞMAN'a

Son olarak, bugüne kadar hayatımın her anında yanımda olan, gösterdiğim emeklere en yakından şahitlik eden, desteklerini hiçbir zaman esirgemeyen çok değerli aileme teşekkür ederim.



İÇİNDEKİLER

	<u>Sayfa</u>
ÖZET	vi
ABSTRACT	vii
TEŞEKKÜR	viii
İÇİNDEKİLER	ix
SİMGELER ve KISALTMALAR DİZİNİ	xi
ŞEKİLLER DİZİNİ	xii
TABLolar DİZİNİ	xiii
1. GİRİŞ	1
1.1. Problemin Tanımı	4
1.2. Çalışmanın Amacı	8
1.3. Metodoloji	9
1.4. Literatür Özeti	11
2. YENİ NESİL HAREKETLİLİK	21
3. YOLCU TALEP TAHMİNİNDE MEKANSAL ZEKA	25
3.1. Yapay Zekâ	27
3.2. Makine Öğrenmesi Süreci	29
3.3. Özellik Seçimi	31
3.3.1. Pearson Korelasyon	33
3.3.2. Ardışık İleri Yönde Seçim	35
3.3.3. Ardışık Geri Yönde Seçim	35
3.4. Tahmin Teknikleri	35
3.4.1. Karar Ağaçları	38
3.4.2. Topluluk Öğrenmesi	39
3.4.2.1. Torbalama Yaklaşımına Dayalı Rastgele Orman	41
3.4.2.2. Güçlendirme Yaklaşımına Dayalı eXtreme Gradient Boosting (XGBoost)	44
3.5. Hiper Parametre Ayarlama	51
3.6. Model Doğrulama	52
3.7. Model Performans Değerlendirme Metrikleri	54
4. UYGULAMA	56
4.1. Çalışma Bölgesinin Belirlenmesi	56
4.2. Veri Setlerinin Temini ve Hazırlığı	59
4.3. Mekânsal Analizlerin Gerçekleştirilmesi	64
4.3.1. Ağ Analizleri	65
4.3.1.1. Hizmet Alanı ve En Yakın Tesis Analizi	66
4.4. Özellik Seçimi Sonuçları	70
4.4.1. Pearson Korelasyon Sonuçları	70
4.4.2. Ardışık İleri Yönde Seçim ve Ardışık Geri Yönde Seçim Sonuçları	75
4.5. Yolcu Talep Tahmin Modellerinin Kurulması	79
4.6. Model Performanslarının Karşılaştırılması	83
4.7. Faktör Önem Düzeylerinin Sıralanması	85

4.8. Seçilen Tahmin Modelinin Mevcut ve Proje Aşamasındaki İstasyonlara Uygulanması	87
5. SONUÇLAR VE ÖNERİLER	97
KAYNAKLAR	104
ÖZGEÇMİŞ	112
TEZ METNİNDEN TÜRETİLEN YAYINLAR/SUNUMLAR	113



SİMGELER ve KISALTMALAR DİZİNİ

AdaBoost	: Adaptive Boosting
AdaBoost.R	: Adaptive Boosting Regressor
ARIMA	: Autoregressive Integrated Moving Average
BIC	: Bayesian Information Criteria
BILSTM	: Bidirectional Long Short-Term Memory
CART	: Classification and Regression Trees
CBS	: Coğrafi Bilgi Sistemleri
ÇDR	: Çoklu Doğrusal Regresyon
DVM	: Destek Vektör Makineleri
DVR	: Destek Vektör Regresyon
GPS	: Global Positioning System
IoT	: Internet of Things
LASSO	: Least Absolute Shrinkage and Selection Operator
LoRa	: Long Range
LSTM	: Long Short-Term Memory
LSTMNN	: Long Short-Term Memory Neural Network
MAPE	: Mean Absolute Percentage Error
MDI	: Mean Decrease Impurity
RFID	: Radio Frequency Identification
RMSE	: Root Mean Square Error
SARIMA	: Seasonal Autoregressive Integrated Moving Average
WMAPE	: Weighted Mean Absolute Percentage Error
XGBoost	: Extreme Gradient Boosting
5G	: Fifth Generation of Cellular Technology

ŞEKİLLER DİZİNİ

	<u>Sayfa</u>
Şekil 1.1: Çalışma iş akışı.	10
Şekil 2.1: Yeni nesil hareketlilik.	22
Şekil 3.1: Yapay zekâ ve alt kümeleri.	28
Şekil 3.2: Makine öğrenmesi süreci.	30
Şekil 3.3: Rastgele Orman çalışma prensibi.	43
Şekil 3.4: XGBoost çalışma prensibi.	47
Şekil 3.5: K-Katlı Çapraz Doğrulama çalışma prensibi.	53
Şekil 4.1: Çalışma bölgesi.	57
Şekil 4.2: Gebze-Darica Metrosu ve M4 Metro Uzatma hattı.	59
Şekil 4.3: Marmaray ve M4 Metro istasyonları için 2023 yılı günlük ortalama yolcu talebi.	62
Şekil 4.4: Sabiha Gökçen Havalimanı 2023 yılı Nisan ayında gerçekleşen hava olayları.	63
Şekil 4.5: Ağ tabanlı ve Öklid mesafeye dayalı hizmet alanı karşılaştırması.	66
Şekil 4.6: Ağ tabanlı 800 metrelik istasyon hizmet alanları.	67
Şekil 4.7: Ağ tabanlı 1600 metrelik istasyon hizmet alanları.	68
Şekil 4.8: En yakın tesis analizi çıktısı.	69
Şekil 4.9: Pearson Korelasyon fazlalık yaklaşımına göre hazırlanan matris.	72
Şekil 4.10: Pearson Korelasyon fazlalık ve ilgililik yaklaşımına göre hazırlanan matris.	74
Şekil 4.11: Ardışık İleri Yönde Seçim durma noktası tespiti.	77
Şekil 4.12: Ardışık Geri Yönde Seçim durma noktası tespiti.	78
Şekil 4.13: Torba dışı hata grafiği ile Rastgele Orman ağaç sayısının belirlenmesi.	81
Şekil 4.14: Grafik üzerinden model karşılaştırma.	84
Şekil 4.15: XGBoost faktör önem düzeyi sıralaması.	86
Şekil 4.16: Marmaray ve M4 Metro istasyonları 2023 yılı gerçekleşmiş ve 2024 yılı tahmin edilen günlük ortalama yolcu talebi.	88
Şekil 4.17: Hafta içi ve hafta sonu yolcu talebindeki dalgalanmalar.	89
Şekil 4.18: Marmaray ve M4 Metro istasyonları için 2024 yılı günlük yolcu tahmini.	90
Şekil 4.19: Gebze-Darica Metro hattı ve M4 Metro Uzatma hattı istasyonları günlük ortalama yolcu talebi.	92
Şekil 4.20: Marmaray ve M4 Metro istasyonları 2023 yılı gerçekleşmiş yolcu talepleri ve farklı ulaşım modu faktörü ile ilişkisi.	93
Şekil 4.21: Farklı ulaşım modu faktörü ve yolcu talebi ilişkisi.	94
Şekil 4.22: Tüm istasyonlar için 2024 yılı günlük yolcu talep tahmin sonuçları.	95

TABLÖLER DİZİNİ

	<u>Sayfa</u>
Tablo 1.1: Literatürdeki çalışmaların sınırlılıkları.	19
Tablo 3.1: XGBoost hiper parametrelerinin değer aralıkları ve varsayılan değerleri.	50
Tablo 4.1: Çalışma veri kümesi.	60
Tablo 4.2: Günlük yolcu sayısı verisi.	61
Tablo 4.3: Mekânsal faktörler veri kaynakları.	64
Tablo 4.4: Mekânsal faktörler ve kullanıldıkları analiz yöntemi.	69
Tablo 4.5: Mekânsal analizler sonucunda oluşturulan veri seti.	70
Tablo 4.6: Faktör isimleri için belirlenen kod adları.	71
Tablo 4.7: Pearson Korelasyon fazlalık yaklaşımı ile faktör seçimi.	73
Tablo 4.8: Pearson Korelasyon fazlalık ve ilgililik yaklaşımı ile faktör seçimi.	75
Tablo 4.9: Ardışık İleri Yönde Seçim sonucunda elde edilen alt küme.	77
Tablo 4.10: Ardışık Geri Yönde Seçim sonucunda elde edilen alt küme.	79
Tablo 4.11: Rastgele Orman için belirlenen hiper parametre değerleri.	81
Tablo 4.12: XGBoost hiper parametreleri ve İzgara Arama yöntemiyle test edilen değerler.	82
Tablo 4.13: İzgara Arama sonuçlarına göre seçilen XGBoost hiper parametre değerleri.	82
Tablo 4.14: Model performanslarının karşılaştırması.	83
Tablo 4.15: Marmaray ve M4 Metro istasyonları için gerçekleştirilen 2024 yılı günlük yolcu talep tahminleri.	88
Tablo 4.16: M4 Metro Uzatma hattı ve Gebze-Darica Metro hattı istasyonları için gerçekleştirilen günlük yolcu talep tahminleri.	91
Tablo 4.17: Proje aşamasındaki istasyonlar için 2024 yılı günlük ortalama yolcu sayıları.	91

1. GİRİŞ

Küresel ölçekte yüksek nüfus artışı ve kentleşme hızı, içerisinde yaşadığımız şehirleri farklı yönlerden etkilemektedir. Günümüzde küresel nüfusun yarısından fazlası şehirlerde yaşamaktadır. Nüfus artışı ve kentleşme eğiliminin bu şekilde devam etmesi durumunda 2050 yılına kadar kentsel nüfusun günümüzdeki kentsel nüfusun iki katına çıkması beklenmektedir. Bu senaryoda, içerisinde bulunduğumuz yüzyılı yarılacağımızda toplam küresel nüfusun yaklaşık %70'inin şehirlerde yaşaması beklenmektedir [Web 1, 2025]. Kentleşme ile birlikte kalabalıklaşan bu şehirler, günümüzde dünya enerjisinin üçte ikisini tüketmekte ve küresel sera gazı emisyonlarının 3'te ikisinden fazlasını üretmektedir. Şehir nüfusundaki ciddi artış ve büyüyen şehirler, ulaşım talebi artırmaktadır [Cyril vd., 2018]. Ulaşım talebin artması sonucu artan bireysel araç sahipliği oranları, sera gazı emisyonları gibi çevresel etkiler başta olmak üzere ekonomik ve sosyal açıdan şehirleri, içerisindeki şehir sakinlerini ve çevreyi olumsuz yönde etkilemektedir. Küresel ölçekte incelendiğinde toplam karbon emisyonlarının yaklaşık %25'lik bir kısmını ulaşım sektörü oluşturmaktadır [Hussain vd., 2023]. Günümüz şehirleri, kentsel çevreyi korumak ve iyileştirmek amaçlarıyla hayati öneme sahip olan ulaşımın sürdürülebilirliği için mücadele vermektedir [Brčić vd., 2018].

Şehirleri çevresel, sosyal ve ekonomik yönlerden olumsuz etkileyen yüksek nüfus ve kentleşme eğilimi neticesinde kentsel ulaşım konusunda daha yenilikçi çözümler geliştirilmektedir. Şehirlerin çevresel ve ekonomik açıdan belirli bir farkındalığa erişmesi ile yenilikçi çözümlerin geliştirilmesi fikrinin ortaya çıkışı klasik hareketlilik kavramını şekillendirmektedir. Yenilikçi fikirlerin gelişmesinde ve hareketlilik kavramının geliştirilmesinde Bilgi ve İletişim Teknolojileri alanının önemi oldukça yüksektir. Gelişen Bilgi ve İletişim Teknolojileri alanı ile birlikte şehirlerde akıllı çözümler mümkün hale gelmekte ve akıllı şehir kavramını ortaya çıkarmaktadır. Bir akıllı şehir, genel altyapı açısından son derece gelişmiş, yüksek teknolojiye sahip, güvenilir iletişim sistemleri ve akıllı güvenlik çözümleriyle desteklenen kentsel bölgeleri ifade etmektedir [Vakula ve Raviteja, 2017]. Akıllı şehir kavramı, akıllı çevre, akıllı atık yönetimi, akıllı sağlık ve akıllı ulaşım gibi unsurları içermektedir. Bu kavramlar arasındaki akıllı ulaşım, Nesnelerin İnterneti (Internet of Things-IoT)

alanındaki gelişmeler neticesinde trafik sıkışıklığının azaltılması ve karbon emisyonlarının önemli ölçüde düşürülmesi gibi birçok fayda sağlamaktadır. Akıllı şehir kapsamında akıllı ulaşım uygulamalarının geliştirilmesi sonucunda hareketlilik kavramı gelişmektedir. Değişim ve gelişim göstererek kullanıcı odaklı, çevreci, ekonomik, dijital ve böylece daha sürdürülebilir ulaşımı destekleyen kentsel hareketlilik kavramı, yeni nesil hareketlilik olarak ifade edilebilmektedir. Yeni nesil hareketlilik kavramı içerisinde elektrikli araçlar, otonom araçlar, mikro hareketlilik ve paylaşımlı hareketlilik ulaşım çözümleri bulunmaktadır [Alam vd., 2024]. Paylaşımlı hareketlilik kavramı içerisindeki toplu taşıma sistemleri kent sakinlerini toplu şekilde taşıyabilme yetenekleri ile trafik sorunlarının çözümünde çevreci, ekonomik ve sosyal anlamda erişilebilir bir alternatif sunmaktadır.

Toplu taşıma sistemleri, büyüyen ve kalabalıklaşan şehirlerde kentsel erişilebilirliği artırırken yüksek kapasitesi ile kent sakinlerinin verimli ve düşük maliyetlerle taşınmasını sağlamaktadır. Bu özellikleri ile yerel yönetimler tarafından trafik ve ulaşım sorunlarının önüne geçilmesi ve yolcuların verimli şekilde taşınması amacıyla ön plana çıkarılmaktadır [Egercioğlu ve Doğan, 2016]. Toplu taşıma sistemlerinde hizmet veren araçların büyük bir bölümü kara tabanlı olmakta ve genel kent trafiği ile aynı yolu kullanmaktadır. Genel kent trafiği ile aynı yolu kullanan otobüs gibi taşıtlar erişilebilirlik oranlarını artırsa da mevcut kent trafiğinden etkilenmeleri büyük bir dezavantajdır. Bu yüzden yüksek trafik sıkışıklığı sorunlarıyla baş etmekte zorlanan yerel yönetimler için, toplu taşıma sistemlerinin verimliliğinin artırılması amacıyla ayrılmış yollara sahip toplu taşıma çözümleri önemli hale gelmiştir [Fouracre vd., 2003]. Ayrılmış yola sahip olan toplu taşıma çözümleri, mevcut şehir trafiğinden etkilenmemesi sebebiyle düzenli sefer aralıklarına sahip olmaktadır. Ayrıca demiryolu altyapısını kullanan metro ve banliyö treni gibi sistemler düzenli sefer aralıklarının yanında yüksek yolcu kapasitesi ve elektrik enerjisine dayalı çalışmalarından dolayı çevreci olma özelliklerine sahip olmaktadır. Sahip oldukları özellikleri bakımından hızlı toplu taşıma olarak adlandırılan bu sistemlere yerel yönetimler tarafından önemli yatırımlar yapılmaktadır. Bu sistemlerin sürdürülebilirliği, akıllı çözümlere dayalı mekânsal kararlara bağlı olmaktadır.

Gün geçtikçe gelişen IoT teknolojilerine bağlı olarak veri toplayan aygıt sayısında artış meydana gelmiştir. Veri üreten aygıtların sayısındaki artış sebebiyle büyük veri kalıpları meydana gelmektedir. Toplu taşıma sistemlerinde akıllı kart teknolojisinin

kullanımıyla seyahat ve yolcu verileri toplanabilmekte ve depolanabilmektedir **[Bagchia ve White, 2005]**. Bu sistemlerden elde edilen büyük veri kalıpları içerisinde bulunan coğrafi referanslı veriler, akıllı çözümlere dayalı mekânsal kararların verilmesini sağlamaktadır. Mekânsal kararların verilmesinde, coğrafi referanslı verilerin işlenmesi, analizi ve görselleştirilmesinde Coğrafi Bilgi Sistemleri (CBS) yaygın biçimde kullanılmaktadır **[Choi, 2023]**. Ancak coğrafi referanslı veri miktarının hızlı artışı sonucunda mekânsal analizlerde CBS teknikleri çeşitli güçlüklerle karşılaşmaktadır. Büyük veri kalıplarıyla gerçekleştirilecek mekânsal uygulamalarda yapay zekâ teknikleri CBS teknikleriyle birlikte kullanılabilir. Yapay zekâ teknikleri ile coğrafi referanslı büyük veri kalıplarına yönelik veri madenciliği işlemleri verimli bir şekilde gerçekleştirilebilmektedir **[Witten vd., 2011]**. CBS ve yapay zekâ tekniklerinin beraber kullanımı, insan ve çevre etkileşimlerini akıllı algoritmalar yardımıyla çözmeyi amaçlayan bir çalışma alanı olan mekânsal zekâ alanını ortaya çıkarmıştır **[Gao, 2021]**. Toplu taşıma sistemlerinden elde edilen akıllı kart verilerinin ve mekânsal zekâ yöntemlerinin toplu taşıma sistemlerine yönelik verilecek mekânsal kararlarda kullanımı, bu sistemlerin daha sürdürülebilir olmasına katkı sağlamaktadır.

Toplu taşıma sistemlerine yönelik verilecek mekânsal kararlar içerisinde istasyon konumlarının ve güzergahlarının belirlenmesi yer almaktadır. Başarılı bir toplu taşıma sistemi, gerçekleşecek yolcu talebinin verimli biçimde karşılanması ile mümkün hale gelmektedir. Burada ifade edilen yolcu talebi, bir ulaşım modunu kullanmak üzere akıllı kart teknolojisi ile istasyonlara geçiş yapan yolcu sayısını ifade etmektedir. İlgili istasyon için gerçekleşen yolcu talebinin düşük olması durumunda yapılan yatırım karşılık bulamayabilirken, çok yüksek talep durumu ise yolcuların verimsiz bir seyahat yapmalarına sebebiyet verebilmektedir. En ideal durum istasyona olan talebin yeteri kadar olması durumunda ortaya çıkmaktadır. Bu sebeplerden dolayı yolcu talebinin tahmin edilmesi verilecek mekânsal kararlarda oldukça önemli bir yer tutmaktadır. Ayrıca toplu taşıma sistemlerinin operasyonel planlaması ve stratejik geliştirmeler için mevcut istasyonların yolcu talebinin değişen koşullara göre değerlendirilmesi gerekmektedir. Değişen koşullara göre yolcu talebindeki değişikliklerin tahmin edilmesi, akıllı kart verilerinin ve mekânsal zekâ tekniklerinin kullanımı gerekmektedir. Gelecekteki toplu taşıma sistemi yolcu talebinin doğru şekilde tahmini,

alıřma alanını doęru temsil eden ve yolcu talebiyle iliřkili faktörlerin doęru řekilde deęerlendirilmesine baęlı olmaktadır [Cyril vd., 2018].

1.1. Problemin Tanımı

Büyüyen řehirlerde artan ulaşım talebi, beraberinde ulaşım ile ilgili çeřitli sorunları meydana getirmektedir. Ulaşım talebinin karřılanması amacıyla toplu taşıma sistemlerine önemli yatırımlar yapılmaktadır. Ancak bu yatırımların, yolcu talepleri ve mekânsal faktörler dikkate alınmadan yapılması hem kaynak israfına yol açmakta hem de sistemin etkinlięini düşürmektedir. Bahsi geen toplu taşıma sistemleri arasında yer alan demiryolu altyapısını kullanan sistemlerin etkin yönetimi, operasyonel verimlilik ve yolcu hizmeti memnuniyeti açısından kritik öneme sahiptir. Bu doęrultuda akıllı kart teknolojisi ile istasyonlara geiř yapan yolcuları ifade eden yolcu talebine yönelik tahminlerin elde edilmesi, ekonomik kalkınma ve kaynak tahsisi üzerinde önemli bir etkiye sahiptir. Bu kapsamda gerekleřtirilen yolcu talep tahmini, yolcu talebinden kaynaklı hizmetin arzının doęru planlanması kapsamında mekânsal karar süreçlerinde bilimsel bir temel saęlayarak kritik bir rol oynamaktadır [Nar ve Arslankaya, 2022], [Feng vd., 2021].

Yolcu talep tahmini alıřmaları için yolcu ve toplu taşıma araçlarına yönelik verilerin toplanması bir ön kořul oluřturmaktadır. Akıllı kart teknolojisinin kullanımı ile yolcuların gerekleřtirmiş oldukları seyahatlere ait güzergâh ve zaman gibi veriler toplanabilmekte ve ilgili güzergahların gelecekteki planlamaları açısından önemli katkılar sunmaktadır [Vakula ve Raviteja, 2017]. Akıllı kart verileri ile yolcuların ilgili istasyon için geiř verileri saatlik ve günlük gibi zaman aralıklarında toplanabilmekte ve böylece istasyon bazında yolcu talebi hesaplanabilmektedir. Toplanan veriler ile günün belirli bir saatindeki veya belirli bir gündeki yolcu sayısındaki artış veya azalışlarla ilgili bilgiler elde edilebilmektedir [Boreiko ve Teslyuk, 2016]. Elde edilen bu veriler ışığında toplu taşıma sistemlerinde gerekleřmiş olan gemiş yolcu taleplerinin kullanımıyla ileriye dönük yolcu talebi tahminleri gerekleřtirilebilmektedir.

Yolcu talebine yönelik gerekleřtirilen alıřmalarda yolcuların seyahat kararlarının birçok dış etkenden etkilendięi gösterilmiştir. Haftanın günü, hava durumu ve buna benzer zamansal anlamda deęiřiklikler sonucu yolcu talebinde dalgalanmalar

yaşanmaktadır. Zamansal faktörlere dayalı olarak yolcu talebinin tahmin edilmesine yönelik literatürde çokça çalışma bulunmaktadır [Liu vd., 2020], [Toque vd., 2017]. Ancak yolcu talebi zamansal faktörlere dayalı olarak değiştiği gibi ilgili istasyon için temel yolcu talebi, istasyonun kendisine has bazı özelliklere ve istasyonun yakın çevresine yönelik mekânsal faktörlere dayalı olarak da değişmektedir. Bir yolcunun yılın belirli bir günü belirli bir toplu taşıma sistemine ait istasyonu kullanmasına yönelik verdiği kararda haftanın günü veya hava durumu gibi faktörler etkili olduğu kadar bir yolcunun o istasyonu kullanmış olmasında istasyon çevresindeki belirli kentsel ilgi noktalarının da etkisi bulunmaktadır. Bu yüzden bir yolcu talebinin hem zamansal açıdan hem de mekânsal açıdan değerlendirilmesi gerekmektedir. Ancak literatürdeki çalışmalarda mekânsal faktörler büyük ölçüde göz ardı edilmiştir. Mekânsal verilerin de veri setine dahil edilmesi ve tahmin modellerinde kullanılması ile modelin eğitildiği istasyonlar dışında da uygulanabilir bir model geliştirilebilecektir. Bu amaç doğrultusunda coğrafi referanslı verilerden anlamlı bilgilerin çıkarılması konusunda çeşitli analizleri gerçekleştirme yeteneğine sahip olan CBS kullanımı gerekmektedir. CBS yazılımları sayesinde yolcu talebini oluşturan mekânsal faktörlere ait veriler üzerinde çeşitli analizler gerçekleştirilip tahmin modellerinde kullanılabilmektedir [Choi, 2023]. Mekânsal verilerin tahmin modellerinde kullanılması ile temel yolcu talebinin farklı konumlardaki istasyonlar için değişiminin yorumlanması mümkün hale gelmektedir. Ancak, büyük veri yapılarının işlenmesi ve analizinde CBS yazılımları tek başına yetersiz kalabilmektedir. Ayrıca yolcu talep tahmini çalışmasında CBS aracılığıyla üretilen çıktıların tahmin modellerinde kullanımı için Yapay zekâ teknikleri önemli hale gelmektedir [Gao, 2021]. Bu yüzden yolcu talebinin tahmin edilmesinde CBS ve yapay zekâ tekniklerinin birlikte kullanımıyla ifade edilen mekânsal zekâ yaklaşımları, mekânsal faktörlerin değerlendirilmesinde kritik bir rol oynayacaktır. CBS ortamında mekânsal faktörlere yönelik gerçekleştirilen analiz çıktılarının tahmin modellerinde kullanımında yapay zekâ teknikleri yüksek doğruluklar elde edilebilecektir.

Yolcu talebine yönelik gerçekleştirilen tahminler, kapsadıkları zaman dilimine göre kısa, orta ve uzun vadeli olacak şekilde genellikle üçe ayrılmaktadır. Kısa vadeli tahminler gerçek zamanlı verilerin değerlendirilmesi ile 5 dakikadan 30 dakikaya kadar olan zaman dilimlerini kapsamaktadır. Bu yaklaşım ile araçların kalabalık seviyelerine yönelik çıktılar elde edilebilmektedir. Orta vadeli tahminler, 30 dakikadan

birkaç saate kadar olan zaman dilimlerini kapsamaktadır. Uzun vadeli tahminler ise genellikle birkaç gün sonrasına yönelik olarak yapılabileceği gibi bir yıl sonrası gibi süreler için de yapılabilmektedir [Torrepadula vd., 2024], [Toque vd., 2017]. Benzer şekilde bir yolcu talep tahmin çalışması, belirli bir araç, durak veya istasyon ve güzergâh dikkate alınarak gerçekleştirilebilmektedir. Belirli bir araç için yapılan yolcu sayısı tahmin çalışmalarında araç içi sensör teknolojilerinden yararlanılmaktadır. Elde edilen tarihsel kalabalık verileri ile bir sonraki durak veya istasyon için kaç kişinin biniş yapacağı tahmin edilebilmektedir. Belirli bir durak veya istasyon için yapılan yolcu sayısının tahmin çalışmalarında, dikkate alınan durak veya istasyondaki geçmiş yolcu sayısı verileri kullanılarak sonraki zaman dilimleri için ilgili durak veya istasyon için yolcu sayısı tahmin edilmektedir. Belirli bir güzergaha yönelik yolcu sayısı tahmini çalışmalarında araçlara veya duraklara yerleştirilen sensörlerden elde edilen veriler tüm bir rota için bir araya getirilmektedir. Birleştirilen tarihi veriler, aynı rota için gelecekteki yolcu sayısının tahmini için kullanılmaktadır [Torrepadula vd., 2024]. Yolcu talebi tahmini, tramvay, otobüs, tren ve metro gibi ulaşım modları için gerçekleştirilebilmektedir. Bu ulaşım modları içinden demiryolu altyapısını kullanan sistemlerde geçmiş yolcu sayısına yönelik verilerin toplanması diğer ulaşım modlarına kıyasla daha kolay ve güvenilirdir. Bu verilerin kullanımı doğrultusunda metro ve tren gibi ulaşım modları için uzun vadeli tahmin modellerinin kurulması ile proje aşamasında bulunan istasyonların çekeceği talep miktarı tahmin edilebilecektir.

Yolcu talep tahmini çalışmalarında istatistiksel, zaman serisi ve makine öğrenmesi tabanlı yaklaşımlar yaygın biçimde kullanılmaktadır. Çoklu Doğrusal Regresyon (ÇDR), Otoregresif Entegre Hareketli Ortalama (Autoregressive Integrated Moving Average-ARIMA), Mevsimsel Otoregresif Entegre Hareketli Ortalama (Seasonal Autoregressive Integrated Moving Average-SARIMA), Destek Vektör Regresyon (DVR), Rastgele Orman ve Yapay Sinir Ağları yöntemlerine yönelik yolcu talep tahmini çalışmaları bulunmaktadır [Kendra vd., 2024], [Cyril vd., 2018], [Milenkovic vd., 2016], [Li vd., 2020], [Toque vd., 2017], [Liyanage vd., 2022].

Toplu taşıma sistemlerinde yolcu talebine yönelik yapılan çalışmalar irdelendiğinde doğrusal olmayan ilişkileri açıklamadaki gücü ile ön plana çıkan makine öğrenmesi yöntemlerinin kullanımının son yıllarda artış gösterdiği görülmüştür. Kurulan tahmin modellerinde girdi olarak, yolcu talebindeki günlük veya saatlik anlamda dalgalanmaları yansıtan zamansal faktörler detaylı biçimde ele alınmıştır. Ancak,

istasyonlar için temel yolcu talebini oluşturacak kentsel ilgi noktalarına ve demografik yapıya yönelik faktörlerin tahmin modellerine dahil edilmedikleri görülmüştür. Bu yüzden zamansal faktörlerin kullanıldığı ancak mekânsal faktörlerin değerlendirilmediği çalışmalarda üretilen tahmin modelleri ancak modelin eğitildiği istasyonlar için ileriye dönük yolculuk talebindeki dalgalanmaları tahmin edebilmiştir. Ayrıca yolcu talebine yönelik yapılan tahmin çalışmalarında karmaşık veri setlerini açıklamadaki gücüyle vurgulanan makine öğrenmesi tekniklerinin zamansal ve mekânsal faktörleri oluşturan veri setlerindeki performansı ölçülemediği görülmüştür. Bu yönleriyle oluşturulan yolcu talep tahmin modellerinin modelin eğitilmiş olduğu istasyonların dışında farklı mekânsal karakteristiğe sahip istasyonlarda uygulanabilirliği oldukça düşük olmaktadır. Bu yüzden mevcut istasyonlar için ileriye dönük tahminler gerçekleştirilebilmiş olsa da kentsel ölçekte proje hazırlığında bulunan toplu taşıma sistemlerine yönelik istasyonlar için gerçekleştirilecek temel yolcu talebinin tahmin edilmesi konusunda mekânsal karar destek sağlayamamaktadır. Bu sebeplerden dolayı hem zamansal hem de mekânsal faktörlerin birlikte değerlendirildiği modeller ile farklı konumlarda bulunan istasyonlar için de uygulanabilir bir model oluşturulabilecektir. Ayrıca yapılan yolcu talep tahmini çalışmalarında kullanılan veri seti için bir özellik seçimi uygulamasının çoğunlukla es geçildiği görülmektedir. Bunun sebeplerinden biri olarak veri setlerinin kısıtlı sayıda faktör içermesi gösterilebilmektedir. Oysa özellik seçimi yöntemlerinin makine öğrenmesi modelleri üzerindeki doğruluk artırıcı etkisini gösteren çalışmalar bulunmaktadır [Cai vd., 2018], [Ratnasingam ve Munoz-Lopez, 2023], [Hassan ve Kaabouch, 2024]. Çok sayıda zamansal ve mekânsal faktör barındıran bir veri seti için özellik seçimi uygulamasının gerçekleştirilmesi model doğruluğunu artıracak ve hesaplama maliyetini önemli ölçüde düşürebilecektir. Son olarak kullanılan makine öğrenmesi yönteminin içermiş olduğu ve model doğruluğunu artırabilen hiper parametre değerlerinin veri setine uygun bir şekilde ayarlanması kurulacak modellerin tahmin doğruluğunu artıracaktır. Tüm bu yaklaşımlarla hem yüksek tahmin performansı gösteren hem de proje aşamasında bulunan farklı mekânsal karakteristiğe sahip yerlerde uygulanabilir yolcu talep tahmin modelleri kurulabilecektir.

1.2. Çalışmanın Amacı

Son yıllarda artan nüfus ve kentleşme hızı kentsel ulaşımın verimli bir şekilde sağlanmasının önünde çeşitli engeller oluşturmaktadır. Çevresel, sosyal ve ekonomik açılardan verimli bir kentsel ulaşım, kent sakinlerinin ulaşım ihtiyaçlarının toplu bir şekilde gerçekleştirilmesini mümkün kılan toplu taşıma sistemlerine yönelik yatırımlarda güçlü mekânsal kararların verilmesini gerektirmektedir. Mevcut ve planlanan toplu taşıma sistemlerine yönelik gelecek yolcu talebinin tahmin edilmesi, güçlü mekânsal kararların verilmesini ve toplu taşıma sistemlerin daha sürdürülebilir olmasına katkı sağlayacaktır. Mekânsal faktörlere ait coğrafi referanslı verilerin gelişmiş CBS analizlerinin kullanımı ile anlamlı hale getirilmesi gerekmektedir. Oluşturulan zengin zamansal ve mekânsal veri setlerinin karmaşık veri setleriyle başa çıkmadaki yeteneği ile öne çıkan makine öğrenmesi metotlarının kullanımını gerektirmektedir. Bu kapsamda öncelikle makine öğrenmesi metotlarında yaşanabilen çoklu doğrusalılık sorununun önlenmesi ve böylece model performansının artırılması amacıyla özellik seçimi yöntemlerinin uygulanması gerekmektedir. Ayrıca makine öğrenmesi yöntemlerinin kendine has ve veri setine göre farklılık gösterebilen hiper parametrelerinin optimal değerlerinin tespiti gerekmektedir.

Böylelikle bu çalışmanın amacı; mevcut veya planlanan toplu taşıma istasyonlarında yolcu talebinin yüksek doğrulukla belirlenmesi için zamansal ve mekânsal faktörlerin, CBS ve makine öğrenmesi tekniklerinin kullanımını kapsayan bir metodoloji sağlamaktır. Bu kapsamda tez çalışmasında hedeflenenler;

- Günlük olarak yolcu talebindeki dalgalanmalara sebep olabilecek zamansal faktörlerin ve bir istasyon için temel yolcu talebini oluşturacak mekânsal faktörlerin tespiti,
- Çeşitli kentsel ilgi noktalarını ve farklı demografik yapıların irdelenmesini mümkün kılan, çalışma bölgesinin ve toplu taşıma sistemine ait örnek istasyonların seçilmesi,
- Tahmin modellerinin eğitimi için gerekli olan geçmiş yolcu sayısı veri setinin temini ve istasyon bazlı günlük toplam yolcu talebini gösterecek biçimde ön hazırlığı,
- Tespit edilen faktörlerin çalışma bölgesi de dikkate alınarak açık veri kaynaklarından ve ticari yazılım çözümlerinden temin edilmesi ve mekânsal

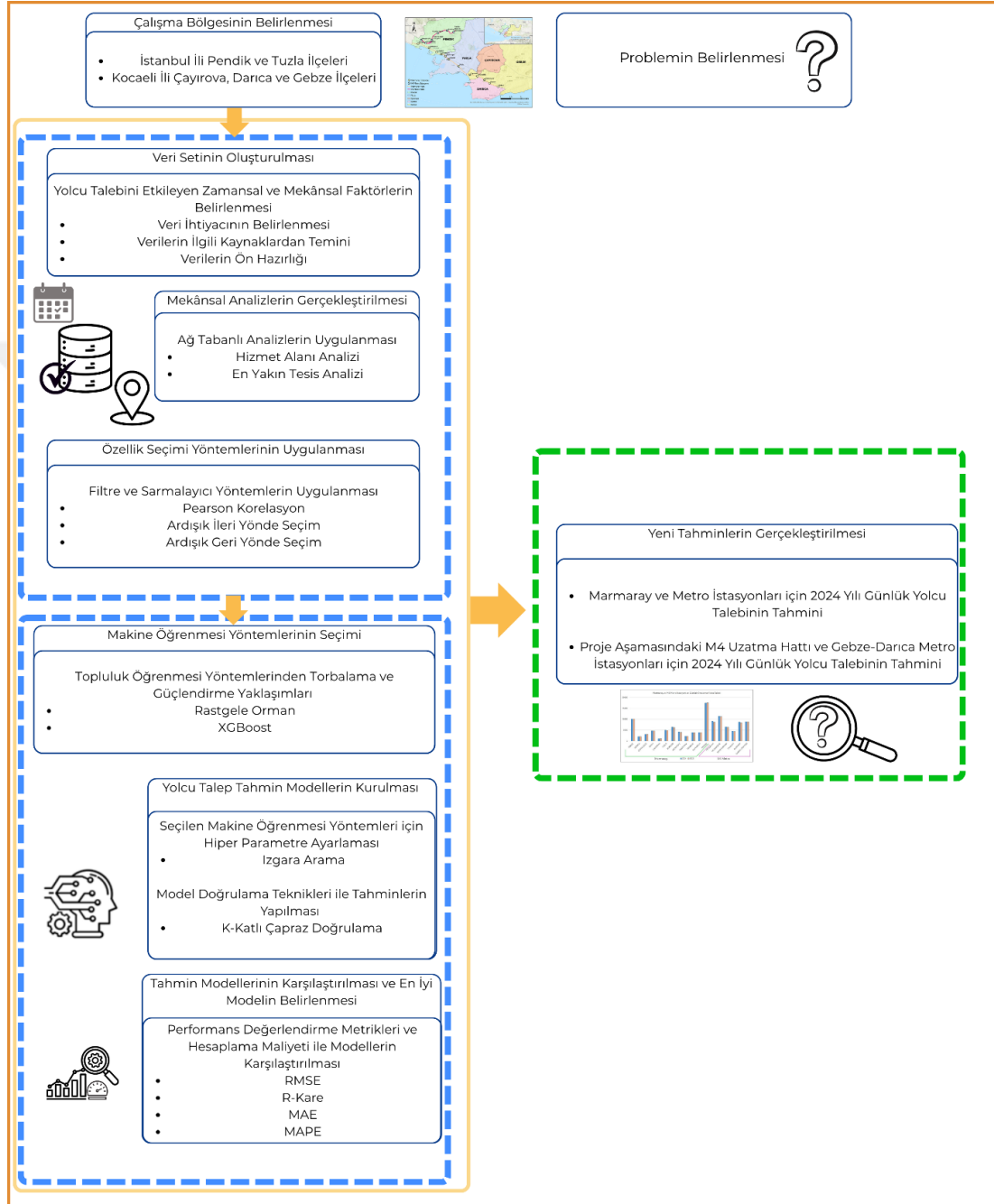
faktörler için temin edilen coğrafi referanslı verilerin CBS analiz teknikleriyle uygulamada kullanılabilir hale getirilmesi,

- Oluşturulan zengin veri seti için uygun makine öğrenmesi tekniğinin seçimi ve veri setine özellik seçimi yöntemlerinin uygulanması ve faktör alt kümelerinin oluşturulması,
- En yüksek tahmin doğruluğunu oluşturan faktör alt kümesinin ve hiper parametre değerlerinin tespiti ile yüksek tahmin doğruluğu ve düşük hesaplama maliyetine sahip makine öğrenmesi yöntemine ait modelin belirlenmesi,
- Oluşturulan yüksek doğrulukta tahmin modeli kullanılarak, modelin eğitildiği istasyonlar için ve çalışma bölgesi sınırları içinde proje aşamasında bulunan istasyonlar için günlük yolcu talebinin tahmin edilmesi olacak şekilde belirlenmiştir.

1.3. Metodoloji

Gerçekleştirilen tez çalışmasının amacı doğrultusunda, öncelikle literatür taraması ile toplu taşıma sistemlerinde yolcu talebinin tahminine yönelik gerçekleştirilmiş geçmiş çalışmalar incelenerek yolcu talep tahmini çalışmalarındaki eksiklikler ve geliştirilebilecek yönler tespit edilmektedir. Çalışma kapsamına üretilmek istenen tahmin modelinin farklı bölgeler içerisinde bulunan istasyonlar için de uygulanabilir olması amacıyla çalışma bölgesi zengin ve çeşitli mekânsal faktörleri barındırması dikkate alınarak belirlenmektedir. Belirlenen çalışma bölgesine yönelik zamansal ve mekânsal faktörlerin tespiti ve gerekli verilerin temini gerçekleştirilmektedir. Temin edilen mekânsal verilerin tahmin modellerinde kullanımı için mekânsal zekâ kapsamında mekânsal analizler uygulanmaktadır. Oluşturulan veri seti için özellik seçimi yöntemleri uygulanarak faktör alt kümeleri elde edilmekte ve karşılaştırılmaktadır. Veri setinin yapısı da dikkate alınarak makine öğrenmesi yöntemleri seçilmekte ve seçilen makine öğrenmesi yöntemleri için ayarlanabilir hiper parametre değerlerinin orijinal veri kümesi ve her bir faktör alt kümesi için parametre ayarlama yöntemleri ile optimal değerlerinin tespiti yapılmaktadır. Model doğrulanması amacıyla model doğrulama tekniklerinin kullanımıyla modeller oluşturulmaktadır. Model performans değerlendirme metrikleri ve hesaplama maliyetleri göz önüne alınarak yüksek performansa sahip tahmin modeli belirlenmektedir. Belirlenen tahmin modeli, modelin eğitildiği istasyonlar için ve

alışma bölgesi sınırları içinde proje aşamasındaki istasyonlar için gnlk yolcu talebinin tahmini amacıyla alıştırılmaktadır. Şekil 1.1, tez alışması için oluşturulan iş akış şemasını ifade etmektedir.



Şekil 1.1: alışma iş akışı.

1.4. Literatür Özeti

Tez çalışmasında zamansal ve mekânsal faktörlerden oluşan veri kümesi, CBS ve yapay zekâ tekniklerinin kullanımıyla toplu taşıma sistemlerine yönelik uzun vadeli yolcu talep tahmin modellerinin oluşturulması amaçlanmıştır. Bu kapsamda literatürde bulunan otobüs, tramvay, metro ve tren gibi sistemlere yönelik kısa ve uzun vadeli yolcu talep tahmin çalışmaları incelenmiştir. Gerçekleştirilen çalışmalarda kullanılan veri kümeleri, makine öğrenmesi algoritmaları ve kurulan modellerin performanslarına yönelik çıktılar dikkate alınmıştır.

Milenkovic vd., 2016 yılında yaptıkları çalışmada, Sırbistan sınırları içerisinde demiryolu ağında yolcu akışını analiz etmişlerdir ve uygulanabilir bir tahmin modeli önermişlerdir. Çalışmada gerçekleşmiş geçmiş aylık yolcu sayısı verileri kullanılmış ve mevsimselliği dikkate almak amacıyla SARIMA yöntemini kullanmışlardır. Tahmin modeli Ocak 2004 ve Haziran 2014 dönemini kapsayacak şekilde 126 aylık veri ile kurulmuştur. İlk 114 aylık gözlem modelin kurulumu için ve geriye kalan 12 aylık gözlem ise SARIMA modelinin test edilmesi amacıyla dikkate alınmıştır. Sırbistan için demiryolu yolcu taşımacılığında güçlü mevsimsellik desenlerinin olduğunu ve tüm çalışma dönemi boyunca sabit olmayan bir trend gösterdiğini belirtmişlerdir. Çalışmada varyans-ortalama grafikleri oluşturulmuş ve ortalamanın ciddi biçimde değiştiğini ve varyansta artış olduğunu tespit etmişlerdir. Varyansı stabilize etmek amacıyla veriler logaritmik dönüşüme tabi tutulmuş ve elde edilen veri seti kullanılmıştır. En iyi modelin tespiti amacıyla farklı SARIMA modelleri kurulmuş ve en uygun model normalize edilmiş Bayesci Bilgi Kriteri (Bayesian Information Criteria-BIC) değerine göre belirlenmiştir. En iyi model normalize edilmiş BIC metriğinde 7.056 ve Ortalama Mutlak Yüzdesel Hata (Mean Absolute Percentage Error-MAPE) metriğinde 4.176 değerlerini almıştır. Geliştirilen SARIMA modelinin Sırbistan demiryollarında aylık yolcu akışlarının tahmini için kullanılabileceği değerlendirilmiştir. Ayrıca yolcu istatistikleri incelendiğinde yüksek yolcu akışlarının yaz tatilinin sona ermesi ve okul ve üniversite faaliyetlerinin başlaması ile gerçekleştiğini eklemişlerdir. Gelecekteki araştırmalarda, demografik bilgilerinde kullanımıyla daha verimli tahmin modelleri kurmayı amaçladıklarını belirtmişlerdir [Milenkovic vd., 2016].

Toque vd., 2017 yılında yaptıkları çalışmada, akıllı kart verilerini kullanarak çok modlu ulaşım sistemlerinde her bir istasyona giren veya her bir duraktan binen yolcu sayısının tahminine yönelik makine öğrenmesi yöntemlerine dayalı bir tahmin modeli kurmuşlardır. Çalışmada bahsedilen çok modlu ulaşım sistemleri içerisinde otobüs, tramvay ve tren bulunmaktadır. Çalışmada hem uzun vadeli hem de kısa vadeli tahmin modelleri geliştirmişlerdir. Bu amaç doğrultusunda Rastgele Orman, Uzun-Kısa Süreli Bellek (Long Short-Term Memory-LSTM), Sinir Ağları ve Takvim Modelleri kullanmışlardır. Uzun vadeli tahmin modelleri bir yıl sonrası için gerçekleştirilirken kısa vadeli tahminler, tren ve tramvay durakları için gelecek 15 dakika ve otobüsler için gelecek 30 dakika için gerçekleştirilmiştir. Kısa vadeli tahminlerle ulaşım arzının gerçek zamanlı taleple eşleştirmenin işletmeler adına faydalı olacağını vurgulamışlardır. Ayrıca kullanıcılar açısından bu tahminlerin seyahat planlamalarında kalabalıklık düzeyleri hakkında fikir sahibi olabileceklerini belirtmişlerdir. Veri seti, Paris Metropol alan içerisindeki önemli bir iş bölgesinde bulunan toplamda 145 istasyon ve durak için toplanmış geçmiş iki yıllık yolcu verisinden oluşmaktadır. Uzun vadeli tahminler için yolcu sayısını etkileyen zamansal faktörler olarak okul dönemleri, okulların tatil olduğu günler, resmî tatiller, pazartesi cumartesiye kadar günler, tatil dönemi iş günleri, dört günlük tatiller belirlenmiştir. Çalışmada her bir istasyon veya durak için ayrı bir model oluşturulmuş ve tahminler geçmiş verilerde ilgili gün türü için gözlemlenen ortalama yolcu sayısı üzerinden yapılmıştır. Çalışmada kullanılan Rastgele Orman algoritması ile uzun vadeli tahminler gerçekleştirilirken yöntemde özel hiper parametreler, 2014 yılı verileri ile Izgara Arama yöntemi kullanılarak %80 eğitim ve %20 doğrulama olacak şekilde çapraz doğrulama tekniği ile optimize edilmiştir. Bu doğrultuda Rastgele Orman için ağaç sayısı 100 olarak belirlenmiştir. Kısa vadeli tahminlerde LSTM yöntemi uygulanmış ve Rastgele Orman ile karşılaştırılmıştır. Karşılaştırma 2015 yılı için toplanan test veri seti üzerinden yapılmıştır. Uzun vadeli tahminlerde Rastgele Orman algoritmasına dayalı modelin gayet iyi performans gösterdiği ortaya konmuştur. Kısa vadeli tahminlerde Tren istasyonları ve tramvay durakları için elde edilen tahmin doğruluğunun otobüs duraklarına göre daha yüksek olduğu ve en iyi sonuçların Rastgele Orman modelleri tarafından sağlandığı gözlemlenmiştir. Ancak LSTM tahmin modelinin de rekabetçi sonuçlar sunduğunu tespit etmişlerdir. Ayrıca, tahmin modellerinin istasyon veya durakların mekânsal yönünü dikkate almadığını belirterek

ideal bir tahmin modeli için bu konunun araştırılması gerektiğini belirtmişlerdir [Toque vd., 2017].

Cyril vd., 2018 yılında yaptıkları çalışmada, toplu taşıma sistemlerinden birisi olan otobüsler için aylık yolcu talebinin tahmin edilmesine yönelik tahmin modelleri oluşturmuşlardır. Çalışmada değerlendirilen otobüslerin ilçeler arası yolcu taşıyan otobüsler olduğunu belirtmişlerdir. Bu kapsamda çalışma alanı 6 bölgeye bölünmüş ve ilçe merkezleri yolculukların başlangıç ve varış noktası olarak kabul edilmiştir. Yolculukların ilçe merkezlerine doğru tek yönlü biçimde yapıldığı varsayılmıştır. Bu sistemler için yolcu talebinin tahmininde 2010 Ocak ve 2012 Aralık tarihleri arasındaki yolcu verilerini ve tek değişkenli ARIMA yöntemini kullanmışlardır. Yolcu verileri, elektronik bilet makinesi teknolojisinden elde edilmiştir. Tahmin edilen değerler 2013 yılında gözlemlenen gerçek talep değerleri ile karşılaştırılmıştır. Oluşturulan model, 2014, 2015, 2016 ve 2017 yılları için gerçekleştirilmiştir. Geliştirilen tahmin modelinin R^2 değerinin 0.96 olarak tespit edildiğini belirtmişlerdir. Sonuçların, yalnızca geçmiş yolcu talebi verilerini kullanan zaman serisi ARIMA yöntemiyle kurulan modelin, birbirine bağlı bölgeler için ve kısa vadeli tahminlerde doğru sonuçlar verdiğini belirtmişlerdir [Cyril vd., 2018].

Liu vd., 2020 yılında yaptıkları çalışmada, hava durumu koşullarının yolcu akışına etkisinin yaygın biçimde incelendiğini ancak geçmiş hava durumu verilerinin yolcu akışı tahminini nasıl daha iyi hale getirileceği konusuna yönelik çalışmaların sınırlı olduğunu belirtmişlerdir. Çalışmalarında derin Uzun Kısa Süreli Bellek Sinir Ağı (Long Short-Term Memory Neural Network-LSTMNN) kullanılarak saatlik metro yolcu akışı tahmin modeli geliştirmişlerdir. Taipei metrosuna ait Ana istasyon için saatlik giriş ve çıkış verilerini elde etmişlerdir. Taipei Ana istasyonunun seçilme sebebi olarak bölgedeki en yoğun metro olduğunu belirtmişlerdir. Çalışmada hava durumu etkilerinin kısa vadeli metro yolcu akışı tahminine olan kapsamlı etkisi tartışılmıştır. Bu kapsamda sıcaklık, yağış, bağıl nem ve rüzgâr hızı olmak üzere dört adet hava durumu değişkeni kullanılmıştır. İlgili hava durumu verileri Taipei Ana istasyonuna en yakın konumdaki gözlem istasyonundan saatlik olarak temin edilmiştir. Hava durumu dışında yılın ayı, haftası, günü ve tatil günleri faktörleri kullanılmıştır. Sonuçlar incelendiğinde, hava durumu faktörünün yolcu akışına etki eden önemli bir faktör olduğu sonucuna varılmıştır. Ayrıca, belirli bir istasyonda yağmurlu günlerde daha doğru tahmin sonuçları için bir saat önceki sıcaklık ve rüzgâr hızı faktörlerinin

en önemli iki hava durumu faktörü olduğu bulunmuştur. Çalışmada ulusal tatil günlerinde ve hafta sonlarında yapılan günlük yolculukların hafta içi yolculuklara kıyasla yağmurlu hava koşullarından daha fazla etkilendiği sonucuna ulaşmışlardır. Bu durumu, ulusal tatil ve hafta sonlarında seyahat planlarının zorunlu olmaması ve insanların yağmurlu günlerde seyahat planlarını değiştirme veya iptal etme olasılığının daha yüksek olması olarak yorumlamışlardır. Tahmin modeli performansının değerlendirilmesi amacıyla Kök Ortalama Kare Hatası (Root Mean Square Error-RMSE) ve MAPE metrikleri kullanılmıştır. Derin LSTM-NN tahmin modeli tek katmanlı LSTM-NN ve Rastgele Orman modellerinin elde etmiş olduğu sonuçlar ile karşılaştırılmıştır. Deneysel sonuçlar, derin LSTM-NN modelinin RMSE için 490.77 ve MAPE için 0.0547 değerleri ile en yüksek doğruluğa ulaşan model olduğunu göstermiştir. Çalışmada uygun hava durumu faktörlerinin kullanımı ile yüksek doğrulukta LSTM-NN tahmin modellerinin kurulabileceği sonucu ortaya çıkmıştır [Liu vd., 2020].

Zhou ve Tang, 2020 yılında yapmış oldukları çalışmada, kentsel raylı sistemlerin işletilmesinde tatil günlerindeki yolcu akışının tahmin edilmesinin öneminden bahsetmişlerdir. Yolcu talebinin yüksek doğrulukta tahmini ile daha etkin bir işletme yönetiminin ve organizasyon planlarının oluşturulabileceğini vurgulamışlardır. Farklı senaryolar için farklı tahmin modellerinin kullanılması gerektiğinden ve bu çalışma kapsamında tatil günlerine özel tahmin gerçekleştirecek bir tahmin modeli kurmuşlardır. Geçmiş tatil günleri ve öncesindeki cumartesi günlerine ait geçmiş verileri kullanmışlardır. Yolcu akışı örnek verisi her gün sabah 5'ten gece 0'a kadar, zaman çözünürlüğü 1 saatlik bir veri seti kullanmışlardır. İstasyonlardaki yolcu talebinin tahmininde iyi bir genelleme yeteneğine sahip olduğunu belirttikleri Destek Vektör Makineleri (DVM) yöntemini kullanmışlardır. Çalışmanın çıktısı ise saatlik olarak tatilin ilk gününe ait gelen yolcu akışı verileri olmuştur. Tahmin verileri gerçek veriler ile kıyaslanmış ve analiz sonucunda tahmin modelinin doğruluğu test edilmiştir. Sonuçlar incelendiğinde yolcu akışına yönelik tahmin hatasının %5'in altında olduğunu ve bu sebeple modelin yüksek doğruluğa ulaşarak kentsel raylı sistemlerin işletiminde önemli karar destek sağlayacağını belirtmişlerdir [Zhou ve Tang, 2020].

Li vd., 2020 yılında yaptıkları çalışmada otobüsler için yolcu akışının tahmin edilmesinin kaynak tahsisi, rota planlaması ve sefer sıklığı ayarlamaları için kritik

olduğunu vurgulamışlardır. Çalışmalarında kümeleme yöntemleri kullanılarak kümeye dayalı zaman aralıkları oluşturulmuştur. Oluşturulan her küme için DVR kullanarak yolcu akışı tahmininde bulunulmuştur. Bu kapsamda DVR parametrelerinin belirlenmesinde Parçacık Sürü Optimizasyonu kullanılmıştır. Çalışma kapsamında Çin'in Guangzhou kentindeki 15 numaralı otobüs hattına ait 2014 yılı içerisindeki 126 günlük yolcu verilerini kullanmışlardır. Kullanılan yolcu verileri her gün için saat 6.00'dan 23.59'a kadar toplamda 16 zaman aralığından oluşmaktadır. Verileri incelediklerinde otobüs hattına ait yolcu akışının döngüsel olduğunu ve her hafta benzer şekilde gerçekleştiğini fark etmişlerdir. Ancak ulusal gün için bu durumun farklılaştığını ve normal günlere kıyasla yolcu akışı bariz bir şekilde daha az olduğunu eklemişlerdir. Yolcu akışı veri kümesi %80 eğitim ve %20 test verisi olarak ikiye ayrılmıştır. Çalışmalarında önerdikleri model, dört aylık bir süreye ait gerçek yolcu verileriyle test edilmiştir. Çalışma sonunda oluşturulan tahmin modelinin özel etkinlik günleri ve hava durumu gibi dış faktörler nedeniyle oluşan kesintileri kapsamlı şekilde tahmin edemediğini bu yüzden bu faktörlerin daha sonraki çalışmalarda değerlendireceklerini belirtmişlerdir [Li vd., 2020].

Baro ve Khoudjai, 2021 yılında yaptıkları çalışmada, yolcu sayısının tahmini için makine öğrenmesi yöntemlerinden topluluk öğrenmesi yaklaşımı içerisindeki torbalama türü olan Rastgele Orman algoritmasını kullanmışlardır. Çalışmada Rastgele Orman algoritmasının mekânsal ve zamansal faktörlere duyarlılığı incelenmiştir. Hizmet verilen istasyon ve tren güzergahı çalışmadaki mekânsal faktörleri oluşturmuştur. Yılın günü ve tatil günleri faktörleri ise zamansal faktörleri oluşturmuştur. Belirlenmiş olan mekânsal ve zamansal faktörlerin tahmin performansına etkilerini incelemişlerdir. Aynı veri seti ile yapılan önceki çalışmalarda kullanılan eXtreme Gradient Boosting (XGBoost), Rastgele Orman ve Yapay Sinir Ağlarına dayalı modellerin performanslarının birbirine yakın olduğunu ve bu yüzden düşük hesaplama maliyeti, yorumlanabilirliği ve daha karmaşık yöntemlere kıyasla oldukça yakın doğrulukta performans göstermesi sebebiyle Rastgele Orman algoritması kullanmışlardır. Modeller Python ortamında 2016 yılı üzerinden eğitilirken 2017 yılı test verisi olarak kullanılmıştır. Çalışmada öncelikle zamansal ve mekânsal faktörlere yönelik ayrı ayrı modeller oluşturmuşlardır. Daha sonrasında hem zamansal hem de mekânsal faktörleri içeren ortak bir model oluşturmuşlardır. Rastgele Orman hiper parametreleri için Izgara Arama ve model doğrulama için Çapraz

Doğrulama metotlarını kullanmışlardır. Zamansal ve mekânsal faktörlerin birlikte değerlendirilmiş olduğu tahmin modelinin RMSE ve Ağırlıklı Ortalama Mutlak Yüzde Hata (Weighted Mean Absolute Percentage Error-WMAPE) model performans değerlendirme metriklerine göre daha yüksek doğrulukta sonuç verdiğini görmüşlerdir. Böylece hem zamansal hem de mekânsal faktörlerin birlikte yer aldığı veri setlerinin Rastgele Orman ile yapılacak tahmin modelleri için uygulanabilirliğini ifade etmişlerdir [**Baro ve Khoudjai, 2021**].

Jie vd., 2021 yılında yapmış oldukları çalışmada, Tayvan'da en sık kullanılan toplu taşıma araçlarından biri olan Taipei metrosunun günlük yolcu tahminlerinin gerçekleştirilmesini amaçlamışlardır. Çalışmada 2018 ile 2019 yılları arasındaki günlük giriş ve çıkış hacmini kullanmışlardır. Çalışma bölgesi tayfun ve deprem gibi afetlerin yaşandığı bir yerde yer almaktadır. Tayfun ve deprem afetlerinin yaşandığı günlerde yolcu sayısında önemli bir azalma yaşandığını görmüşlerdir. Bu sebeple değerlendirmeye alınan dış faktörlerden birisi afet faktörü olmuştur. Çalışmada kullanılan dış faktörlerden bir diğeri Google Arama verileri olmuştur. Bu kapsamda Google Trend üzerinden belirli konulara dayalı olarak anahtar kelimelerin arama sıklıkları toplanmıştır. Dikkate alınan son faktör ise takvimsel değişkenler olmuştur. Çalışmada belirlenen faktörlere ait veri setleri dört farklı kombinasyon şeklinde modele dahil edilmiştir. İlk veri seti takvimsel değişkenlerin tek başına bulunduğu veri setidir. İkinci veri seti takvimsel ve afet faktörlerinin bulunduğu veri seti olmuştur. Üçüncü veri seti takvimsel değişkenler ve Google Arama verilerini barındıran veri seti olmuştur. Son ve dördüncü veri seti ise bu üç faktörün bir arada yer aldığı veri seti olmuştur. Belirlenen faktörler ile Doğrusal Regresyon, En Az Mutlak Küçültme ve Seçme Operatörü (Least Absolute Shrinkage and Selection Operator-LASSO), Rastgele Orman ve ARIMA yöntemleriyle tahmin modelleri oluşturmuşlardır. Tahmin modelleri, RMSE ve MAPE ölçütleri ile karşılaştırılmıştır. Karşılaştırma sonucunda en kötü performansı Doğrusal Regresyon gösterirken en iyi performansı Rastgele Orman algoritması göstermiştir. En yüksek doğruluğu veren veri seti tüm faktörlerin bir arada değerlendirildiği dördüncü veri seti olmuştur. Google Arama verilerinin model doğruluğunu gerçekten artırabileceğini bulmuşlardır. Gelecek çalışmalar için elde edilen sonuçların mekânsal ve zamansal açıdan farklı istasyonlar için genellenebilir olup olmadığının değerlendirileceğini belirtmişlerdir [**Jie vd., 2021**].

Kusonkhum vd., 2022 yılında yaptıkları çalışmada, zaman içindeki yolcu talebindeki değişimi tahmin edebilecek makine öğrenmesi modeli geliştirmeyi amaçlamışlardır. Çalışmayı Tayland'daki Hızlı Transit Mor hattına ait 16 istasyon için gerçekleştirmişlerdir. Her istasyon için 15 dakikalık periyotlarda yolcu verileri toplanmıştır. Çalışmada Yapay Sinir Ağları, Rastgele Orman ve Karar Ağaçları gibi algoritmaları kullanmışlar ve oluşturulan tahmin modellerini karşılaştırmışlardır. Çalışmalarında hava durumunu da içeren zamansal bir veri seti kullanmışlar ve mekânsal faktörleri dikkate almamışlardır. Zamansal faktörler olarak; tarih, gün, ay, zaman dilimi, resmî tatiller, hafta sonları ve hava durumu faktörleri kullanılmıştır. Bu zamansal faktörler 2017 ve 2019 tarihleri arası dikkate alınarak toplanmıştır. Bu çalışmada regresyon yönteminin doğruluk oranının düşük olduğunu belirterek sınıflandırma yaklaşımını seçtiklerini belirtmişlerdir. Ancak regresyon için kullanılan yöntemin Doğrusal Regresyon olduğunu belirtmişlerdir. Bu amaç doğrultusunda yolcu sayılarına yönelik kümeler oluşturmuşlardır. Verilerin %80'i eğitim ve %20'si test olacak şekilde ikiye ayrılmıştır. Yapay Sinir Ağlarının bu çalışma kapsamında yüksek doğrulukta sonuçlar verdiğini ifade etmişlerdir. Çalışma sonucunda hava durumu ile ilişkili yağmur faktörünün model doğruluğunu artırdığını görmüşlerdir **[Kusonkhum vd., 2022]**.

Nar ve Arslankaya, 2022 yılında yaptıkları çalışmada demiryolu toplu taşıma sistemine yönelik yolcu talep tahmini ele alınmıştır. Çalışmada Yenikapı–Kirazlı metro hattına ait istasyonlar ve 2019 yılına ait geçmiş yolcu verileri kullanılmıştır. Bu çalışmada hem hat hem de istasyon bazında 2020 yılı için tahminler üretilmiştir. Hem hat hem de istasyon için yolcu talep tahminin tek çalışmada yapmalarını literatüre katkı olarak sunmuşlardır. İlk olarak hat bazında yolcu talebi tahmini regresyon analizi ve basit ortalama gibi istatistiksel yöntemlerle yapılmış ve sonuçlar ise MAPE metriği ile karşılaştırılmıştır. Çalışmanın ikinci adımında istasyon bazında yolcu talep tahmini Python programı aracılığıyla Yapay Sinir Ağları ve makine öğrenmesi algoritmalarından Doğrusal Regresyon, Rastgele Orman, Polinomsal Regresyon ve DVR ile gerçekleştirilmiştir. Tahminler hafta içi ve hafta sonu günleri için ayrı ayrı gerçekleştirilmiştir. Modellerin karşılaştırılmasında BIAS, MSE, MAPE ve RMSE metrikleri kullanılmıştır. Çalışma sonucunda hat bazında en iyi tahmin modelinin basit ortalama ile yapıldığı ve istasyon bazında ise en başarılı sonuçların Karar Ağaçları ile elde edildiği görülmüştür **[Nar ve Arslankaya, 2022]**.

Liyanage vd., 2022 yılında yapmış oldukları çalışmada, talebe dayalı toplu taşıma hizmetlerinin sürdürülebilirliği için kamu toplu taşıma sistemlerinde kısa vadeli yüksek doğrulukta tahminlerin gerçekleştirilmesi gerektiğini vurgulamışlardır. İşletmeler için önemli olan kısmın gelecek seyahat planının nerede ve ne zaman olacağının öngörülmesi ve böylece işletmelerin ulaşım tarifelerini hızlı bir şekilde ayarlamaları olduğunu belirtmişlerdir. Avustralya'nın Melbourne şehri bu araştırmanın çalışma sınırlarını oluşturmuştur. Bu çalışmada akıllı kart teknolojisinin sağlamış olduğu 2018 yılı bir aylık gerçek kullanım verilerini kullanarak otobüs yolcu talebinde yapay zekâ tabanlı derin öğrenme modeli önermişlerdir. Çalışmada ilgili bölge için en yoğun 18 otobüs hattı ve 1781 otobüs durağına ait veriler kullanılarak LSTM ve Uzun Kısa Süreli Belleğe Sahip Çift Yönlü Sinir Ağları (Bidirectional Long Short-Term Memory-BILSTM) derin öğrenme yöntemlerine dayalı gelecek 15, 30 ve 60 dakikaya yönelik tahminler yapan modeller üretilmiştir. Önceki çalışmalarda bu tahminlerin simülasyon verileri kullanılarak gerçekleştirildiğini, ancak bu çalışmada gerçek verilerin kullanımının literatüre katkı olarak sunmuşlardır. Kart verileri istatistiksel olarak incelendiğinde hafta içi günlerinin okul ve işe gitme sebepleriyle yoğun günler olduğu ve hafta içleri ile hafta sonlarının her hafta neredeyse aynı döngüsel varyasyonu izlediği sonucuna erişilmiştir. Ayrıca hafta sonlarında yolcu talebinin daha az dalgalandığını görmüşlerdir. Bu çalışmada LSTM ve BILSTM yöntemleri aynı veri seti ile beş geleneksel derin öğrenme yöntemine dayalı oluşturulmuş tahmin modelleri ile MAPE, MAE ve RMSE metrikleri ile karşılaştırılmıştır. Eğitim seti %70 olarak belirlenmiş ve model parametre değerlerinin kalibre edilmesi amacıyla kullanılmıştır. Test veri seti ise %30 olarak ayarlanmış ve modelin doğrulanması ve tahmin doğruluğunun test edilmesi için kullanılmıştır. Çalışma sonucunda BILSTM modelinin yolcu talebinin tahmininde %90'ın üzerinde doğruluğa ulaşarak diğer modellere kıyasla daha doğru sonuçlara ulaştığını göstermişlerdir [Liyanage vd., 2022].

Literatürdeki yolcu talebi tahminine yönelik çalışmalar **Tablo1.1**'de özet halde sunulmuştur. Çalışmalar incelendiğinde yolcuların bir istasyon veya duraktan ilgili ulaşım moduyla seyahat kararlarının haftanın günü, hava durumu ve buna benzer çeşitli zamansal etkenlerden etkilendiği gösterilmiştir. Bu kapsamda yolcu talebindeki dalgalanmaların tahminine yönelik çok sayıda çalışma bulunmaktadır. Ancak ölçeklenebilir bir yolcu talep tahmin modeli için yolcu talebindeki dalgalanmalar

kadar istasyon özellikleri ve çevresindeki mekânsal faktörlerle ilişkili temel yolcu talebinin de istasyon bazlı olarak tahmin edilebilmesi gerekmektedir. Bu doğrultuda istasyonun özelliklerine yönelik faktörler değerlendirilmiş olsa da istasyonun çevresiyle ve özellikle hizmet alanıyla ilişkili kentsel ilgi noktaları ve demografik yapı gibi mekânsal faktörlerin değerlendirilmediği görülmüştür. Bu durum, mevcut yolcu talep tahmin modellerinin uygulanabilirliği yalnızca modelin eğitilmiş olduğu istasyonlarla sınırlı kalmasına sebebiyet vermektedir. Ayrıca, son çalışmalarda karmaşık veri setlerini açıklamadaki başarısıyla sıklıkla tercih sebebi olmaya başlayan Rastgele Orman gibi makine öğrenmesi yöntemlerinin kullanımında artış yaşanmıştır. Ancak bu yöntemlerin hem zamansal hem de mekânsal faktörlerden oluşan veri setlerindeki performansları yeterince test edilmemiştir.

Tablo 1.1: Literatürdeki çalışmaların sınırlılıkları.

Yazarlar ve Yayın Adı	Yöntem	Veri Seti	Eksiklikler
Milenkovic vd., 2016, SARIMA Modelling Approach for Railway Passenger Flow Forecasting	SARIMA	Zamansal	Mekânsal faktörlerle ilişkili olarak istasyonlar arası farklılıklar göz ardı edilmiştir.
Toque vd., 2017, Short Long Term Forecasting of Multimodal Transport Passenger Flows with Machine Learning Methods	Rastgele Orman LSTM Sinir Ağları	Zamansal	Her istasyon için ayrı model kurulmuş olsa da bu istasyonların coğrafi özellikleri modele entegre edilmemiştir.
Cyril vd., 2018, Modelling and Forecasting Bus Passenger Demand using Time Series Method	ARIMA	Zamansal	Yalnızca geçmiş yolcu verisi kullanılmış, dışsal faktörler modele entegre edilmemiştir.
Liu vd., 2020, Impacts of Weather on Short-Term Metro Passenger Flow Forecasting Using a Deep LSTM Neural Network	LSTMN	Zamansal	Model sadece tek bir istasyon için geliştirildiğinden genellenebilirlik sınırlı kalmıştır.

Tablo 2.1 (devam): Literatürdeki çalışmaların sınırlılıkları.

Zhou ve Tang, 2020, Forecast of Urban Rail Transit Passenger Flow in Holidays Based on Support Vector Machine Model	DVM	Zamansal	Model yalnızca tatil günlerine özel geliştirildiğinden zamansal sınırlılıklar içermektedir.
Li vd., 2020, Forecasting Bus Passenger Flows by Using a Clustering-Based Support Vector Regression Approach	DVR	Zamansal	Mekânsal faktörlere yer verilmemiştir.
Baro ve Khoudjai, 2021, Passenger Flow Forecasting on Transportation Network: Sensitivity Analysis of The Spatiotemporal	Rastgele Orman	Zamansal Mekânsal	Mekânsal faktörler yalnızca istasyonun kendisi ve ilgili güzergâh bilgisi gibi çok kısıtlı biçimde ele alınmıştır.
Jie vd., 2021, Forecasting Daily MRT Passenger Flow in Taipei Based on Google Search Queries	Doğrusal Regresyon LASSO Rastgele Orman ARIMA	Zamansal	Modelde mekânsal yayılım ve istasyonlar arası farklılıklar detaylı işlenmemiştir.
Kusonkhum vd., 2022, Using a Machine Learning Approach to Predict the Thailand Underground Train's Passenger	Yapay Sinir Ağları Rastgele Orman Karar Ağaçları	Zamansal	İstasyonlar arası mekânsal ilişkiler değerlendirilmemiştir.
Nar ve Arslankaya, 2022, Passenger demand forecasting for railway systems	Yapay Sinir Ağları Doğrusal Regresyon Rastgele Orman Polinomsal Regresyon DVR	Zamansal	İstasyon tahminlerinde karşılaştırma yapılmış ancak mekânsal ilişkiler incelenmemiştir.
Liyanage vd., 2022, AI-Based Neural Network Models for Bus Passenger Demand Forecasting Using Smart Card Data	LSTM BILSTM	Zamansal	Model kısa bir zamansal veriye dayandığından mevsimsel analizleri sınırlamıştır.

2. YENİ NESİL HAREKETLİLİK

Günümüzde her geçen gün artan nüfus ve yüksek kentleşme hızıyla birlikte şehirlerdeki yolculuk talepleri artış göstermektedir. Büyüyen şehirlerde yaşayan nüfus, yolculuk talebini karşılamak amacıyla bireysel araç sahipliği oranını artırmaktadır. Şehirlerde artan bireysel araç sahipliği şehirlere ağır sosyal ve çevresel baskılar yüklemiş ve şehirlerde sürdürülebilirliğin gerçekleştirilmesini zorlaştırmıştır [Li vd., 2019]. Şehirlerin hem çevresel hem de ekonomik açıdan belirli bir farkındalığa ulaşmasının ardından bir başlangıç noktasından bir varış noktasına olan seyahati tanımlayan hareketlilik kavramı şekillenmektedir. Bilgi ve İletişim Teknolojileri alanında yaşanan gelişmeler neticesinde hareketlilik kavramı değişmekte ve gelişmektedir. Böylece Nesnelerin İnterneti (IoT), kablosuz iletişim teknolojileri, yapay zekâ ve yeni nesil teknolojilerin kullanımıyla ulaşım teknolojilerinin akıllı hale getirilmesi ile değişen hareketlilik kavramı yeni nesil hareketlilik olarak ifade edilebilmektedir.

Sensörler, gömülü sistemler ve mobil cihaz gibi teknolojilerin yerel bir ağa veya internete bağlı olması IoT çağını başlatmıştır. IoT kavramının temeli, çevrede bulunan çeşitli nesne ve objelerin, Radyon Frekansı ile Tanımlama (RFID), sensörler ve mobil cihazlar gibi kablosuz iletişim teknolojilerinin birbirleri ile etkileşime geçmesine dayanmaktadır [Atzori vd., 2010]. Etkileşim, Wi-Fi ve Bluetooth gibi teknolojiler ile kısa menzilli veya Uzun Mesafe (Long Range-LoRa) teknolojisi ve 5G gibi ağlar aracılığı ile geniş menzilli olabilmektedir [Zantalis vd., 2019]. Bu etkileşim, küresel dinamikleri önemli şekilde değiştirme potansiyeline sahiptir [Rastogi vd., 2023].

Günümüzde sayısı gün geçtikçe artan IoT sensörlerinin gerçek zamanlı olarak veri toplaması ve veri paylaşması yoluyla büyük veri yapılarını meydana getirmesi, çevremizi algılama ve çevreyle kurmuş olduğumuz etkileşim biçimini şekillendirmektedir [Rastogi vd., 2023]. IoT ile toplanan veriler her zaman için kaliteli olmamakla birlikte verilerin analizi, anlamlı bilgilerin çıkarılması ve akıllı uygulamaların gerçekleştirilmesi gerekmektedir. Ayrıca büyük verilerde bu işlemlerin gerçekleştirilmesi geleneksel veri tabanları oldukça zor bir hale gelmiştir. Bu yüzden yapılandırılmış ve yapılandırılmamış büyük verilerin işlenmesi ve analizi gibi işlemler için özel tekniklere ihtiyaç duyulmaktadır. Bu kapsamda büyük veri yapıları içerisinde anlamlı bilgilerin elde edilmesi yapay zekânın alt dalı olan makine

öğrenmesi algoritmaları ile mümkün olmaktadır [Rastogi vd., 2023], [Zantalis vd., 2019]. Böylece, IoT sensör ağlarının ve makine öğrenimi gibi akıllı algoritmaların kullanımı büyük çevresel verinin toplanması, birleştirilmesi ve değerlendirilmesini sağlamaktadır. Bu açıdan özellikle topluluk öğrenmesi yöntemiyle çalışan makine öğrenmesi algoritmaları büyük verilerle başa çıkmada önemli katkılar sunmaktadır [Zantalis vd., 2019]. IoT, otomasyon, verimlilik, üretkenlik ve bağlantılılığı desteklerken makine öğrenmesi, anlamlı desenlerin çıkarılması ve tahminlerin oluşturulmasını desteklemektedir. Bu entegrasyon, kentlerde akıllı uygulamaları mümkün kılarak akıllı şehir kavramını oluşturmaktadır [Rastogi vd., 2023].

Şehirlerde yoğunlaşan nüfus, kentsel ortamlardaki hareketlilik kavramını akıllı şehir bağlamında en önemli konulardan birisi haline getirmektedir [Faria vd., 2017]. Ayrıca gün geçtikçe otomobiller, trenler, otobüsler ve bisikletler ile yol ve yayalar giderek daha fazla sensör teknolojisine sahip olması hareketlilik kavramını geliştirmektedir. Yeni nesil hareketlilik kavramı akıllı hareketlilik ile de ilişkilendirilebilmektedir. Elektrikli araç, otonom araç, mikro hareketlilik ve paylaşımlı hareketlilik kavramları hareketliliğin yeni nesil teknolojilerle geliştirilmesini ifade etmektedir (Şekil 2.1) [Alam vd., 2024].



Şekil 2.1: Yeni nesil hareketlilik.

Elektrikli araçlar, bataryalı elektrikli araçlar ve hibrit elektrikli araçlar olarak ikiye ayrılabilir. Bataryalı elektrikli araçlar, bataryada depolamış oldukları elektriğin elektrik motorlarını çalıştırması yoluyla hareket etmektedirler. Hibrit elektrikli araçlar, içten yanmalı motorlu araçların kullanmış oldukları yakıt tiplerini ve elektrik enerjisini beraber kullanmaktadırlar [Chan vd., 2010]. Otonom araçlar, sürüş esnasında güvenlik açısından kritik olan bazı işlevlerin yerine getirilmesinde sürücüye destek olan ve teknolojisine bağlı olarak insansız bir şekilde tüm sürüşü gerçekleştirebilen araçlardır [Kidmose, 2025]. Elektrikli araçlar ve otonom araçlar, şehirlerdeki ulaşımı daha çevreci ve güvenli yapma potansiyeline sahiptirler. Ancak artan bireysel araç sahipliğinin sebep olduğu trafik sorunu karşısında etkili bir çözüm sunmamaktadır. Bu kapsamda şehirlerdeki özellikle kısa mesafeler için trafik sorununu çözme potansiyeline sahip mikro hareketlilik çözümleri ön plana çıkmaktadır. Mikro hareketlilik, şehirlerde ilk ve son kilometre yolculuklarını şekillendiren bisiklet ve skuter gibi bir dizi hafif ve küçük araçları ifade etmektedir [Abduljabbar vd., 2021]. Ancak mikro hareketlilik araçlarının bireysel kullanılması kaynak kullanımı açısından çok fazla verimli ve esnek olmamaktadır. Ayrıca kalabalık şehirlerde kitleler haline ulaşımın sağlanması trafik sorunlarının çözümünde büyük öneme sahiptir. Bu yüzden birden fazla kullanıcının belirli araçları paylaşımlı olarak kullanması ön plana çıkmaktadır. Paylaşımlı hareketlilik olarak ifade edilen bu durum, birden fazla kullanıcıya esnek ve kullanışlı ulaşım imkânı sağlarken farklı ulaşım modlarını birleştiren hibrit bir yapı da sağlamaktadır. [Martinez vd., 2024], [Alam vd., 2024]. Burada paylaşılan bir filo olabileceği gibi seyahatin kendisi de paylaşılabilmektedir. Bir filo olması durumunda hizmet bisiklet veya skuter gibi aracın kendisi olur ve yolcu bu aracı kullanarak yolculuğunu tamamlar. Seyahatin kendisinin paylaşılması ise paylaşımlı yolculuk gibi bir hizmete bir yolcu olarak erişildiği durumları ifade etmektedir [Martinez vd., 2024]. Paylaşımlı hareketlilik kapsamında bulunan ve halka açık biçimde kamu yararı gözetilerek sağlanan toplu taşıma sistemleri yüksek yolcu kapasiteleri ile ön plana çıkmaktadır.

Birçok şehir, trafik kaynaklı kirliliğin ve sıklıkla artma tehdidiyle karşı karşıya kaldığında, bireysel araçlardan toplu taşıma araçlarına geçişi teşvik eden politika ve stratejiler benimsemiştir. Büyüyen şehirlerde kişisel araçlarla karşılanması güç olan erişilebilirlik oranları toplu taşıma araçları ile daha kolay giderilebilmektedir. Toplu taşıma araçları, sahip oldukları yüksek kapasitelerini kullanarak insanları daha

verimli ve düşük maliyetle taşıma özelliklerine sahip olmaktadırlar. Toplu taşıma sistemleri, otobüs, metrobüs, tramvay, tren, banliyö tren, metro ve vapur gibi şehir sakinlerinin ortak biçimde kullandığı araçları içermektedir. Toplu taşıma araçlarının çok büyük bir kısmı kara tabanlı araçlardan oluşmaktadır. Yüksek trafik sıklığı sebebiyle büyük şehirlerde toplu taşıma verimliliğinin artırılması noktasında araçların ayrılmış yollara sahip olması oldukça önemli hale gelmiştir. Bu sebeple raylı sistemler, kara tabanlı sistemlere kıyasla daha çevreci olmaları ve ayrılmış yola sahip olmaları gibi özellikleri ile şehirlerde toplu taşımada tercih sebebi olmaktadır. Metro ve banliyö trenleri raylı sistemi kullanan toplu taşıma araçlarına örnek olarak verilebilmektedir. Metrolar tamamen ayrılmış ve hemzemin olmayan hatlar kullanarak yüksek hız ve sefer sıklığı gibi hizmetler vermektedirler. Banliyö trenleri ise çoğunlukla hemzemin düzeyde ve genellikle geniş bir demiryolu ağının bir parçası olmakla birlikte kendilerine ayrılmış yol ağını kullanmaktadırlar. Bu toplu taşıma araçları genellikle banliyöleri şehir merkezlerine bağlama hizmeti veren araçlar olarak anılmakta ve sahip oldukları özellikleri bakımından hızlı toplu taşıma sistemleri olarak ifade edilmektedir. Bu sistemlerin sürdürülebilir olması, akıllı kart gibi teknolojilerin kullanımını gerektirmektedir **[Fouracre vd., 2003]**.

Akıllı kart teknolojisi ile toplu taşımada, ödeme kolaylığı bakımından zamandan tasarruf sağlanmakta ve yolcular daha kolay seyahat etmektedirler. Ayrıca farklı ulaşım modlarında da kullanımı sayesinde esneklik sağlanmaktadır **[Welde, 2012]**. Akıllı kart ile yolcuların kullanmış oldukları, toplu taşıma araçları, güzergahları ve seyahat zamanları IoT destekli olarak takip edilebilmekte ve depolanabilmektedir **[Vakula ve Raviteja, 2017]**, **[Zantalis vd., 2019]**. Bu kapsamda akıllı kart verilerinin makine öğrenmesi metotlarıyla anlamlı hale getirilmesi ile toplu taşıma sistemlerine yönelik tahmin ve projeksiyonlar elde edilebilmekte ve güzergahların planlanması ve geliştirilmesi mümkün olmaktadır **[Rastogi vd., 2023]**, **[Vakula ve Raviteja, 2017]**.

3. YOLCU TALEP TAHMİNİNDE MEKANSAL ZEKA

Günümüzde kişisel bilgisayarların ve kablosuz iletişim araçlarının yaygın kullanımıyla çok fazla veri üreticisi ortaya çıkmıştır. Bu yüzden her gün çeşitli sektörlerde yapılan işlemlere ait tarih ve kişisel veriler hakkında verilerin tutulması ile çok büyük miktarda veriler meydana gelmektedir. Bu sebeplerle günümüz büyük veri çağı olarak ifade edilmektedir [Alpaydın, 2014]. Bu büyük veriler, çok zengin coğrafi referanslı bilgiler taşımaktadır. Küresel Konumlama Sistemi (Global Positioning System-GPS) izleri, uzaktan algılama görüntüleri, konuma dayalı sosyal medya verileri, binalar, yollar, arazi kullanımı ve nüfus dağılımı gibi verilerin her biri coğrafi referanslı verilere örnek olarak verilebilmektedir [Choi, 2023].

Coğrafi referanslı verilerin toplanması, analizi ve görselleştirilmesi konusunda CBS kullanılmaktadır. CBS'nin kullanımı ile karar vericilerin mekânsal desenler ve ilişkileri değerlendirerek daha doğru kararlar almasına imkân sağlanmaktadır. Artan veri miktarı ve gelişen bilgisayar teknolojileri, CBS'nin şehir planlaması, çevre yönetimi ve ulaşım gibi çeşitli alanlardaki önemini giderek artmaktadır. Ancak, mekânsal verilerin miktarı katlanarak artmaya devam ettikçe günümüzdeki CBS teknikleri büyük veri kümelerini işleme ve analiz etme kısmında çeşitli zorluklarla karşılaşmaktadır [Choi, 2023].

Günümüzde gelişmekte olan Yapay Zekâ alanı, mekânsal araştırmalar için yeni fırsatlar meydana getirmektedir [Gao, 2021]. Yapay zekâ, genellikle insan zekâsı gerektiren öğrenme, problem çözme ve karar verme gibi görevleri yerine getirebilen akıllı makinelerin geliştirilmesini kapsamaktadır [Choi, 2023]. Yapay zekâ tekniklerinin mekânsal anlamda kullanımı yeni bir konu değildir. 1980'li yıllarda Smith (1984), yapay zekânın mekânsal problemlerin çözümündeki potansiyel rolünü aktarmıştır. Yapay zekâ teknikleri, birey ve grupların mekânsal karar alma davranışlarının modellenmesi, mekânsal veri analizi ve mekânsal karar verme faaliyetleri açısından son derece uygulanabilir araçlar ve ayrıntılı modeller geliştirmek açısından elverişli kabul edilmektedir. Geliştirilen bu modeller ile geleneksel mekânsal karar alma yaklaşımlarıyla yakalanması güç olgulara odaklanılmasına imkân tanınmaktadır [Smith, 1984].

Böylece CBS ve yapay zekâ tekniklerinin coğrafi referanslı veriler üzerinde uygulanması sonucunda mekânsal zekâ kavrama ortaya çıkmaktadır. Bu kavram, mekânsal akıl yürütme ve coğrafi olgular ile dinamikleri hakkında keşif süreçlerini taklit eden akıllı bilgisayar programları geliştirmeyi, insan ve çevre etkileşimleri ile ilgili problemleri mekânsal bağlamlar çerçevesinde çözmeyi amaçlayan bir çalışma alanı olarak değerlendirilmektedir [Gao, 2021]. Yapay zekânın alt dalı olan makine öğrenmesi ve mekânsal araçların kullanımı sonucunda verilerden desenler elde edilebilmekte ve tahminler oluşturulabilmektedir. Makine öğrenmesinin bir alt dalı olan derin öğrenme ile görüntü, video, nokta bulutu ve metin gibi verilerden bilgi çıkarımı otomatize edilmektedir.

Gün geçtikçe daha fazla sayıda sensör, mobil cihazlar ve IoT aygıtları aracılığıyla dijital izler oluşturulmaktadır. Kullanıcıların dijital yaşamlarının fiziksel etkinlikleri ile etkileşime geçmeleri durumunda ortaya çıkan dijital izler, insan dinamiklerinin daha iyi anlaşılması için yapay zekâ teknikleri tarafından kullanılmaktadır. Dijital izler, mekânsal ve zamansal özellikleri içeren çok boyutlu verilerdir. Bu verilerin kullanımı ile kent planlama ve hareketlilik yapılarının incelenmesi ve çeşitli çözümlerin geliştirilmesinde başarıyla uygulanmaktadır. Çeşitli aygıtlardan elde edilen bu geçmiş verilerin CBS ve yapay zekâ tekniklerinin kullanımı ile ileriye dönük güçlü tahmin modelleri kurulabilmektedir [Torrepadula vd., 2024].

Toplu taşıma sistemlerinde akıllı kart verilerinin kullanılmaya başlanması ile seyahatler ve yolcular hakkında yüksek kaliteli veriler üretilmektedir [Bagchia ve White, 2005]. Toplu taşıma sistemlerinde yolcu talebi tahmin problemi, şehir sakinlerinin ihtiyaçları doğrultusunda dinamik biçimde değişmekte olduğundan bazı önemli zorlukları barındırmaktadır. Bu zorlukların üstesinden gelinmesinde yoğun olarak yapay zekâ teknikleri kümesinde yer alan makine öğrenmesi yöntemleri kullanılmaktadır [Torrepadula vd., 2024]. Makine öğrenmesi, geçmiş verilerden öğrenerek tahmin modelleri oluşturmada sıklıkla kullanılmaktadır. Makine öğrenmesi algoritmaları ile oluşturulan tahmin modelleri toplu ulaşım alanında talepler ve talebi etkileyen faktörler arasındaki ilişkiyi açıklamak için güçlü araçlar olarak ortaya çıkmaktadır [Nar ve Arslan, 2022]. Mekânsal karar verme ve kentsel planlamada önemli olan bir rol oynayan yolcu talebine yönelik güçlü bir tahmin modeli, karmaşık modelleme teknikleri ve büyük verinin kullanımı ile kurulmaktadır.

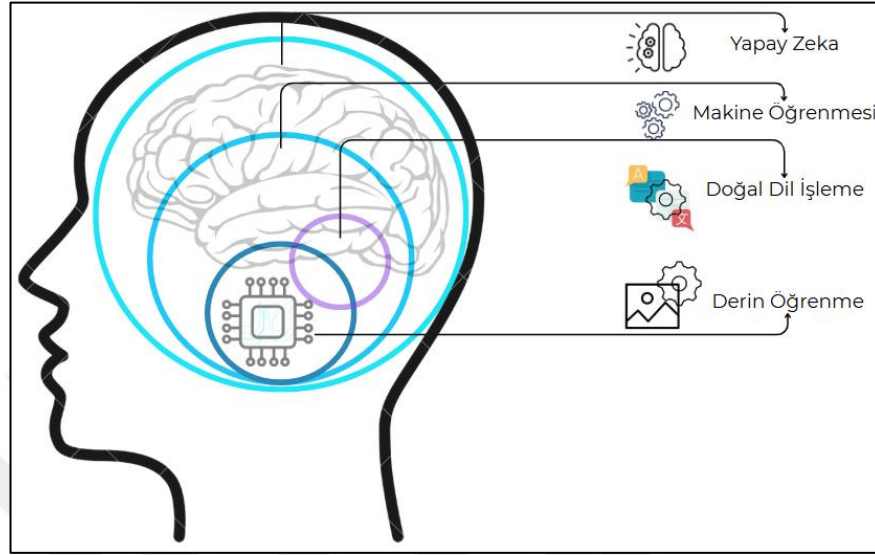
3.1. Yapay Zekâ

Her geçen gün artan veri miktarı, verilerden anlamlı bilginin çıkarılması süreçlerini güçleştirmektedir. İnsan davranışları, alışkanlıkları ve geri bildirimleri hakkında gerçek zamanlı üretilen veriler veri tabanlarında toplanarak büyük veri kümelerini oluşturmaktadır [Kim ve Shin, 2016]. Büyük veri kümelerinden kalıpların keşfedilmesi süreci veri madenciliği olarak ifade edilmektedir. Veri madenciliği, ekonomik avantajların sağlanması, tahminleme yapılması ve karar destek sistemlerinin güçlendirilmesi gibi süreçler için kritik hale gelmiştir. İstatistik, tahmin ve makine öğrenmesi gibi çeşitli disiplinlerle ilişkilendirilmekte olan veri madenciliği, karmaşık veri yapılarını anlamlandırma yoluyla yeni ve faydalı bilgilerin ortaya çıkarılmasını amaçlamaktadır. Tüm bu işlemlerin gerçekleştirilebilmesi kapsamında verilerin incelenmesi ve anlamlı örüntülerin tespit edilebilmesi için bilgisayar ve algoritmalara ihtiyaç duyulmaktadır [Alpaydın, 2014], [Witten vd., 2011].

Büyük verilerden anlamlı bilgilerin çıkarılmasında ihtiyaç duyulan bilgisayar ve algoritmaların geliştirilmesi doğrultusunda yapay zekâ alanı son yıllarda yaygın biçimde kullanılmaktadır. Ancak yapay zekâ araştırmaları 1950'li yıllara kadar dayanan ve ilk olarak sembolik akıl yürütmeyi içeren problemleri çözebilecek bilgisayar programları geliştirme girişimleriyle başlamıştır. Günümüzde yapay zekâ, öğrenme, problem çözme yeteneği ve karar verme gibi mantık ve algılama yeteneği gerektiren karmaşık görevleri zeki varlıkları taklit ederek gerçekleştirebilen bilgisayarları veya algoritmaları ifade etmektedir. Yapay zekâ araştırmaları bilişsel ve mühendislik yaklaşımları olarak ikiye ayrılabilir. Bilişsel (Cognitive) yaklaşım, insan bilişsel sistemini bilgisayar olarak gören bir benzetmeye dayanmakta ve doğal dil işleme ve görüntü işleme bu yaklaşım içerisinde yer almaktadır. Mühendislik yaklaşımı, bilişsel süreçleri taklit edebilen ve yorumlama, teşhis, izleme, tahmin, planlama ve tasarım gibi işlevleri gerçekleştirebilen makineler geliştirmektir [Smith, 1984].

Yapay zekâ alanı, makine öğrenmesi, derin öğrenme ve doğal dil işleme gibi alt dallara sahiptir (Şekil 3.1). Makine öğrenmesi, makinelerin verileri incelemesini, verilerden öğrenmesini, anlamlı örüntülerin ortaya çıkarılmasını ve anlamlı bilgilerin edinmesini sağlayan teknikleri ifade etmektedir [Alpaydın, 2014]. Makine öğrenmesinin büyük veri kalıplarındaki bilgilerin keşfedilmesindeki yeteneği ile başarılı bir veri

madenciliği süreci gerçekleştirilebilmektedir. Makine öğrenmesi ile gerçekleştirilecek çalışmalarda, problemin ve verinin yapısına uygun olarak denetimli, denetimsiz, yarı-denetimli ve pekiştirmeli öğrenme şeklinde dörde ayrılabilir [Sarker, 2021].



Şekil 3.1: Yapay zekâ ve alt kümeleri.

Denetimli öğrenmede veri setindeki bir girdi için hangi çıktı değeri karşılık gelmekte bilinmektedir. Daha açık bir ifade ile belirli girdilerden elde edilmesi gereken çıktıların tanımlandığı yani etiketli verilerin bulunduğu durumlarda kullanılmaktadır. Buradaki amaç eğitim kümesindeki girdi ve çıktı çiftlerinin kullanımıyla çıktı değerine en yakın sonuçları veren modelin kurulmasıdır. Bu yaklaşımla genellikle tahmin çalışmaları yürütülmektedir. Denetimsiz öğrenmede ise, eğitim kümesindeki girdilere karşılık bir çıktı değeri bulunmamaktadır. Yani, etiketlenmemiş veri kümeleri analiz edilmektedir. Bu yaklaşımda genellikle eğitim veri kümesi uygun alt kümelere ayrılmaktadır [Nilsson, 1998]. Yarı denetimli öğrenme, denetimli ve denetimsiz öğrenmenin bir kesişimini ifade etmektedir. Bu yaklaşımda hem etiketli hem de etiketsiz veriler üzerinde çalışılmaktadır. Veri setindeki verilerin bir kısmının etiketli ve bir kısmının da etiketsiz olması durumunda tercih edilmektedir. Dördüncü yaklaşım pekiştirmeli öğrenme ise, yazılım ajanlarının ve makinelerin belirli bir çevrede en uygun davranışı otomatik olarak değerlendirmesine olanak tanıyan bir makine öğrenmesi algoritması türüdür. Bu öğrenme türü, ödül veya ceza temelinde çalışmaktadır. Çalışma prensibi, çevresiyle etkileşim kuran bir ajanın belirlenmiş hedef doğrultusunda belirli görevleri yerine getirmesidir. Gerçekleştirilen görevlere bağlı olarak ajana ödül tanımlanmakta

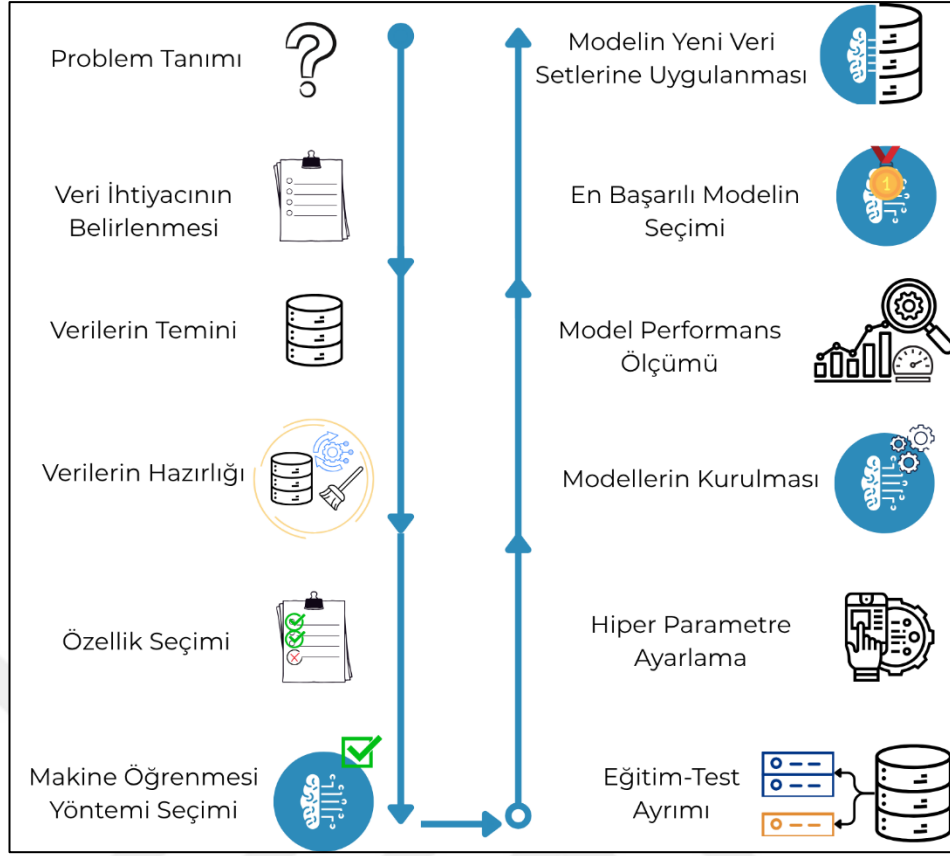
ve çevreden toplanan bilgiler doğrultusunda ödülü artırmak veya riski en aza indirmek amaçlanmaktadır [Yatawatta, 2024], [Sarker, 2021].

Derin öğrenme, büyük ve karmaşık veri kümelerini işlemek için sinir ağı tabanlı yaklaşımları kullanmaktadır ve makine öğrenmesi alanının bir alt alanıdır [Choi, 2023]. Yapay sinir ağları, çok sayıda bireysel hesaplama düğümünden oluşmaktadır. Bazı durumlarda modeller yalnızca ikili yanıt üretmekte iken bazı durumlarda ise daha karmaşık çıktılar elde etmek mümkündür [Titterington, 2009]. Sinir ağlarında giriş ve çıkış katmanları arasında gizli katmanlar bulunmaktadır. Sinir ağları ve derin öğrenme arasındaki temel fark bu gizli katmanların sayısıdır. Sinir ağlarında bir veya iki adet gizli katman olabilirken, derin öğrenmede çok sayıda gizli katman bulunabilmektedir. Derin öğrenme, çok çeşitli görevleri yerine getirebilmek amacıyla birçok uygulama alanında kullanılmaktadır. Görüntü tanıma, nesne tespiti, görüntü tamamlama, metin sınıflandırma, konuşma tanıma ve konuşma üretme alanlarında derin öğrenme oldukça popülerdir.

Makine öğrenmesi alanının bir alt dalı olan doğal dil işleme, bilgisayarların metin şeklindeki verileri ve insan dilini anlamasına, değerlendirmesine ve anlam çıkarmasına yardımcı olması amacıyla geliştirilmiş bir yöntemdir. Doğal dil işleme, genel olarak kural tabanlı veya makine öğrenmesi tabanlı tekniklerle gerçekleştirilebilmektedir. Makine öğrenmesi tabanlı tekniklerde, sınıflandırma veya regresyon amaçları için eğitimli veri setlerinden öğrenme modelleri oluşturmayı içermektedir. Kural tabanlı teknikleri hesaplamalı model eğitimi yoluyla değil, belirli anlamsal kuralları içeren yöntemleri ifade etmektedir [Khalef vd., 2021].

3.2. Makine Öğrenmesi Süreci

Günümüzde artan veri miktarı, gelişen bilgisayar teknolojileri ve yapay zekâ teknikleri sayesinde insanlar tarafından çözülmesi çok güçleşen problemlerin makineler tarafında çözülmesi mümkün hale gelmiştir. Büyük veri kalıplarından öğrenerek ve algoritmalar yardımıyla anlamlı çıktılar elde edilmesi olarak ifade edilen makine öğrenmesi süreci, belirli bir takım iş adımlarına sahiptir (Şekil 3.2).



Şekil 3.2: Makine öğrenmesi süreci.

Problemin tanımlanması makine öğrenmesinin ilk adımını oluşturmaktadır. Anlaşılır ve doğru biçimde yapılmamış bir problem tanımı diğer süreçleri de olumsuz yönde etkileyecektir. Problem tanımının net yapılması, ihtiyaç duyulacak verilerin probleme uygun belirlenmesi aşamasına önemli bir katkı sağlayarak çıktıların doğruluğuna doğrudan etki etmektedir. İhtiyaç duyulan verilerin tespitinin ardından istenen verilerin nasıl toplandığının ve nerelerde depolandığının tespiti ve temini gerçekleştirilmektedir. Bu aşamada verinin kalitesi, miktarı ve veri formatı model performansını etkileme potansiyeline sahiptir. Temin edilen veri setinin yapısı incelenerek temel düzeyde fikir sahibi olunmakta ve verilerin hazırlığı sürecine geçilmektedir. Bu kısımda, veri setinin makine öğrenmesi tekniklerinde uygulanabilir hale getirilmesi yer almaktadır. Öncelikle, verilerin farklı kaynaklardan gelebileceği düşünüldüğünde verilerin birleştirilmesi gerekmektedir. Veri seti içerisindeki eksik, tekrarlı ve gereksiz verilerin tespiti ve ardından veri türlerinin dönüşümü gibi süreçler uygulanmaktadır. Veri hazırlığı sürecinin ardından makine öğrenmesi teknikleri ile oluşturulacak tahmin modellerinin başarısında önemli bir etkisi bulunan özellik seçimi süreci uygulanmaktadır. Özellik seçimi, veri seti içerisinde bulunan ve modellerde

girdi değişken olarak kabul edilecek verilerin ilgililiğinin ve fazlalığının analizine yönelik bir alt veri seti elde edilmesi sürecidir [Cai vd., 2018]. Özellik seçimi ile veri seti boyutu azaltılabilmekte, modelin doğruluğu ve genellenabilirliği artırılabilir. Problem çözümünün gerçekleştirilme biçiminin regresyon veya sınıflandırma olmasıyla ilişki olarak ve veri setinin yapısı da dikkate alınarak makine öğrenmesi algoritması belirlenmektedir. Seçilen algoritma ile tahmin modellerinin kurulmasından önce veri seti eğitim ve test veri seti olarak ikiye ayrılmakta ve böylece model, eğitim veri seti ile verilerdeki ilişkileri öğrenirken test veri seti ile doğruluğu analiz edilebilmektedir. Daha sonrasında model performansını doğrudan etkileme potansiyeline sahip makine öğrenmesi yöntemi hiper parametrelerinin veri setlerine uygun biçimde ayarlanması gerekmektedir. Bu aşamada eğitim verisi üzerinden farklı hiper parametre değerleri denenerek en yüksek doğruluğa ulaştıran hiper parametre kombinasyonu belirlenmektedir. Daha sonraki süreçte en iyi performansı gösteren hiper parametreler kullanılarak doğrulama teknikleri ile modeller çalıştırılmaktadır. Doğrulama teknikleri ile oluşturulan modeller ile aşırı uyum ve yetersiz uyum sorunlarının önüne geçilebilmektedir. Modelin eğitimi tamamlandıktan sonra modelin görmemiş olduğu test verileri üzerinden model performansı ve hesaplama maliyetleri değerlendirilmekte ve en yüksek doğruluğa sahip model belirlenmektedir. Son aşamada, eğitilmiş en ideal model yeni veriler üzerinde tahminler gerçekleştirmek için kullanılmaktadır. Bu süreçlerin başarılı şekilde gerçekleştirilmesi ile büyük miktardaki geçmiş veriler işlenerek tahmin problemleri çözülebilmekte ve karar destek sistemlerinin güçlendirilmesi sağlanmaktadır [Witten vd., 2011], [Alpaydın, 2014].

3.3. Özellik Seçimi

Makine Öğrenimi yöntemlerinin başarısını etkileyen birçok etken bulunmaktadır. Günümüz büyük veri çağında teorik olarak daha fazla özelliğin (faktör) değerlendirilmesi daha güçlü modeller oluşturmamızı sağlaması gerekirken pratikte bu durum değişebilmektedir. Bir veri kümesinde yer alan ilgisiz ve gereksiz özellikler aşırı öğrenme sorununa neden olarak modellerin performansını düşürebilmektedir. Bir model ilgisiz ve gereksiz özelliklere karşı dayanıklılık gösterebilse de bu özelliklerin varlığı depolama ve hesaplama maliyetlerini artırmakta ve yorumlanabilirliği sınırlandırmaktadır [Alsahaf vd., 2022]. Bu sebeple büyük veri kümesi içerisinde ilgili ve gerekli özelliklerin seçilmesi işlemini ifade eden özellik seçimi süreci

uygulanmaktadır. Bu çalışmada özellik seçimi ile kastedilen veri seti içerisindeki mekânsal ve zamansal faktörlerin seçimi şeklinde olmaktadır.

Özellik seçimi, makine öğrenmesinde model oluşturma aşamasında belirli bir özellik seçme yöntemine göre orijinal özellik kümesinden bir alt küme elde etme süreci olarak tanımlanan kritik bir süreçtir. Özellik seçimi sürecindeki temel amaç, en anlamlı özelliklerin seçimi yapılırken alakasız, gereksiz ve gürültülü olanların elenmesi ve böylece en anlamlı özellikler alt kümesi elde etmektir. Ayrıca elde edilen alt küme ile veri boyutu anlamlı şekilde küçültülmüş olmaktadır. Modellerin yorumlanabilirliğinin artırılması, öğrenme süresinin kısaltılması, doğruluğun iyileştirilmesi ve modelin genellebilirliğini artırarak aşırı uyum riskinin azaltılması gibi çeşitli nedenlerden dolayı makine öğrenmesi tekniklerinin uygulandığı çalışmalarda yaygın biçimde kullanılmaktadır [Ratnasingam ve Munoz-Lopez, 2023], [Cai vd., 2018]. Özellik seçimi yöntemleri öğrenme yöntemleriyle ilişkilerine göre Filtre (Filter), Sarmalayıcı (Wrapper) ve Gömülü (Embedded) modeller olarak sınıflandırılmaktadır [Cai vd., 2018].

Filtreleme tabanlı yöntemler, bir veri kümesindeki özellikler arasındaki korelasyon veya bağımlılığı ölçmek için istatistiksel ölçümler kullanmaktadır. Özelliklerin seçiminde çeşitli istatistiksel testler kullanılarak elde edilen katsayılar kullanılmaktadır [Algherairy vd., 2022]. Filtre yöntemleri, makine öğrenmesi algoritmasından bağımsız çalışmakta ve bu yüzden aşırı uyum riski daha azdır. Diğer yöntemlere göre bir diğer ve en büyük avantajlarından birisi çok daha az hesaplama gücü gerektirmesidir. Bu özelliği sayesinde çok büyük verilere kolayca ölçeklenebilmektedir. Mevcut Filtre yöntemleri, tek değişkenli (univariate) veya çok değişkenli (multivariate) olarak kategorilere ayrılmaktadır. Tek değişkenli yöntemler her özelliği ayrı ayrı test ederken, çok değişkenli yöntemler aynı anda bir özellik alt kümesini dikkate almaktadır. Hızlı ve kolay uygulanabilir olmaları sebebiyle χ^2 testi (Ki-kare testi), öklid mesafesi ve Pearson Korelasyon gibi tek değişkenli yöntemler yüksek boyutlu veri setlerinin kullanıldığı çalışmalarda oldukça yaygın biçimde uygulanmaktadır [Pudjihartono vd., 2022].

Sarmalayıcı yöntemler, tüm alt kümeleri değerlendirerek en uygun özellik alt kümesini seçmek için makine öğrenmesi algoritmasını kullanmaktadır [Algherairy vd., 2022]. Özellik kümelerinin makine öğrenmesi algoritmasının yaptığı tahmine dayalı değerlendirilmesi yapılmaktadır [Yu ve Liu, 2004]. Yöntemde alt kümelerin

değerlendirilmesinin hem avantajları hem de dezavantajları vardır. Avantajı, kullanıcının en optimum eşiği veya seçilen özellik sayısını belirlemesine gerek olmamasıdır. Dezavantajı ise veri seti içerisinde hangi özelliklerin nispeten daha önemli olduğunun belli olmamasıdır [Pudjihartono vd., 2022]. Ayrıca, eğer yüksek boyutlu bir veri seti mevcutsa yoğun hesaplama gerektirmektedir. Bu nedenle diğer yöntemlere kıyasla daha yavaştır [Algherairy vd., 2022]. Sarmalayıcı yöntemlerde, Ardışık ileri yönde Seçim (Sequential Forward Selection) ve Ardışık Geri Yönde Seçim (Sequential Backward Selection) gibi yöntemler öne çıkmaktadır [Taheri ve Tucci, 2024].

Gömülü yöntemler, özellik seçimi sürecini doğrudan makine öğrenmesi algoritmasının eğitimi sırasında gerçekleştirmektedir [Algherairy vd., 2022]. Böylece özellik alt kümesinin aranması ve modelin eğitimi tek bir adımda birleşmiş olmaktadır. Karar ağacı tabanlı algoritmalarda özellik önemi, Ortalama Safsızlık Azalması (Mean Decrease Impurity-MDI) gibi metriklerle göre sıralanmaktadır. Bu yöntemlerde özellikler arası etkileşimlere izin verilmektedir. Bu yöntemde gereksiz özellikler otomatik şekilde ortadan kaldırılmadıkları için daha öncesinde gereksiz özelliklerin filtrelenmesi gerekmektedir [Pudjihartono vd., 2022]. Gömülü yöntem olarak Rastgele Ormanlar ve Gradyan Artırma Ağaçları gibi karar ağacı tabanlı yöntemler ile yapılan özellik seçimi örnek verilebilir [Taheri ve Tucci, 2024].

Tez çalışması kapsamında Filtre yöntemlerinden Pearson Korelasyon ve Sarmalayıcı yöntemlerden Ardışık ileri yönde Seçim ve Ardışık Geri Yönde Seçim yöntemleri ele alınmıştır.

3.3.1. Pearson Korelasyon

Veri kümesi içerisindeki faktörlerin birbirleriyle ve hedef değişken arasındaki doğrusal ilişkiyi ölçmek için parametrik testi kullanan en popüler Filtre yöntemlerinden birisi Pearson Korelasyon'dur. Bu yöntem, normal dağılıma sahip sürekli verilerde çok iyi biçimde çalışmaktadır [Otchere vd., 2022]. Pearson Korelasyon yönteminde faktörler arasındaki doğrusal ilişkinin gücünü ve yönünü gösteren ρ katsayısı hesaplanmaktadır. Bir faktörde yaşanan değişimin başka bir faktörde yaşanan değişimle ne kadar ilişkili olduğunu ölçmektedir. Hedef değişkenle yüksek mutlak korelasyona sahip özelliklerin tahmin problemleri için uygun özellikler

olduğu kabul edilmektedir. Diğer taraftan faktörlerin birbirleri arasında yüksek korelasyona sahip olmaları durumunda tahmin edilmesi istenen hedef değişken üzerindeki etkileri dağıtmakta ve modeli olumsuz etkilemektedir [Aksangür vd., 2022]. Bu yüzden korelasyon katsayısı hem faktörler arasında hem de faktörler ile hedef değişken arasında incelenmelidir. Yöntem, doğrusal ilişkileri dikkate aldığı için faktörler arasında doğrusal olmayan ilişkileri yakalamada kötü olacağı unutulmamalıdır [Taheri ve Tucci, 2024].

Veri kümesinde bulunan X ve Y olarak ifade edilen iki faktör arasındaki Pearson Korelasyon katsayısı eşitlik (3.1)'teki gibi hesaplanmaktadır [Taheri ve Tucci, 2024].

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \quad (3.1)$$

Eşitlikte bulunan $\text{cov}(X,Y)$, X ve Y faktörleri arasındaki kovaryansı ifade etmektedir. σ_X ve σ_Y sırasıyla X ve Y faktörlerinin standart sapmalarını ifade etmektedir. Eşitlik (3.2)'de X ve Y faktörler arasındaki kovaryansın hesabı gösterilmiştir [Taheri ve Tucci, 2024].

$$\text{cov}(X,Y) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \quad (3.2)$$

Eşitlikte bulunan $\bar{X}$ ve $\bar{Y}$ sırasıyla X ve Y faktörlerinin ortalama değerlerini ifade etmektedir. Eşitlikte bulunan n , veri noktalarının sayısını ifade etmektedir. İki faktör arasındaki kovaryans değeri elde edildikten sonra bu faktörlerin standart sapmalarına oranlanması ile iki faktör arasındaki Pearson Korelasyon katsayısı $\rho_{X,Y}$, elde edilmektedir [Taheri ve Tucci, 2024]. Korelasyon katsayısı -1 ile +1 arasında bir değer almaktadır. Katsayının +1'e yaklaşması iki faktör arasında pozitif yönde güçlü bir doğrusal ilişki olduğu anlamı çıkmaktadır. Eğer katsayı tam olarak 1 ise iki faktör arasında pozitif yönde kusursuz bir doğrusal ilişki olduğu anlamı taşımaktadır. Katsayının -1'e yaklaşması iki faktör arasında negatif yönde güçlü bir doğrusal ilişki olduğu anlamı çıkmaktadır. Eğer katsayı tam olarak -1 ise iki faktör arasında negatif yönde kusursuz bir doğrusal ilişki olduğu anlamı taşımaktadır. Katsayının sıfır olması ise iki faktör arasında doğrusal bir ilişkinin olmadığı anlamına gelmektedir.

3.3.2. Ardışık İleri Yönde Seçim

Ardışık İleri Yönde Seçim yönteminde, algoritma öncelikle boş bir faktör kümesi ile başlar ve ardından tahmin doğruluğunu en iyi şekilde artıran faktör aranır. Tahmin doğruluğunu en iyi şekilde artıran faktör belirli bir ölçüt fonksiyonuna göre değerlendirilmektedir. Tahmin doğruluğunu en iyi artıran faktör başlangıçta boş olan faktör kümesine eklenmektedir. Daha sonrasında ilk adımda elde edilen faktör kümesi ile birlikte diğer faktörler arasından tahmin doğruluğunu en iyi şekilde artıran faktör seçilmektedir. Bu işlem kalan özelliklerin de benzer şekilde değerlendirilmesi ile belirli bir iyileşmenin sağlanmadığı noktaya devam etmektedir. Bu yöntemde alakasız faktörlerin veya gürültünün kaldırılması ile hesaplama verimliliği artırılmakta ve modelin aşırı uyum riski azaltılmaktadır [Aboudi ve Benhlime, 2016].

3.3.3. Ardışık Geri Yönde Seçim

Ardışık Geri Yönde Seçim yöntemi, Ardışık İleri Yönde Seçim yöntemine benzer şekilde ancak arama ters yönde gerçekleştirilmektedir. Algoritma öncelikle veri kümesinde bulunan bütün faktörleri içeren bir faktör kümesi ile başlar ve ardından tahmin doğruluğunu en az artıran faktörleri teker teker çıkartmaktadır. Bu işlem, belirli bir iyileşmenin sağlanmadığı noktaya kadar devam etmektedir. Ardışık İleri Yönde Seçim yönteminde olduğu gibi en iyi faktör alt kümesi elde etme garantisi sunmamaktadır. Ancak çözüm bulmaya yönelik hızlı bir yakınsama sağlamaktadır [Aboudi ve Benhlime, 2016].

3.4. Tahmin Teknikleri

Bir tahmin problemi, geçmiş veriler ile gelecekteki durumu tahmin etme sürecini ifade etmektedir. Bu problemin çözümünde büyük veri kümelerinden kalıpların keşfedilmesi ve anlamlı çıktıların elde edilmesi süreci izlenmektedir. Bu kapsamda, yapay zekâ alanında algoritma ve istatistiksel yöntemlerle tahmin modellerinin oluşturulmasını sağlayan makine öğrenmesi teknikleri kullanılmaktadır [Sison vd., 2021]. Tahmin problemlerinde, bir dizi girdi değişken ile çıktı değişkenden oluşan bir veri kümesi kullanılmaktadır. Problemin amacı, geçmiş verileri kullanarak girdi değişkenler ile çıktı değişkenler arasındaki ilişkiyi öğrenmek ve bu ilişkiyi en iyi şekilde temsil eden tahmin modelini geliştirmektir. Tahmin edilen değerler ile gerçek

değerler arasındaki fark, bir hata veya kayıp fonksiyonu ile ölçülmektedir **[Friedman, 2001]**.

Tahmin probleminde hedeflenen çıktının yapısına yönelik olarak sınıflandırma ve regresyon gibi tahmin yaklaşımları bulunmaktadır. Sınıflandırma, hedeflenen çıktının kategorik olması durumunda kullanılmaktadır. Veri setindeki girdi değişkenleri, çıktı değişkeninin hangi sınıfa ait olacağının tahmininde kullanılmaktadır. Sınıflandırma sonuçları, veri setinde az sayıda değişken bulunması durumunda, sayısal yorumlamaya kıyasla daha kolay anlaşılabilen grafiksel görselleştirmelerle ifade edilebilmektedir **[Jaiswal ve Samikannu, 2017]**. Sınıflandırma problemleri, ikili sınıflandırma, çoklu sınıf sınıflandırması ve çoklu etiket sınıflandırması şeklinde üçe ayrılmaktadır. İkili sınıflandırma, “doğru ve yanlış” veya “evet ve hayır” gibi iki sınıf etiketine sahip sınıflandırma görevlerini ifade etmektedir. İki sınıftan daha fazla sınıf etiketi problemlerinde ise çoklu sınıf sınıflandırma kullanılmaktadır. Bir örneğin birden fazla sınıf veya etiketle ilişkilendirildiği problemlerde ise çoklu etiket sınıflandırması kullanılmaktadır. Regresyon ise, çıktı değişkeninin sürekli veri olması durumunda kullanılmaktadır. Veri setindeki girdi değişkenleri, sürekli yapıdaki çıktı değişkenini tahmin etmeye çalışmaktadır. Sınıflandırma ile regresyon arasındaki en önemli fark, çıktı değişkeninin sınıflandırma için kategorik ve regresyon için sürekli yapıda olmasıdır **[Sarker, 2021]**.

Tahmin problemlerinin çözümünde sınıflandırma yaklaşımı için özelleşmiş K-En Yakın Komşuluk ve Naive Bayes gibi yöntemler ve regresyon yaklaşımları için özelleşmiş ÇDR ve ARIMA gibi yöntemler bulunmaktadır **[Jiang ve Xu, 2022]**, **[Nar ve Arslankaya, 2021]**. Hem regresyon hem de sınıflandırma için kullanılabilir yöntemlere, DVM, Karar Ağaçları ve Karar Ağacı tabanlı Rastgele Orman ve eXtreme Gradient Boosting (XGBoost) gibi teknikler örnek verilebilmektedir **[Gatera vd., 2023]**, **[Teimourian vd., 2024]**.

Regresyon yöntemine dayalı yolcu talep tahmini çalışmalarında istatistiksel, zaman serisi ve makine öğrenmesi tabanlı yaklaşımlar yaygın biçimde kullanılmaktadır. ÇDR gibi istatistiksel yöntemler yolcu talebini doğrusal ilişkilerle tahmin etmeye çalışmasından dolayı hesaplama maliyeti düşük yöntemlerden olmaktadır. ARIMA yöntemi, zaman serilerine dayalı bir yöntemdir ve veri dağılımının oldukça düzenli olduğu senaryolarda özellikle uzun vadeli tahminler için başarılı bir yöntemdir. ARIMA yönteminin mevsimsel versiyonu olan SARIMA ise, haftalık veya yıllık

dalgalanmalar gibi yolcu akışı verilerinde doğal olarak bulunan periyodik desenleri yakalayabilmektedir. ARIMA yönteminin uygulanabilirliği, sadece durağan, doğrusal ve tek değişkenli zaman serileri ile sınırlı olmaktadır. Yolcu talebinin çok sayıda faktörden etkilenebilmesi sebebiyle, yalnızca geçmiş verilere dayalı yapılan tahminler düzensiz dalgalanmaları açıklayamamaktadır. Bu nedenle düzensiz faktörlerin değerlendirilmesi gerekmektedir [Feng vd., 2021]. Bu tür kısıtlamalar sebebiyle ARIMA gibi yöntemlerin kullanımı gün geçtikçe azalmıştır [Torrepadula vd., 2024]. ARIMA ve SARIMA yöntemlerine kıyasla hesaplama açısından karmaşık olmalarına rağmen, makine öğrenimine dayalı modeller yapılandırılabilir olup, zaman serilerinin ötesinde farklı veri türlerini girdi olarak işleme yeteneğine sahiptirler [Sison vd., 2021]. Bu sebeple doğrusal olmayan verilerle başa çıkabilen yöntemlerin kullanımına yönelik çalışmalara yönelim yaşanmıştır. DVR veya DVM, doğrusal olmayan verilerde ve çok boyutlu değişkenlerin bulunduğu durumlar karşısında iyi başa çıkabilen basit yöntemlerdendir [Torrepadula vd., 2024]. Yöntemde girdiler, doğrusal bir fonksiyona dayalı biçimde yüksek boyutlu bir uzaya yansıtılarak, hedef değişkeni belirli bir tolerans sınırı içerisinde tahmin eden bir model oluşturmayı amaçlamaktadır [Bargam vd., 2024]. Bu yöntemde optimal bir hiper düzleme göre tahmin yapılmaktadır. Rastgele Orman ise tahmin edici olarak adlandırılan birçok ağacın paralel biçimde eğitildiği ve elde ettikleri tahmin sonuçlarının birleşimiyle nihai tahminin elde edildiği topluluk öğrenmesi yöntemidir. Çok sayıda tahmin edicinin kullanımıyla tek tahmin edici bulunan modellere kıyasla daha yüksek doğruluk elde etmeyi amaçlamaktadır [Breiman, 2001]. Bu yönleriyle karmaşık veri setlerinde sıklıkla tercih edilen yöntemlerdendir. Yapay Sinir Ağları, çok boyutlu ve doğrusal olmayan verilerle başa çıkma yetenekleri sayesinde yolcu talebinin tahminine yönelik çalışmalarda kullanılan yöntemlerdendir [Torrepadula vd., 2024]. Bu sebeplerden dolayı doğrusal olmayan algoritmaların kullanıldığı tahmin modelleri karmaşık veri setlerinin bulunduğu durumlarda zamansal ve mekânsal faktörlerin etkisiyle değişen yolcu taleplerini daha doğru bir şekilde açıklayabilecektir. Bu yöntemler arasında bulunan topluluk öğrenmesi türlerinden torbalama yaklaşımına dayanan Rastgele Orman algoritması, derin öğrenme yöntemlerine kıyasla daha düşük hesaplama maliyeti, yüksek yorumlanabilirliği ve daha karmaşık yaklaşımlara göre bile iyi performans göstermesi gibi sebeplerle sıklıkla tercih edilmektedir. Özellikle kullanılan veri kümesinin karmaşık mekânsal ve zamansal bağımlılıkları içermesi ve bu bağımlılıkların özellik mühendisliği yoluyla temsil edilebildiği uygulamalar için

oldukça uygundur [Baro ve Khouadjia, 2021]. Rastgele Orman algoritmasında bulunan ağaçlar, paralel bir biçimde eğitilmekte ve nihai tahmin bu tahmin sonuçlarının birleşimi yoluyla elde edilmektedir. Topluluk öğrenmesi yöntemleri içerisinde bulunan bir diğer algoritma XGBoost, Rastgele Orman algoritmasına benzer şekilde çalışmaktadır. Rastgele Orman algoritmasında her ağaç birbirinden bağımsız şekilde eğitilmekteyken XGBoost algoritmasında her bir ağaç bir önceki ağacın hatalarını dikkate alarak yeni tahminler gerçekleştirmektedir. Bu özelliği ile XGBoost, hataları minimize etmeye çalıştığı için genellikle Rastgele Orman algoritmasına kıyasla genelde daha doğru sonuçlar vermektedir. Buna karşın Rastgele Orman algoritmasında ayarlanması gereken hiper parametre sayısı, XGBoost algoritmasındaki hiper parametre sayısına göre oldukça düşüktür. Tüm bu özellikleri ile XGBoost daha doğru sonuçlar verebilirken Rastgele Orman daha düşük hesaplama maliyeti sağlamaktadır. Hem Rastgele Orman hem de XGBoost insanlar tarafından yüksek yorumlanabilirlik sağlayan karar ağaçlarına dayalı algoritmalar olduklarından dolayı veriler hakkında bilgi edinme ve çıktının kabul edilebilirliğinin doğrulanması noktasında avantaj sağlamaktadır [Dietterich, 2003].

3.4.1. Karar Ağaçları

Hem regresyon hem de sınıflandırma için kullanılabilen Karar Ağaçları, Ağaç benzeri dallanan bir yapıya sahip olmalarından dolayı Karar Ağaçları olarak adlandırılmaktadırlar. Karar Ağaçları, tek değişkenli ve çok değişkenli olarak sınıflandırılabilir. Tek değişkenli bir ağaçta, her iç düğümde test yalnızca girdi değişkenlerinden birisini kullanmaktadır. Çok değişkenli bir ağaçta ise, bir karar düğümünde tüm girdi değişkenleri kullanılabilir ve bu nedenle daha genel bir yapıya sahiptir [Alpaydın, 2014].

Bir karar ağacı düğümler ve yapraklardan oluşmaktadır. Kök düğüm en üst kısımda bulunurken yapraklar en alt kısımda bulunmaktadır [Dietterich, 2003]. Ağaç, öğrenme sürecinde verinin doğasındaki karmaşıklığa bağlı biçimde büyür, dallar ve yapraklar eklenmektedir. Bir girdi verildiğinde her düğümde bir test uygulanır ve sonuca bağlı olarak bir dallanma seçilmektedir. Bu süreçte, kök düğümden başlayarak tüm eğitim verisi ile en iyi bölünme noktası aranmaktadır. Bu bölünme özelliğın sayısal veya ayrık olması durumuna göre farklı sayıda olmaktadır. Daha sonrasında her bölünme için alt özellik kümeleri elde edilmektedir. Bu işlem özyinelemeli olarak

devam etmektedir. Daha fazla bölünmenin gerekmediği noktada yaprak düğüme ulaşılmaktadır. Bir bölmenin kalitesi safsızlık ölçüsüyle değerlendirilmektedir. Yaprak düğüme ulaşıldığında elde edilen değer çıktı olarak kabul edilmektedir. Yaprakta yazılı olan çıktı, sınıflandırma durumunda bir sınıf etiketi iken regresyon durumunda sayısal (sürekli) bir değer olmaktadır [Alpaydın, 2014].

Karar Ağaçlarına yönelik birçok algoritma önerilmiştir. Bu algoritmalarından birisi hem regresyon hem de sınıflandırma görevleri için çözüm sunan Sınıflandırma ve Regresyon Ağaçları (Classification and Regression Trees-CART) algoritmasıdır. Breiman ve ark. (1984) tarafından önerilen CART, ilk başta çok büyük bir ağaç oluşturmaktadır. Daha sonrasında ağacın tahmin performansını artırmak amacıyla tahmin performansına katkı sağlamayan dalları bir budama kriterine göre budamaktadır. Tüm bu süreç boyunca her düğüm için bölmeyi seçerken genel düğüm safsızlığını maksimum düzeyde azaltacak şekilde karar verilmektedir [Grömping, 2009]. Sınıflandırma ağacında, bir bölmenin (split) kalitesi safsızlık (impurity) ölçütü ile değerlendirilirken regresyon ağacında, ortalama kare farkı gibi uygun bir ölçüt ile değerlendirilmektedir. CART, parametrik olmayan bir yöntem olmasıyla ilişkili olarak girdi verilerin değerlerinin dağılımı hakkında daha öncesinden varsayımında bulunulmasına gerek bulunmamaktadır. Ayrıca oluşturulan modellerin kolay yorumlanabilir yapıları sıklıkla tercih edilme sebeplerindendir [Alpaydın, 2014].

Ancak Karar Ağaçları gibi tek öğrenici ile çalışmakta olan algoritmalar ile oluşturulan tahmin modellerinde genelleme doğruluğu açısından yüksek varyansa ve yüksek yanlılığa sahip olmaktadır. Yanlılık, öğrenme yaklaşımının ortalama tahminin hedeften ne kadar saptığının bir ölçütüyken varyans, aynı büyüklükte farklı eğitim veri setleri için öğrenme yaklaşımının tahminlerinin ne kadar değişkenlik gösterdiğini ifade etmektedir. Karar Ağaçlarının avantajlı yapısını kullanabilmek ve dezavantajlarını azaltmak için çeşitli yaklaşımlar ve teknikler geliştirilmiştir. Bu amaç doğrultusunda, birden fazla öğreniciyi birleştiren topluluk öğrenmesi metotları uygulanmaktadır. Böylece tahmin modellerinin performansını artırılabilir [Yao, 2009].

3.4.2. Topluluk Öğrenmesi

Topluluk öğrenimi, birden fazla öğreniciyi birleştiren ve genellikle denetimli makine öğrenmesi görevlerinde kullanılan bir terimdir. Topluluk öğrenmesi yöntemleriyle, tek

bir öğrenici üreten algoritmaların yaşamış oldukları istatistiksel, hesaplama ve temsil problemlerinin üstesinden gelinebilmektedir [Sagi ve Rokach, 2018], [Dietterich, 2002].

İstatistiksel problem, öğrenme algoritmasının mevcut eğitim verilerinin miktarı için çok büyük bir alanda uygun çözümü araması durumunda ortaya çıkmaktadır. Bu durumda eğitim verisinde aynı doğruluğu veren birkaç farklı çözüm bulunabilir ve öğrenme algoritmasının tahmin gerçekleştirmek için bunlardan birini seçmesi gerekmektedir. Seçilen modelin gelecekteki veri noktalarını iyi tahmin etmeme riski bulunmaktadır. Bu risk, birden fazla modelin tahminlerinin birleştirilmesi ile azaltılabilmektedir. Hesaplama problemi, öğrenme algoritmasının en iyi çözümü seçememe riskinde ortaya çıkmaktadır. Sinir ağları ve karar ağacı gibi algoritmalarda en iyi çözümün seçilmesi hesaplama açısından oldukça zor olduğundan dolayı sezgisel yöntemler uygulanmaktadır. Ancak yerel minimumlarda takılı kalma riskleri taşımaktadırlar. Bunun yerine birkaç farklı yerel minimumun ağırlıklı bir kombinasyonu tahmin için yanlış yerel minimum değeri seçme riskini azaltılabilmektedir. Temsil problemi, algoritmanın model havuzunun içerisinden ilişkiyi en iyi açıklayan çözümün bulunamadığı durumda ortaya çıkmaktadır. Bazı durumlarda modellerin ağırlıklı toplamı temsil edilebilecek çözüm çeşitliliğini artırmaktadır [Dietterich, 2002]. Farklı yaklaşımların ağırlıklı birleşimi sayesinde algoritma daha doğru bir yaklaşım oluşturabilmektedir.

Topluluk öğrenmesinde temel öğrenici olarak adlandırılan bir öğrenici, bir dizi etiketlenmiş örneği girdi olarak alan ve bu örnekleri genelleştiren bir sınıflandırıcı veya regresyon modeli üreten algoritmadır. Üretilen modeller kullanılarak etiketlenmemiş yeni örnekler için tahminler gerçekleştirilmektedir. Bir topluluk öğrenicisi karar ağacı, yapay sinir ağları ve doğrusal regresyon modeli gibi herhangi bir makine öğrenmesi algoritması türünde olabilmektedir. Bir topluluk öğrenmesinin temel varsayımı, oluşturulan birden fazla modelin birleştirilerek, tek öğrenicinin yapacağı hatanın telafi edilmesidir. Böylece topluluğun genel tahmin performansının tek bir öğrenici kullanılarak yapılan tahmin performansından daha iyi olması beklenmektedir. Topluluk öğrenmesi yöntemleri, karar ağaçları gibi hesaplama maliyetleri düşük öğreticilerle oluşturulması durumunda oldukça verimli çalışmaktadır. Bu sebeple yapılan araştırmaların büyük bir kısmı karar ağacı tabanlı topluluklar oluşturmaya odaklanmıştır [Sagi ve Rokach, 2018], [Dietterich, 2002].

Torbalama (Bagging) ve Güçlendirme (Boosting) en popüler topluluk öğrenmesi algoritmalarındandır. Torbalama, eğitim setinden yerine koyma yöntemiyle rastgele alt örneklemeler oluşturmaktadır. Bunun sonucunda her biri orijinal veri seti ile aynı boyutta tahmin ediciler oluşturulmaktadır. Regresyon probleminde tahmin edicilerin tahminlerinin ortalaması alınırken, sınıflandırmada nihai karar için çoğunluk oylaması dikkate alınmaktadır. Güçlendirme ise veri setinin dağılımını önceden oluşturulan öğrencilerin performansına göre uyarlamalı olarak değiştirmekte ve her tahmin edicinin tahminleri performansı ile orantılı ağırlıklar almaktadır. Regresyon probleminde bu modellerin ağırlıklı ortalaması hesaplanırken, sınıflandırmada her modelin performansına göre ağırlıklandırılmış oylama dikkate alınmaktadır [Kotsiantis ve Kanellopoulos, 2012].

3.4.2.1. Torbalama Yaklaşımına Dayalı Rastgele Orman

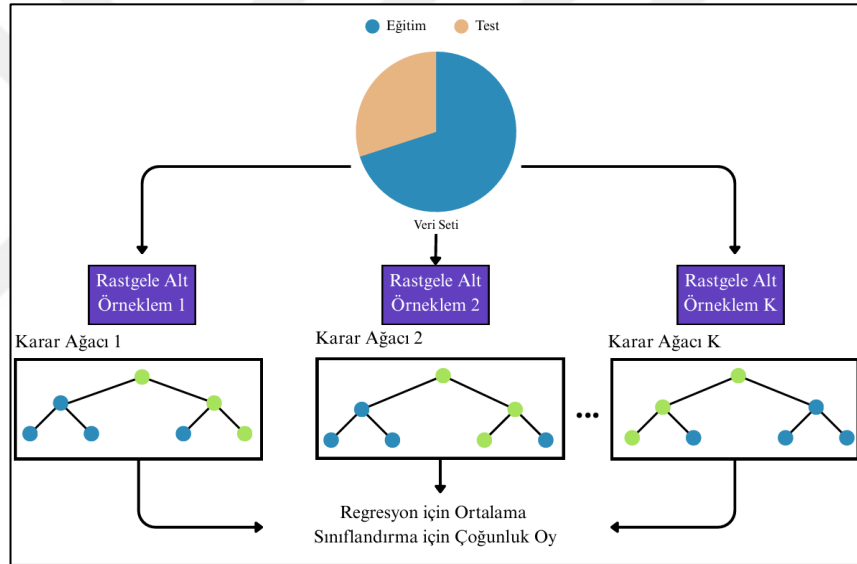
Topluluk Öğrenmesi başlığında bahsedildiği gibi bir örneklem kümesi içerisindeki girdi değişkenleri kullanılarak çıktı değişkeninin tahmin edilmesi istendiğinde bazı zorluklar ortaya çıkmaktadır. Böyle bir durumda tahmin edici yüksek varyans sebebiyle iyi performans gösteremez. Bu sebepten dolayı, çok sayıda örneklem kümesi ve tahmin edici kullanımı ile yüksek varyansın düşürülmesi sağlanarak daha yüksek doğrulukta tahmin edicilerin oluşturulması gerekmektedir.

Bu amaç doğrultusunda Torbalama diye ifade edilen “Bagging” yöntemi uygulanmaktadır. Torbalama yöntemi, istatistiksel veya makine öğrenmesi modellerinin doğruluğunu artırmak amacıyla aşırı öğrenmeyi önleyen ve “Bootstrap aggregation” olarak da geçen bir topluluk öğrenme yöntemidir [Breiman, 1996]. “Bootstrap” ifadesi, $\{(y_i, x_i), i = 1, \dots, n\}$ ile ifade edilen örneklem kümesinden rastgele biçimde yerine koyarak n boyutunda alt örneklemeler oluşturulmasını ve bu alt örneklemelerden bağımsız tahmin ediciler oluşturulmasını ifade etmektedir. Bu yöntemde, alt örneklemelerin orijinal örneklem kümesi ile aynı boyutta olmasının sebebi, her bir tahmin edici için yeterli miktarda örnek sağlanmasıdır. Bir örneklem kümesi, x_i ile ifade edilmekte olan girdi değişkenleri ve bu değişkenlere karşılık gelen y_i çıktı değişkenleri içermektedir. Alt örnekleme, yerine koyma yöntemiyle yapıldığı zaman aynı tahmin edici ile eğitim yapılırken bir girdi ve çıkış değişkeni çiftini (y_i, x_i) ifade eden orijinal örneklerin bir kısmı birden çok kez yer alabileceği gibi bazı örnekler eğitime hiç dahil edilmeyebilir. Her bir örnekleme dayalı olarak $\hat{y}_l = \hat{f}_l(x_1)$ şeklinde

tahmin ediciler eğitilmektedir. Tahmin ediciler bağımsız bir şekilde eğitilmeleri dolayısıyla her bir tahmin edici farklı hesaplama birimleri kullanılarak paralel şekilde eğitilebilmektedir [Sagi ve Rokach, 2018]. Böylece K sayıda tahmin ediciyi içeren $(\hat{f}_1, \hat{f}_2, \dots, \hat{f}_K)$ bir topluluk oluşturulmaktadır. K sayıda tahmine edici ile K adet tahmin çıktısı $(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_K)$ elde edilmektedir. Torbalama için “aggregation” ifadesi alt örneklem ile oluşturulmuş olan bağımsız tahmin edicilerin toplu şekilde değerlendirilmesini ifade etmektedir [Breiman, 1996], [Sagi ve Rokach, 2018]. Tahmin edicilerin eğitiminin ardından sınıflandırma problemlerinde çoğunluk oy dikkate alınırken, regresyon probleminde tahmin çıktılarının ortalaması nihai tahmini oluşturmaktadır [Özbayrak vd., 2024]. Torbalama bu özellikleri ile topluluk öğrenmesinde özellikle yüksek varyansla başa çıkmakta oldukça etkili bir yaklaşım sunmaktadır. Topluluk öğrenmesi metotlarından Rastgele Orman, torbalama yaklaşımını karar ağaçlarından yararlanarak uygulayan en popüler makine öğrenmesi metotlarından birisidir.

Rastgele Orman, $h(x; \theta_k)$ ($k = 1, \dots, K$) şeklinde ifade edilen çok sayıda karar ağacından oluşturulmaktadır. Breiman (2001) tarafından ifade edilen Rastgele Orman, CART (Sınıflandırma ve Regresyon Ağaçları) yapısına dayalı olarak genelleme performansını artırmak amacı doğrultusunda torbalama yöntemiyle çalışmaktadır [Grömping, 2009]. Burada ifade edilen x , uzunluğu n olan girdi değişkenleri temsil etmektedir. Burada ifade edilen θ_k , her ağacın yapısında ve eğitiminde kullanılan bağımsız ve özdeş dağılımlı rastgele vektörleri ifade etmektedir. Rastgele Orman algoritmasında farklı ağaçların artık hataları arasında düşük korelasyon olması ve bireysel ağaçların tahmin hatalarının düşük olması gibi iki ana hedef bulunmaktadır. Bireysel ağaç hatalarının düşük tutulması amacıyla ağaçlar maksimum derinliğe kadar büyütülmektedir. Artık hata korelasyonunun düşük olması amacıyla her ağaç, eğitim setinden yerine koyma örneklemeyle oluşturulan bir örneklem üzerinden büyütülmektedir. Her ağaç, orijinal eğitim setinden yerine koymalı olarak seçilmiş farklı alt örneklemle eğitilmektedir. Her düğümde bölme işlemi yapılırken tüm değişkenler yerine n sayıda girdi değişkeni arasından belirlenen rastgele m değişken değerlendirilmektedir [Breiman, 2001, Segal, 2004]. Sınıflandırma probleminde K adet karar ağacının oy çokluğu alınmakta iken regresyon probleminde tahmin çıktılarının ortalama değeri alınmaktadır (Şekil 3.3).

Rastgele Ormanlar tahminlerin oluşturulmasında aşırı uyum sorununa karşı dayanıklı olma özelliğiyle etkili bir yöntemdir. Doğru türde rastgelelik eklenmesi, onları hem doğru sınıflandırıcılar hem de doğru regresyon modelleri haline getirmektedir. Ayrıca, temel tahmin edicilerin gücü ve aralarındaki korelasyonlara dayanan çerçeve, rastgele ormanın tahmin yeteneği hakkında içgörü sağlamaktadır [Breiman, 2001]. Rastgele Orman algoritmasının genelleme yeteneğinin performans analizi, torbalamanın doğasında yer alan yerine koyma yönteminde dışarda bırakılan örneklemeler üzerinden yapılabilmektedir. Seçilmeyerek dışarıda bırakılan bu örnekler, torba dışı (out-of-bag) olarak ifade edilmektedir. Torba dışı örnekler ağaçların açıklanamayan varyanslarını tahmin etmek amacıyla kullanılmakta ve böylece tahmin modelinin performansının değerlendirilmesi mümkün olmaktadır [Lin, 2021], [Breiman, 2001].



Şekil 3.3: Rastgele Orman çalışma prensibi.

Rastgele Orman, her bir ağacın yapısını kontrol eden çeşitli hiper parametreler içermektedir. Ormanın büyüklüğü ile ilgili ağaç sayısı ve her bölünmede aday bölünme değişkenleri olarak dikkate alınan değişken sayısı gibi hiper parametrelerin optimize edilmesi model performansını olumlu biçimde etkilemektedir. Ancak hangi hiper parametrelerin rutin olarak ayarlaması gerektiği konusunda net bir kılavuz bulunmamaktadır [Probst vd., 2019].

Rastgele Orman hiper parametrelerinden her bir ağaç büyütülürken bölme işlemi sırasında rastgele seçilen aday değişkenlerin sayısını ifade eden *mtry* parametresinin değerinin düşük olması daha farklı ve daha düşük korelasyonlu ağaçlar

oluşturmaktadır. Bu durum toplulaştırma (aggregating) sırasında daha iyi bir istikrar sağlamaktadır. Ancak düşük *mtry* değeri aynı zamanda ağaçların, rastgele çekilen küçük bir aday değişken kümesinden seçilen optimal olmayan değişkenlere dayanarak inşa edilmesinden dolayı ortalama olarak daha kötü performans göstermesine de yol açabilmektedir. Bu yüzden bir denge kurulması gerekmektedir. Çeşitli yazılım ortamlarında varsayılan değer olarak sınıflandırma problemleri için $\sqrt{p}$ ve regresyon problemleri için $p/3$ olarak ayarlanmıştır. Buradaki p değeri girdi değişkenlerinin sayısını ifade etmektedir [Probst vd., 2019].

Temel hiper parametrelerden bir değeri olan ağaç sayısı, klasik anlamda ayarlanabilir (tuneable) bir parametre değildir. Birçok çalışmada ağaç sayısını artırmanın çeşitli performans ölçütlerine göre değerlendirildiğinde modeli daha iyi hale getirdiği gözlemlenmiştir. Yakınsama hızı ve dolayısıyla en iyi performansı elde etmek için gerekli ağaç sayısı, veri setinin özelliklerine bağlıdır. Diğer taraftan ağaçlar birbirinden bağımsız olarak eğitildiği için ağaç sayısının artması hesaplama süresini artırmaktadır. Çok sayıda çalışmada ilk 100 ağacın en büyük performans artışını sağladığı tespit edilmiştir. Birçok çalışmada doğruluğun azalmadığı ve belirli bir artış göstermediği bir ağaç sayısının dikkate alınması amacıyla yakınsama davranışı incelenmektedir. Bu sebeple, artan ağaç sayısına karşılık performansı gösteren torba dışı (out-of-bag) hata grafikleri ile incelenebilmektedir [Probst vd., 2019].

3.4.2.2. Güçlendirme Yaklaşımına Dayalı eXtreme Gradient Boosting

Bir örneklem kümesi içerisindeki girdi değişkenleri kullanılarak çıktı değişkeninin tahmin edilmesi istendiğinde tahmin edici yüksek varyans ve yanlılık sebebiyle iyi performans gösterememektedir. Yüksek varyansın giderilmesinde torbalama yöntemi başarılı sonuçlar verse de hem yüksek varyans hem de yanlılık sorununun giderilmesinde farklı bir yaklaşım olarak Güçlendirme (Boosting) yaklaşımına sahip yöntemler uygulanmaktadır.

Güçlendirme, bir öğrenme algoritmasının performansını artırmaya yönelik geliştirilen bir topluluk öğrenmesi yaklaşımıdır [Freund ve Schapire, 1996]. Freund ve Schapier (1997) tarafından sınıflandırma problemleri için her yeni sınıflandırıcının önceki oluşturulan sınıflandırıcıların performansını dikkate alarak oluşturulduğu ardışık bir algoritma olan Adaptive Boosting (AdaBoost) algoritmasını tanıttıktan sonra

güçlendirme yöntemleri büyük ilgi görmeye başlamıştır [Gey ve Poggi, 2006]. Böylece hem sınıflandırma hem de regresyon problemleri için kullanılan bir dizi yöntem geliştirilmiştir [Zhang vd., 2009].

Güçlendirme yönteminde tahmin ediciler ardışık biçimde eğitilmektedir. Torbalama yönteminde olduğu gibi ilk tahmin edici, orijinal eğitim kümesinden yerine koyarak seçilen n boyutundaki örneklemeler üzerinden eğitilmektedir. Torbalama yönteminde her örneğin seçilme olasılığı eşit dereceye sahipken, Güçlendirme yönteminde belirli bir örneğin, belirli bir modelin eğitim kümesinde yer alma olasılığı, önceki modellerin o örnek üzerindeki performansına bağlı kalmaktadır [Drucker, 1997]. Bu kapsamda, çıktı değişkeni tahmin için kullanılan fonksiyonun tahmininde eğitim veri setinin örnekleri için $\{w_t^i\}_{i=1}^n$ ağırlık veri seti kullanılmaktadır. Bu süreçte, önceki modellerin hatalarına odaklanılarak ağırlıkların ayarlanması gerçekleştirilmekte ve son olarak birleştirilmiş model, eğitim verilerinin tahmin kalitesine göre ağırlıklandırılmış fonksiyonların doğrusal bir kombinasyonu olmaktadır [Borra ve Ciaccio, 2002]. Güçlendirme ilk başta sınıflandırma problemi olarak ortaya çıkarılmış olsa da regresyon problemleri için de çeşitli algoritmalar geliştirilmiştir. Genel olarak, Adaptive Boosting Regressor (Adaboost.R) ve gradyan tabanlı olacak şekilde iki farklı yaklaşım ele alınmaktadır [Kotsiantis ve Kanellopoulos, 2012].

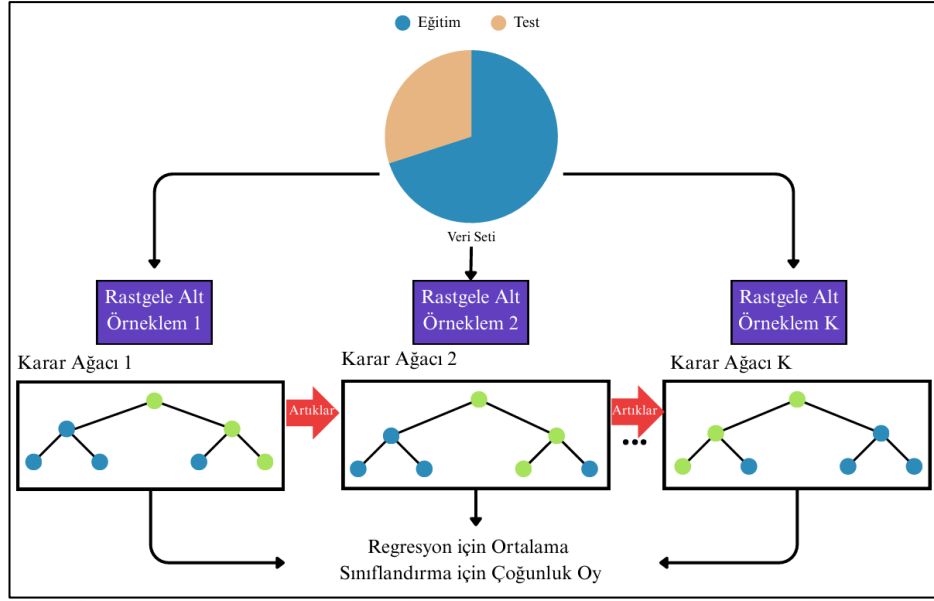
Drucker (1997) tarafından regresyon problemleri için önerilen Güçlendirme metodoloji Adaboost algoritmasının bir modifikasyonudur ve Adaboost.R olarak geçmektedir. Bu yöntem, regresyon ağacını temel tahmin edici olarak kullanarak kötü tahmin edilen örneklerin ağırlıklarını artırırken, iyi tahmin edilen değerlerin ağırlıklarını azaltmaktadır [Zhang vd., 2009], [Drucker, 1997]. Algoritma öncelikle tüm eğitim örneklerine aynı ağırlıkları $w_i^0 = 1$ ($i = 1, \dots, n$) atamaktadır. Eğitim örnekleminden n örnek Torbalama yönteminde olduğu gibi yerine koyma metoduyla seçilmektedir. Ancak bu yöntemdeki fark, i 'inci örneğin seçilme olasılığı o iterasyona kadar yapılmış olan hatalara bağlı bir ağırlığın atanması ile hesaplanmaktadır [Borra ve Ciaccio, 2002]. Bu her bir örneğin seçilme olasılığının örneğe atanan ağırlığa bağlı olduğunu göstermektedir.

Daha sonra, çekilen örnekleme dayalı olarak bir tahmin fonksiyonu $h_t(x)$, uyarlanmaktadır. Bu fonksiyon, eğitim verisi üzerinden yapılan tahminleri temsil etmektedir. Sonrasında eğitim örneklerinin tamamı üzerinde her örnek için göreceli mutlak tahmin hataları hesaplanmaktadır. Tüm örneklerin ağırlıkları da hesaba

katılmasıyla kayıpların ortalaması alınarak öğrenici performansı hesaplanmaktadır. Hesaplama sonuçlarına göre her bir örneğin ağırlıkları güncellenmektedir [**Borra ve Ciaccio, 2002**]. Son olarak, bir örneğin tahmini, her bir temel tahmin edicinin çıktılarından ağırlıklı ortalama olarak inşa edilmektedir [**Zhang vd., 2009**].

Diğer bir yaklaşım ise Adaboost algoritmasını kayıp fonksiyonu üzerinden bir gradyan iniş algoritması olarak yorumlanan Friedman (2001) tarafından geliştirilen Gradient Boosting Machine (GBM) algoritmasıdır. Yöntemin her bir işlem adımında önceki fonksiyondaki hatalar dikkate alınarak her yeni fonksiyon bu hataları düzeltecek biçimde eğitilmektedir. Bu yaklaşımda bir kayıp fonksiyonu kullanılmaktadır. GBM algoritması, ileriye dönük aşamalı bir prosedür kullanmaktadır. Problemin çözümünde öncelikle her adımda bir önceki adımda elde edilen tahmin fonksiyonunun negatif gradyanı ile en paralel olan fonksiyon seçilmektedir. Gradyan, mevcut parametrelerin hangi yönde değiştirilmesi gerektiğini göstermekte ve bu yönde ne kadar ilerlenmesi gerektiğini bir doğrultu arama işlemiyle belirlemektedir. Bu adımda modelin hatasını azaltmak için en uygun fonksiyon belirlenmeye çalışılmaktadır. Bu koşul, en küçük kareler yöntemi ile elde edilmektedir [**Friedman, 2001**], [**Borra ve Ciaccio, 2002**]. Açıklanan bu süreç her iterasyonda hataları minimize eden yeni bir modelin eklenmesi ile belirli bir iterasyon veya durma noktasına kadar devam etmektedir. Böylece başlangıçta zayıf tahmin edicilerin birleşimi ile güçlü bir tahmin modeli kurulmaktadır.

Bu kapsamda XGBoost, topluluk öğrenmesinde bulunan ve Güçlendirme yaklaşımını kullanan metotlardan birisidir. XGBoost, hem sınıflandırma hem de regresyon çalışmalarında kullanılabilen ağaç güçlendirme için ölçeklenebilir bir yöntemdir (**Şekil 3.4**). Yöntemin başarısı, birçok makine öğrenmesi ve veri madenciliği yarışmalarında geniş çaplı bir şekilde tanınmıştır. Özellikle 2015 yılında Kaggle web sayfasında yayınlanan yarışmalarda kazanan çözümlerin çoğunluğunda XGBoost algoritmasının kullanıldığı görülmektedir. Mağaza satış tahmini, müşteri davranış tahmini, reklam tıklama oranı tahmini, tehlike risk tahmini gibi çalışma alanlarında başarılı sonuçlar veren bu yöntemin başarısının arkasında yatan en önemli etken, çeşitli senaryolara sağladığı ölçeklenebilir olma özelliğidir. XGBoost, mevcut popüler çözümlerden çok daha hızlı çalışabilmenin yanında dağıtık veya bellek kısıtlı ortamlarda ölçeklenebilmektedir. Yüksek ölçeklenebilirliği, bazı önemli sistemsel ve algoritmik optimizasyona dayanmaktadır [**Chen ve Guestrin, 2016**].



Şekil 3.4: XGBoost çalışma prensibi.

XGBoost, örnek ağırlıklarını yaklaşık ağaç öğrenmesi sırasında işleyebilmeyi mümkün kılan teorik olarak doğrulanmış bir ağırlıklı çeyrek değer kestirimi prosedürü ve en iyi modelin keşfi sürecini hızlandıran paralel ve dağıtık hesaplama tekniklerini kullanmaktadır. Ayrıca çok büyük örnek içeren verilerin masaüstü bilgisayarlarda işlenebilmesini olağan kılan çekirdek dışı hesaplama yeteneğine sahiptir [Chen ve Guestrin, 2016].

XGBoost algoritmasının temel amacı, eşitlik (3.3)'te ifade edilen kayıp fonksiyonu ve düzenleme terimini içeren aşağıdaki amaç fonksiyonunu minimize etmektir.

$$L(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \quad (3.3)$$

Formülde l , tahmin edilen değer $\hat{y}_i$ ile gerçek değer y_i arasındaki farkı ölçen türevlenebilir ve konveks bir kayıp fonksiyonunu ifade etmektedir. Eşitlik (3.4)'de detayı verilen ikinci terim Ω , düzenleme terimi olarak geçmekte ve modelin karmaşıklığını kontrol edip cezalandırmaktadır. Formülde bulunan γ , her yaprak için ceza katsayısını, T yaprak sayısını, λ yaprak skorlarının büyüklüğünü cezalandıran katsayıyı ve w , ilgili yaprağın skorunu ifade etmektedir.

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda ||w||^2 \quad (3.4)$$

Düzenleme terimi, öğrenilen ağırlıkların daha düzgün olmasını sağlayarak modelin aşırı öğrenmesini önlemeye katkıda bulunur. Düzenlenmiş amaç fonksiyonu, sezgisel olarak daha basit ve yüksek tahmin gücüne sahip fonksiyonlar kullanan bir model seçme eğilimindedir. Düzenleme parametresinin sıfır olması durumunda amaç fonksiyonu geleneksel gradyan ağaç Güçlendirme yöntemine dönüşmektedir.

Eşitlik (3.4)'de bulunan formül, parametre ve fonksiyonlar içeren bir ağaç topluluğu modelidir. Bu yüzden klasik matematiksel yöntemlerle optimize etmek mümkün değildir. Bunun yerine model eklemeli şekilde eğitilmektedir. Eşitlik (3.5)'te bulunan $y_i^{(t)}$, t 'inci iterasyonundaki i 'inci örneğin tahminidir. Eşitliğe f_t , eklenerek aşağıdaki amaç fonksiyonunun minimize edilmesi gerekmektedir. Bunun anlamı, modeli en fazla iyileştiren f_t 'nin eşitlik (3.3)'e dayanarak aç gözlü biçimde eklendiği anlamına gelmektedir.

$$L(t) = \sum_{i=1}^n l(y_i - y_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (3.5)$$

Amaç fonksiyonunun daha hızlı biçimde çözümü amacıyla türevi alınmakta ve eşitlik (3.5)'teki t adımındaki basitleştirilmiş amaç fonksiyonu kullanılmaktadır. Buradaki g_i ve h_i , kayıp fonksiyonunun birinci ve ikinci dereceden türevleridir.

$$L(t) = \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (3.6)$$

J yaprağına ait örnekler kümesi $I_j = \{i \mid q(x_i) = j\}$ şeklinde tanımlanmaktadır. Böylece t 'inci ağacın değeri için eşitlik (3.6) aşağıdaki eşitlik (3.7) ve ardından eşitlik (3.8)'de olduğu gibi tekrar yazılabilir.

$$\tilde{L}^{(t)} = \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (3.7)$$

$$\sum_{j=1}^T \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \quad (3.8)$$

Eşitlik (3.8)'te j 'inci yaprağın optimal ağırlığını ifade eden w_j hesaplanmaktadır. Ardından karşılık gelen optimal değer eşitlik (3.10)'da hesaplanmaktadır. Eşitlik (3.10), bir ağaç yapısının kalitesini ölçmek için puanlama fonksiyonu olarak kullanılabilir.

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} (h_i + \lambda)} \quad (3.9)$$

$$\tilde{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^T \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} (h_i + \lambda)} + \gamma T \quad (3.10)$$

Tüm olası ağaçlar (q) için bu değerleri hesaplamak zordur. Bunun yerine tek bir yapraktan başlayıp ağaca dallar ekleyerek iteratif biçimde çalışan açgözlü bir algoritma kullanılmaktadır. Sol ve sağ düğümlerin örnek kümelerini I_L ve I_R olarak ifade eder ve $I = I_L \cup I_R$ olduğunu varsayıldığında bölme sonrası kayıp azalması eşitlik (3.11)'da gösterildiği şekilde ifade edilmektedir. Her karar ağacı bölünmesinde kaybı minimize etmeye bir diğer ifadeyle bilgi kazancını maksimize etmeye çalışmaktadır. Bu formül bölme adaylarının değerlendirilmesi amacıyla kullanılmaktadır.

$$L_{\text{split}} = \frac{1}{2} \left[\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in I_R} h_i + \lambda} + \frac{\left(\sum_{i \in I} g_i\right)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma \quad (3.11)$$

XGBoost modelinin performansını en iyi seviyelere çıkartmak için çeşitli model parametrelerinin doğru şekilde ayarlanması gerekmektedir. Hiper parametre ayarı yapılmadığı takdirde model varsayılan değerler ile çalışacaktır. Veri kümelerinin kendine has bazı özelliklere sahip olabileceğinden dolayı bu özelliklerin dikkate alınmasında hiper parametre ayarlaması önemli bir işlem adıdır [Ryu vd., 2020]. Doğru şekilde ayarlanmış parametreler sayesinde model, aşırı öğrenme ve aşırı karmaşıklık gibi modelin doğruluğunu ve performansını etkileyebilecek sorunlar önlenmiş olur. Aşırı öğrenme, modelin gürültülü verileri öğrenmeye başlaması ve bunların anlamlı olgular olarak kabul etmesi durumunda yaşanmaktadır [Parsa vd., 2020].

XGBoost için *eta*, *gamma*, *max_depth*, *min_child_weight*, *colsample_bytree*, *subsample*, *nround* parametreleri modelin başarısı için öne çıkan ve dikkatlice ayarlanması gereken parametrelerdendir. Hiper parametreler ayarlanırken kendilerine özgü değer aralıklarına ve varsayılan değerlerinin de dikkate alınması gerekmektedir (**Tablo 3.1**). Her bir ekleme adımından sonra yeni özelliklerin ağırlıkları doğrudan elde edilebilir ve öğrenme oranını ifade eden *eta* bu özelliklerin ağırlıklarını küçülterek daha temkinli bir süreç oluşturulmasını sağlamaktadır. Kayıp fonksiyonu içindeki ağaç düğümlerinin doğru şekilde bölünebilmesi için gerekli olan kayıp azalmasını belirten *gamma* değerinin büyük olması algoritmanın daha temkinli çalışmasına katkı sağlamaktadır. Bir ağacın alabileceği maksimum derinliği ifade eden *max_depth* ne kadar büyükse model belirli bir örneğe ait özel ilişkileri o kadar fazla öğrenmektedir. Bu durum modeli daha karmaşık hale getirebilmekte ve aşırı öğrenme riskini artırmaktadır. Belirli bir düğüm için gerekli olan gözlemlerin ağırlıklarının toplamının minimum değeri *min_child_weight* ile ifade edilmektedir. Ağaç bölme adımı, örnek ağırlıkları toplamı bu parametre değerinden küçük olan bir yaprak düğüm oluşturursa, model bu düğümden daha fazla bölme yapmaktan vazgeçer. Bu değerinin büyük olması algoritmanın daha temkinli çalışacağını ifade etmektedir. Her ağacın oluşturulmasında kullanılacak alt örnekleme oranı *colsample_bytree* ile ifade edilmektedir. Her ağaç için büyütülmeye önceki rastgele örneklenecek eğitim verisinin oranı, *subsample* ile ifade edilmektedir. Bu parametrenin doğru ayarlanması aşırı öğrenme sorununun engellenmesini sağlamaktadır. En yüksek gradyan artırmalı ağaç sayısı *nrounds* ile ifade edilmektedir [Ryu vd., 2020], [Tarwidi vd., 2023].

Tablo 3.1: XGBoost hiper parametrelerinin değer aralıkları ve varsayılan değerleri.

Hiper Parametre	Değer Tipi	Değer Aralığı	Varsayılan Değer
nrounds	Tam sayı	$(0, \infty)$	100
max_depth	Tam sayı	$[0, \infty)$	6
eta	Ondalık sayı	$[0,1]$	0.3
gamma	Ondalık sayı	$[0,\infty)$	0
colsample_bytree	Ondalık sayı	$(0, 1]$	1
min_child_weight	Tam sayı	$[0,\infty]$	1
subsample	Ondalık Sayı	$(0,1]$	1

3.5. Hiper Parametre Ayarlama

Makine öğrenmesi algoritmaları çeşitli hiper parametreler barındırmaktadır. Bu hiper parametrelerin optimal değerleri, modelde kullanılan veri setine göre değişmektedir. Farklı veri setlerine göre farklı değerler alabilen bu hiper parametrelerin optimal değerlerinin elde edilmesi model performansına artırabilmektedir. Model içerisinde denenen farklı hiper parametrelerin optimal olup olmadığının kararı ise daha önceden belirlenmiş performans ölçütüne göre kararlaştırılmaktadır [Probst vd., 2019].

Hiper parametre ayarlama süreci, farklı hiper parametre değerlerini içeren kümeler arasından tahmin modelinde en iyi tahmin sonuçlarını verecek olanın seçilmesidir. Ancak çok sayıda hiper parametre ayarlaması hesaplama süresi açısından büyük bir zorluk oluşturmaktadır. Hiper parametrelerin ayarlanması işleminde dikkate edilmesi gereken bir kavram da aşırı uyumdur. Parametre değerleri, eğitim verilerine aşırı uyum sağlama eğiliminde olmaktadır. Eğitim verilerine aşırı uyum sağlayan parametreler eğitim verisinde modelin iyi sonuçlar vermesini sağlarken farklı bir veride kötü sonuçlar vermektedir. Optimal olmayan hiper parametre değerlerinin seçiminden ve aşırı uyum riskinden kaçınmak amacıyla çeşitli yöntemler uygulanmaktadır [Probst vd., 2019].

Hiper Parametre Ayarlama sürecinde Izgara Arama (Grid Search), Rastgele Arama (Random Search), Bayesian Optimizasyon (Bayesian Optimization) ve Genetik Algoritma (Genetic Algorithm) gibi çeşitli yöntemler uygulanmaktadır. Bu yaklaşımlardan Izgara Arama (Grid Search), öğrenme süreci esnasında olası hiper parametre değerlerinin sistematik biçimde taranması ve tüm sürecin otomatik şekilde gerçekleştirilmesini sağlamaktadır. Bu süreç esnasında her bir hiper parametreye bir dizi değer verilmektedir. Tüm olası hiper parametre değer kombinasyonları yinelenerek modeller eğitilmektedir. Eğitilen modeller ile tahminler yapılmakta ve değerlendirme ölçütleri ile karşılaştırılmaktadır. Böylece en yüksek doğruluklu modeli oluşturan hiper parametreler belirlenmiş olmaktadır. Izgara Arama yönteminin önemli avantajlarından birisi hiper parametre ayarlarının bağımsız olması sayesinde paralel hesaplama yapma kabiliyetinin bulunmasıdır [Jiang ve Xu, 2022].

3.6. Model Doğrulama

Makine öğrenmesinde modelin görülmemiş verilerde kötü genelleme yapması durumunu ifade eden aşırı uyum sorunuyla karşılaşılması çok olasıdır. Bu durumda tahmin modeli, eğitim verisinde iyi performans göstermekte iken test verisinde kötü performans gösterebilmektedir. Aşırı uyum, eğitim veri setinin çok küçük veya çok fazla gürültü barındırması ve hiper parametrelere bağlı sebeplerden dolayı ortaya çıkabilmektedir [Ying, 2019]. Hiper parametrelerin optimizasyonu ve tahmin modellerinin aşırı uyum riskinin değerlendirilmesi için üretilen tahmin modellerinin doğrulanması gerekmektedir.

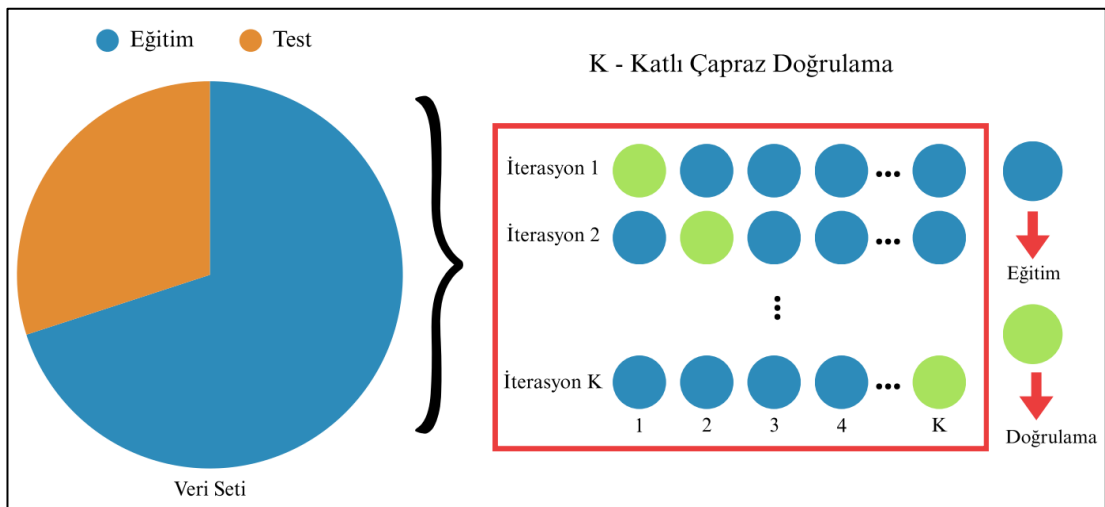
Veri kümesinde bulunan verilerin bir kısmının ayrılması modelin doğrulanması amacıyla izlenen yöntemlerdendir. Veri ayırma yönteminde modelin doğrulanması için verilerin belirli bir kısmı modelin eğitimi amacıyla ayrılmakta iken belirli bir kısmı da modelin doğruluğunun test edilmesi için ayrılmaktadır. Mühendislik uygulamalarında verilerin üçte birinin doğrulama için ayrılması ve kalan üçte ikisinin eğitim için kullanılması yaygın bir yaklaşımdır. Eğitim ve doğrulama kümelerindeki verilerin temsili olması için veri kümesinin dengeli biçimde ayrılması oldukça önemlidir. Verilerin ayrılması işleminin her kümede dengeli şekilde veri bulunması amacıyla uygulanması Katmanlı-Ayrırma (Stratified-Holdout) olarak adlandırılmaktadır [Feng vd., 2005].

Bu yöntem ile seçilen belirli bir örneklemin neden olabileceği yanlılığı azaltmanın daha genel bir yolu, eğitim ve doğrulama sürecinin farklı rastgele örnekler ile birden çok kez tekrarlanmasıdır. Farklı yinelemelerle elde edilen hata oranlarının ortalaması genel bir hata oranı elde edilmesini sağlamaktadır. Bu şekilde gerçekleştirilen model doğrulama süreci, Çapraz Doğrulama (Cross-Validation) olarak bilinmektedir [Feng vd., 2005].

Çapraz doğrulama tekniğinde aday modeller eğitim veri seti üzerinde eğitilmekte ve dışarıda bırakılmış olan doğrulama veri seti üzerinden hatalar hesaplanmaktadır. Model karmaşıklığının artmasıyla beraber eğitim hatası azalmaktadır. Doğrulama setindeki hata ise belirli bir karmaşıklık seviyesine kadar azalmakta ve belirli bir seviyeden sonra azalmamaya başlamaktadır. Ayrıca veri içinde gürültü bulunuyorsa hata oranı artış bile gösterebilmektedir. Bu sebeplerden dolayı optimal karmaşıklık seviyesini gösteren bir dirsek noktası bulunmalıdır [Alpaydın, 2014].

Çapraz doğrulamada doğrulama setleri aynı zamanda test amaçlı olarak da kullanılmaktadır. Bu yüzden doğrulama seti fiilen kullanılan verilerin bir parçasıdır. Tüm işlemlerin ardından belirli bir algoritmaya karar verip hatasının raporlanması gerektiğinde ayrı bir test verisinin kullanılması gerekmektedir. Bu test veri setinin modelin eğitimi aşamasında hiç kullanılmamış olması ve yeterli büyüklükte olması gerekmektedir. Bu sebeple ilk başta veri setinin bir kısmı test verisi olarak ayrılmalı ve geriye kalan veri seti çapraz doğrulama için eğitim ve doğrulama olarak ayrılmalıdır. Böylece eğitim seti belirli bir öğrenme algoritması ve model yapısı için parametre optimizasyonu için kullanılırken, doğrulama veri seti ise öğrenme algoritmasının hiper parametrelerini veya model yapısını optimize etmek için kullanılmaktadır. Son olarak test veri seti kullanılarak kullanılacak nihai model belirlenmektedir [Alpaydın, 2014].

Çapraz Doğrulama yöntemleri üç genel kategoriye ayrılır: K-Katlı Çapraz Doğrulama (K-Fold Cross Validation), Birini Dışarıda Bırak (Leave-One-Out) ve Yeniden Örnekleme (Bootstrap) [Feng vd., 2005]. Bu tekniklerden K-Katlı Çapraz Doğrulama en yaygın kullanıma sahip olan tekniklerden birisidir. Yöntemde, veri kümesi rastgele biçimde K eşit parçaya bölünmektedir. Eğitim ve doğrulama çiftleri oluşturmak amacıyla K parçadan birisi doğrulama seti olarak ayrılır ve geri kalan K-1 parça birleştirilerek eğitim seti oluşturulur. Bu işlem K kez tekrarlanmaktadır (Şekil 3.5). K değeri genellikle 2 ile 10 arasında seçilmektedir. Tüm çapraz doğrulama prosedürünün birkaç kez tekrarlanması sonucunda elde edilen çıktıların ortalaması ile tahminin varyansı azaltıldığı için güvenilir sonuçlar vermektedir [Probst vd., 2019].



Şekil 3.5: K-Katlı Çapraz Doğrulama çalışma prensibi.

3.7. Model Performans Değerlendirme Metrikleri

Makine öğrenmesi metotlarında regresyon tabanlı tahmin modelleri sürekli veri yapısındaki hedef değişkenin değerini gerçek değerden belli ölçülerde farklı tahmin etmektedir. Bu sebeple geliştirilen tahmin modellerinin performansının değerlendirilmesi büyük önem taşımaktadır. Tahmin modellerinin performansı, gerçek değer ile tahmin edilen değer arasındaki farkı ifade eden artıklar üzerinden hesaplama yapan performans değerlendirme metrikleri ile değerlendirilmektedir [Asthana vd., 2023]. Makine öğrenmesinde Kök Ortalama Kare Hatası (Root Mean Square Error-RMSE), R^2 , Ortalama Mutlak Hata (Mean Absolute Error-MAE) ve Ortalama Mutlak Yüzdesel Hata (Mean Absolute Percentage Error-MAPE) sıklıkla kullanılan model performans değerlendirme metriklerindendir. Bir tahmin probleminde n örneklem büyüklüğündeki veri seti için y_i , gerçek değerleri, $\hat{y}_i$, tahmin değerlerini ve $\bar{y}_i$, gerçek değerlerin ortalama değerini ifade etmektedir.

RMSE, n örneklem büyüklüğündeki veri seti için tahmin edilen değerler ile gerçek değerler arasındaki farkların standart sapmasını ifade etmekte ve eşitlik (3.12)'teki gibi hesaplanmaktadır [Asthana vd., 2023].

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3.12)$$

R^2 performans metriği, geliştirilen modellerin karşılaştırılması için oldukça basit ama etkili bir yöntemdir. R^2 , $[-\infty, 1]$ aralığında değerler alabilmektedir. En iyi sonuç 1'dir ve tahminler ile hedef değişken değerlerinin tam anlamıyla örtüştüğünü ifade etmektedir. Değerin sıfıra yaklaşması modelin uyumunun ve güvenilirliğinin azaldığına işaret etmektedir. R^2 değeri, eşitlik (3.13)'de gösterildiği şekilde hesaplanmaktadır [Segovia vd., 2023].

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (3.13)$$

MAE, bir gözlemin tahmini ile gerçek değer arasındaki farkın büyüklüğünü ifade etmektedir. MAE, tüm gözlem grubundaki hataların ortalamasıdır ve regresyona

dayalı makine öğrenmesi algoritmaları için bir kayıp fonksiyonu olarak kullanılarak hataların nicel bir ölçüsünü ifade etmektedir [Asthana vd., 2023]. Her bir tahmin hatası, eşitlik (3.14)'te gösterildiği şekliyle ilgili örnek için gerçek değer ile tahmin edilen değer arasındaki farktır.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3.14)$$

MAPE, regresyon problemleri için kullanılan önemli bir değerlendirme metriğidir. Bu metrik, göreceli hatalara duyarlı olmayı amaçlamaktadır. MAPE, tahmin edilen ve gerçek değerler arasındaki tüm mutlak yüzde hatalarının ortalamasıdır. Bu metrik hesaplanırken farkın gerçek bir değere bölünmesi gerekmektedir. Gerçek değerlerin 0'a yakın veya 0 olması durumunda MAPE skorunun 0'a bölme hatası alacağı anlamına gelmektedir. Bu nedenle 0 veya 0'a yakın değerlerin bulunduğu veri setlerinde uygulanması önerilmemektedir [Segovia vd., 2023]. MAPE, eşitlik (3.15)'te gösterildiği şekilde hesaplanmaktadır.

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \quad (3.15)$$

4. UYGULAMA

4.1. Çalışma Bölgesinin Belirlenmesi

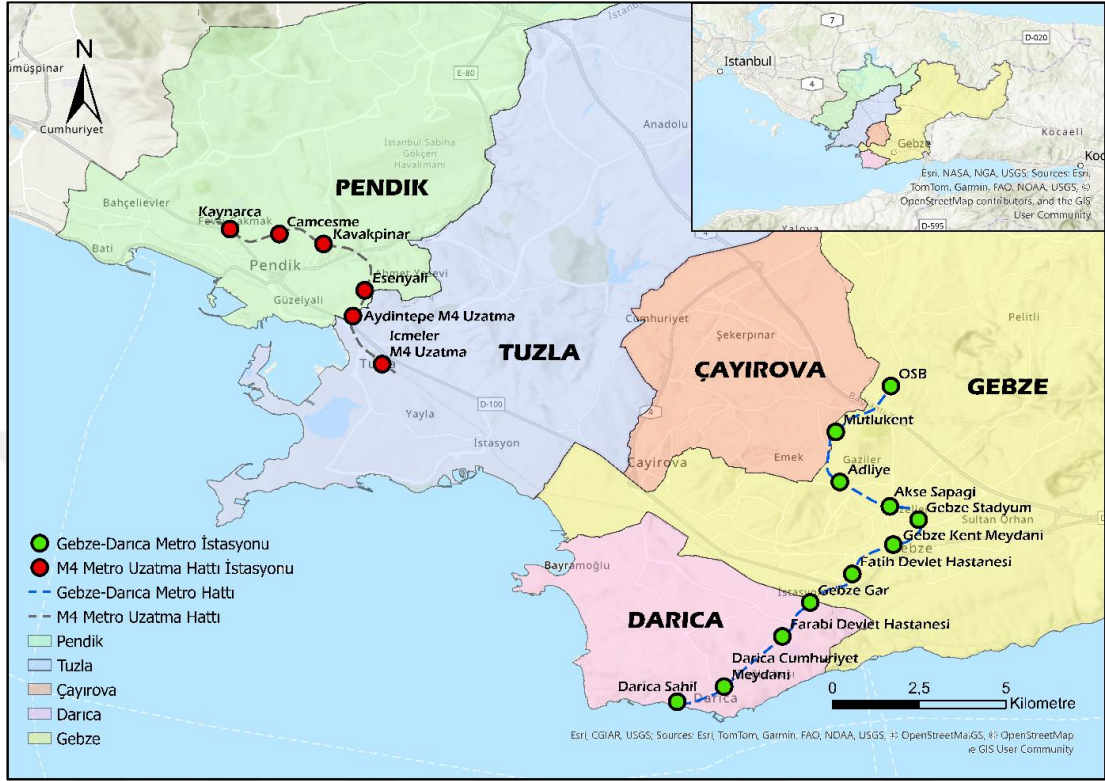
Çalışma bölgesi olarak İstanbul ili sınırları içerisinde yer alan Pendik ve Tuzla ilçeleri ile Kocaeli ili sınırları içerisinde yer alan Çayirova, Darıca ve Gebze ilçeleri seçilmiştir (**Şekil 4.1**). İstanbul, yaklaşık 15 milyonluk nüfusu ile Türkiye'nin en kalabalık şehri durumundadır [**Web 2, 2025**]. Ayrıca, karayolu, raylı ve deniz ulaşım hizmetleri ile geniş bir toplu taşıma ağına sahiptir. Pendik ve Tuzla ilçeleri İstanbul'un en güneydoğusunda kalan ilçeleridir. Pendik İlçesinin, adrese dayalı nüfus kayıt sistemi 2024 verilerine göre nüfusu 749.356 kişi olarak belirtilmiştir [**Web 2, 2025**]. Pendik, içerisinde Sabiha Gökçen Havaalanı, denizyolu ve yüksek hızlı tren, metro ve Marmaray gibi demiryolu bağlantıları bulundurmasıyla ulaşım açısından çeşitliliğe sahip bir ilçedir [**Web 3, 2025**]. Tuzla İlçesinin, adrese dayalı nüfus kayıt sistemi 2024 verilerine göre nüfusu 301.400 kişi olarak belirtilmiş ve İstanbul sınırlarında nüfus artış hızı en yüksek 3'üncü ilçedir [**Web 2, 2025**]. Asya ve Avrupa'yı birleştiren bir konumda Marmara bölgesinin Doğu kısmında bulunan Kocaeli, 12 ilçeden oluşmakta ve yaklaşık 2 milyonluk nüfusa sahiptir [**Web 4, 2025**]. Kocaeli, İstanbul, Sakarya, Bursa ve Yalova illerine komşu durumundadır. Kocaeli sınırları içerisinde yer alan Çayirova, Darıca ve Gebze ilçeleri Marmara Bölgesi'nin doğusunda, İzmit Körfezi'nin kuzey kesiminde yer almaktadır. Çayirova ilçesinin adrese dayalı nüfus kayıt sistemi 2024 verilerine göre nüfusu 157.503, Darıca ilçesinin 231.442 ve Gebze ilçesinin 411.800 kişi olarak belirtilmiştir [**Web 2, 2025**]. Bu ilçeler arasından Gebze, Marmara bölgesinin en büyük ikinci ilçesi konumundadır [**Web 5, 2025**]. Bu ilçeler Anadolu'nun İstanbul'a ve Avrupa'ya bağlantı konumunda bulunmaktadır ve sınırları içerisinde Yüksek Hızlı Tren ve Marmaray demiryolu hattı geçmektedir. Marmaray demiryolu hattı, İstanbul Avrupa yakasında bulunan Halkalı'dan sefere başlamakta ve boğazı aşarak Gebze istasyonunda son bulmaktadır [**Web 4, 2025**].

istasyonunda son bulmaktadır. Toplam sefer süresi 52 dakika olan hattın yolcu taşıma kapasitesi saatlik 70.000 olarak belirtilmiştir [Web 6, 2025]. Marmaray banliyö tren hattı, ilk olarak 2013 yılında hizmet vermeye başlamış ve günümüzde istasyon ağını genişleterek Gebze-Halkalı arasında 76 km'lik bir uzunlukta 43 istasyon ile hizmet vermektedir. Asya kıtasını Avrupa kıtasına bağlayan özelliği ile ön plana çıkan hattın 1,4 km'lik bir kısmı Boğaz'ın altından tüp tünelden geçmektedir. Gebze-Halkalı arası sefer sıklığı 15 dakika olmakla birlikte toplam sefer süresi 108 dakika olan banliyö tren hattı günlük 287 sefer düzenlemektedir [Web 7, 2025].

M4 metro hattına ait PENDİK M4, TAVŞANTEPE, FEVZİ ÇAKMAK – HASTANE, YAYALAR – ŞEYHLİ, KURTKÖY ve SABİHA GÖKÇEN HAVALİMANI istasyonları ve Marmaray banliyö treni hattına ait GEBZE, DARICA, OSMANGAZİ, FATİH-GTÜ, ÇAYIROVA, TUZLA, İÇMELER, AYDINTEPE, GÜZELYALI, TERSANE, KAYNARCA ve PENDİK istasyonları çalışma bölgesi sınırları içerisinde bulunmaktadır. Toplamda 6 M4 Metro istasyonu ve 12 Marmaray banliyö treni istasyonu çalışma kapsamında kullanılacak istasyonlar olarak belirlenmiştir.

Çalışma kapsamında üretilecek yolcu talep tahmin modelleri 2024 yılına yönelik mevcut istasyonlar için ve proje aşamasında bulunan Gebze-Darica ve M4 Metro Uzatma hatlarına ait istasyonlara uygulanması hedeflenmektedir (Şekil 4.2). Proje aşamasında bulunan metrolardan Darica ve Gebze sınırları içerisinde bulunan ve Çayırova sınırına yakın bir doğrultudan geçmekte olan Gebze-Darica Metro hattı, Darica sahil kısmından başlayarak Gebze Organize Sanayi Bölgesi'nde son bulmaktadır. Metro hattı, 15,4 km uzunluğunda ve toplamda 11 istasyon içermektedir. DARICA SAHİL, DARICA CUMHURİYET MEYDANI, FARABİ DEVLET HASTANESİ, GEBZE GAR, FATİH DEVLET HASTANESİ, GEBZE KENT MEYDANI, GEBZE STADYUM, AKSE SAPAĞI, ADLİYE, MUTLUKENT ve OSB istasyonlarından oluşan metro hattı, Gebze ve Darica ilçeleri arasında karayolu bağlantısına alternatif bir tercih oluşturması ve Gebze Marmaray istasyonu ile entegrasyon amaçlı olarak inşa edilmektedir. Proje için 2024 itibarıyla %86 oranında fiziki ilerleme oranı rapor edilmiştir. [Web 8, 2025]. Proje aşamasındaki bir diğer metro hattı, M4 Kadıköy-Tavşantepe hattının devamı şeklinde toplamda 6 istasyon bulunacak şekilde inşa edilmektedir. KAYNARCA, ÇAMÇEŞME, KAVAKPINAR, ESENYALI, AYDINTEPE M4 UZATMA ve İÇMELER M4 UZATMA istasyonlarından oluşan hattın tamamlanması halinde mevcut M4 aktarmaları ile

Kaynarca Merkez istasyonunda Sabiha Gökçen Havalimanı istikametine ve Pendik Yüksek Hızlı Tren garı istikametine aktarma yapılabilecektir [Web 9, 2025].



Şekil 4.2: Gebze-Darica Metrosu ve M4 Metro Uzatma hattı.

4.2. Veri Setlerinin Temini ve Hazırlığı

Toplu taşıma sistemlerinde yolcu talebi, vatandaşların ihtiyaçlarındaki dinamik yapıdan dolayı değişiklik göstermektedir. Bu nedenle, haftanın günü, mevsim, tatil günleri ve hava koşulları gibi zamansal ve istasyon çevresindeki kentsel ilgi noktaları ve nüfus gibi çeşitli mekânsal dış faktörlerden güçlü bir şekilde etkilenmektedir [Torrepadula vd., 2024]. Yolcu talebindeki günlük dalgalanmaların yorumlanabilmesi için zamansal ve baz yolcu talebinin yorumlanabilmesi için mekânsal faktörleri içeren bir veri seti hazırlanmıştır. İlk aşamada istasyon konumlarına ve güzergahlarına ait konumsal veriler, açık veri kaynağı OpenStreetMap (OSM) üzerinden temin edilmiştir. Zamansal faktörler tahmin modelinin kurulmasında 01.01.2023-31.12.2023 tarihleri aralığında olacak şekilde ve tahminin gerçekleştirilmek istendiği 01.01.2024-31.12.2024 tarihleri aralığında olacak şekilde açık veri kaynaklarından elde edilmiştir. Mekânsal faktörler ise hem açık veri

kaynaklarından hem de ticari yazılım çözümlerinden elde edilmiştir. Veri kümesi içerisinde 34 adet zamansal ve 48 adet mekânsal olmak üzere toplamda 82 faktör bulunmaktadır (**Tablo 4.1**).

Tablo 4.1: Çalışma veri kümesi.

Zamansal Faktör	Mekânsal Faktör	
- Haftanın Günleri (Pazartesi - Pazar) - Hafta sonu / Hafta içi - Yılın Ayları (Ocak-Aralık) - Mevsimler (Kış-Sonbahar) - Yaz Tatili - Üniversite Ders Dönemi - Eğitim-Öğretim Takvimi - Hava Muhalefeti Olan Günler - Bayram (Milli ve Dini) Günleri - Resmî Tatil Günleri	Kentsel İlgi Noktası	Demografik Yapı
	- Şehir Dışı Aktarma İstasyonu - Hastane - Devlet Hastanesi - Okul - Üniversite - Otel - Sanayi Merkezi - Restoran - Alışveriş Merkezi - Çarşı Merkezi - Sahil - Vapur - Havalimanı - Hızlı Ulaşım Modu - Otobüs Durağı - Taksi Durağı - Sağlık Merkezi - Eğitim Merkezi - Kurum Binası - Rekreasyon Alanı - Dini Tesis - Farklı Ulaşım Modu - Eğlence Tesis - Yeme İçme Alanı - Spor Tesis	- Nüfus - Nüfus Yoğunluğu - 15-29 Yaş Aralığında Nüfus - 15-29 Yaş Aralığında Nüfus Oranı - Hane Halkı Büyüklüğü - Ortalama Hane Halkı Büyüklüğü - Alım Gücü İndeksi - İşsizlik Oranı

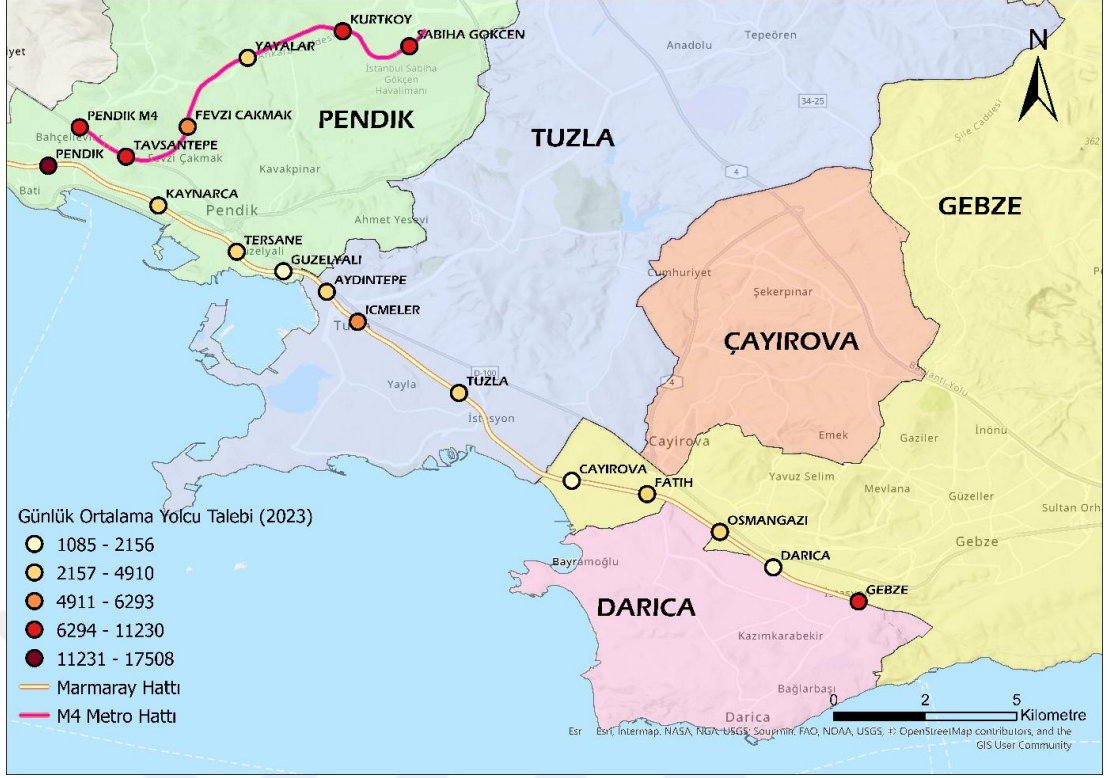
Çalışma kapsamında gerçekleştirilecek tahmin için bir denetimli öğrenme metodu kullanılacağından mekânsal ve zamansal faktörlere karşılık gelen çıktı değerlerine ihtiyaç duyulmaktadır. Oluşturulacak tahmin modeli için 2023 yılına ait istasyon bazlı geçmiş yolcu biniş verileri kullanılmıştır. 2023 yılına ait M4 Metro ve Marmaray banliyö treni istasyonları için geçmiş yolcu verileri İstanbul Büyükşehir Belediyesi Açık Veri Portalı üzerinden temin edilmiştir [**Web 10, 2024**]. Veri seti içerisinde her bir kayda ait özgün numara, tarih bilgisi, gün bilgisi, saat bilgisi, ulaşım aracının

kullandığı yol tipi, kullanılan hat, yolcuların giriş yaptığı istasyon bilgisi ve ilgili saat için belirli bir istasyona giriş yapan toplam yolcu sayısı verileri bulunmaktadır. Veri setinde bulunan yolcu sayısı verisi her bir istasyona giriş yapan yolcuların saatlik olarak toplam sayısını vermektedir. Ancak bu çalışmada gerçekleştirilecek tahmin günlük yolcu talebine yönelik olduğu için saatlik bazda tutulmakta olan yolcu verileri Excel ortamında günlük olacak şekilde düzenlenmiştir (**Tablo 4.2**).

Tablo 4.2: Günlük yolcu sayısı verisi.

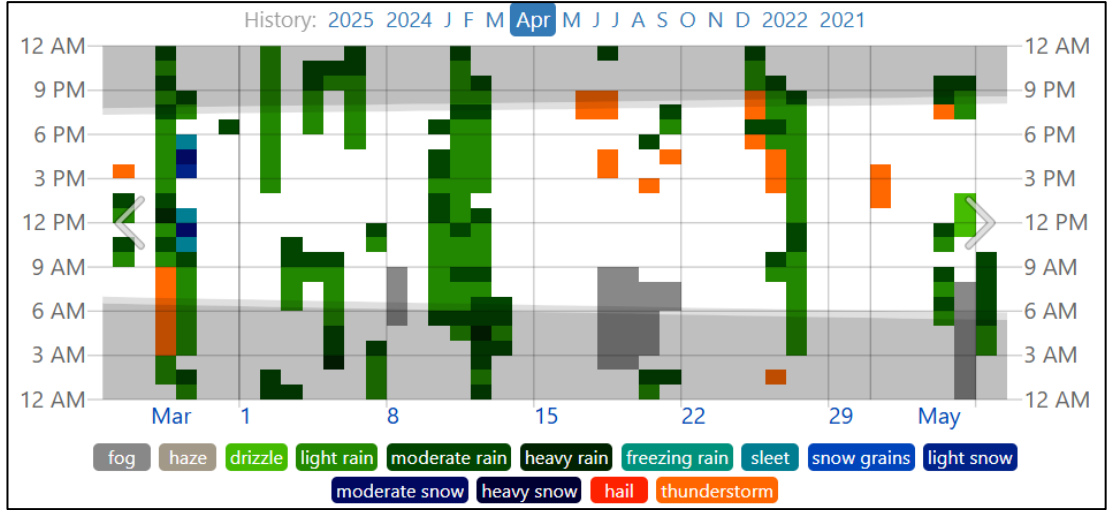
id	tarih	gun	yol_tipi	hat	istasyon	yolcu_sayisi
1	2023-01-01	PAZAR	RAYLI	Halkalı-GEBZE	GEBZE	12104
2	2023-01-02	PAZARTESI	RAYLI	Halkalı-GEBZE	GEBZE	10228
3	2023-01-03	SALI	RAYLI	Halkalı-GEBZE	GEBZE	9088
⋮	⋮	⋮	⋮	⋮	⋮	⋮
6570	2023-12-31	PAZAR	RAYLI	KADIKOY-SABIHAGOKCEN HAVALIMANI	SABIHA GOKCEN	9357

Günlük toplam yolcu talebini içeren veri seti, her bir istasyondaki ortalama günlük talebi yansıtacak şekilde harita üzerinde gösterilmiştir (**Şekil 4.3**). Böylece 2023 yılı için istasyonların günlük ortalama ne kadar yolcu çektikleri kolay bir şekilde yorumlanabilir hale getirilmiştir. Oluşturulan haritada M4 Metro hattına ait istasyonların Marmaray banliyö treni hattına ait istasyonlara kıyasla daha çok yolcu çektiği görülmektedir. Ancak istasyon odaklı bakıldığında en çok yolcu çeken istasyonun Marmaray banliyö treni hattına ait PENDİK istasyonu olduğu görülmektedir.



Şekil 4.3: Marmaray ve M4 Metro istasyonları için 2023 yılı günlük ortalama yolcu talebi.

Zamansal faktörlere ilişkin veriler, hem 2023 hem de 2024 yılı dikkate alınarak açık veri kaynaklarından toplanmıştır. Toplanan zamansal verilerden 2023 yılına ait olanlar tahmin modelinin eğitilmesinde ve 2024 yılına ait olanlar yeni tahminlerin yapılmasında kullanılmıştır. Yaz tatili faktörü, ilk ve ortaöğretim eğitim takvimleri dikkate alınarak Türkiye Millî Eğitim Bakanlığı web sayfasından elde edilmiştir [Web 11, 2025], [Web 12, 2025]. Üniversite ders dönemi faktörü için istasyon noktalarına en yakın devlet üniversitesi olan Gebze Teknik Üniversitesi akademik takvimi dikkate alınmıştır [Web 13, 2025], [Web 14, 2025]. Hava muhalefeti faktörü için weatherspark web sayfasından Sabiha Gökçen Havalimanı meteoroloji istasyonundan elde edilen 2023 ve 2024 yıllarına ait günlük veriler arasından yoğun kar yağışının, yoğun yağmur yağışının ve fırtınanın yaşanmış olduğu günler ilgili grafikler aracılığıyla tespit edilmiştir [Web 15, 2025]. Şekil 4.4'te Sabiha Gökçen Havalimanı meteoroloji istasyonuna yönelik weatherspark web sayfasından elde edilen örnek bir grafik çıktısı görülmektedir. Bayram günleri ve resmî tatillere ait veriler, takvim bilgilerinin paylaşılmış olduğu web sayfasından elde edilmiştir [Web 16, 2025].



Şekil 4.4: Sabiha Gökçen Havalimanı 2023 yılı Nisan ayında gerçekleşen hava olayları.

Mekânsal faktörlerin bir kısmı açık veri kaynaklarından elde edilmiş ve bir kısmı Esri firması üzerinden temin edilmiştir (**Tablo 4.3**). Mekânsal faktörler genel itibarıyla kentsel ilgi noktaları ve demografik yapı özelliklerini içermektedir. Kentsel ilgi noktası faktörleri arasında bulunan Şehir Dışı Aktarma İstasyonu faktörü, bir istasyonun otobüs veya Yüksek Hızlı Tren aracılığıyla şehir dışına aktarma yapıp yapılamayacağı dikkate alınarak veri kümesine dahil edilmiştir. Şehir Dışı Aktarma İstasyonu faktörü dışındaki tüm kentsel ilgi noktası faktörleri istasyonun çevresiyle ilişkilidir. Kentsel ilgi noktası faktörlerinin bir kısmı için her bir istasyonun hizmet alanı alanındaki ilgili faktörün sayısı dikkate alınırken bir diğer kısmı için yürüme mesafesine dayalı mesafeler dikkate alınmıştır. Demografik yapıya ait veriler içinse yine istasyon hizmet alanları dikkate alınmıştır.

Tablo 4.3: Mekânsal faktörler veri kaynakları.

Mekânsal Faktör	Veri Kaynağı
- Şehir Dışı Aktarma İstasyonu, Hastane, Devlet Hastanesi, Okul, Üniversite, Otel, Sanayi Merkezi, Alışveriş Merkezi, Çarşı Merkezi, Sahil, Vapur, Havalimanı, Hızlı Ulaşım Modu, Otobüs Durağı, Taksi Durağı, Restoran	OpenStreetMap
- Sağlık Merkezi, Eğitim Merkezi, Kurum Binası, Rekreasyon Alanı, Dini Tesis, Farklı Ulaşım Modu, Eğlence Tesis, Yeme İçme Alanı, Spor Tesis, Nüfus, Nüfus Yoğunluğu, 15-29 Yaş Aralığında Nüfus, 15-29 Yaş Aralığında Nüfus Oranı, Hane Halkı Büyüklüğü, Ortalama Hane Halkı Büyüklüğü, Alım Gücü İndeksi, İşsizlik Oranı	Esri

Zamansal faktörlerin makine öğrenmesi yöntemlerinde kullanımı için veri hazırlığı aşamasında ikili (boolean) ve one-hot encoding yöntemleri izlenmiştir. Mekânsal faktörlerin veri setine dahil edilmesinde ise mekânsal analiz tekniklerine dayalı sürekli (continuous) veri tipindeki çıktılar kullanılmıştır.

4.3. Mekânsal Analizlerin Gerçekleştirilmesi

Yolcu talep tahmini için kurulacak tahmin modellerin doğruluğunun yüksek ve yorumlanabilir olması büyük ölçüde veri setine dayalıdır. Toplu taşıma odaklı bir yolcu talep tahmininde hem zamansal hem de mekânsal faktörlerin birlikte değerlendirilmesi talepte yaşanan günlük değişimlerin ve istasyon özelinde baz yolcu talebinin yüksek doğrulukla tahmininde oldukça önemlidir. Toplu taşıma sistemlerine ait istasyonlar için yolcu talebini oluşturan mekânsal faktörlerin veri setine doğru şekilde dahil edilmesi her bir istasyona olan genel yolcu talebinin anlaşılmasını kolaylaştırmaktadır. Bu doğrultuda mekânsal faktörler için elde edilmiş veri setine mekânsal analizlerin uygulanması gerekmektedir.

CBS, nokta, çizgi ve çokgen tipindeki coğrafi verileri toplama, kaydetme ve işleme yetenekleri ile mekânsal problemlerin çözümünde büyük bir rol oynamaktadır. Mekânsal problemlerin çözümü ve daha iyi mekânsal kararların verilebilmesi için çeşitli CBS yazılımları bulunmaktadır. Bu kapsamda oldukça yaygın bir biçimde kullanılan Esri firmasına ait ticari yazılım ArcGIS, coğrafi verileri entegre biçimde

kullanarak, verilerin yönetilmesi, analiz edilmesi ve haritalanması için önemli yetenekleri sunmaktadır [Web 17, 2025]. Çalışma kapsamında mekânsal faktörlere ait coğrafi verilere çeşitli coğrafi analizlerin uygulanması ve görselleştirilmesi amaçları doğrultusunda ArcGIS yazılımı içerisindeki ağ analizleri kullanılmıştır.

4.3.1. Ağ Analizleri

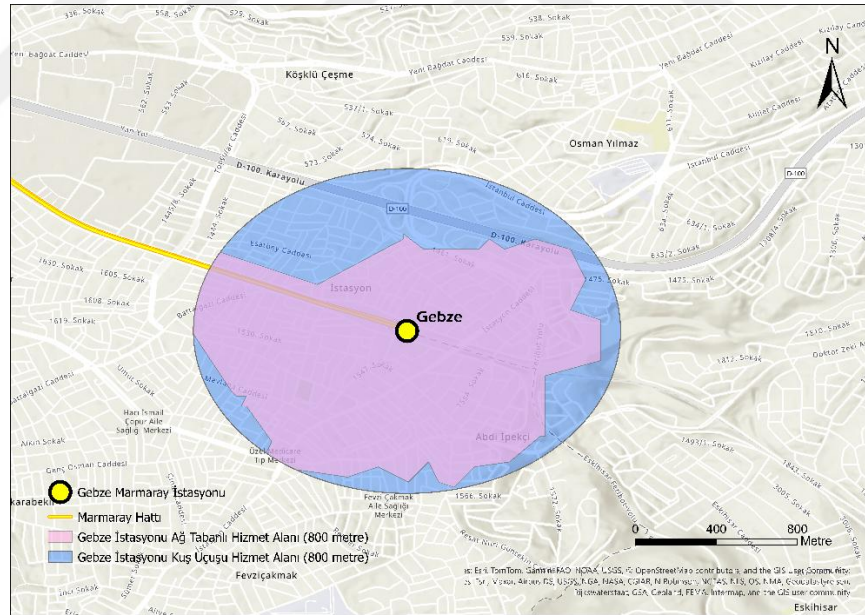
İnsanlar, taşıtlar ve kaynaklar ağlar boyunca hareket etmektedir. Bu ağlar, araba ve kamyon gibi taşıtlar için karayolunu, uçaklar için uçuş rotalarını ve petrol gibi kaynaklar için boru hatlarını ifade etmektedir. Tüm bu hareketler bir ağ yapısı ile modellenerek iki nokta arasındaki en kısa yolun bulunması gibi çeşitli ağ analizleri gerçekleştirilebilmektedir. Burada ifade edilen ağ, noktalardan ve bu noktaları birbirine bağlayan rotaları temsil eden kenarlardan oluşan bir sistemdir. Ağ sistemi sayesinde yol bağlantıları kenarlar olarak ifade edilebilmekte ve bu sayede bir noktadan başka bir noktaya gerçekleştirilecek seyahatlerde gerçek sokak yapıları kullanılabilir. Bu özelliği ile klasik metotlarda kullanılan Öklid mesafesi yönteminde kıyasla gerçek seyahat mesafesi ve gerçek seyahat süresi hesaplanabilmektedir. CBS yazılımlarında ağ analizi, tanımlanan bir ağ sistemi üzerinde gerçekleştirilen matematiksel analizleri ifade etmektedir. ArcGIS yazılımı içerisinde bulunan ağ analizleri aracı ile rota oluşturma, en yakın tesis, hizmet alanı, konum tahsisi ve başlangıç-varış matrisi analizleri gerçekleştirilebilmektedir [Web 18, 2025].

Rota oluşturma analizi ile, bir başlangıç noktasından bir veya birden fazla varış noktasına olan en iyi rotanın oluşturulması sağlanabilmektedir. En iyi rota, belirli bir günün belirli bir saatine göre geçerli trafik koşullarının da dikkate alınabilmesi ile en hızlı şekilde gidilebilmesini sağlayan rota olabileceği gibi seyahat mesafesini en fazla kısaltan rota da olabilmektedir. En yakın tesis analizi, olaylar ile tesisler arasındaki en uygun rotaları göstermekte ve seyahat maliyetlerini raporlamaktadır. Hizmet alanı analizi, belirli bir noktadan belirli bir seyahat maliyetinde erişilebilir alanların oluşturulmasını sağlamaktadır. Konum tahsisi analizi, belirli bir mal ve hizmet sağlayan tesisler ile bunları tüketen bir dizi talep noktasının bulunması durumunda tesislerin talep noktalarını en verimli şekilde karşılayacak şekilde konumlandırılmasını sağlamaktadır. Başlangıç-varış matrisi analizi, ağ üzerinde birden fazla başlangıç noktasından birden fazla varış noktasına en düşük maliyetli rotaların oluşturulmasını

ve maliyetlerinin raporlanmasını sağlamaktadır [Web 19, 2025]. Çalışma kapsamında mekânsal faktörlere ait elde edilen coğrafi veriler için istasyon noktaları dikkate alınarak hizmet alanı ve en yakın tesis analizleri uygulanmıştır.

4.3.1.1. Hizmet Alanı ve En Yakın Tesis Analizi

Hizmet alanı analizi, bir noktadan belirli bir süre veya mesafe için ne kadarlık bir alana erişilebileceğinin bulunmasını sağlamaktadır. Toplu taşıma sistemlerine ait istasyonlar için hizmet alanlarının oluşturulmasında ağ analizlerinin kullanılması, yürüme veya araçla yapılacak seyahatler için klasik yöntemlerdeki Öklid mesafe yerine gerçek yol bağlantılarının kullanılmasını sağlamaktadır. Gerçek yol bağlantılarının kullanılması analize girilen mesafe veya süre için yol ağı boyunca erişilebilen yolları kapsayan bir hizmet alanı elde edilmesini sağlamaktadır [Web 19, 2025]. Şekil 4.5'te yol bağlantıları kullanılarak yapılan hizmet alanı ve Öklid mesafesine dayalı oluşturulan tampon bölge arasındaki fark karşılaştırılmıştır.



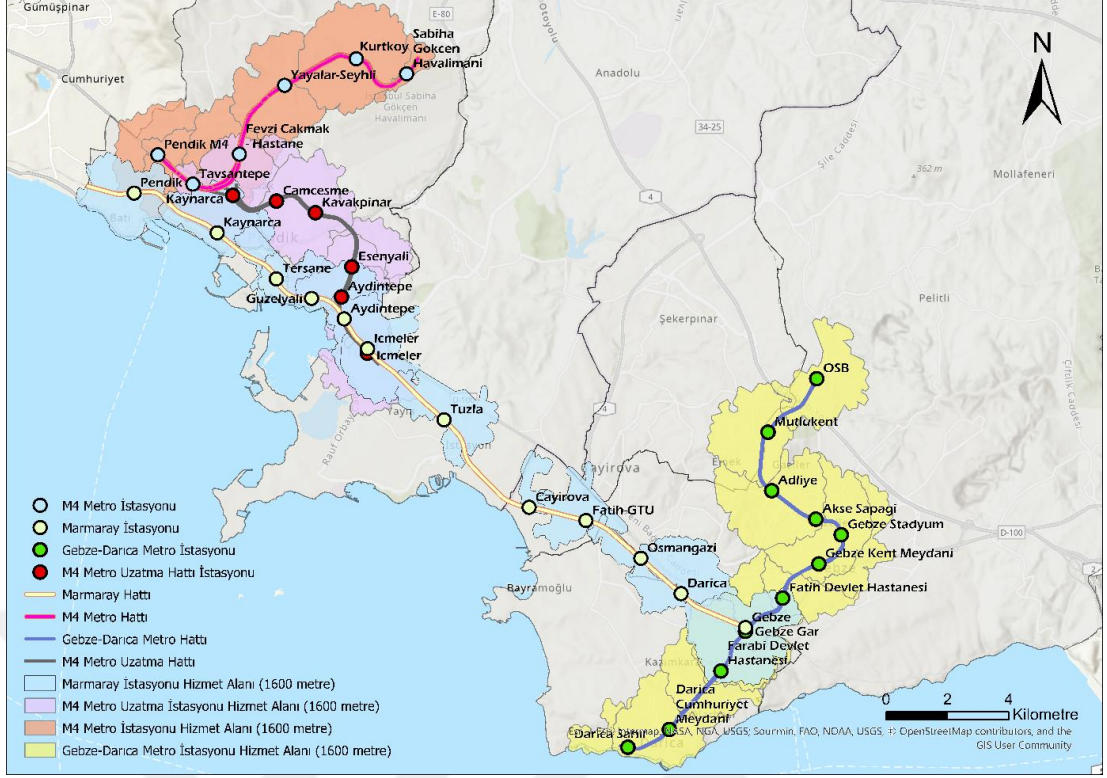
Şekil 4.5: Ağ tabanlı ve Öklid mesafeye dayalı hizmet alanı karşılaştırması.

Her bir istasyon için hizmet alanları, çalışma sınırları içerisindeki her bir metro ve banliyö treni istasyonları için uygulanmıştır. Hizmet alanı, metro ve banliyö trenleri gibi hızlı toplu taşıma araçları için 800 metrelik bir alan olarak kabul edilmektedir [Ryus, 2013]. Bu çalışmada öncelikle Şekil 4.6'da görüldüğü üzere her bir istasyon için 800 metrelik hizmet alanları oluşturulmuştur. Daha sonrasında yolcu talebinin

daha geniş bir hizmet alanı için de yorumlanabilmesi amacıyla her bir istasyon için 1600 metrelik hizmet alanları **Şekil 4.7**'de görüldüğü şekilde oluşturulmuştur. Çalışmada hizmet alanı analizi, her bir istasyon için çalıştırılmış ve çokgen biçimindeki hizmet alanı içerisindeki ilgili kentsel ilgi noktasının sayısı ve demografik yapı bilgileri veri setine eklenmiştir.

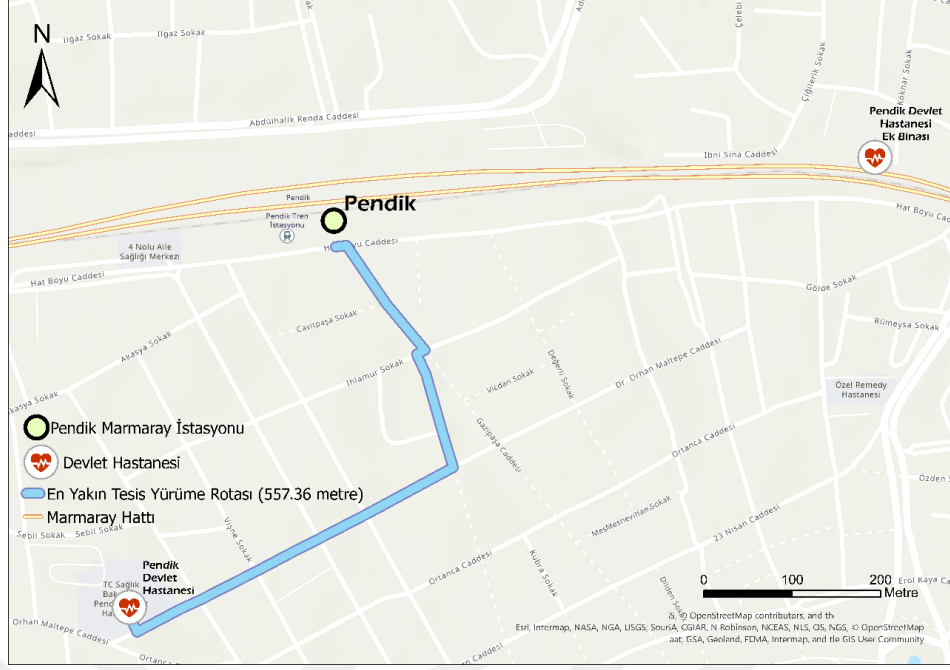


Şekil 4.6: Ağ tabanlı 800 metrelik istasyon hizmet alanları.



Şekil 4.7: Ağ tabanlı 1600 metrelik istasyon hizmet alanları.

Çalışmada kullanılan bir diğer ağ analizi, olaylar ile tesisler arasındaki seyahat maliyetini ölçerek birbirlerine en yakın olanları eşleştiren en yakın tesis analizidir. En yakın tesis analizinde kaç adet tesisin bulunacağı ve seyahat yönünün tesise doğru mu yoksa tesisten olaya doğru mu olacağı belirtilebilmektedir. Ayrıca tesisler aranırken belirli bir seyahat maliyeti sınırı gibi kısıtlar belirlenebilmektedir. Bu analiz ile olaylar ile tesisler arasındaki en uygun rotalar ve rotalara ait seyahat raporları üretilmektedir [Web 19, 2025]. Çalışmada istasyon noktaları bir olayı temsil etmekte iken kentsel ilgi noktalarını içeren mekânsal faktörlerin her birisi tesisleri temsil etmektedir. Her bir istasyon için en yakın tesis analizleri Şekil 4.8'de gösterildiği şekilde bir istasyonun çevresinde bulunan hastanelerden en yakın olanına yürüme yoluyla bir rota oluşturulmakta ve mesafesi çıktı olarak verilmektedir. Çalışmada ilgili noktaya olan kilometre biriminde yürüme mesafesi kullanılmıştır.



Şekil 4.8: En yakın tesis analizi çıktısı.

Tablo 4.4'te mekânsal faktörler içerisindeki kentsel ilgi noktaları ve demografik yapıya ait verilerin hangi analiz için kullanılmış olduğu verilmiştir. Hizmet alanı ve en yakın tesis analizleri sonucunda elde edilen veri setinin sütunlarında mekânsal faktörün adı ve satırlarda mekânsal faktör için analiz sonucu elde edilen sayısal ifadeler yer almaktadır (**Tablo 4.5**).

Tablo 4.4: Mekânsal faktörler ve kullanıldıkları analiz yöntemi.

Hizmet Alanı Analizi			En Yakın Tesis Analizi
Kentsel İlgi Noktası	Demografik Yapı		Kentsel ilgi Noktası
Otobüs Durağı, Taksi Durağı, Sağlık Merkezi, Eğitim Tesisi, Kurum Binası, Rekreasyon Alanı, Dini Tesis, Farklı Ulaşım Modu, Eğlence Tesisi, Restoran, Yeme İçme Alanı, Spor Tesisi	Nüfus, Hane Halkı, Ortalama Hane Halkı, Büyüklüğü, Nüfus Yoğunluğu, Alım Gücü, İndeksi, 15-29 Yaş Aralığında Nüfus, 15-29 Yaş Aralığında Nüfus Oranı, İşsizlik Oranı		Hastane, Devlet Hastanesi, Okul, Üniversite, Otel, Sanayi Merkezi, Restoran, Alışveriş Merkezi, Çarşı Merkezi, Sahil, Vapur, Havalimanı ve Hızlı Ulaşım Modu

Tablo 4.5: Mekânsal analizler sonucunda oluşturulan veri seti.

	Farklı Ulaşım Modu (800 metre hizmet alanı)	Nüfus (1600 metre hizmet alanı)	...	Hastane (metre)
Gebze	2	44077		1.5
Darica	1	60953		2.6
⋮	⋮	⋮	⋮	⋮
Yayalar	2	56578		2.1

4.4. Özellik Seçimi Sonuçları

Toplu taşıma sistemleri için gerçekleştirilecek tahmin modelleri makine öğrenmesi metotları ile oluşturulacaktır. Bu sebeple makine öğrenmesi yöntemlerinin başarısını etkileyen etkenlerden birisi olan ilgisiz ve fazlalık olan özelliklerin (faktörlerin) elenmesi yoluyla alt küme oluşturulması gerekmektedir. Bu sayede makine öğrenmesi metotları ile oluşturulan tahmin modellerindeki çoklu doğrusallık sorununun önüne geçilebilmektedir. Çalışma kapsamında mekânsal ve zamansal toplamda 82 faktör içeren veri kümesi içerisinde özellik seçimi metotları ile alt kümeler elde edilmiştir. Özellik seçimi için filtreleme metotlarından Pearson Korelasyon ve Sarmalayıcı metotlardan Ardışık İleri Yönde Seçim ve Ardışık Geri Yönde Seçim uygulanmıştır.

4.4.1. Pearson Korelasyon Sonuçları

Çalışma kapsamında veri setinde bulunan faktörlerin birbirleriyle ve çıktı değişkenle olan ikili doğrusal ilişkilerinin ölçülmesi amacıyla Pearson Korelasyon kullanılmıştır. Pearson Korelasyon yöntemi ile faktörler arasından fazlalık olanlar ile çıktı değişkenle düşük korelasyona sahip ilgisiz faktörler elenmiştir. Bu kapsamda korelasyon matrisi *R* yazılımı ortamında oluşturulmuştur. Oluşturulan matriste sütunlar ve satırlar faktörlerden ve çıktı değişken olan yolcu sayısından oluşmaktadır. Matriste bulunan her bir hücre -1 ile +1 arasında değişen korelasyon katsayısı değerlerini içermektedir. Ancak matrisin yorumlanabilmesinin daha kolay olması açısından korelasyon katsayısı değerleri mutlak değer şeklinde tekrar düzenlenerek her bir hücre için 0 ile +1 arasında değişen korelasyon katsayıları elde edilmiştir. Ayrıca faktörlerin isimlerinin matris üzerinde gösteriminin daha kolay olması amacıyla faktörler **Tablo 4.6**'da gösterilen kod adları ile ifade edilmiştir.

Tablo 4.6: Faktör isimleri için belirlenen kod adları.

f1 = yolcu sayısı f2 = pazartesi f3 = salı f4 = çarşamba f5 = perşembe f6 = cuma f7 = cumartesi f8 = pazar f9 = hafta içi f10 = ocak f11 = şubat f12 = mart f13 = nisan f14 = mayıs f15 = haziran f16 = temmuz f17 = ağustos f18 = eylül f19 = ekim f20 = kasım f21 = aralık f22 = kış f23 = ilkbahar f24 = yaz f25 = sonbahar f26 = üniversite ders dönemi f27 = eğitim-öğretim takvimi f28 = yaz tatili f29 = hava muhalefeti olan günler f30 = milli bayram f31 = ramazan ayı f32 = dini bayram f33 = bayram f34 = resmî tatil f35 = şehir dışı aktarma durağı	f36 = hastane (yürüme mesafesi) f37 = devlet hastane (yürüme mesafesi) f38 = okul (yürüme mesafesi) f39 = üniversite (yürüme mesafesi) f40 = otel (yürüme mesafesi) f41 = sanayi merkezi (yürüme mesafesi) f42 = restoran (yürüme mesafesi) f43 = alışveriş merkezi (yürüme mesafesi) f44 = çarşı merkezi (yürüme mesafesi) f45 = sahil (yürüme mesafesi) f46 = vapur (yürüme mesafesi) f47 = havalimanı (yürüme mesafesi) f48 = hızlı ulaşım modu (yürüme mesafesi) f49 = otobüs durağı (800 metre hizmet alanı) f50 = taksi (800 metre hizmet alanı) f51 = sağlık merkezi (800 metre hizmet alanı) f52 = eğitim merkezi (800 metre hizmet alanı) f53 = kurum binası (800 metre hizmet alanı)	f54 = rekreasyon alanı (800 metre hizmet alanı) f55 = dini tesis (800 metre hizmet alanı) f56 = farklı ulaşım modu (800 metre hizmet alanı) f57 = eğlence merkezi (800 metre hizmet alanı) f58 = restoran (800 metre hizmet alanı) f59 = yeme içme alanı (800 metre hizmet alanı) f60 = spor tesisi (800 metre hizmet alanı) f61 = nüfus (800 metre hizmet alanı) f62 = hane halkı (800 metre hizmet alanı) f63 = ortalama hane halkı büyüklüğü (800 metre hizmet alanı) f64 = nüfus yoğunluğu (800 metre hizmet alanı) f65 = alım gücü indeksi (800 metre hizmet alanı) f66 = 15-29 yaş arası nüfus (800 metre hizmet alanı) f67 = 15-29 yaş arası nüfus oranı (800 metre hizmet alanı) f68 = işsizlik oranı (800 metre hizmet alanı) f69 = sağlık merkezi (1600 metre hizmet alanı)	f70 = eğitim merkezi (1600 metre hizmet alanı) f71 = kurum binası (1600 metre hizmet alanı) f72 = rekreasyon alanı (1600 metre hizmet alanı) f73 = dini tesis (1600 metre hizmet alanı) f74 = farklı ulaşım modu (1600 metre hizmet alanı) f75 = eğlence merkezi (1600 metre hizmet alanı) f76 = restoran (1600 metre hizmet alanı) f77 = yeme içme alanı (1600 metre hizmet alanı) f78 = spor tesisi (1600 metre hizmet alanı) f79 = nüfus (1600 metre hizmet alanı) f80 = hane halkı (1600 metre hizmet alanı) f81 = nüfus yoğunluğu (1600 metre hizmet alanı) f82 = 15-29 yaş aralığında nüfus (1600 metre hizmet alanı) f83 = 15-29 yaş aralığında nüfus oranı (1600 metre hizmet alanı)
---	---	--	---

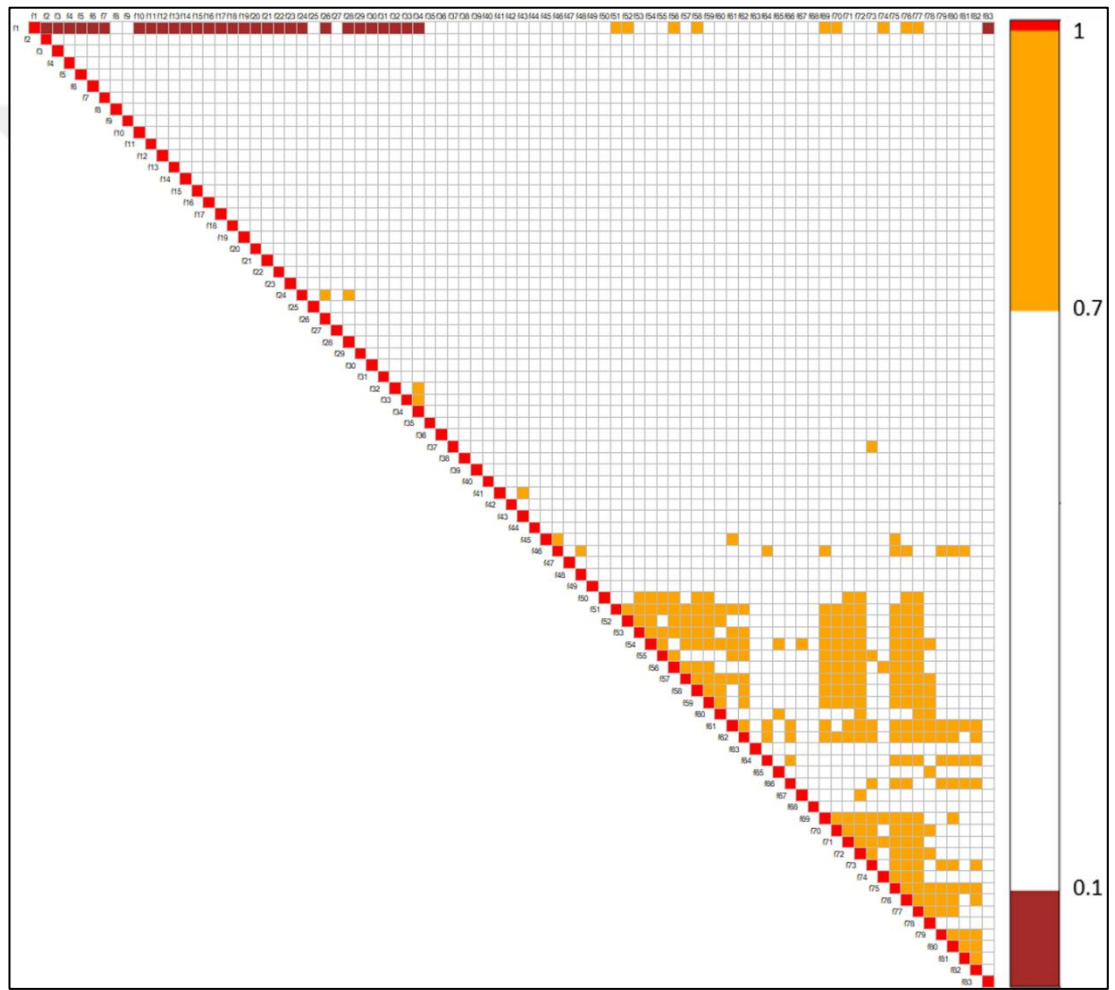
Faktörlerin fazlalık olup olmadığının kararının verilmesinde güçlü ilişkiyi ifade edecek belirli bir eşik değeri belirlenmektedir. Pearson Korelasyon ile yapılan özellik seçimi çalışmalarında 0.5, 0.6 ve 0.7 gibi farklı eşik değerleri kullanılmıştır [Şişman ve Aydınoglu, 2023], [Aksangür vd., 2022], [Ghandeharioun vd., 2023]. Bu çalışmada faktörlerin fazlalığının testi amacıyla 0.7 eşik değeri olarak belirlenmiştir. Korelasyon katsayısı 0.7 ve üzeri olan her iki faktörden bir tanesi manuel biçimde elimine edilmiştir. Eliminasyon için oluşturulan korelasyon matrisi **Şekil 4.9**'da

Tablo 4.7: Pearson Korelasyon fazlalık yaklaşımı ile faktör seçimi.

Zamansal Faktör (31 faktör)	Mekânsal Faktör (20 faktör)
Haftanın Günleri (Pazartesi-Pazar)	Şehir Dışı Aktarma Durağı
Hafta içi	Hastane (yürüme mesafesi)
Yılın Ayları (Ocak-Aralık)	Devlet Hastanesi (yürüme mesafesi)
Mevsimler (Kış, İlkbahar ve Sonbahar)	Okul (yürüme mesafesi)
Yaz Tatili	Üniversite (yürüme mesafesi)
Üniversite Ders Dönemi	Otel (yürüme mesafesi)
Eğitim-Öğretim Takvimi	Sanayi Merkezi (yürüme mesafesi)
Hava Muhalefeti Olan Günler	Restoran (yürüme mesafesi)
Milli Bayram Günleri	Çarşı Merkezi (yürüme mesafesi)
Dini Bayram Günleri	Sahil (yürüme mesafesi)
Ramazan Ayı Günleri	Havalimanı (yürüme mesafesi)
Bayram Günleri (Milli ve Dini)	Hızlı Ulaşım Modu (yürüme mesafesi)
	Otobüs Durağı (800 metre hizmet alanı)
	Ortalama Hane Halkı Büyüklüğü (800 metre hizmet alanı)
	Alım Gücü İndeksi (800 metre hizmet alanı)
	15-29 Yaş Aralığında Nüfus (800 metre hizmet alanı)
	15-29 Yaş Aralığında Nüfus Oranı (800 metre hizmet alanı)
	İşsizlik Oranı (800 metre hizmet alanı)
	Farklı Ulaşım Modu (1600 metre hizmet alanı)
	15-29 Yaş Aralığında Nüfus Oranı (1600 metre hizmet alanı)

Pearson Korelasyon ile özellik seçimi için ikinci bir yaklaşım olarak, fazlalık ve ilgililiğin birlikte değerlendirilmesi amacıyla bir özellik seçimi uygulanmıştır. Bu

kısımda hem korelasyon değeri 0.7'den büyük faktör çiftlerinden birisi elenmiş hem de çıktı değişken olan yolcu sayısı ile düşük korelasyona sahip faktörler elimine edilmiştir. Düşük korelasyon için eşik değer olarak korelasyon katsayısı 0.1 olarak kabul edilmiştir. Böylece hem fazlalık hem de ilgililiğin bir arada değerlendirildiği **Şekil 4.10**'da bulunan korelasyon matrisi meydana gelmiştir. Excel ortamına aktarılan matris üzerinde yolcu sayısı ile belirlenen eşik değerin altında olan faktörler tamamen veri setinden çıkarılmıştır. Ardından fazlalık değerlendirmesi daha önce yapıldığı şekliyle tekrar edilmiştir.



Şekil 4.10: Pearson Korelasyon fazlalık ve ilgililik yaklaşımına göre hazırlanan matris.

Özellik seçimi sonucunda 82 faktör arasından 24 faktör ile yeni bir alt küme elde edilmiştir (**Tablo 4.8**). Oluşturulan bu alt küme ile hem fazlalık olan faktörler elenmiş hem de yolcu sayısı ile düşük korelasyon gösteren faktörler kümeden çıkarılmıştır.

Tablo 4.8: Pearson Korelasyon fazlalık ve ilgililik yaklaşımı ile faktör seçimi.

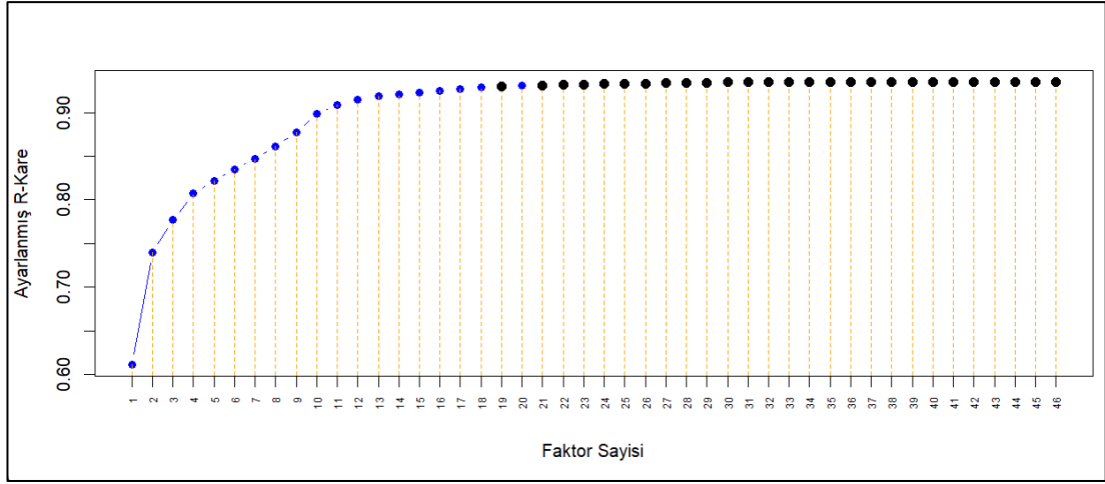
Zamansal Faktör (4 faktör)	Mekânsal Faktör (20 faktör)
Pazar	Şehir Dışı Aktarma Durağı
Hafta İçi	Hastane (yürüme mesafesi)
Sonbahar	Devlet Hastanesi (yürüme mesafesi)
Eğitim-Öğretim Takvimi	Okul (yürüme mesafesi)
	Üniversite (yürüme mesafesi)
	Otel (yürüme mesafesi)
	Sanayi Merkezi (yürüme mesafesi)
	Restoran (yürüme mesafesi)
	Çarşı Merkezi (yürüme mesafesi)
	Sahil (yürüme mesafesi)
	Havalimanı (yürüme mesafesi)
	Hızlı Ulaşım Modu (yürüme mesafesi)
	Otobüs Durağı (800 metre hizmet alanı)
	Ortalama Hane Halkı Büyüklüğü (800 metre hizmet alanı)
	Alım Gücü İndeksi (800 metre hizmet alanı)
	15-29 Yaş Aralığında Nüfus (800 metre hizmet alanı)
	15-29 Yaş Aralığında Nüfus Oranı (800 metre hizmet alanı)
	İşsizlik Oranı (800 metre hizmet alanı)
	Farklı Ulaşım Modu (1600 metre hizmet alanı)
	15-29 Yaş Aralığında Nüfus Oranı (1600 metre hizmet alanı)

4.4.2. Ardışık İleri Yönde Seçim ve Ardışık Geri Yönde Seçim Sonuçları

Özellik seçiminde, Pearson Korelasyon gibi Filtre metotlarının kullanılması çok az hesaplama maliyeti gerektirmesi gibi avantajlı bir durumun yanında faktörler arasındaki yalnızca doğrusal ilişkilerin dikkate alınması gibi bir dezavantaja sahiptir. Ayrıca Pearson Korelasyon yöntemiyle elde edilen alt kümede faktörlerin tahmin

doğruluğuna olan katkıları göz ardı edilebilmektedir. Bu sebeple faktörlerin tahmin doğruluğuna olan katkılarının da değerlendirilebilmesi amacıyla, makine öğrenmesi algoritmalarını kullanarak özellik seçimi yapan Sarmalayıcı metotlardan Ardışık İleri Yönde Seçim ve Ardışık Geri Yönde Seçim metotlarının kullanımı gerçekleştirilmiştir.

Ardışık İleri Yönde Seçim ve Ardışık Geri Yönde Seçim ile özellik seçimi, R yazılımı ortamında ileri ve geri yönde özellik seçimi yapılabilmesini sağlayan *leaps* kütüphanesi kullanılarak gerçekleştirilmiştir. Ardışık İleri Yönde Seçim için boş bir alt küme oluşturulmuş ve her adımda Çoklu Doğrusal Regresyon yönteminde Ayarlanmış R^2 değerine göre model performansını en çok artıran faktörün alt kümeye eklenmesi sağlanmıştır. Metot uygulanırken toplamda 36 faktörün çok yüksek doğrusal ilişki barındırdığı dolayısıyla orijinal kümeden çıkarıldığı uyarısı alınmış ve bu sebeple geri kalan 46 faktör üzerinden özellik seçimi uygulanmıştır. İlk faktör eklendikten sonra diğer faktörler arasından ilk faktörün de değerlendirilmesiyle model doğruluğunu en çok artıran ikinci faktörün alt kümeye eklenmesi sağlanmıştır. Bu şekilde teker teker tüm faktörlerden model doğruluğunu en çok artıranlar sırayla eklenmiştir. Bu yöntemde model doğruluğu için Ayarlanmış R^2 (Adjusted R-squared) metriği kullanıldığından dolayı her yeni bir faktörün eklenmesi durumunda model doğruluğu daha da artıyor gibi görünmektedir. Yeni faktörler eklendikçe anlamlı bir doğruluk artışı yaşanmıyorsa eklenen her yeni faktör modelin aşırı uyum yapmasına sebep olabilmektedir. Bu sebeple Ayarlanmış R^2 ' için doğruluk artışının 0.001'den küçük olması durumunda özellik seçimi algoritmasının durdurulması sağlanmıştır. **Şekil 4.11**'de 46 faktör üzerinden gerçekleştirilen özellik seçiminde 21'inci faktörde algoritma durdurulmuştur. Böylece 82 faktör arasından 21 faktörden oluşan yeni bir alt küme oluşturulmuştur.



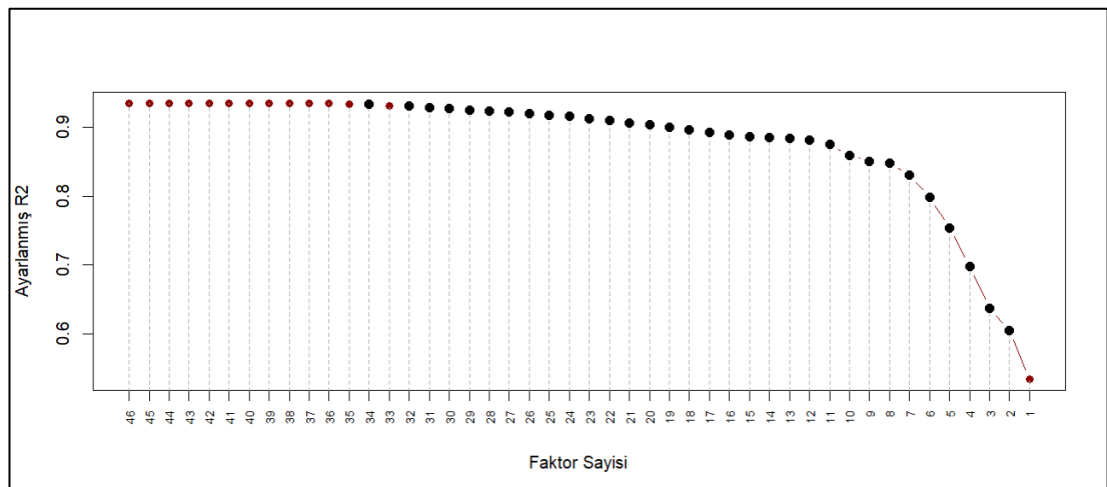
Şekil 4.11: Ardışık İleri Yönde Seçim durma noktası tespiti.

Ardışık İleri Yönde Seçim ile yapılan özellik seçimi sonucunda **Tablo 4.9**'da belirtilen faktörleri içeren alt küme elde edilmiştir. Ayrıca Tablo 1'de belirtilen adım numaraları faktörlerin boş veri kümesine eklenme sıralarını göstermektedir. İlk adımda boş bir veri kümesine eklenen Devlet Hastanesi faktörü, Doğrusal Regresyon yöntemine göre Ayarlanmış R^2 metriği değerini en fazla artıran faktör anlamını taşımaktadır.

Tablo 4.9: Ardışık İleri Yönde Seçim sonucunda elde edilen alt küme.

Adım	Eklenen Faktör	Adım	Eklenen Faktör
1	Devlet Hastanesi (yürüme mesafesi)	12	Restoran (1600 metre hizmet alanı)
2	Otobüs Durağı (800 metre hizmet alanı)	13	Üniversite Ders Dönemi
3	15-29 Yaş Aralığında Nüfus Oranı (1600 metre hizmet alanı)	14	Ramazan Ayı
4	Farklı Ulaşım Modu (800 metre hizmet alanı)	15	Restoran (yürüme mesafesi)
5	Pazar	16	Hane halkı (1600 metre hizmet alanı)
6	Hızlı Ulaşım Modu (yürüme mesafesi)	17	Rekreasyon (800 metre hizmet alanı)
7	Taksi Durağı (800 metre hizmet alanı)	18	Eğitim Öğretim Takvimi
8	Çarşı Merkezi (yürüme mesafesi)	19	Milli Bayramlar
9	Okul (yürüme mesafesi)	20	Cuma
10	Sonbahar	21	Şubat Ayı
11	Hava Muhalefeti Olan Günler		

Ardışık Geri Yönde Seçim ile tüm faktörlerin arasından model doğruluğunu en az şekilde artıran faktörün veri kümesinden çıkarılması yoluyla özellik seçimi uygulanmıştır. Ardışık İleri Yönde Seçim ile arasındaki temel fark Ardışık Geri Yönde Seçim metodunda tüm faktörlerin bulunduğu bir veri seti ile başlanmaktadır. Her adımda çoklu doğrusal regresyon yönteminde Ayarlanmış R^2 değerine göre model performansını en az artıran faktör veri setinden çıkarılmıştır. Ardışık Geri Yönde Seçim algoritması çalıştırıldığında Ardışık İleri Yönde Seçim yönteminde olduğu gibi toplamda 36 faktör, yüksek doğrusal ilişkiye sahip olmasından dolayı veri setinden çıkarılmış ve 46 faktör üzerinden özellik seçimi uygulanmıştır. Tüm faktörlerin bulunması belirli bir doğruluk sağlıyor olsa da aşırı uyum riski taşımaktadır. Algoritma çalıştırıldığında ilk faktör elendikten sonra diğer faktörler arasından en az doğruluk artıran faktör elenmiş ve bu şekilde teker teker tüm faktörlerden model doğruluğunu en az artıranlar sırayla çıkarılmıştır. Her bir faktör çıkarıldığında model doğruluğu belirli ölçüde azalmaktayken genellenebilirliğinin artırılması için optimum alt kümeye ulaşmak istenmiştir. Bu amaç doğrultusunda Ayarlanmış R^2 metriğine yönelik bir durma noktası belirlenmiştir. Ayarlanmış R^2 için doğruluk azalışının 0.001'in üstünde olması durumunda özellik seçimi algoritmasının durdurulması sağlanmıştır. Böylece **Şekil 4.12**'de 46 faktör üzerinden gerçekleştirilen özellik seçimi sonucunda doğruluğun eşik değere göre düşmeye başladığı noktadan itibaren orijinal veri setindeki 82 faktör arasından 33 faktör ile yeni bir alt küme oluşturulmuştur (**Tablo 4.10**).



Şekil 4.12: Ardışık Geri Yönde Seçim durma noktası tespiti.

Tablo 4.10: Ardışık Geri Yönde Seçim sonucunda elde edilen alt küme.

Adım	Seçilen Faktör	Adım	Seçilen Faktör
1	Restoran (yürüme mesafesi)	18	Şehir Dışı Aktarma Durağı
2	Devlet Hastanesi (yürüme mesafesi)	19	Üniversite (yürüme mesafesi)
3	Otel (yürüme mesafesi)	20	Okul (yürüme mesafesi)
4	Havalimanı (yürüme mesafesi)	21	Otobüs Durağı (800 metre hizmet alanı)
5	Sanayi Merkezi (yürüme mesafesi)	22	Şubat Ayı
6	Hızlı Ulaşım Modu (yürüme mesafesi)	23	Ramazan Ayı
7	Hastane (yürüme mesafesi)	24	Hava Muhalefeti Olan Günler
8	Alışveriş Merkezi (yürüme mesafesi)	25	Temmuz Ayı
9	Vapur (yürüme mesafesi)	26	Ağustos Ayı
10	Taksi Durağı (800 metre hizmet alanı)	27	Haziran Ayı
11	Çarşı Merkezi (yürüme mesafesi)	28	Mart Ayı
12	Cuma	29	Ocak Ayı
13	Pazartesi	30	Mayıs Ayı
14	Salı	31	Eylül Ayı
15	Perşembe	32	İlkbahar
16	Çarşamba	33	Nisan
17	Cumartesi		

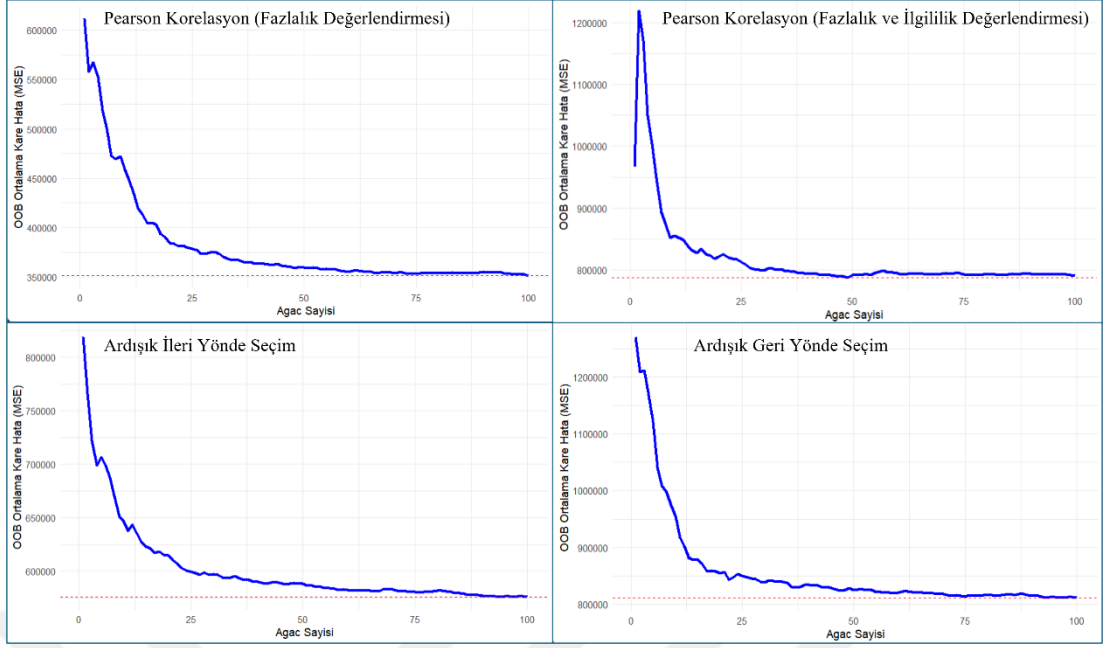
4.5. Yolcu Talep Tahmin Modellerinin Kurulması

Çalışmada tahmin modellerinin kurulumunda kullanılan veri seti, 18 istasyona ait 2023 yılı boyunca kaydedilmiş günlük yolcu talebi ile zamansal ve mekânsal faktörleri içermektedir. Veri seti toplamda 6570 gözlemden oluşmaktadır. Buradaki her bir gözlem, belirli bir istasyona ait belirli bir tarihte tek bir satırda kaydedilen verileri temsil etmektedir. Her gözlemde; tarih, istasyon adı, her bir faktöre karşılık gelen sayısal değerler ve ilgili gün için toplam yolcu sayısı yer almaktadır. Bu gözlemler kullanılarak makine öğrenmesi metotları ile tahmin modelleri kurulmuştur. Kurulacak tahmin modellerinin aşırı uyum ve yetersiz uyum arasında bir denge kurabilmesi ve daha sonrasında tahmin modelinin test edilmesi amacıyla veri seti gözlemlerinin %70'i

eğitim ve %30'u test olacak şekilde iki parçaya ayrılmıştır. Bu sayede veri setinden rastgele seçilen 4599 gözlem modellerin eğitimi için kullanılırken 1971 gözlem ile kurulan tahmin modellerinin test edilmesini sağlamıştır.

Tahmin modelleri, torbalama yaklaşımını kullanan Rastgele Orman ve güçlendirme yaklaşımını kullanan XGBoost makine öğrenmesi algoritmalarına dayalı olarak *R* yazılımı ortamında *randomForest* ve *xgboost* kütüphaneleri kullanılarak kurulmuştur. Hem Rastgele Orman hem de XGBoost algoritmaları ile tahmin modelleri kurulurken tahmin doğruluğuna etki eden birtakım hiper parametrelerin doğru bir şekilde ayarlanması gerekmektedir. Bu algoritmaların sahip oldukları hiper parametre değerleri kullanılan veri setine göre değişiklik göstermektedir. Bu sebeple çalışma kapsamında tüm faktörlerin bulunduğu orijinal veri kümesi ve özellik seçimi sonucu elde edilen 4 veri alt kümesi için farklı hiper parametre değerleri belirlenmiştir.

Çalışmada kullanılan Rastgele Orman için *R* yazılımı ortamında her bölünmede rastgele seçilecek faktör sayısı *mtry* ve ağaç sayısı *ntree* hiper parametreleri bulunmaktadır ve bu hiper parametreler çalışma kapsamında manuel biçimde ayarlanmıştır. Her bölünmede rastgele seçilecek faktör sayısı hiper parametresi, regresyon çalışmalarında varsayılan olarak da kabul edilen şekliyle kullanılan faktör sayısının 3'te biri olacak şekilde ayarlanmıştır. Ağaç sayısı ise doğası gereği ayarlanabilir bir hiper parametre olmadığından dolayı manuel biçimde her bir veri kümesi için torba dışı (out-of-bag) hata grafikleri incelenerek ayarlanmıştır. Öncelikle %70 eğitim verisi üzerinden faktör sayısına göre belirlenmiş *mtry* değeri ve 100 *ntree* değeri ile Rastgele Orman modeli çalıştırılmıştır. **Şekil 4.13**'te her bir alt küme için Rastgele Orman modelinin çalıştırılması sonucunda MSE metriğine göre elde edilen torba dışı hata grafikleri görülmektedir. Grafikler incelendiğinde ağaç sayısının 100 ağaca yaklaşmasıyla birlikte yaklaşık düz bir hata seviyesinde ilerlediği ve daha fazla iyileşme eğilimi göstermediği görülmektedir. Tüm faktörleri içeren veri kümesi için de benzer bir grafik oluşmuştur. Bu sebeple ağaç sayısının tüm veri setleri için 100 olarak belirlenmesine karar verilmiştir.



Şekil 4.13: Torba dışı hata grafiği ile Rastgele Orman ağaç sayısının belirlenmesi.

Orijinal veri kümesi ve özellik seçimi sonucu oluşturulan alt veri kümeleri içerisindeki faktör sayısı dikkate alınarak belirlenen her bölünme için seçilen rastgele faktör sayısı *mtry* ve torba dışı hata grafiklerinin incelenmesiyle belirlenen ağaç sayısı *ntree* parametreleri her bir veri kümesi için **Tablo 4.11**'de belirtildiği şekilde Rastgele Orman modeline dahil edilmiştir.

Tablo 4.11: Rastgele Orman için belirlenen hiper parametre değerleri.

	Tüm Faktörleri İçeren Veri Kümesi	Pearson Korelasyon (Fazlalık Değerlendirmesi)	Pearson Korelasyon (Fazlalık ve İlgililik Değerlendirmesi)	Ardışık İleri Yönde Seçim	Ardışık Geri Yönde Seçim
<i>ntree</i> (ağaç sayısı)	100	100	100	100	100
<i>mtry</i> (her bölünmede rastgele seçilen faktör sayısı)	27	17	8	7	11

XGBoost, Rastgele Orman metoduna kıyasla model doğruluğuna etki eden daha fazla sayıda ayarlanabilir hiper parametre içermektedir. Bu yüzden hiper parametrelerin model doğruluğunu artıracak şekilde belirlenebilmesi için hiper parametre ayarlama metodlarından Izgara Arama yöntemi kullanılmıştır. Izgara Arama için farklı hiper parametrenin her birisi için test edilmesi istenen 3 farklı değer tanımlanmıştır (**Tablo 4.12**).

Tablo 4.12: XGBoost hiper parametreleri ve Izgara Arama yöntemiyle test edilen değerler.

Hiper Parametre	Test Edilen Değerler
<i>nrounds</i>	50, 100, 150
<i>max_depth</i>	4, 6, 8
<i>eta</i>	0.05, 0.1, 0.3
<i>gamma</i>	0, 0.1, 0.3
<i>Colsample_bytree</i>	0.6, 0.8, 1
<i>min_child_weight</i>	1, 5, 10
<i>Subsample</i>	0.7, 0.8, 1

Hiper parametre değerleri, kullanılan veri kümesine göre değişebileceği için tüm veri kümeleri için ayrı bir Izgara Arama uygulanmıştır. XGBoost algoritmasının hiper parametreleri, %70 eğitim verisi üzerinde Izgara Arama ve 5-Katlı Çapraz Doğrulama teknikleri kullanılarak ayarlanmıştır. Hiper parametre değerlerinin oluşturduğu modeller RMSE metriğine göre değerlendirilmiş ve en iyi sonucu veren hiper parametre değerlerine yönelik çıktı elde edilmiştir. Izgara Arama sonucunda belirlenen hiper parametre değerleri **Tablo 4.13**'te belirtilmiştir.

Tablo 4.13: Izgara Arama sonuçlarına göre seçilen XGBoost hiper parametre değerleri.

	Tüm Faktörleri İçeren Veri Kümesi	Pearson Korelasyon (Fazlalık Değerlendirmesi)	Pearson Korelasyon (Fazlalık ve İlgililik Değerlendirmesi)	Ardışık İleri Yönde Seçim	Ardışık Geri Yönde Seçim
<i>nrounds</i>	150	100	50	150	150
<i>max_depth</i>	4	6	8	4	8
<i>eta</i>	0.3	0.3	0.3	0.3	0.1
<i>gamma</i>	0.5	0.1	0.1	0	0.5
<i>Colsample_bytree</i>	0.6	0.8	1	0.6	1
<i>min_child_weight</i>	10	10	1	1	10
<i>Subsample</i>	1	1	0.7	1	0.7

Rastgele Orman ve XGBoost için her bir veri kümesi için hiper parametre değerlerinin belirlenmesinin ardından her bir veri kümesi ve ilgili hiper parametre değeri ile model yapısı oluşturulmuştur. Tahmin modellerinin doğrulamasının yapılabilmesi amacıyla model doğrulama tekniklerinden K-Katlı Çapraz Doğrulama tekniği K sayısı 5 olacak biçimde uygulanmıştır. Böylece %70 eğitim verisi olarak ayrılan veri seti 5 eşit parçaya bölünmüş ve doğrulama tekniği, her çalışmada 5 eşit parçadan birisini doğrulama seti ve geri kalan parçaları eğitim seti olarak ayırmıştır. Bu işlem toplam 5 kez gerçekleştirilerek seçilen tek bir eğitim verisinin neden olabileceği yanlışlık önlenmiştir. Çapraz Doğrulama ile oluşturulan modeller, ilk başta ayrılan %30 test verisi ile model performanslarının karşılaştırılması için hazır duruma getirilmiştir.

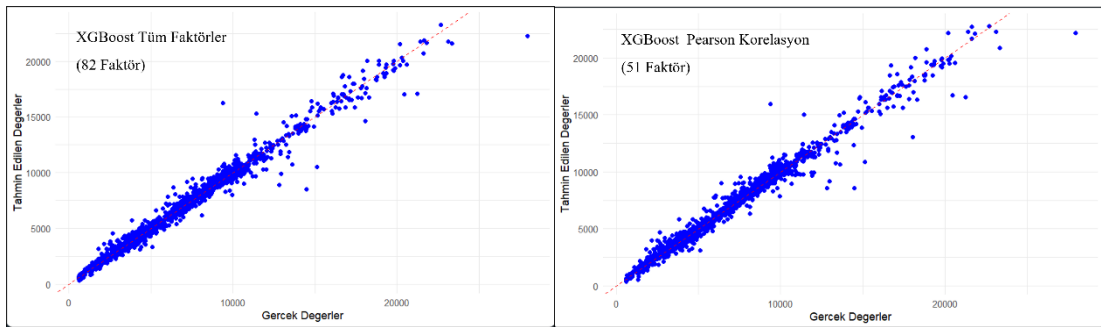
4.6. Model Performanslarının Karşılaştırılması

Tüm faktörler içeren orijinal veri kümesi ve özellik seçimi sonucu elde edilen 4 farklı alt küme ile oluşturulan Rastgele Orman ve XGBoost modelleri 5-Katlı Çapraz Doğrulama sonucunda performans metrikleri ile karşılaştırılmıştır (**Tablo 4.14**).

Tablo 4.14: Model performanslarının karşılaştırması.

Veri Kümesi	Rastgele Orman				XGBoost			
	RMSE	R^2	MAE	MAPE	RMSE	R^2	MAE	MAPE
Tüm Faktörleri İçeren Veri Kümesi (82 Faktör)	676 .1179	0 .9719	333 .2539	6 .0984	525 .8118	0 .9829	286 .2245	5 .4982
Pearson Korelasyon (Fazlalık Değerlendirmesi) (51 Faktör)	685 .1648	0 .9711	335 .7318	6 .1517	550 .6307	0 .9813	285 .0134	5 .3978
Pearson Korelasyon (Fazlalık ve İlgililik Değerlendirmesi) (24 Faktör)	989 .6264	0 .9402	572 .7163	10 .4325	937 .1401	0 .9460	508 .8471	9 .2552
Ardışık İleri Yönde Seçim (21 Faktör)	861 .1779	0 .9546	485 .7235	8 .6852	791 .4647	0 .9614	454 .4231	8 .2314
Ardışık Geri Yönde Seçim (33 Faktör)	980 .3073	0 .9411	594 .4743	10 .9411	715 .2274	0 .9686	381 .4294	7 .0214

Tahmin modelleri karşılaştırıldığında tüm veri kümelerinde XGBoost algoritmasının RMSE, R^2 , MAE ve MAPE metriklerinde Rastgele Orman algoritmasına kıyasla çok daha iyi sonuçlar verdiği görülmektedir. Bu yüzden veri kümelerinin değerlendirilmesinde referans yöntem XGBoost kabul edilmiştir. Tüm faktörlerin bulunduğu orijinal veri kümesi ve 4 farklı özellik seçimi veri alt kümesi XGBoost modeli baz alınarak karşılaştırılmıştır. Orijinal veri kümesinin özellik seçimi alt kümelerine kıyasla daha iyi sonuç verdiği görülmektedir. Ancak Pearson Korelasyon fazlalık değerlendirmesine yönelik alt veri kümesine yönelik modelin orijinal veri kümesine yönelik model ile yakın doğruluğu verdiği ve dolayısıyla her iki veri kümesinin de kullanılabilirliği söz konusu olmuştur. Pearson Korelasyon fazlalık değerlendirmesine yönelik özellik seçiminde orijinal veri kümesinde 82 olan faktör sayısı 51'e indirilmiş ve üstelik orijinal veri kümesine oldukça yakın ve yüksek bir doğruluk sağlanmıştır. Ayrıca orijinal veri kümesine yönelik XGBoost algoritmasına kıyasla modelin eğitim süresi kapsamında hesaplama maliyeti olarak yaklaşık %21 oranında bir kazanç sağlamıştır. Ek olarak her iki modelin test verisi olarak ayrılmış veriler üzerindeki tahminlerinin grafik üzerinde gösterimi ile performanslarının yorumlanması sağlanmıştır (**Şekil 4.14**). **Şekil 4.14**'te bulunan grafiklerde yatay eksen gerçek yolcu sayılarını ifade etmekte iken dikey eksen tahmin edilen yolcu sayılarını ifade etmektedir. Grafik üzerinde bulunan mavi noktalar gerçek ile tahmin değerlerinin ilişkisini göstermektedir. Grafikler incelendiğinde her iki modelin yolcu sayısı tahminlerinin oldukça yakın olduğu görülebilmektedir. Böylece hem veri setinde bulunan faktör sayısında düşüşe gidilerek hesaplama maliyetleri azaltılmış hem de model doğruluğu büyük ölçüde korunmuştur. Bu sebeplerden dolayı en ideal model, 51 faktör içeren veri kümesi ile kurulan XGBoost modeli olarak belirlenmiştir.

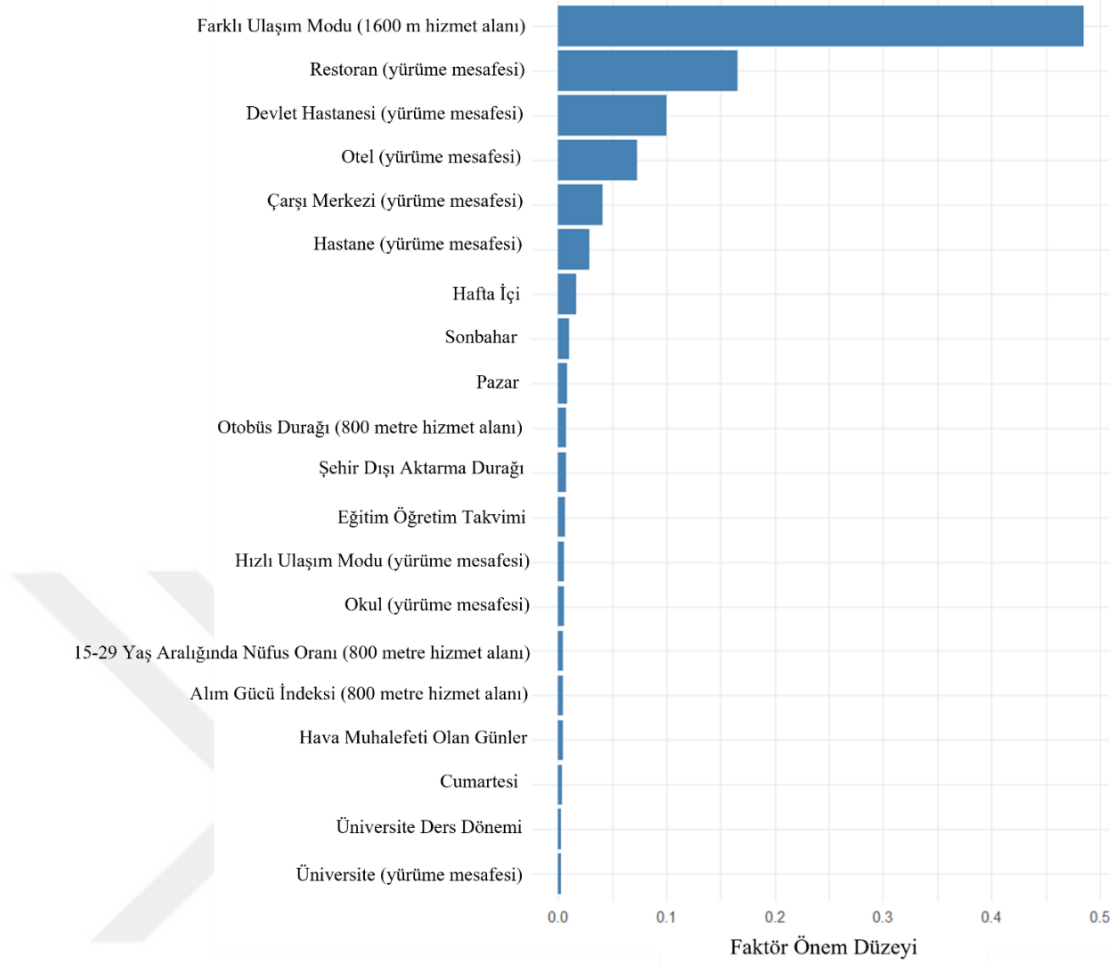


Şekil 4.14: Grafik üzerinden model karşılaştırma.

İdeal modele yönelik performans değerlendirme metrikleri detaylı incelendiğinde XGBoost için günlük ortalama hatalı tahmin edilen yolcu sayısını ifade eden RMSE metriği yaklaşık 550 olarak bulunmuştur. Bu ifade, yapılan tahminlerin ortalama olarak yaklaşık +/- 550 yolcu civarında gerçek değere yaklaştığı şeklinde gerçekleştiğini söylemektedir. RMSE metriği, büyük hatalara karşı yüksek duyarlı olduğundan dolayı MAE metriğine kıyasla daha yüksek değerler alabilmektedir. Faktörlerin yolcu sayısını açıklamadaki gücünü ifade eden R^2 değeri yaklaşık 0.98 olarak hesaplanmıştır. Gerçek ve tahmin edilen yolcu sayıları arasındaki farkların mutlak değerlerinin ortalamasını ifade eden MAE ölçütü yaklaşık 285 olarak bulunmuştur. Bu değer, tahmin edilen yolcu sayısının ortalama olarak gerçek değerden ne kadarlık bir sapma yaptığıнын göstergesidir. MAE metriğinin sonuçlarının RMSE metriğine kıyasla daha düşük çıkma sebebi büyük hataların daha az cezalandırılmasından kaynaklanmaktadır. Yolcu sayısına yönelik tahminlerin ve gerçek değerler arasındaki tüm mutlak hataların ortalamasını ifade eden MAPE, yaklaşık %5 olarak hesaplanmıştır. Bu ifade yapılan tahminlerin yüzdesel olarak ne kadar sapma gösterdiğinin göstergesidir. Bu kapsamda ideal model, mekânsal ve zamansal faktörlerin tahmin modelindeki önem düzey sıralamalarının belirlenmesinde ve 2024 yılı için mevcut ve proje aşamasındaki istasyonlarda tahmin gerçekleştirilmesi amacıyla baz model olarak kullanılmaktadır.

4.7. Faktör Önem Düzeylerinin Sıralanması

Çalışma kapsamında Pearson Korelasyon fazlalık değerlendirmesine dayalı özellik seçimine dayalı olarak kurulan XGBoost tahmin modeli faktörlerin önem düzeylerinin sıralanması amacıyla kullanılmıştır. Mekânsal ve zamansal faktörlerin model doğruluğuna ve dolaylı olarak yolcu talebiyle olan ilişkilerin irdelenmesi amacıyla XGBoost modeli içerisinde bulunan faktör önem düzeyi sıralaması fonksiyonu kullanılmıştır. XGBoost içerisinde bulunan *xgb.importance()* fonksiyonu içerisindeki *Gain* metriği baz alınarak faktörlerin modele olan katkıları yukarıdan aşağı azalan biçimde sıralanmıştır (**Şekil 4.15**). Görselleştirmenin daha kolay olması amacıyla önem düzeyi en yüksek ilk 20 faktör gösterilmiştir. *Gain* metriği, karar ağacında bir faktörün her bölünmedeki sağladığı ortalama bilgi kazancını ifade etmektedir.



Şekil 4.15: XGBoost faktör önem düzeyi sıralaması.

Şekil 4.15 incelendiğinde mekânsal faktörlerin zamansal faktörlere kıyasla modelin doğruluğunu artırmada daha etkili olduğu görülmektedir. Mekânsal faktörlerin bu özelliği, erişilebilirliğin yolcu talebi tahmininde daha belirleyici bir rol oynadığını ifade etmektedir. Faktör önem düzeyi sıralaması incelendiğinde 1600 metrelik hizmet alanı içerisindeki farklı ulaşım modları faktörünün diğer faktörlere kıyasla büyük bir farkla ilk sırada olduğu görülmektedir. Bu faktör, her bir istasyon oluşturulmuş yürüme mesafesi cinsinden 1600 metrelik hizmet alanları içerisindeki otobüs durağı, paylaşımlı bisiklet istasyonu ve taksi durağı gibi ulaşım modlarının sayısını ifade etmektedir. Faktör önem düzeyi sıralamasında ikinci sırada restoran faktörü bulunmaktadır. Bu faktör her istasyon için istasyonun çevresinde bulunan en yakın restorana olan yürüme mesafesinin kilometre cinsinden ifadesidir. Üçüncü sırada, istasyon çevresindeki en yakın devlet hastanesine olan yürüme mesafesini ifade eden devlet hastanesi faktörü yer almaktadır. Dördüncü, beşinci ve altıncı sıralarda sırasıyla

otel, arşı merkezi ve hastane (zel ve devlet) faktrleri yer almaktadır. Bu faktrler yine benzer ekilde en yakın olanına yrme mesafesini ifade etmektedir. Faktr nem dzeyi sıralamasına bakıldığında ilk sıralarda meknsal faktrlerin yer aldığı ve meknsal faktrler ierisinde yrme mesafesine dayalı bir hizmet alanı oluřturulup ierisinde ilgili faktrn miktarını ifade eden faktrlerden farklı ulařım modları ilk sırada yer almıřtır. Bir istasyondan ilgili faktr iin en yakın olanına yrme mesafesini ifade eden faktrlerin de nemli bir kısmı st sıralarda yer almaktadır. Bu durum yrme mesafesine dayalı faktrlerin modelin doėruluėunu artırmada nemli derecede etkili olduėunu gstermektedir. Zamansal faktrler, meknsal faktrlere kıyasla daha dřk ancak faktre gre deėiřmekle birlikte anlamlı katkılar sunmuřtur. Zamansal faktrler ierisindeki hafta ii olma durumunun model doėruluėunu en nemli dzeyde artıran zamansal faktr olduėu grlmektedir. Hafta ii faktrnn hemen ardından sonbahar ve pazar gn faktrleri yer almaktadır. Faktr nem dzeyi sıralaması genel hatlarıyla incelendiėinde hem meknsal hem de zamansal faktrlerin modelde farklı seviyelerde doėruluėu artırıcı katkılar sundukları grlmřtr.

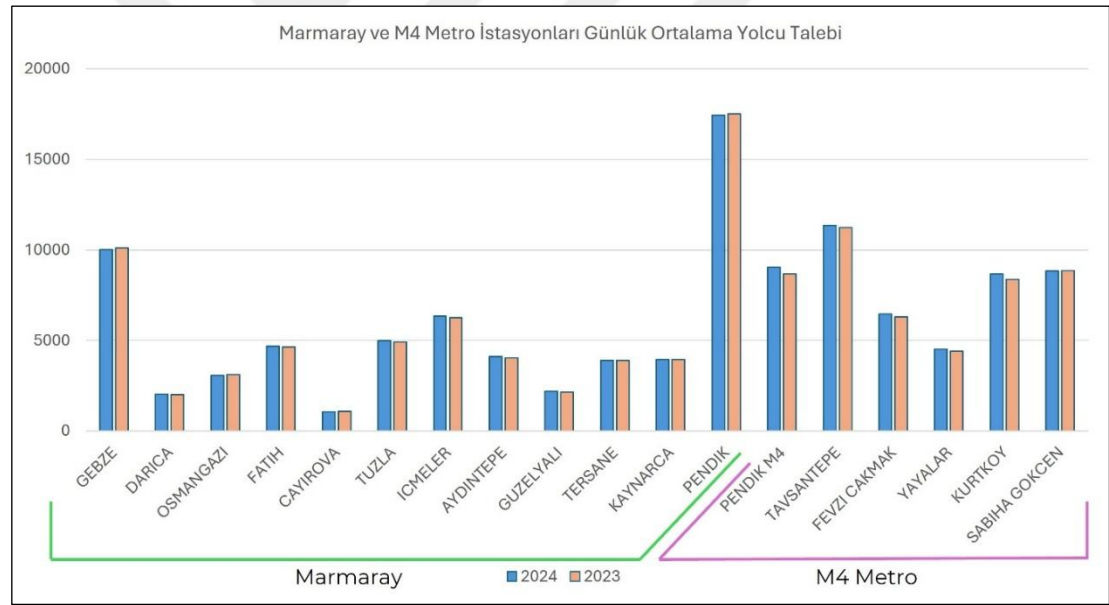
4.8. Seilen Tahmin Modelinin Mevcut ve Proje Ařamasındaki İstasyonlara Uygulanması

Oluřturulan tahmin modelleri arasından Pearson Korelasyon fazlalık deėerlendirmesine ynelik alt kmeye dayalı ayarlanmış hiper parametre deėerlerine sahip XGBoost modeli 2024 yılına ynelik tahminlerin oluřturulacaėı model olarak seilmiřtir. Yeni tahminlerin oluřturulmasında kullanılacak veri kmesi ierisinde 31 adet zamansal ve 20 adet meknsal faktr bulunmaktadır. Yapılacak tahmin 2024 yılı iin gerekleřtirileceėi iin zamansal faktrleri ieren veri seti tekrardan dzenlenmiřtir. Yeni tahminler, modelin eėitilmiş olduėu Marmaray ve M4 Metro istasyonları ve henz proje ařamasında bulunan ancak istasyonların nereye konumlandırılacaėı bilinen M4 Metro Uzatma hattı ve Gebze-Darıca Metro hattına ait toplamda 35 istasyon iin gerekleřtirilmiřtir. Yeni istasyonların eklenmesiyle birlikte 2024 yılı iin oluřturulan yeni veri seti toplamda 12810 gzlem iermektedir. ncelikle 2024 yılı iin Marmaray ve M4 Metro hattı istasyonlarına ynelik gerekleřtirilen tahmin sonuları incelenmiřtir. Tahmin sonularına ynelik rnek ıktı yapısı **Tablo 4.15**'te verilmiřtir.

Tablo 4.15: Marmaray ve M4 Metro istasyonları için gerçekleştirilen 2024 yılı günlük yolcu talep tahminleri.

Tarih	İstasyon	Yolcu Sayısı
2024-01-01	GEBZE	10138
2024-01-02	GEBZE	9361
⋮	⋮	⋮
2024-12-30	SABIHA GOKCEN	11454
2024-12-31	SABIHA GOKCEN	10093

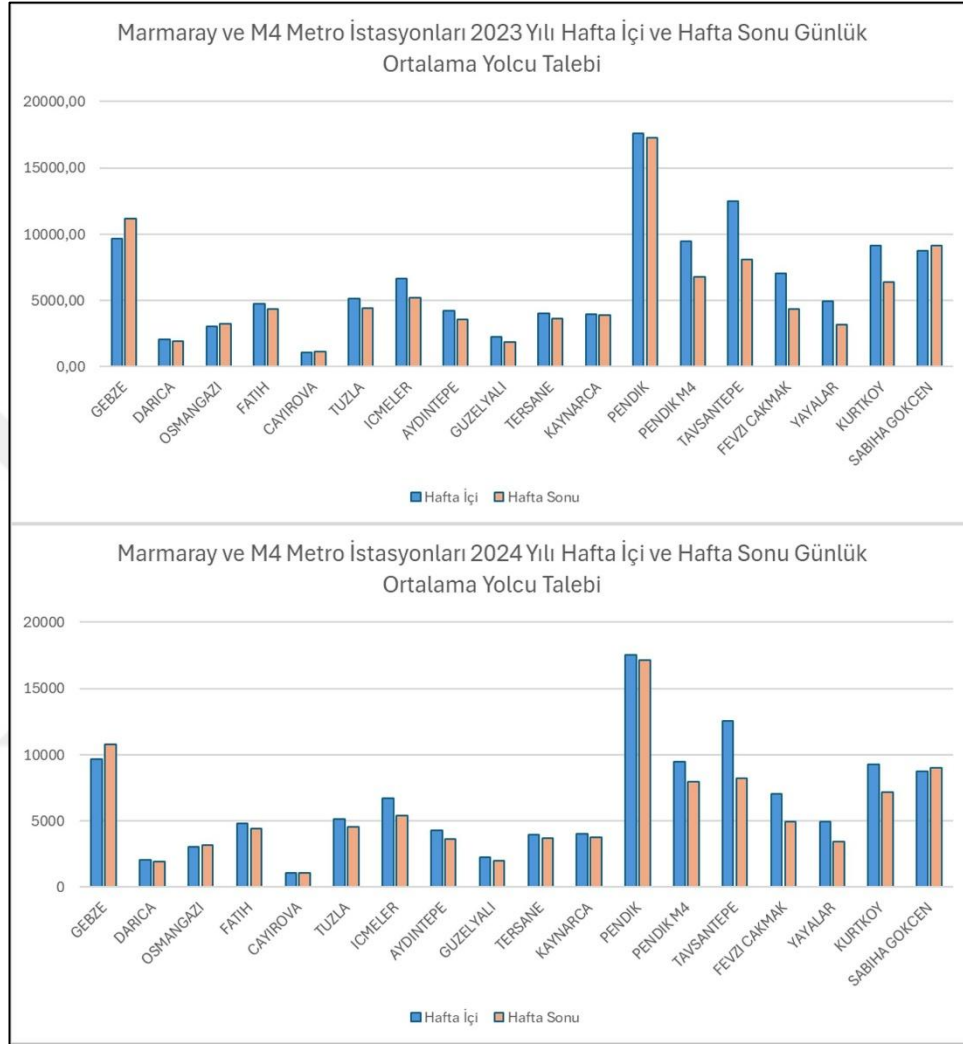
Marmaray ve M4 Metro istasyonlarına ait 2023 yılı günlük yolcu talep verileri mevcut olduğundan dolayı 2024 yılı için gerçekleştirilen tahminler ile kıyaslama gerçekleştirilmiştir. **Şekil 4.16**'da istasyon bazında 2023 yılı için gerçekleşmiş günlük ortalama yolcu talebi ve 2024 yılı için model tarafından tahmin edilen günlük ortalama yolcu talebi verilmiştir.



Şekil 4.16: Marmaray ve M4 Metro istasyonları 2023 yılı gerçekleşmiş ve 2024 yılı tahmin edilen günlük ortalama yolcu talebi.

Marmaray ve M4 Metro istasyonları için gerçekleşmiş ve bir yıl sonrası için tahmin edilen değerler incelendiğinde yolcu talebinin günlük ortalamasında büyük değişiklikler yaşanmadığı ve istasyonların temel yolcu talebinin yakın seviyelerde tahmin edildiği görülmektedir. Bu durum genel yolcu talebini oluşturan mekânsal faktörlerin bir yıl içerisinde anlamlı bir değişime uğramamasıyla ilişkilendirilebilir. Bu sebeple temel yolcu talebinin incelenmesinin ardından dalgalanmaları görebilmek

amacıyla zamansal faktörlerin yolcu talebi üzerindeki etkisinin incelenmesi gerekmektedir. Bu amaç doğrultusunda modele en yüksek katkıyı sağlayan hafta içi faktörünün yolcu talebindeki etkisi incelenmiştir (Şekil 4.17).

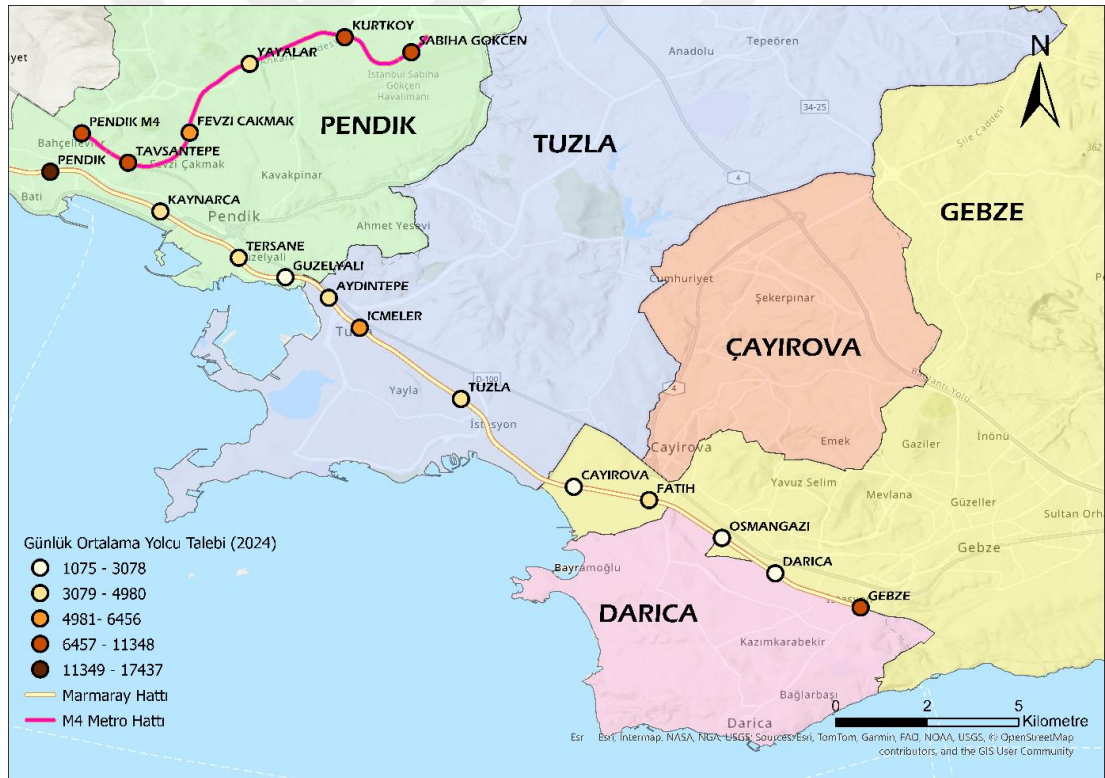


Şekil 4.17: Hafta içi ve hafta sonu yolcu talebindeki dalgalanmalar.

Şekil 4.17’de hafta içi ve hafta sonu tarihlerinde 2023 yılı için gerçekleşmiş ve 2024 yılı için tahmin edilen ortalama yolcu talebi gösterilmiştir. 2023 yılı için hafta içi ve hafta sonu yolcu talebinde yaşanan farklılıklar istasyon bazlı incelendiğinde birçok istasyonda hafta içi ortalama yolcu talebinin hafta sonu ortalamasına kıyasla daha yüksek çıktığı görülmektedir. Tahmin modeli ile oluşturulan 2024 yılı için incelendiğinde de 2023 yılına benzer şekilde dalgalanmaların yaşanmış olduğu görülmüştür. Bu sonuçlar doğrultusunda tahmin modelinin 2023 için yolcu dalgalanmalarını iyi bir şekilde öğrenebildiği ve 2024 için yansıtılabildiği

görülmektedir. Bu durum, tahmin modelin yolcu dalgalanmalarını güçlü bir şekilde yakalayıp yansıtabildiğinin göstergesidir.

Gerçekleştirilen tahminlerin istasyon bazlı biçimde çalışma sınırları içerisindeki konumlarıyla ilişkilendirilebilmesi amacıyla **Şekil 4.18**'de bulunan harita oluşturulmuştur. Harita incelendiğinde, M4 Metro hattına ait istasyonların Marmaray hattı istasyonların geneline kıyasla daha çok yolcu talebi çektiği görülmektedir. Ancak istasyonların geneline bakıldığında en çok günlük yolcu talebinin PENDİK Marmaray istasyonunda gerçekleştiği görülmektedir. PENDİK istasyonu Yüksek Hızlı Tren bağlantısı bulunan ve diğer istasyonlara kıyasla en fazla kentsel ilgi noktasına sahip olma özelliği taşımaktadır. Bu özellikleri ile PENDİK istasyonunun yüksek yolcu talebi çekmesi beklenen bir durumdur. En düşük yolcu talebi ise GÜZELYALI, ÇAYIROVA, OSMANGAZİ ve DARICA istasyonlarında meydana gelmiştir.



Şekil 4.18: Marmaray ve M4 Metro istasyonları için 2024 yılı günlük yolcu talep tahmini.

Çalışmada elde edilen yüksek doğruluklu model ile yapılan bir diğer tahmin henüz proje aşamasında bulunan M4 Metro Uzatma hattı ve Gebze-Darica Metro hattı istasyonlarına uygulanmıştır. Tahmin modeli, M4 Metro Uzatma hattına ait 6 istasyon

ve Gebze-Darica Metro hattına ait 11 istasyon olmak üzere toplamda 17 istasyon için uygulanmıştır. Günlük bazda 2024 yılı için gerçekleştirilen tahmin çıktısının örneği **Tablo 4.16**'da verilmiştir.

Tablo 4.16: M4 Metro Uzatma hattı ve Gebze-Darica Metro hattı istasyonları için gerçekleştirilen günlük yolcu talep tahminleri.

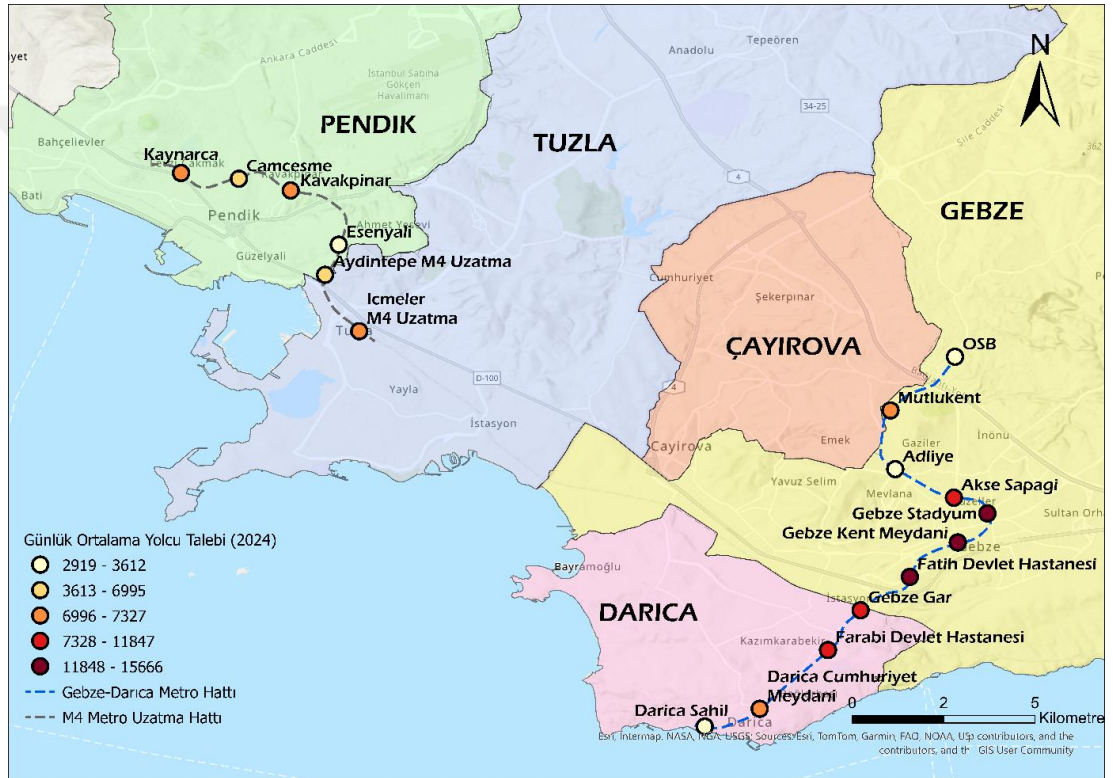
Tarih	İstasyon	Yolcu Sayısı
2024-01-01	KAYNARCA	7113
2024-01-02	KAYNARCA	7694
⋮	⋮	⋮
2024-12-30	OSB	3734
2024-12-31	OSB	3595

İstasyon bazlı gerçekleştirilen 2024 yılı için günlük yolcu talebi tahminleri sonucunda her bir istasyon için günlük yolcu sayılarının ortalama değerleri **Tablo 4.17**'de görülmektedir. Tablo incelendiğinde M4 Metro Uzatma hattı istasyonları içerisinde bulunan KAVAKPINAR istasyonunun en yüksek günlük ortalama yolcu talebine sahip olacağı öngörülmektedir. Gebze-Darica Metro hattı istasyonları için incelendiğinde ise FATİH DEVLET HASTANESİ istasyonunun en yüksek yolcu talebine sahip olacağı öngörülmektedir. Bu tahmin sonuçlarına göre Gebze-Darica Metro hattı istasyonlarının daha yüksek yolcu talebi çekeceği sonucu elde edilmiştir.

Tablo 4.17: Proje aşamasındaki istasyonlar için 2024 yılı günlük ortalama yolcu sayıları.

PROJE	İSTASYON	Günlük Ortalama Yolcu Talebi (2024)
M4 Metro Uzatma İstasyonları	İÇMELER	7155
	AYDINTEPE	6995
	KAYNARCA	7216
	CAMÇESME	6717
	ESENYALI	3570
	KAVAKPINAR	7306
	ADLIYE	2919
Gebze-Darica Metro İstasyonları	AKSE SAPAĞI	11847
	DARICA CUMHURİYET MEYDANI	7327
	DARICA SAHİL	3612
	FARABİ DEVLET HASTANESİ	9727
	FATİH DEVLET HASTANESİ	15666
	GEBZE GAR	10005
	GEBZE KENT MEYDANI	14219
	GEBZE STADYUM	15536
	MUTLUKENT	7235
	OSB	3378

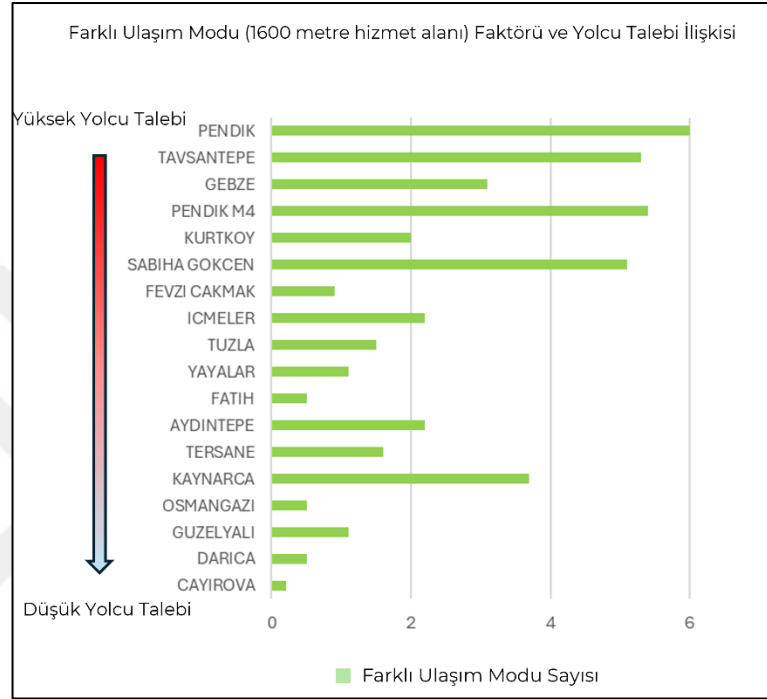
Üretilen yüksek doğruluklu tahmin modelinin proje aşamasındaki istasyonlar için uygulanmasının ardından CBS ortamında yolcu talebi haritası oluşturulmuştur. **Şekil 4.19**'da bulunan haritada her bir istasyonun 2024 yılı için tahmin edilen günlük yolcu sayılarının ortalaması yer almaktadır. Haritada her bir istasyon, hangi yolcu talebi aralığında bulunuyorsa o talebi temsil eden bir renk almaktadır. Oluşturulan harita incelendiğinde Gebze-Darica Metro hattı istasyonlarının genel anlamda M4 Metro Uzatma hattı istasyonlarına kıyasla daha yüksek yolcu talebi çekeceği öngörülmektedir.



Şekil 4.19: Gebze-Darica Metro hattı ve M4 Metro Uzatma hattı istasyonları günlük ortalama yolcu talebi.

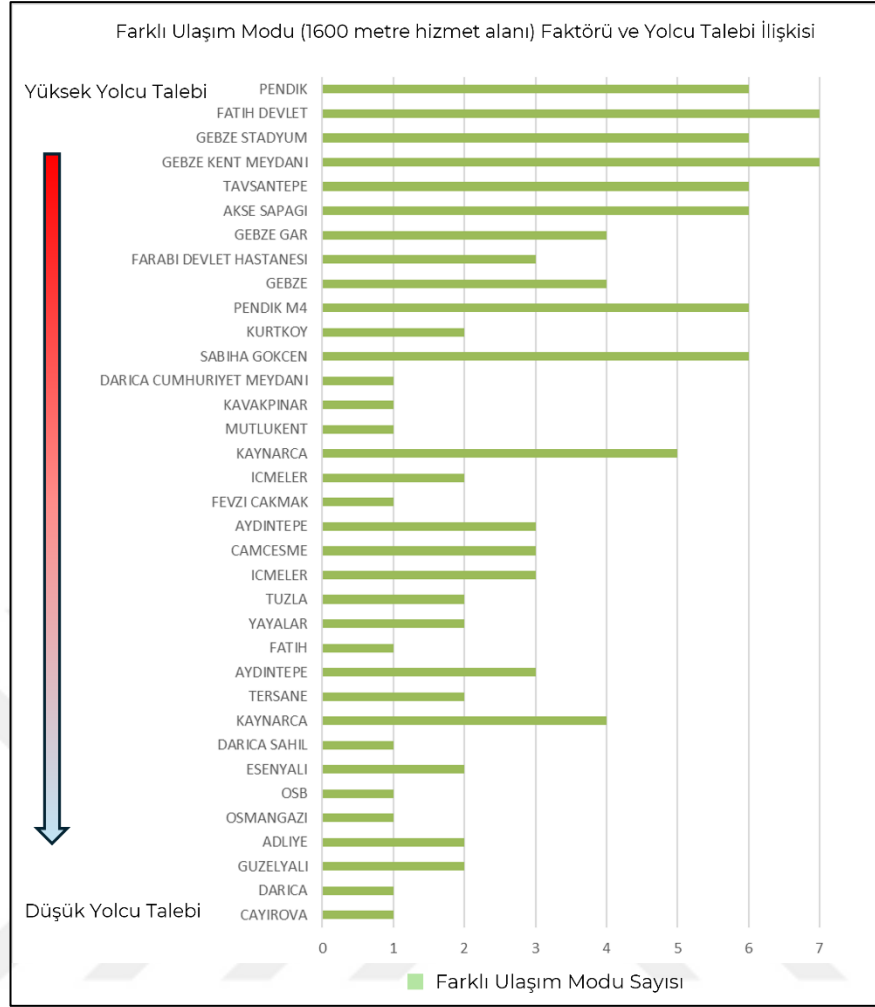
Bir toplu taşıma istasyonu için temel yolcu talebini oluşturan faktörlerin mekânsal faktörler olduğu bilinmektedir. Mekânsal faktörlerin yolcu talebi ile olan ilişkisinin incelenmesi amacıyla tahmin modeline bilgi kazancı açısından en yüksek katkıyı sağlayan 1600 metrelik hizmet alanında farklı ulaşım modu faktörü kullanılmıştır. Bu faktör, her bir istasyon için yürüme mesafesine dayalı 1600 metrelik hizmet alanı sınırları içerisinde bulunan otobüs, bisiklet, taksi vb. farklı ulaşım modlarının sayısını ifade etmektedir. Öncelikle tahmin modelinin eğitilmiş olduğu istasyonlarda 2023 yılı için gerçekleşen günlük ortalama yolcu talebi miktarı ile ilgili faktörün her bir istasyon

için almış olduğu değerler kıyaslanmıştır. Bu ilişkinin incelenmesi amacıyla her istasyon için hizmet alanı içerisindeki farklı ulaşım modu sayısını ifade eden sütun grafiği oluşturulmuştur (**Şekil 4.20**). Grafikte yorum kabiliyetinin artırılması amacıyla modelin eğitildiği istasyon adları 2023 yılı için gerçekleşmiş yolcu talebinin yüksek olması durumundan düşük olması durumuna göre yukarıdan aşağıya sıralanmıştır.



Şekil 4.20: Marmaray ve M4 Metro istasyonları 2023 yılı gerçekleşmiş yolcu talepleri ve farklı ulaşım modu faktörü ile ilişkisi.

Modelin eğitildiği istasyonlar için incelendiğinde bir istasyonun hizmet alanında farklı ulaşım modu sayısının artması durumunda günlük ortalama yolcu talebi de genellikle artış göstermektedir. Bir istasyon için farklı ulaşım modlarının çeşitliliği, o istasyonun erişilebilirliğini artırmakta ve böylece yüksek yolcu talebi yaşanmaktadır. Bu açıdan istasyon çevresindeki ulaşım modlarının çeşitliliği o istasyon için oluşacak yolcu talebini artırması sebebiyle dikkate alınması gereken başlıca mekânsal faktörlerden olmuştur. Aynı mekânsal faktör hem modelin eğitildiği hem de proje aşamasındaki istasyonlar için 2024 yılına yönelik günlük yolcu talep tahmin sonuçları ile değerlendirilmiştir (**Şekil 4.21**). Grafikte istasyon adları, tahmin edilen günlük ortalama yolcu sayısına göre azalan şekilde yukarıdan aşağıya doğru sıralanmıştır.



Şekil 4.21: Farklı ulaşım modu faktörü ve yolcu talebi ilişkisi.

Çalışma alanı sınırları içerisindeki tüm istasyonlar için oluşturulan grafik incelendiğinde ulaşım modu çeşitliliğinin yüksek olduğu istasyonlarda tahmin edilen yolcu talebi de yüksek olmaktadır. Daha önce modelin eğitildiği istasyonlar için gözlemlenen yüksek ulaşım modu çeşitliliği ve yüksek yolcu talebi ilişkisi buradaki tahminlere de yansımış bulunmaktadır. Bu sonuçlarla birlikte, mevcut istasyonlar ile eğitilen tahmin modelinin, istasyon hizmet alanı içerisindeki ulaşım modu çeşitliliği gibi mekânsal faktörlerle yolcu talebi arasındaki ilişkiyi iyi biçimde öğrenebildiği, mevcut ve proje aşamasındaki istasyonlarda ileriye yönelik tahminlerde uygulayabildiği görülmüştür.

Çalışmada üretilen 2024 yılına ait günlük ortalama yolcu talep tahminleri hem mevcut hem de proje aşamasındaki istasyonlar için elde edilmiştir. Tüm hatlar için tahmin sonuçlarının birlikte yorumlanabilmesi amacıyla **Şekil 4.22'**de bulunan harita oluşturulmuştur.



Şekil 4.22: Tüm istasyonlar için 2024 yılı günlük yolcu talep tahmin sonuçları.

Şekil 4.22’de bulunan harita incelendiğinde özellikle Gebze-Darica Metro istasyonlarının yüksek düzeyde yolcu talebi çekeceği öngörülmektedir. Fatih Devlet Hastanesi, Gebze Kent Meydanı ve Gebze Stadyum istasyonları en yüksek yolcu talebini çekmesi beklenen istasyonlardır. Bu istasyonlardan Fatih Devlet Hastanesi, bir devlet hastanesinin hemen yanında bulunması özelliği ile ön plana çıkmaktadır. Gebze Kent Meydanı ise çarşı merkezi olarak kabul edilen bir bölgede bulunması, farklı ulaşım modlarına aktarma yapılabilir olması ve etrafında çok zengin kentsel ilgi noktası barındırması gibi özellikleri ile ön plana çıkmaktadır. Gebze Stadyum ise yine benzer şekilde zengin kentsel ilgi noktası barındırması ve rekreasyon alanı çeşitliliği ile ön plana çıkmaktadır. M4 Uzatma hattına ait istasyonlar için tahmin edilen günlük ortalama yolcu talebinin çevresindeki M4 Metro ve Marmaray istasyonlarına benzer şekilde olduğu görülmüştür. M4 Metro Uzatma hattı istasyonları incelendiğinde konum olarak mevcuttaki M4 ve Marmaray hatlarına ait istasyonlar arasında kalarak bir açıdan mevcuttaki bu iki hattı birbirine bağlama özelliği barındırmaktadır. M4 Uzatma hattı istasyonlarından İcmeler istasyonu için bakıldığında aynı hattın diğer istasyonlarına kıyasla daha çok yolcu talebi çekeceği

yönünde tahmin gerçekleştirilmiştir. Bunun sebepleri arasında İçmeler Marmaray istasyonu ile aktarma istasyonu özelliği barındırması olarak yorumlanabilir. Henüz proje aşamasında bulunan istasyonlar için daha önce gerçekleşmiş bir yolcu istatistiği bulunmadığından dolayı tahminlerin doğruluğu test edilemese de bölgenin barındırmış olduğu özelliklerin incelenmesi doğrultusunda tahmin sonuçlarının gerçeğe yakın sonuçlar vermiş olabileceği ve karar vericiler tarafından dikkate alınabilecek sonuçlar içerdiği düşünülmektedir.



5. SONUÇLAR VE ÖNERİLER

Küresel ölçekte yaşanan nüfus artışı ve yüksek kentleşme hızı sonucunda kentsel ulaşım çeşitli zorluklarla karşı karşıya kalmaktadır. Yaşanan zorlukların üstesinden gelmek amacıyla kent sakinlerinin hareketlilik kavramı şekillenmektedir. Çevresel, sosyal ve ekonomik açıdan farkındalığın oluşması ile değişim gösteren hareketlilik kavramı, Bilgi ve İletişim Teknolojileri alanında yaşanan gelişmeler sonucunda gelişim göstererek yeni nesil hareketlilik kavramını ortaya çıkarmaktadır. Yeni nesil hareketlilik kapsamında bulunan toplu taşıma sistemleri, kent sakinlerin toplu bir halde yüksek verimde taşınmasına olanak tanımaktadır. Toplu taşıma sistemleri içerisinde bulunan ve yüksek kapasite avantajının yanında kendine ait demiryolu altyapısını kullanması sayesinde düzenli sefer aralıklarına sahip olan metro ve banliyö treni gibi sistemler, kentsel ulaşımın verimli bir şekilde yönetilmesi amacıyla öne çıkmaktadır. Yerel yönetimler tarafından ciddi yatırımların yapıldığı bu sistemlerin sürdürülebilirliği büyük ölçüde mekânsal kararlara bağlı olmaktadır. Metro ve banliyö treni gibi sistemler için kurulan istasyonlarda meydana gelen yolcu talebinin yeterli seviyede olması, yapılan yatırımların verimliliğinin ve sürdürülebilirliğinin sağlanması açısından büyük önem taşımaktadır. Sürdürülebilir bir toplu taşıma için bir istasyonun mevcut yolcu talebinin bilinmesi ile ileriye dönük yolcu talebi dalgalanmalarının tahmini ve proje aşamasındaki istasyonlar için temel yolcu talebinin yüksek doğrulukla tahmin edilebilmesi gerekmektedir. Yolcu talebindeki dalgalanmaların yüksek doğrulukla tahmini için zamansal faktörlerin ve temel yolcu talebinin tahmininde ise istasyon ve çevresine ait özelliklere yönelik mekânsal faktörlerin değerlendirilmesi gerekmektedir. Bu yüzden CBS ve yapay zekâ yetkinliklerinin kullanıldığı mekânsal zekâ yaklaşımına dayalı metodoloji uygulanmalıdır.

Gerçekleştirilen tez çalışmasında, mevcut ve proje aşamasındaki istasyonlar için yüksek doğrulukta ve ölçeklenebilir bir yolcu talep tahmin modelinin kurulmasına yönelik metodoloji geliştirilmiştir. Bu doğrultuda öncelikle metro ve banliyö treni gibi ulaşım sistemleri bulunan, kentsel ilgi noktası çeşitliliği açısından zengin ve farklı demografik yapıları barındıran bir çalışma bölgesi seçilmiştir. Bu özellikleri karşılayan İstanbul ve Kocaeli illerine ait Tuzla, Pendik, Çayirova, Darıca ve Gebze ilçeleri çalışma bölgesi olarak ve çalışma bölgesi sınırları içerisinde bulunan M4

Metrosuna ait 6 istasyon ve Marmaray banliyö trenine ait 12 istasyon yolcu talebine yönelik tahmin modelinin eğitilmesinde kullanılmıştır. Hem metro hem de banliyö trenine ait istasyonların kullanımı ile farklı sistemlerin farklı istasyonlarındaki yolcu talebinin yorumlanması sağlanmıştır. Zamansal farklılıkları içeren 34 adet zamansal ve kentsel ilgi noktaları ve demografik yapıyı içeren 48 adet mekânsal faktöre ait veriler çalışma bölgesi sınırları içerisinde olacak şekilde açık veri kaynaklarından ve ticari yazılım çözümlerinden elde edilmiştir. Çalışma kapsamında mekânsal faktörler arasında bulunması beklenen istasyon çevresindeki otopark varlığı faktörü veri temini aşamasında yaşanan sorunlar sebebiyle eklenememiştir. Bu sebeple bu çalışmada otopark varlığının yolcu talebiyle olan ilişkisi incelenememiştir.

Mekânsal faktörlerin veri setine dahil edilmesi amacıyla istasyon noktalarına CBS yazılımı ortamında ağ tabanlı analiz tekniklerinden hizmet alanı ve en yakın tesis analizleri uygulanmıştır. Her bir istasyon için oluşturulan 800 ve 1600 metrelik hizmet alanları içerisindeki ilgili kentsel ilgi noktası sayısı ve demografik yapı bilgileri veri setine eklenmiştir. Bu işlemlerin sonucunda hem zamansal hem de mekânsal faktörleri içeren veri seti elde edilmiştir. Çalışma kapsamında yolcu talebinin tahmini makine öğrenmesi yöntemlerinden denetimli öğrenme yaklaşımına dayalı olarak gerçekleştirileceği için her bir istasyon için 2023 yılı için geçmiş yolcu talebi verileri açık veri kaynağından elde edilmiştir. Çalışma bölgesi sınırları içerisindeki M4 Metro ve Marmaray istasyonları için saatlik yolcu talebi veri seti günlük yolcu talebini verecek şekilde düzenlenmiştir. Toplamda 82 faktörden oluşan veri seti ve yolcu talepleri makine öğrenmesi yöntemleri ile yolcu talebinin tahmin edilmesinde girdi olarak kullanılmıştır.

Makine öğrenmesi yöntemlerine dayalı oluşturulacak tahmin modellerinin performansının artırılması amacıyla faktörler arasından ilgisiz ve gereksiz faktörlerin atılmasını sağlayan ve böylece veri alt kümeleri oluşturan özellik seçimi yöntemleri uygulanmıştır. Çalışma kapsamında karşılaştırmalı olarak sunulabilmesi amacıyla Filtre yöntemlerinden Pearson Korelasyon ve Sarmalayıcı yöntemlerden Ardışık İleri Yönde Seçim ve Ardışık Geri Yönde Seçim yöntemleri kullanılmıştır. Pearson Korelasyon yönteminde öncelikle fazlalık değerlendirmesine dayalı daha sonrasında hem fazlalık hem de ilgililik değerlendirmesine dayalı iki farklı yaklaşım uygulanmıştır. Bunun sonucunda hem Filtre hem de Sarmalayıcı yöntemlere dayalı 4 farklı alt küme oluşturulmuştur.

Yolcu talebinin, oluşturulan veri kümeleri ile tahmini için karmaşık veri yapılarını açıklamadaki gücü ile ön plana çıkan makine öğrenmesi tekniklerinden Rastgele Orman ve XGBoost algoritmalarının kullanımına karar verilmiştir. Her iki algoritma için ayarlanması gereken hiper parametre değerleri bulunmaktadır. Rastgele Orman için iki hiper parametre manuel biçimde ayarlanırken XGBoost için 7 hiper parametre değerinin en iyi değerleri için Izgara Arama metodu uygulanmıştır. Rastgele Orman hiper parametrelerinden ağaç sayısı, torba dışı hata grafiği incelenerek ve her bölünmede rastgele seçilecek faktör sayısı ilgili veri kümesindeki faktör sayısının üçte biri olacak şekilde ayarlanmıştır. XGBoost için hiper parametre değerleri, her bir hiper parametre için belirlenen 3 farklı değere dayalı olarak Izgara Arama yöntemiyle belirlenmiştir. Izgara Arama metodu her bir veri kümesi için K-Katlı Çapraz Doğrulama ile çalıştırılmış ve bu sayede veri kümesine dayalı olarak en yüksek doğruluğu veren hiper parametre değerleri belirlenmiştir.

Özellik seçimi sonucu oluşan veri kümeleri ve orijinal veri kümesine yönelik olarak belirlenmiş hiper parametre değerleri kullanılarak yolcu talep tahmin modelleri kurulmuştur. Modeller, açık kaynak kodlu *R* yazılımı üzerinde gerçekleştirilmiştir. Modelleri doğrulanması amacıyla model doğrulama tekniklerinden birisi olan K-Katlı Çapraz Doğrulama tekniği uygulanmıştır. K-Katlı Çapraz Doğrulama sonucunda tahmin modelleri model performans değerlendirme metrikleri ile karşılaştırılmıştır. Karşılaştırma sonucunda tüm metriklerde XGBoost tahmin modelinin daha iyi sonuç verdiği görülmüştür. Bu yüzden veri kümelerinin karşılaştırılmasında kullanılacak model olarak belirlenmiştir. Veri kümeleri karşılaştırıldığında orijinal veri kümesinin en iyi sonucu verdiği görülmektedir. Ancak Pearson Korelasyon fazlalık değerlendirmesine yönelik veri kümesine ait sonuçların orijinal veri kümesine yönelik sonuçlara oldukça yakın olduğu görülmektedir. Bu yüzden en ideal modelin seçimi için hesaplama maliyetlerinin karşılaştırılması sonucunda Pearson Korelasyon fazlalık değerlendirmesine yönelik veri kümesinin %21 daha düşük hesaplama maliyeti gerektirdiği görülmüştür. Ayrıca 82 faktör sayısının 51'e düşmesi ile toplanması gereken veri miktarında da düşüşe gidilmiştir. Bu özelliklerin değerlendirilmesi ile en ideal model olarak Pearson Korelasyon fazlalık değerlendirmesine yönelik veri kümesinin kullanıldığı XGBoost tahmin modeli seçilmiştir.

XGBoost modeli ile faktör önem düzeyleri sıralanarak modele en çok bilgi kazancını sağlayan faktörlerin belirlenmesi sağlanmıştır. Oluşturulan grafik incelendiğinde ilk

sırada 1600 metrelik hizmet alanında bulunan farklı ulaşım modu sayısı yer almaktadır. Bu sonuç ile bir istasyonun çevresindeki ulaşım modu çeşitliliğine ait verinin bir tahmin modelinde bulunmasının tahmin sonuçlarının doğruluğunu artırıcı bir etkiye sahip olacağı sonucuna varılmıştır. Ayrıca en önemli ilk 20 faktör arasında çoğunlukla mekânsal faktörlerin bulunduğu görülmüştür. Özellikle en yüksek öneme sahip ilk 6 faktörün mekânsal olduğu ve bu yüzden yolcu talebi modeli oluşturulurken mekânsal faktörlerin önemle değerlendirilmesi gerektiği sonucuna varılmıştır. Bu sonuçlar neticesinde bir istasyondaki yolcu talebini oluşturan faktörlerin çevresiyle ilişkili kentsel ilgi noktalarına bağlı olduğu sonucuna ulaşılmıştır. Zamansal faktörlerden ise hafta içi, sonbahar ve Pazar faktörleri ön plana çıkmıştır. Genel hatlarıyla incelendiğinde hem mekânsal hem de zamansal faktörlerin modelde farklı seviyelerde doğruluğu artırıcı katkılar sundukları görülmüştür. Yolcu talebinin istasyonun mekânsal çevresiyle daha çok şekillendiği, zamansal anlamda ise hafta içi etkisinin ön plana çıktığı görülmüştür.

XGBoost tahmin modeli daha sonrasında M4 Metro, Marmaray istasyonları ve proje aşamasında bulunan M4 Metro Uzatma ve Gebze-Darica Metro istasyonları için 2024 yılına ait günlük yolcu talebinin tahmin edilmesinde kullanılmıştır. Modelin eğitilmiş olduğu istasyondaki zamansal dalgalanmaları iyi öğrenip yeni tahminlere ne kadar yansıtılabildiğinin incelenmesi amacıyla Marmaray ve M4 Metro istasyonlarının 2023 mevcut ve 2024 tahmin edilen yolcu talebi değerleri grafik üzerinde karşılaştırılmıştır. Karşılaştırma amacıyla faktör önem düzeyi sıralamasında zamansal faktörler arasında en üst sırada bulunan hafta içi faktörü kullanılmıştır. Hafta içine yönelik oluşturulan grafik incelendiğinde 2023 yılı için görülen yolcu talebindeki dalgalanmalar 2024 yılı için doğru şekilde yansıtılabildiği görülmüştür. Ayrıca her bir istasyon için temel yolcu talebinin karşılaştırılması amacıyla öncelikle 2023 yılında gerçekleşen istasyon bazlı yolcu talebi faktör önem düzeyi sıralamasında en üst sırada bulunan 1600 metrelik hizmet alanında bulunan farklı ulaşım modu faktörü ile incelenmiştir. Hizmet alanı içerisinde ulaşım modu çeşitliliğin yüksek olduğu istasyonlarda gerçekleşmiş olan günlük yolcu talebi miktarının da genellikle yüksek olduğu görülmüştür. Böylece yüksek ulaşım modu çeşitliliği ile yüksek yolcu talebinin yaşanmasının beklenen bir durum olduğu sonucuna varılmıştır. Bu sonuçlar doğrultusunda proje aşamasındaki istasyonlar için tahmin edilen yolcu talebi ve istasyon hizmet alanındaki farklı ulaşım modu sayısı beraber incelenmiştir. İnceleme sonucunda ulaşım modu çeşitliliği yüksek

olan istasyonlar için tahmin edilen günlük ortalama yolcu talebinin de yüksek olduğu görülmüştür. Tüm bu değerlendirmeler sonucunda XGBoost tahmin modelinin eğitildiği istasyonlarda meydana gelen yolcu talebi dalgalanmalarını ve temel yolcu talebini iyi öğrenebildiği ve bunu proje aşamasında bulunan istasyonlarda iyi uygulayabildiği çıkarımı yapılmıştır. Bu sonuçlar doğrultusunda çalışmada kullanılan veri kümesinin ve makine öğrenmesi yönteminin farklı bölgelerde bulunan mevcut veya proje aşamasındaki istasyonlar için de uygulanabilir bir yolcu talep tahmin modeli olduğu düşünülmektedir.

Gerçekleştirilen tez çalışması, yeni nesil hareketlilik kapsamında toplu taşıma sistemleri için mekânsal zekâ yöntemlerinin kullanımı doğrultusunda ölçeklenebilir yolcu talep tahminine yönelik bir metodoloji sunmaktadır. Tez çalışmasının literatüre temel katkısı, yolcu talebinin tahmin edilmesinde sıklıkla kullanılan zamansal faktörlerin çalışma bölgesine has takvimsel değişikler ile genişletilmesi, mevcut çalışmalarda çok kısıtlı bir şekilde değerlendirilen mekânsal faktörlerin istasyon bazında kentsel ilgi noktaları ve demografik yapının CBS analizleri ile üretilmesi, yapay zekâ içerisindeki gelişmiş makine öğrenmesi metotları ile tahmin modellerinin kurulmasından önce zamansal ve mekânsal faktörleri içeren zengin veri kümesine karşılaştırmalı bir şekilde özellik seçimi uygulanması, ideal veri kümesine yönelik oluşturulan modellerin proje aşamasındaki istasyonlar için ileriye dönük tahmin modellerinin kurulması ve bu özellikleriyle birlikte ölçeklenebilir tahmin modellerinin oluşturulmasına yönelik metodoloji sunmasıdır.

Tez çalışması ile elde edilen sonuçlar, toplu taşımanın sürdürülebilirliği dikkate alınarak değerlendirildiğinde yolcu talebi tahmin çalışmalarının verimli bir şekilde gerçekleştirilebilmesi için öneriler sunulmuştur:

- Yolcu talebindeki dalgalanmaların ve temel yolcu talebinin yüksek doğrulukta tahmini için hem zamansal hem de mekansal faktörlerin bölgenin özellikleri de dikkate alınarak tespiti, temini ve hazırlığı gerekmektedir.
- Mekânsal faktörlerin CBS yazılımı ortamında ağ tabanlı hizmet alanı ve en yakın tesis analizleri ile anlamlı hale getirilmesi gerçek yürüme mesafelerinin dikkate alınması açısından önemlidir.

- Yüksek doğrulukta tahminler için karmaşık ilişkileri açıklamadaki gücü ile ön plana çıkan XGBoost algoritmasının veri kümesine uygun hiper parametreler ile kullanımı gerekmektedir.
- Zamansal ve mekânsal faktörler açısından büyük veri kümesiyle tahmin modeli geliştirilmesinde Pearson Korelasyon gibi özellik seçimi yöntemlerinin uygulanması ile model performansının artırılması gerekmektedir.
- Makine öğrenmesi yöntemleri ile oluşturulan tahmin modellerinin K-Katlı Çapraz Doğrulama teknikleri ile doğrulanması modelin geçerliliği açısından önemlidir.
- Oluşturulan tahmin modellerinin performansları performans değerlendirme metrikleri ile karşılaştırılmalı ve hesaplama maliyetleri değerlendirilmelidir.
- Çalışmada kullanılan faktörlerin yolcu talep tahmin modeline sağladıkları doğruluk artırıcı katkı ortaya konmalı ve ileriki çalışmalara yönelik olarak veri gereksinimlerinin tespiti gerekmektedir.
- Çalışma kapsamında otopark verisine ulaşılamadığından, mekânsal faktörler arasında otopark faktörüne yer verilmemiştir. Gelecek çalışmalarda, istasyon çevresindeki otopark varlığının yolcu talebi üzerindeki etkisinin belirlenebilmesi amacıyla bu faktörün yolcu talep tahmin modellerine dâhil edilerek model performansına katkısının test edilmesi önerilmektedir.
- Daha sonraki çalışmalarda özellik seçimi yöntemlerinden Gömülü yöntemlerin model doğruluğunu artırıcı etkisi test edilebilir.
- Günlük olarak değerlendirilen yolcu talep tahmini saatlik olarak da değerlendirilip günün farklı zaman aralıklarında yolcu talebinin değişimi incelenebilir.
- İleriki çalışmalarda bir istasyondaki yolcu talebinin istasyon çevresindeki paylaşımlı bisiklet ve skuter gibi ulaşım modları ile olan ilişkisi ortaya çıkarılabilir.
- Çalışmanın çıktıları doğrultusunda, mekânsal faktörlerin yolcu talebi üzerindeki etkisi dikkate alınarak, gelecekteki çalışmalarda proje aşamasındaki metro ve banliyö hatlarında alternatif istasyonlar arasında en ideal yolcu talebi

potansiyeline sahip istasyonların belirlenmesine yönelik çalışmalar gerçekleştirilebilir.

- Gelecek çalışmalar için kümeleme metotlarının yolcu talebi tahmin doğruluğu üzerindeki etkisi test edilebilir.
- Yolcu biniş verilerinin kullanımı ile gerçekleştirilen bu çalışma doğrultusunda ileriki çalışmalarda hem biniş hem iniş verisinin mevcudiyeti durumunda derin öğrenme metotlarına yönelik çalışmalar gerçekleştirilebilir.



KAYNAKLAR

- Abduljabbar R. L., Liyanage S., Dia H., (2021), "The role of micro-mobility in shaping sustainable cities: A systematic literature review", *Transportation Research Part D: Transport and Environment*, 92, 102734.
- Aksangür İ., Eren B., Erden C., (2022), "Evaluation of data preprocessing and feature selection process for prediction of hourly pm10 concentration using long short-term memory models", *Environmental Pollution*, 311, 119973.
- Alam T., Gupta R., Ahamed N. N., Ulah A., Almaghthwi A., (2024), "Towards sustainable iot-based smart mobility systems in smart cities", *GeoJournal*.
- Alpaydın E., (2014), "Decision Trees", MIT Press, 213–238.
- Alsahaf A., Petkov N., Shenoy V., Azzopardi G., (2021), "A framework for feature selection through boosting", *Expert Systems with Applications*, 187, 115895.
- Asthana P., Mishra S., Gupta N., Derawi M., Kumar A., (2023), "Prediction of student's performance with learning coefficients using regression based machine learning models", *IEEE Access*, PP, 1–1.
- Atzori L., Iera A., Morabito G., (2010), "The internet of things: A survey", *Computer Networks*, 54(15), 2787–2805.
- Bagchi M., White P., (2005), "The potential of public transport smart card data", *Transport Policy*, 12, 464–474.
- Bargam B., Boudhar A., Kinnard C., Bouamri H., Nifa K., Chehbouni A., (2024), "Evaluation of the support vector regression (svr) and the random forest (rf) models accuracy for streamflow prediction under a data-scarce basin in morocco", *Discover Applied Sciences*.
- Baro J., Khouadjia M., (2021), "Passenger flow forecasting on transportation network: sensitivity analysis of the spatiotemporal features", in *2021 International Conference on Data Mining Workshops (ICDMW)*, 734–741, Auckland, New Zealand, 7-10 December.
- Boreiko O., Teslyuk V., (2016), "Structural model of passenger counting and public transport tracking system of smart city", in *2016 XII International Conference on Perspective Technologies and Methods in MEMS Design (MEMSTECH)*, 124–126, Lviv - Polyana, Ukraine, 20-24 April.
- Borra S., Di Ciaccio A., (2002), "Improving nonparametric regression methods by bagging and boosting", *Computational Statistics Data Analysis*, 38(4), 407–420. *Nonlinear Methods and Data Mining*.

Brcic D., Slavulj M., Šojat D., Jurak J., (2018), The role of smart mobility in smart cities, in “5th International Conference on Road and Rail Infrastructure”, 1601–1606, Zadar , Croatia,17-19 May.

Breiman L., (1996), “Bagging Predictors”, Machine Learning, 24, 123–140.

Breiman L., (2001), “Random forests”, Machine Learning, 45, 5–32.

Cai J., Luo J., Wang S., Yang S., (2018), “Feature selection in machine learning: A new perspective”, Neurocomputing, 300, 70–79.

Chan C. C., Bouscayrol A., Chen K., (2010), “Electric, hybrid, and fuel-cell vehicles: Architectures and modeling”, IEEE Transactions on Vehicular Technology, 59(2), 589–598.

Chen T., Guestrin C., (2016), “Xgboost: A scalable tree boosting system”, in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 785–794.

Choi Y., (2023), “Geoai: Integration of artificial intelligence, machine learning, and deep learning with gis”, Applied Sciences.

Cyril A., Mulangi R. H., George V., (2018), “Modelling and forecasting bus passenger demand using time series method”, in 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions (ICRITO), 460–466, Noida, India, 29-31 August.

Dietterich T., (2002), “Ensemble learning”, The handbook of brain theory and neural networks, 110–125.

Dietterich T., (2003), “Machine learning”, Proceedings of the 2nd International Workshop on Multiple Classifier Systems, 1056-1059, Berlin, Heidelberg, 2-4 July.

Drucker H., (1997), “Improving regressors using boosting techniques”, Proceedings of the 14th International Conference on Machine Learning.

Egercioğlu Y., Doğan K., (2016), “Effective sustainable public transportation approaches in İzmir”, Procedia - Social and Behavioral Sciences, 223, 487–492.

El Aboudi N., Benhlila L., (2016), “Review on wrapper feature selection approaches”, in 2016 International Conference on Engineering MIS (ICEMIS), 1–5, Agadir, Morocco, 22-24 September.

Faria R., Brito L., Baras K., Silva J., (2017), “Smart mobility: A survey”, in 2017 International Conference on Internet of Things for the Global Community (IoTGC), 1–8, Madeira Island, City of Funchal,10-13 July.

Feng, C.-X. J., Yu Z.-G. S., Wang, J.-H. J., (2005), “Validation and data splitting in predictive regression modeling of honing surface roughness data”, *International Journal of Production Research*, 43(8), 1555–1571.

Feng F., Zhang J., Liu C., Li W., Jiang Q., (2021), “Short-term railway passenger demand forecast using improved wasserstein generative adversarial nets and web search terms”, *IET Intelligent Transport Systems*, 15(3), 432–445.

Fouracre P., Dunkerley C., Gardner G., (2003), “Mass rapid transit systems for cities in the developing world”, *Transport Reviews*, 23, 299–310.

Freund Y., Schapire R. E., (1996), “Experiments with a new boosting algorithm”, in *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, ICML’96*, Morgan Kaufmann Publishers Inc., 148–156, San Francisco, CA, USA, 3-6 July.

Friedman J., (2001), “Greedy function approximation: A gradient boosting machine”, *The Annals of Statistics*.

Gao S., (2021), “Geospatial Artificial Intelligence (GeoAI)”, Oxford Bibliographies.

Gatera A., Kuradusenge M., Bajpai G., Mikeka C., Shrivastava S., (2023), “Comparison of random forest and support vector machine regression models for forecasting road accidents”, *Scientific African*, 21, e01739.

Gey S., Poggi J.-M., (2006), “Boosting and instability for regression trees”, *Computational Statistics Data Analysis*, 50(2), 533–550.

Ghandeharioun Z., Zendehdel P., Wu W., (2023), “Exploring deep learning approaches for short-term passenger demand prediction”, *Data Science for Transportation*.

Grömping U., (2009), “Variable importance assessment in regression: Linear regression versus random forest”, *The American Statistician*, 63, 308–319.

Hassan M., Kaabouch N., (2024), “Impact of feature selection techniques on the performance of machine learning models for depression detection using eeg data”, *Applied Sciences*, 14, 10532.

Hussain Z., Marcel B., Majeed A., Tsimisaraka R. S. M., (2023), “Effects of transport–carbon intensity, transportation, and economic complexity on environmental and health expenditures”, *Environment, Development and Sustainability: A Multidisciplinary Approach to the Theory and Practice of Sustainable Development*, 26(7), 16523–16553.

Jaiswal J., Samikannu R., (2017), “Application of random forest algorithm on feature subset selection and classification and regression”, in *2017 World Congress on Computing and Communication Technologies (WCCCT)*, 65–68, Tiruchirappalli, Tamil Nadu, India, 2-4 February.

Jiang X., Xu C., (2022), “Deep learning and machine learning with grid search to predict later occurrence of breast cancer metastasis using clinical data”, *Journal of clinical medicine*.

Jie H., Zou H., Xu Q., (2021), “Forecasting daily mrt passenger flow in Taipei based on google search queries”, in 2021 International Symposium on Computer Science and Intelligent Controls (ISCSIC), 46–50, Rome, Italy, 12-14 November.

Kendra M., Blažeková O., Vojteková M., (2024), “Determinants of irregular demand for regional rail passenger services – case study of high tatras in slovakia”, *Transportation*, 1–17.

Khalef R., El-adaway I. H., (2021), “Automated identification of substantial changes in construction projects of airport improvement program: Machine learning and natural language processing comparative analysis”, *Journal of Management in Engineering*, 37(6).

Kidmose B., (2025), “A review of smart vehicles in smart cities: Dangers, impacts, and the threat landscape”, *Vehicular Communications*, 51, 100871.

Kim S., Shin D. H., (2016), “Forecasting short-term air passenger demand using big data from search engine queries”, *Automation in Construction*, 70, 98–108.

Kotsiantis S., Kanellopoulos D., (2012), “Combining bagging, boosting and random subspace ensembles for regression problems”, *International journal of innovative computing, information control: IJICIC*.

Kusonkhum W., Srinavin K., Leungbootnak N., (2022), “Using a machine learning approach to predict the thailand underground train’s passenger”, *Journal of Advanced Transportation*, 2022, 1–15.

Li C., Wang X., Cheng Z., Bai Y., (2020), “Forecasting bus passenger flows by using a clustering-based support vector regression approach”, *IEEE Access*, 8, 19717–19725.

Li Y., Miao L., Chen Y., Hu Y., (2019), “Exploration of sustainable urban transportation development in china through the forecast of private vehicle ownership”, *Sustainability*.

Lin C.-Y., (2021), “Forward stepwise random forest analysis for experimental designs”, *Journal of Quality Technology*, 53(5), 488–504.

Liu L., Chen R.-C., Zhu S., (2020), “Impacts of weather on short-term metro passenger flow forecasting using a deep lstm neural network”, *Applied Sciences*.

Liyanage S., Abduljabbar R., Dia H., Tsai P.-W., (2022), “Ai-based neural network models for bus passenger demand forecasting using smart card data”, *Journal of Urban Management*, 11(3), 365–380.

Makridakis S., Spiliotis E., Assimakopoulos V., Semenoglou A.-A., Mulder G., Nikolopoulos K., (2017), “Statistical, machine learning and deep learning forecasting

methods: Comparisons and ways forward”, *Journal of the Operational Research Society*, 74, 1–20.

Martinez L. M., Pritchard J. P., Crist P., (2024), “Shared mobility’s role in sustainable mobility: Past, present, and future”, *Annual Review of Environment and Resources*, 49(Volume 49, 2024), 191–222.

Milenkovic M., Svadlenka L., Melichar V., Bojovic N., Avramovic Z., (2016), “Sarima modelling approach for railway passenger flow forecasting”, *Transport*, 33, 1–8.

Nar M., Arslankaya S., (2022), “Passenger demand forecasting for railway systems”, *Open Chemistry*, 20(1), 105–119.

Nilsson N., (1998), “Introduction to Machine Learning”, MIT Press (to appear).

Otchere D. A., Ganat T. O. A., Ojero J. O., Tackie-Otoo B. N., Taki M. Y., (2022), “Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions”, *Journal of Petroleum Science and Engineering*, 208, 109244.

Özbayrak F., Foster J. T., Pyrcz M. J., (2024), “Spatial bagging to integrate spatial correlation into ensemble machine learning”, *Computers Geosciences*, 186, 105558.

Parsa A. B., Movahedi A., Taghipour H., Derrible S., Mohammadian A. K., (2020), “Toward safer highways, application of xgboost and shap for real-time accident detection and feature analysis”, *Accident Analysis Prevention*, 136, 105405.

Probst P., Wright M. N., Boulesteix A., (2019), “Hyperparameters and tuning strategies for random forest”, *WIREs Data Mining and Knowledge Discovery*.

Pudjihartono N., Fadason T., Kempa-Liehr A. W., O’Sullivan J. M., (2022), “A review of feature selection methods for machine learning-based disease risk prediction”, *Frontiers in Bioinformatics*.

Rastogi A., Taterh S., Kumar B., (2023), “Internet of things enabled machine learning-based smart systems: A bird’s eye view”, in *Eng. Proc.*, 2023, 36.

Ratnasingam S., Muñoz-Lopez J., (2023), “Distance correlation-based feature selection in random forest”, *Entropy*.

Rocco di Torrepadula F., Napolitano E. V., Di Martino S., Mazzocca N., (2024), “Machine learning for public transportation demand prediction: A systematic literature review”, *Engineering Applications of Artificial Intelligence*, 137, 109166.

Ryu S.-E., Shin D.-H., Chung K., (2020), “Prediction model of dementia risk based on xgboost using derived variable extraction and hyper parameter optimization”, *IEEE Access*, 8, 177708–177720.

Ryus P., Danaher A., Walker M., Nichols F., Carter B., Ellis E., Cherrington L., Bruzzone A., (2013), “Transit capacity and quality of service manual”, 3rd ed.

Sagi O., Rokach L., (2018), “Ensemble learning: A survey”, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8, e1249.

Sarker I., (2021), “Machine learning: Algorithms, real-world applications and research directions”, SN Computer Science.

Segal M., (2003), “Machine learning benchmarks and random forest regression”, Technical Report, Center for Bioinformatics Molecular Biostatistics, University of California, San Francisco.

Segovia J., Toaquiza J., Llanos J., Rivas D., (2023), “Meteorological variables forecasting system using machine learning and open-source software”, Electronics, 12, 1007.

Sison N., Li L., Han M., (2021), “Survey of machine learning and deep learning techniques for travel demand forecasting”, in 2021 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI), 606–613, Atlanta, GA, USA, October 18-21, 2021.

Smith T. R., (1984), “Artificial intelligence and its applicability to geographical problem solving”, The Professional Geographer, 36(2), 147–158.

Şişman S., Aydinoglu A., (2023), “Toplu taşınmaz değerlemesi için coğrafi verinin İşlenmesi ve rastgele orman makine Öğrenmesi tekniğiyle tahmin modelinin geliştirilmesi”, Harita Dergisi.

Taheri N., Tucci M., (2024), “Enhancing regional wind power forecasting through advanced machine-learning and feature-selection techniques”, Energies.

Tarwidi D., Pudjaprasetya S. R., Adytia D., Apri M., (2023), “An optimized xgboost-based machine learning method for predicting wave run-up on a sloping beach”, MethodsX, 10, 102119.

Teimourian A., Rohacs D., Dimililer K., Teimourian H., Yildiz M., Kale U., (2024), “Airfoil aerodynamic performance prediction using machine learning and surrogate modeling”, Heliyon, 10(8), e29377.

Titterington M., (2010), “Neural networks”, WIREs Computational Statistics, 2(1), 1–8.

Toque F., Khouadjia M., Come E., Trepanier M., Oukhellou L., (2017), “Short long term forecasting of multimodal transport passenger flows with machine learning methods”, in 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), 560–566, Yokohama, JAPAN, 16-19 October.

Vakula D., Raviteja B., (2017), “Smart public transport for smart cities”, in 2017 International Conference on Intelligent Sustainable Systems (ICISS), 805-810, Palladam, India, 7-8 December.

Web 1, (2025), <https://www.worldbank.org/en/topic/urbandevelopment/overview>, (Eriřim Tarihi: 02/01/2025).

Web 2, (2025), <https://data.tuik.gov.tr/Bulten/Index?p=Adrese-Dayali-Nufus-Kayit-Sistemi-Sonuclari-2024-53783>, (Eriřim Tarihi: 05/04/2025).

Web 3, (2025), <http://pendik.gov.tr/cografya>, (Eriřim Tarihi: 05/04/2025).

Web 4, (2025), <https://www.kocaeli.bel.tr/sosyo-ekonomik-yapi.html>, (Eriřim Tarihi: 05/04/2025).

Web 5, (2025), <https://www.gebze.bel.tr/bugunku-gebze.html>, (Eriřim Tarihi: 06/04/2025).

Web 6, (2025), <https://www.metro.istanbul/Hatlarimiz/HatDetay?hat=M4>, (Eriřim Tarihi: 07/04/2025).

Web 7, (2025), https://www.tcddtasimacilik.gov.tr/marmaray/tr/gunluk_tren_saatleri, (Eriřim Tarihi: 07/04/2025).

Web 8, (2025), <https://www.gebzedaricametro.com/>, (Eriřim Tarihi: 07/04/2025).

Web 9, (2025), <https://www.metro.istanbul/Hatlarimiz/ProjeHalindekiHatlar?projeInsaat=0&q=28>, (Eriřim Tarihi: 07/04/2025).

Web 10, (2024), <https://data.ibb.gov.tr/>, (Eriřim Tarihi: 15/12/2024).

Web 11, (2025), <https://www.meb.gov.tr/2023-2024-egitim-ogretim-yilina-ait-calisma-takvimi-aciklandi/haber/30337/tr>, (Eriřim Tarihi: 10/05/2025).

Web 12, (2025), <https://www.meb.gov.tr/2024-2025-egitim-ogretim-yili-takvimi-aciklandi/haber/33888/tr>, (Eriřim Tarihi: 10/05/2025).

Web 13, (2025), https://www.gtu.edu.tr/fileman/Files/UserFiles/oidb/AKADEMIK_TAKVIM/2023-2024%20akademik%20takvim%20revize.pdf, (Eriřim Tarihi: 07/04/2025).

Web 14, (2025), https://www.gtu.edu.tr/fileman/Files/UserFiles/oidb/AKADEMIK_TAKVIM/2024-2025%20akademik%20takvim.pdf, (Eriřim Tarihi: 07/04/2025).

Web 15, (2025), <https://weatherspark.com/y/148623/Average-Weather-at-Sabiha-G%C3%B6k%C3%A7en-International-Airport-Turkey-Year-Round>, (Eriřim Tarihi: 07/04/2025).

Web 16, (2025), https://www.takvim.com/2023_takvimi.html, (Erişim Tarihi: 07/04/2025).

Web 17, (2025), <https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview>, (Erişim Tarihi: 01/05/2025).

Web 18, (2025), <https://pro.arcgis.com/en/pro-app/latest/help/analysis/networks/network-analyst-solver-types.htm>, (Erişim Tarihi: 01/05/2025).

Web 19, (2025), <https://pro.arcgis.com/en/pro-app/latest/help/analysis/networks/network-analyst-solver-types.htm>, (Erişim Tarihi: 01/05/2025).

Welde M., (2012), “Are smart card ticketing systems profitable? Evidence from the city of trondheim”, *Journal of Public Transportation*, 15(1), 133–148.

Witten I. H., Frank E., Hall M. A., (2011), Chapter 1 - what’s it all about?, in I. H. Witten E. Frank M. A. Hall, eds, “Data Mining: Practical Machine Learning Tools and Techniques (Third Edition)”, third edition Edition, The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann, Boston, 3–38.

Yao P., (2009), Credit scoring using ensemble machine learning, in “2009 Ninth International Conference on Hybrid Intelligent Systems”, Vol. 3, 244–246, Shenyang, China, 12-14 August.

Yatawatta S., (2024), “Reinforcement learning”, *Astronomy and Computing*, 48, 100833.

Ying X., (2019), “An overview of overfitting and its solutions”, *Journal of Physics: Conference Series*, 1168, 022022.

Yu L., Liu H., (2004), “Efficient feature selection via analysis of relevance and redundancy”, *Journal of Machine Learning Research*, 5, 1205–1224.

Zantalis F., Koulouras G., Karabetsos S., Kandris D., (2019), “A review of machine learning and iot in smart transportation”, *Future Internet*.

Zhang C.-X., Zhang J.-S., Zhang G.-Y., (2009), “Using boosting to prune double-bagging ensembles”, *Computational Statistics Data Analysis*, 53(4), 1218–1231.

Zhou G., Tang J., (2020), “Forecast of urban rail transit passenger flow in holidays based on support vector machine model”, in 2020 5th International Conference on Electromechanical Control Technology and Transportation (ICECTT), 585–589, Nanchang, China, 15-17 May.

ÖZGEÇMİŞ

Ali Doğan GÜMÜŞSOY, 2022 yılında Konya Teknik Üniversitesi Mühendislik ve Doğa Bilimleri Fakültesi Harita Mühendisliği bölümünü bölüm birinciliği ile bitirerek lisans derecesini almıştır. Lisansüstü eğitimine Konya Teknik Üniversitesi Harita Mühendisliği Anabilim Dalı'nda başlamış ve 2024 yılında Gebze Teknik Üniversitesi Ulaşım Teknolojileri Anabilim Dalı Akıllı Şehir ve Ulaşım Teknolojileri Programı'na yatay geçiş yapmıştır. 2024 yılından bu yana Gebze Teknik Üniversitesi Ulaşım Teknolojileri Enstitüsü Araştırma Görevlisi olarak görev yapmaktadır.



TEZ METNİNDEN TÜRETİLEN YAYINLAR/SUNUMLAR

Gümüşsoy A. D., Aydınoglu A. Ç., (2025), “Akıllı Kart Verileri ile Mekansal Analitik ve Makine Öğrenmesi Tekniklerinin Kullanımıyla Toplu Taşıma Yolcu Talep Tahmini”, GTU 9. Lisansüstü Araştırmalar Sempozyumu, Kocaeli, Türkiye, 26-27 Mayıs.

Marmara Belediyeler Birliği (MBB) – Lisansüstü Tez Desteği Programı – 2025 Yüksek Lisans Araştırmacı Desteği.

