T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ VE GENETİK ALGORİTMA TEMELLİ ATM LOKASYONLARININ BELİRLENMESİ

İzzet BİRDEN

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Danışman

Prof. Dr. Banu DİRİ

Temmuz, 2024

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

MAKİNE ÖĞRENMESİ VE GENETİK ALGORİTMA TEMELLİ ATM LOKASYONLARININ BELİRLENMESİ

İzzet BİRDEN tarafından hazırlanan tez çalışması 17.07.2024 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı, Bilgisayar Mühendisliği Programı **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

Prof. Dr. Banu DİRİ
Yıldız Teknik Üniversitesi
Danışman

Jüri Üyeleri

Prof. Dr. Banu DİRİ, Danışman
Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üye. Furkan ÇAKMAK, Üye
Yıldız Teknik Üniversitesi _____

Dr. Öğr. Üye. Öznur ŞENGEL, Üye
İstanbul Kültür Üniversitesi _____

Danışmanım Prof. Dr. Banu DİRİ sorumluluğunda tarafımca hazırlanan “Makine Öğrenmesi ve Genetik Algoritma Temelli ATM Lokasyonlarının Belirlenmesi” başlıklı çalışmada veri toplama ve veri kullanımında gerekli yasal izinleri aldığımı, diğer kaynaklardan aldığım bilgileri ana metin ve referanslarda eksiksiz gösterdiğimi, araştırma verilerine ve sonuçlarına ilişkin çarpıtma ve/veya sahtecilik yapmadığımı, çalışmam süresince bilimsel araştırma ve etik ilkelerine uygun davrandığımı beyan ederim. Beyanımın aksinin ispatı halinde her türlü yasal sonucu kabul ederim.

İzzet BİRDEN

İmza

Biricik Eşim

ve

Oğluma

TEŞEKKÜR

Çalışmalarım boyunca sürekli desteğini hissettiğim ve bana yol gösteren kıymetli danışmanım Sayın Prof. Dr. Banu Diri'ye, iş hayatı ve akademik çalışmalarımın yoğunluğunda her daim işimi kolaylaştıran ve manevi destekleriyle bana güç olan sevgili eşim Şeyma Birden'e, varlığıyla moral bulduğum canım oğlum Kerem'e ve üzerimde her daim emeği olan aileme teşekkürü borç bilirim.

İzzet BİRDEN

İÇİNDEKİLER

KISALTMA LİSTESİ vii
ŞEKİL LİSTESİ ix
TABLO LİSTESİ x
ÖZET xi
ABSTRACT xiii
1 GİRİŞ 1
1.1 Literatür Özeti 3
1.2 Tezin Amacı 10
1.3 Literatüre Katkı 11
1.4 Tez Organizasyonu 11
1.5 Hipotez 12
2 VERİ SETİ 13
2.1 Veri Seti Detayları 13
3 YÖNTEMLER 18
3.1 Özelliklerin Belirlenmesi 18
3.2 Özellik Seçimi 19
3.3 Makine Öğrenmesi ve Regresyon Modelleri 21
3.3.1 Doğrusal Regresyon 21
3.3.2 Rastgele Orman 22
3.3.3 Karar Ağaçları 22
3.3.4 Ridge Regresyon 23
3.3.5 Lasso Regresyon 23
3.3.6 Gradient Boosting 23
3.3.7 Elastic Net Regresyonu 24
3.3.8 SVR (Destek Vektör Regresyonu) 24
3.3.9 K-En Yakın Komşu (KNN Regresyonu) 25
3.4 Model Değerlendirme Metrikleri 25
3.4.1 Ortalama Kare Hata (MSE) 25
3.4.2 Ortalama Mutlak Hata (MAE) 26
3.4.3 R-Kare 26
3.5 Genetik Algoritma ve Özellik Seçimi 26
3.5.1 Popülasyon,Birey ve Genetik Operatörler 27

3.5.2 Uyum (Fitness) Fonksiyonu…… 27
3.5.3 Seçim Operatörü…….. 28
3.5.4 Mutasyon Operatörü…….29
3.5.5 Çaprazlama Operatörü……. 30
3.6 Weka ile Özellik Seçimi 31
3.6.1 Weka'nın Özellik Seçimi Arayüzü……...……... 31
3.7 Weka'da Özellik Seçimi İçin Kullanılan Yöntemler 32
3.7.1 InfoGainAttributeEval…….....……....33
3.7.2 GainRatioAttributeEval……...…….……....33
3.7.3 ReliefFAttributeEval……...…….…...34
3.7.4 CfsSubsetEval……...……..... 34
3.7.5 WrapperSubsetEval……...……...34
3.8 Weka ile Özellik Seçimi Parametreleri.....35
4 DENEYSEL SONUÇLAR 36
4.1 Makine Öğrenmesi Modellerinin Sonuçları ve Karşılaştırılması……...36
4.1.1 Makine Öğrenmesi Parametrelerinin Değiştirilmesiyle Oluşan Model Performansları…….38
4.1.2 Özellik Sayısındaki Değişimin Model Performansına Etkisi …..41
4.2 Genetik Algoritma ile Özellik Seçim Sonuçları…….....45
4.3 Weka ile Özellik Seçimi Sonuçları …….……...47
4.4 Tüm Sonuçların Karşılaştırılması……...……...51
5 SONUÇ 52
KAYNAKÇA 54
A ÖZELLİKLER 57
TEZDEN ÜRETİLMİŞ YAYINLAR 59

KISALTMA LİSTESİ

ALG	Algoritma
AHP	Analytic Hierarchy Process
ATM	Automated Teller Machine
BKM	Bankalararası Kart Merkezi
CFLP	Capacitated Facility Location Problem
DFLP	Dynamic Facility Location Problem
DIFLP	Discrete Facility Location Problem
DVM	Destek Vektör Makinesi
DVR	Destek Vektör Regresyonu
FAHP	Fuzzy Analytic Hierarchy Process
FLP	Facility Location Problem
FLRP	Facility Location and Repositioning Problem
GA	Genetik Algoritma
GIS	Geographical Information Systems
KA	Karar Ağaçları
KNN	K-Nearest Neighbors
LR	Lojistik Regresyon
MABAC	Multi-Attributive Border Approximation Area Comparison
MAE	Mean Absolute Error
MCDM	Multi-Criteria Decision-Making
MPFLP	Multi-Period Facility Location Problem
MSE	Mean Squared Error
MÖ	Makine Öğrenmesi
NFLP	Network Facility Location Problem
NRMSE	Normalized Root Mean Squared Error
PCA	Principal Component Analysis
PSO	Particle Swarm Optimization
RNN	Recurrent Neural Network
RF	Random Forest
RO	Rastgele Orman
STFLP	Static Facility Location Problem
SVR	Support Vector Regression

TDFLP	Time Dependent Facility Location Problem
TOPSIS	Technique for Order of Preference by Similarity to Ideal Solution
QR	Quick Response Code

ŞEKİL LİSTESİ

Şekil 1.1 2024 yılı ATM, POS ve ÖKC sayıları **2**
Şekil 2.1 Ham veri setini içeren .csv dosyası **15**
Şekil 2.2 Özellikler arası korelasyon matrisi **15**
Şekil 2.3 Özellikler arası çift grafik (pair plot) **17**
Şekil 3.1 Ortalama kare hata (MSE) formülü **25**
Şekil 3.2 Ortalama mutlak hata (MAE) formülü **26**
Şekil 3.3 R-Kare formülü **26**
Şekil 3.4 Genel genetik algoritma akışı **30**
Şekil 3.5 Weka'nın özellik seçim arayüzü **32**
Şekil 4.1 Genetik algoritma akışı **46**

TABLO LİSTESİ

Tablo 2.1 Veri setinin bazı özellikler için genel istatistikleri **14**
Tablo 3.2 Özellik seçme yöntemleri ile elde edilen özellik sıralamaları **20**
Tablo 4.1 81 özellik için makine öğrenmesi algoritma başarı değerleri **38**
Tablo 4.2 10 özellik için makine öğrenmesi algoritma başarı değerleri **38**
Tablo 4.3 Rastgele Orman algoritmasının hiper-parametrelere göre başarıları.... **39**
Tablo 4.4 Karar Ağacı algoritmasının hiper-parametrelere göre başarıları **40**
Tablo 4.5 SVR algoritmasının hiper-parametrelere göre başarıları **40**
Tablo 4.6 Gradient Boosting algoritmasının hiper-parametrelere göre başarıları **40**
Tablo 4.7 5 özellik için makine öğrenmesi algoritma başarı değerleri **41**
Tablo 4.8 10 özellik için makine öğrenmesi algoritma başarı değerleri **42**
Tablo 4.9 15 özellik için makine öğrenmesi algoritma başarı değerleri **42**
Tablo 4.10 20 özellik için makine öğrenmesi algoritma başarı değerleri **43**
Tablo 4.11 25 özellik için makine öğrenmesi algoritma başarı değerleri **43**
Tablo 4.12 30 özellik için makine öğrenmesi algoritma başarı değerleri **44**
Tablo 4.13 35 özellik için makine öğrenmesi algoritma başarı değerleri **44**
Tablo 4.14 Genetik Algoritma ile seçilen özellikler **47**
Tablo 4.15 Genetik Algoritma ile seçilen özelliklerin Rastgele Orman algoritmasına verilmesiyle oluşan performans sonuçları **47**
Tablo 4.16 CfsSubsetEval ile seçilen özellikler **48**
Tablo 4.17 WrapperSubsetEvaluation ile seçilen özellikler **48**
Tablo 4.18 InfoGainAttributeEval ile seçilen özellikler **49**
Tablo 4.19 ReliefFAttributeEval ile seçilen özellikler **49**
Tablo 4.20 GainRatioAttributeEval ile seçilen özellikler **49**
Tablo 4.21 Weka'da seçilen özelliklerin Rastgele Orman algoritması ile modellenmesi sonucu oluşan performans sonuçları **50**
Tablo 4.22 Özellik seçim yöntemlerinde kullanılan parametreler **50**
Tablo 4.23 Makine Öğrenmesi, Genetik Algoritma ve Weka'nın performans karşılaştırması **51**

ÖZET

Makine Öğrenmesi ve Genetik Algoritma Temelli ATM Lokasyonlarının Belirlenmesi

İzzet BİRDEN

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Yüksek Lisans Tezi

Danışman: Prof. Dr. Banu DİRİ

ATM'ler, bankacılık hizmetlerinin müşterilere ulaştırılmasında önemli rol oynayan dijital kanallardır. Şube gişesi, mobil şube, internet bankacılığı vb. kanallardan yapılan finansal ve finansal olmayan işlemlerin birçoğu ATM'lerden de sağlanmaktadır. Bu bakımdan, ATM'lerin ulaşılabilir lokasyonlarda olmaları müşteriler açısından önem arz etmektedir. ATM'ler, şube bağımlı ve şube bağımsız olarak konumlandırılırlar. Banka şubelerinde bulunan ATM'lerin lokasyonu, ait olduğu şubeye bağlı olmaktadır. Bu nedenle, çalışma yapılırken dış cephe (offsite) ATM'lerin lokasyon seçimleri ele alınmıştır. ATM lokasyon seçim çalışmalarında, literatürde farklı yöntemler denenmiştir. İstatistiki yöntemler, makine öğrenmesi teknikleri ve özellikle Coğrafi Bilgi Sistemleri (GIS) bu alanda uygulanmaktadır. Bu çalışmada ele alınan problem, gerçek bir dünya problemidir. Veri seti olarak, gerçek ATM lokasyon verilerinden oluşan ve farklı demografik, ekonomik ve coğrafi özellikleri içeren bir veri seti kullanılmıştır. Bu çalışmada, lokasyon tespiti için ilk aşamada farklı makine öğrenmesi ve regresyon modelleri denenmiştir. Model, verilen lokasyon girdilerine göre 0 ile 100 arasında bir puan üretmekte ve bu puanın 100'e yakın olması, o lokasyona ATM kurulmasının bankalar açısından

daha faydalı olacağı anlamına gelmektedir. Makine öğrenmesi modelleri arasında en başarılı sonuç Rastgele Orman modeli ile elde edilmiştir. Makine öğrenmesi çalışmasında 97 özellikten oluşan ham veri seti 81 özelliğe indirgenmiştir. İkinci fazda, Genetik Algoritma kullanılarak, özellik seçimi yapılması sağlanmıştır. Toplamda 11 özellik seçilmiş olup, bu özellikler tekrardan ilk fazda makine öğrenmesinde en fazla başarıyı gösteren model olan Rastgele Orman algoritması ile değerlendirilmiştir. Model performansları R-kare, ortalama kare hatası (MSE) ve ortalama mutlak hata (MAE) metrikleriyle ölçülmüştür. Üçüncü aşamada, Weka uygulaması kullanılarak veri setinden özellikler seçilmiş olup, bu özellikler de ikinci fazdaki gibi Rastgele Orman modeli ile değerlendirilmiştir. İlk fazdaki Rastgele Orman modeli 0,91 R-Kare oranıyla başarı elde ederken, Genetik Algoritma ile özellik seçimleri sonrası bu oranın 0,97'ye, Weka ile özellik seçimleri sonrası 0,93'e yükseldiği görülmüştür. Gelecekteki çalışmalarda, daha büyük veri setlerinin kullanımı, farklı makine öğrenmesi algoritmalarının entegrasyonu ve hiper-parametre optimizasyonlarının denenmesiyle modellerin başarısı artırılabilir.

Anahtar Kelimeler: ATM, makine öğrenmesi, genetik algoritma, weka

ABSTRACT

Determination of ATM Locations Based On Machine Learning And Genetic Algorithm

İzzet BİRDEN

Department of Computer Engineering

Master of Science Thesis

Supervisor: Prof. Dr. Banu DİRİ

ATMs are digital channels that play an important role in delivering banking services to customers. Many of the financial and non-financial transactions made through channels such as branch counters, mobile branches, internet banking, etc. are also available through ATMs. In this respect, it is important for customers that ATMs are located in accessible locations. ATMs are positioned as branch-dependent and branch-independent. The location of ATMs in bank branches depends on the branch to which they belong. Therefore, this study focuses on the location selection of offsite ATMs. In ATM location selection studies, different methods have been tried in the literature. Statistical methods, machine learning techniques and especially Geographic Information Systems (GIS) have been applied in this field. The problem in this study is a real world problem. In this study, different machine learning and regression models are tested in the first stage for location detection. The model produces a score between 0 and 100 based on the given location inputs, and a score closer to 100 means that it is more beneficial for banks to install ATMs in that location. The most successful result among the Machine Learning models was obtained with the Random Forest model. In the machine learning study, the raw data set consisting of 97 features was reduced to 81 features. In the second phase, genetic algorithm were used to select features. A total of 11 features were selected

and these features were again evaluated with the Random Forest algorithm, which was the most successful model in Machine Learning in the first phase. Model performances were measured by R-square, mean square error (MSE) and mean absolute error (MAE) metrics. In the third phase, features were selected from the dataset using the Weka application, and these features were evaluated with the Random Forest model as in the second phase. While the Random Forest model in the first phase achieved a success with an R-square ratio of 0.91, this ratio increased to 0.97 after feature selection with Genetic Algorithm and to 0.93 after feature selection with Weka. In future studies, the success of the models can be improved by using larger data sets, integrating different Machine Learning algorithms, and testing hyper-parameter optimizations.

Keywords: ATM, machine learning, genetic algorithm, weka

1
GİRİŞ

Günümüzde, çoğu sektörde olduğu gibi bankalar açısından da müşteri memnuniyetini artırmak için müşteri odaklılık kavramı ön plana çıkmıştır. Bu kapsamda, müşteri gözünde prestij kazanmak için de müşteri odaklı tasarımlar, müşteriye özel bildirimler ve kampanyalar yapılmaktadır. ATM'ler de bir dijital kanal olarak bu memnuniyet göstergesinin ve prestijin önemli bir kaynağıdır.

50 yıl önce bankalar, yeni müşteriler çekmek ve gelir kaynakları yaratmak için şube dışı dağıtım kanallarının geliştirilmesine yatırım yapmaya başlamıştır. 1960'larda ortaya çıkan bankamatikler (ATM'ler) alternatif hizmet kanallarının ilk girişimi sayılabilir. ATM'lerin 1970'lerde yaygınlaşmasının ardından 1990'larda çağrı merkezleri ortaya çıkmış, 2000'lerden sonra internet ve kablosuz teknolojiler bankacılık sektöründe önemli rol oynar hale gelmiştir. Bu sebeple, hızla değişen müşteri davranışları ve beklentileri, bankaları bu alternatif kanalları sürekli iyileştirme çabalarına girmeye teşvik etmiştir.

ATM'den yapılan işlemlerden bazıları, nakit para çekme, yatırma, fatura işlemleri yapma, hesap özeti ve ekstre görüntüleme, havale/EFT gibi para transferi işlemleri gibi temel bankacılık hizmetleri sayılabilir. Özellikle, nakit para çekme ve yatırma gibi nakit işlemler, şube haricinde, dijital veya diğer alternatif kanallar arasında sadece bankamatikler tarafından yapılabilmektedir. Bu tür işlemlerin bankamatiklerden yapılabiliyor olması, şube iş yükünün önemli ölçüde azalmasını sağlamakta ve müşteri memnuniyetini de artırmaktadır. Banka şubelerinin her lokasyona kurulumu kolay olmayacağından, bankamatiklerin müşterilere daha fazla hizmet verebilmesi ve müşteriler tarafından daha fazla ulaşılabilir olması açısından doğru yerlere konumlandırılmaları elzem olmaktadır. ATM'lerin bir diğer önemli özelliği ise yeniden dolaşım yoluyla nakit maliyetinin azaltılmasında da kritik bir rol oynamaya devam etmesidir. Bu nakit dolaşımı, sadece bankalar veya finansal kurumlar için değil, aynı zamanda dolaşımdaki nakit miktarını azaltarak ekonomi ve toplum için de fayda sağlamaktadır [1].

Bankalararası Kart Merkezi (BKM) verilerine göre, 2024 yılında Türkiye'de Şekil 1.1'de görüleceği üzere 53 bin adet ATM yer almaktadır. ATM'leri sayısı, işlem

hacmi ve değeri, kanalın sürekli büyümesini göstermekte ve önemini kanıtlamaktadır [2].

2024 YILI			
Dönem	POS Sayısı	ATM Sayısı	ÖKC Yazılım Sayısı
OCAK	2.007.006	52.971	2.144.483
ŞUBAT	2.006.434	53.073	2.156.906
MART	2.016.342	53.122	2.190.432
1. DÖNEM	2.016.342	53.122	2.190.432
NISAN	2.026.118	53.260	2.219.879
2024 YILI	2.026.118	53.260	2.219.879

Şekil 1.1 2024 yılı ATM, POS ve ÖKC sayıları

ATM'lerde aynı zamanda temassız ödeme yöntemlerinin yaygınlaşması, QR kod ile nakit çekme ve yatırma gibi işlemlerin yapılması, biyometrik giriş yöntemlerinin entegre edilmesi, görme engelli müşteriler için ATM'lerde işlem yaptırılmasının sağlanması ATM'lere bankalar tarafından önemli ölçüde yatırımların devam ettiğini ve aynı zamanda daha güvenli, işlem sürelerini kısaltan ve müşteri odaklı işlemlerin yapılmasının hedeflendiğini göstermektedir.

ATM'lerin stratejik olarak doğru yerlerde konumlandırılmaları, bankaların hizmet kalitesini artırma ve operasyonel verimliliği artırma açısından büyük bir öneme sahiptir. Bankamatiklerin doğru konumlandırılmasının hem banka hem müşteriler tarafından karşılıklı yarar sağlama durumu vardır. Müşteriler, bankamatiklere ne kadar kolay ulaşırlarsa, aldıkları hizmet neticesinde işlemlerini çok daha kısa sürelerde yapma imkanı bulurlar. Banka açısından da yapılan işlemlerden kar sağlama, müşterilere ATM'ler vasıtasıyla ürün ve hizmetlerini pazarlama ve bu sayede ürün satışlarını artırma gibi yararlar sağladığından bahsedilebilir. Bu nedenle, ATM'lerin ideal konumlarda olmaları iki taraf için de önem arz etmektedir. Bu yüzden bankalar, konumlandırma stratejilerini hassas bir şekilde değerlendirmelidir.

ATM'ler konum itibariyle temelde iki kategoriye ayrılır: On-site ve Off-site ATM'ler. Şube bağımlı ATM'ler, lobi (şubenin içinde bulunan) ve on-site (ön cephe), yani şubenin hemen dışında, duvara monte edilmiş; müşterilerin şube içine girmesine gerek kalmadan işlem yapabilmelerine olanak sağlayan ATM'ler olarak ikiye ayrılmaktadır. Ancak genel çerçeve olarak şube ve şube bağımsız ya da on-site ve off-site ATM'ler olarak ikiye ayrılabilir. Şube kurulumlarının kolay

olmaması ve on-site ATM'lerin de şubelerin kuruldukları lokasyonda bulunma zorunluluğundan dolayı, müşterilere yakınlık ve hizmet verme noktasında kısıtlar bulunmaktadır. Müşteriler, şubeye gitme mecburiyetinde olup, özellikle nüfusun yoğun olduğu lokasyonlarda uzun işlem sıraları beklemektedir. Bu bakımdan, off-site ATM'lerin varlığı, bu sorunları aşmak için büyük önem arz etmektedir. Off-site ATM'ler banka şubelerinin bulunmadığı alışveriş merkezleri, caddeler, havaalanları, tren istasyonları, üniversiteler gibi alanlarda yer almaktadır.

ATM'lerin lokasyonlarının belirlenmesi, çok sayıda özelliğin dikkate alınması gereken karmaşık bir problem olarak öne çıkmaktadır. ATM'lerin en uygun lokasyonlarına karar verirken, göz önünde bulundurulması gereken çok fazla kriter bulunmakta ve bu da sorunun karmaşıklığını artırmaktadır. ATM kurulumu ve satın alınması yüksek maliyetler içerdiğinden, ATM'lerin ilk fazda yanlış konumlandırılıp daha sonra taşınması veya kaldırılması, bankalara maddi anlamda büyük bir külfet getirmektedir. Bu noktada, ATM bilgisi yüksek banka uzmanlarının da yorumları, konumlandırma stratejisi açısından önem arz etmektedir. Son dönemlerde, makine öğrenmesi, hibrit ve matematiksel modeller gibi farklı modern teknik ve yaklaşımlarla ATM lokasyonlarının en uygun biçimde seçilebilmesi hedeflenmektedir.

1.1 Literatür Özeti

ATM lokasyonlarının tespit edilmesine yönelik ilk çalışmalar için banka ya da finans sektörlerindeki lokasyon tespit çalışmalarından ziyade, genel lokasyon tespit çalışmaları ön plana çıkmaktadır. Yapılan bir çalışmada, bir yola en az sayıda batarya şarj istasyonu yerleştirerek, sürücülerin ya da elektrikli araçların en kısa yol üzerinden hedef rotaya gitmesi amacı güdülmüş ve çizge teoremi (Graph Theory)'nde Öklid (Euclid) metriği kullanılmıştır [3]. M.Basset ve arkadaşları (2020) açık deniz rüzgar enerji istasyonlarının konumlarının seçilmesiyle alakalı yürüttükleri çalışmada AHP ve PROMETHEE tabanlı MCDM (Çok Kriterli Karar Verme) yaklaşımı önermişlerdir. Bu çalışmada, alandaki uzmanların görüşlerinden de faydalanarak çalışmayla alakalı problemleri ele almışlardır. Alan seçiminde MCDM tekniklerinin literatürde çok fazla kullanılmadığından bahsetmişler ve kullandıkları MCDM yönteminin çoklu kriterlere sahip bir enerji tesisi için uygun

yer seçiminde etkili olduğunu belirtseler de, bu enerji istasyonlarının saha seçimi konusunda hala birçok kısıtlamanın olduğundan bahsetmişlerdir [4].

Banka şubesi lokasyon analizi çalışmalarında ise N.Cinar ve S.Ahiska yaptıkları çalışmada FAHP (Bulanık Analitik Hiyerarşi Süreci) ve TOPSIS (İdeal Çözüme Benzerliğe Göre Tercih Sıralaması Tekniği) yöntemlerini birlikte kullanmışlardır. Veri setindeki kriterlerin ağırlıklarını farklı alanlardan banka yöneticileri belirlemişlerdir. Bulanık AHP ve TOPSIS yöntemlerinin banka şubesi yeri seçiminde kuruluşlar tarafından bir karar destek sistemi olarak kullanabileceği sonucuna varmışlardır [5].

B.Sennaroglu ve G.Celebi yaptıkları çalışmada gerçek bir dünya problemi olan askeri havaalanı için en uygun yer seçimi problemine odaklanmışlardır. Yapılan çalışma ve karar kriterleri daha önce literatürde görülmediğinden orijinal denebilir. Sadece havalimanı gerekliliklerinden ziyade, çevresel ve sosyal etkiler de göz önünde bulundurulmuştur. Bu yüzden, yaptıkları çalışmada üç farklı çok kriterli karar verme yöntemi uygulamışlardır. Bunlar, analitik hiyerarşi süreci entegreli PROMETHEE (Zenginleştirme Değerlendirmesi İçin Tercih Sıralama Organizasyonu Yönetimi) ve VIKOR'dur. PROMETHEE'nin, genelleştirilmiş kriter türlerini seçmek ve parametrelerin değerini belirlemek için güçlü bir araç olduğu belirtilmiştir. Bu üç yöntemin alternatif olarak sıraladığı lokasyonların ve önerdiği en iyi lokasyonun, COPRAS, MAIRCA ve MABAC yöntemleriyle aynı sonucu ürettiği görülmüştür [6].

A.Arabani ve R.Farahani (2011), FLP (Tesis Yer Seçimi Problemi) ile ilgili kümeleme çalışması yapmışlardır. STFLP (Statik Tesis Yer Seçimi Problemi), CFLP (Sürekli Tesis Yer Seçimi Problemi), DIFLP (Ayrık Tesis Yer Seçimi Problemi) ve NFLP (Ağ Tesisi Konumlandırma Problemleri) olarak kategorize edilmiştir. DFLP (Dinamik Tesis Yer Seçimi Problemi)'yi ise, FLRP (Tesis Yer Seçimi ve Tekrardan Konumlandırma Problemi), MPFLP (Çok Dönemli Tesis Yeri Seçim Problemi) ve TDFLP (Zamana Bağlı Tesis Yeri Seçim Problemi) olarak ayırmışlardır [7].

GIS (Coğrafi Bilgi Sistemleri), lokasyon veya yer seçimi problemlerinde en çok kullanılan yöntemlerden biridir. P.Dandekar ve arkadaşları (2014) yaptıkları çalışmada, GIS ve RS (Uzaktan Algılama)'yi kullanarak ATM lokasyonlarının

belirlenmesini amaçlamışlardır. GIS'in etkili planlama, organize etme ve karar verme konularında önemli bir rol oynadığından bahsetmişlerdir. GIS'in ATM'lerin adresleri açısından alanları ya da tesisleri coğrafi olarak kodlama becerisinden yararlandıklarını ve GIS teknolojisinin ağ analisti aracı yardımıyla çalışma alanındaki herhangi bir konumdan en yakın ATM'yi bulma yeteneğini kanıtladığını belirtmişlerdir [8].

J.Zhou ve arkadaşları (2020), coğrafi bilgi sistemleri ve çok kriterli karar verme yöntemlerini birleştirerek, fotovoltaik enerji üretim sistemi ile elektrikli araç şarj istasyonunun kombinasyonu olan PVCS (Fotovoltaik Şarj İstasyonu)'lerin konum kararlarını vermeyi amaçlamışlardır. Coğrafi bilgi sistemlerinin uygunluk analizi ile yedi alan seçmişlerdir. İkili duyarlılık analizi ve karşılaştırmalı analiz kullanarak modelin kararlığını ve güvenliğini artırmışlardır. Sonuç olarak, farklı durgunluk katsayıları belirleyerek karar tercihlerini ifade etmenin gerekli olduğunu belirtmişlerdir [9].

M.Khan ve N.Yesmin, coğrafi bilgi sistemi ile uydu tabanlı görüntü kullanarak ATM lokasyonlarını tespit etmeye odaklanmışlardır [10].

J.Lozano ve arkadaşları (2016), rüzgar çiftliği yer seçimleri için MCDM (Çok Kriterli Karar Verme) yöntemlerinin bazı bulanık yaklaşımlarını kullanmışlardır. Kriterlerin ağırlıklarını elde etmek için FAHP (Bulanık Mantık Hiyerarşi Süreci) uygularken, alternatifleri değerlendirmek için FTOPSIS (İdeal Çözüme Benzerlik ile Tercih Sıralaması İçin Bulanık Teknik) yöntemini uygulamışlardır. Alternatif ve kriterlerin veri tabanını elde etmek için ise GIS kullanmışlardır. FTOPSIS yöntemiyle 33.290 adet alternatif uygun lokasyon olduğunu analiz etmişlerdir. [11].

Tesis, alan ya da saha yönetiminde değişkenlerin çoğu kesin olarak ölçülebilir ya da sayısallaştırılabilir. Bu neden kısıtlamalar ve amaç fonksiyonu matematiksel olarak kesin bir şekilde ifade edilebilir. Ancak aynı şey banka şubeleri ve ATM'ler için söylenemez çünkü bankacılık diğer hizmetlerden farklı olarak çok çeşitli müşterilere hitap eden ve farklı hizmetlerin sağlanmasını gerektiren büyük ölçüde bir sosyal mühendislik problemidir. Hizmet kalitesine ilişkin değişkenlerin birçoğu, müşteri beklentilerinin ve istenen hizmet kalitesinin karşılanamamasına yönelik toleransının ölçülememesi nedeniyle kesin olarak belirlenememektedir. Bu nedenle, bankacılık müşterilerini sosyal ve ekonomik faktörlere dayalı olarak

kümelemek gerekmektedir [12]. N.Kisore ve B.Koteswaraiah çalışmalarında, ATM müşterilerinin konum tercihlerini belirlerken, değerlerdeki bulanıklıkla başa çıkmak için GDBSCAN (Yoğunluk Tabanlı Kümeleme Algoritması) yerine, kendilerinin adlandırdığı GFDBSCAN (Genelleştirilmiş Bulanık Yoğunluk Tabanlı Kümeleme) yöntemini önermişlerdir. Bu yeni kümeleme versiyonunu, müşterileri temel sosyo-ekonomik parametreler etrafında kümelemek için uygulamışlardır. GFDBSCAN'ın hem mekansal hem de mekansal olmayan bulanıklıkla daha iyi başa çıktığını görmüşlerdir. Aynı zamanda, ATM'lerin hizmet alanlarını optimize etmek amacıyla, ATM'lerin yerleştirileceği ideal konumları belirlemek için Voronoi diyagramlarını kullanmışlardır [12].

Literatürde, ATM alan seçimi ile ilgili farklı yöntemler denenmiştir. Yapılan bir çalışmada, mevcut ATM lokasyonlarındaki iş yükü ve yoğunluğu değerlendirmek ve hangi lokasyonlara ilave ATM kurulacağını belirlemek için kuyruk analizi kullanılmıştır [13]. Bir başka çalışmada [14], model ve algoritmalar geliştirdiği benzin istasyonları ve ATM'leri isteğe bağlı hizmet tesisleri olarak adlandırmıştır. Müşterilerin bu tesisleri nihai varış noktası olarak görmedikleri, bir konumdan diğerine planlı bir seyahatte bu tesislerin yanlarından geçtikleri için bu hizmetten yararlandıkları belirtilmiştir.

M. Aldajani ve H.Alfares çalışmalarında (2009), minimum ATM sayısını ve ATM'lerin konumlarını hesaplayan, iki boyutlu konvolüsyona dayalı yeni bir yöntem önermişlerdir. Bu yeni yöntem ile yaptıkları deneylerde, algoritmanın ATM yerleştirme problemlerini optimuma yakın bir şekilde çözdüğünü göstermişlerdir [15].

ATM'ler için bir diğer önemli faktör ise ATM'lerdeki nakit yönetimidir. Etkin bir nakit yönetim sistemine sahip olmak için ATM'lerde nakit yönetim modeli olmalıdır. Bu model sayesinde, ATM'lerdeki nakit para bitmemeli, ATM'ler müşterilere 7/24 hizmet edebilmelidir. Bu model, ATM'lerin konumlarıyla oldukça bağlantılıdır. Dolayısıyla, ATM konumlarını ve nakit yönetimi birlikte düşünmek gereklidir. Bu konudaki bir çalışmayı M.Genevois ve arkadaşları gerçekleştirmişlerdir. Literatürde daha önce bu iki konuyu birleştiren bir çalışma eksikliği olduğunu belirtmişlerdir. Sonuç olarak, nakit yönetimi yol haritasının, ATM'lerin konumlarına göre daha önceden çizilmesi gerektiğini bildirmişlerdir [16].

Literatürde ATM ve diğer tesislerin lokasyon tespiti için PSO (Parçacık Sürü Optimizasyonu) yöntemi de kullanılmaktadır. Y. Li ve arkadaşları yaptıkları çalışmada, ATM konumlandırma modelinin faktörlerini araştırmışlardır. PSO algoritmasına dayanarak coğrafi bilgi sistemi teknolojisini matematiksel modelle birleştirerek tahmin yapmışlardır. Her bir ATM'nin fiyat, kurulum maliyeti vb. açılardan eş olduklarını kabul etmişlerdir [17].

Lokasyon seçim problemlerinin çözümünde makine öğrenmesi kullanımı son yıllarda daha fazla ilgi görmektedir. Makine öğrenmesi, verilerden anlamlı desen ve ilişkiler çıkarma becerisi sayesinde karmaşık ve çok kriterli yer seçimi ve optimizasyon problemlerinde çözümler sunar. Makine öğrenmesi tekniklerinin kullanıldığı ATM lokasyon tespit çalışmaları, literatürde GIS, PSO veya MCDM yöntemleri kadar sık bulunmamaktadır. Yine de yer seçimi problemlerinde faktör analizi, küme analizi, sinir ağları, regresyon analizi ve korelasyon analizi gibi birçok yöntemden faydalanılmaktadır [18]. K.Bilginol ve arkadaşları, kanonik korelasyon analizi ve sıradan en küçük kareler regresyon analizi kullanarak çalışma yürütmüşlerdir. Korelasyon yöntemi ile ATM yer seçimi kararını hangi faktörlerin etkilediğini analiz etmişlerdir. ATM lokasyonlarını etkileyen sonuç kriterleri ise sıradan en küçük kareler regresyon modeli ile analiz edilmiştir. Sonuç olarak, coğrafi bilgiler kullanılarak nüfus, sosyal medya ve rakip bilgilerine dayalı doğrusal regresyon analizi ile ATM işlemlerini tahmin eden bir altyapı geliştirilmiştir. Ayrıca, mevcut ATM'lerin bulunduğu konumlar dışındaki uygulanabilir konumlar ve ATM'nin 500 metre yarıçapındaki konumlar bulunmuştur [18].

H. Yurdakul ve arkadaşları, ATM lokasyon belirleme problemine, Rastgele Orman (Random Forest) ve XgBoost algoritmalarını farklı hiper-parametreler ile deneyerek karar ağacı temelli bir çözüm sunmuşlardır. ATM lokasyon parametresini müşteri işlem adetleri parametresine bağlı olarak düşünmüşler ve algoritmaları bu parametre özelinde eğitmişlerdir. Aylık ortalama işlem adedi tahmin modeli sonuçlarında XgBoost, 0,77 NRMSE (Normalize Edilmiş Kök Ortalama Karekök Hata) değeriyle en iyi algoritma olmuştur. Varlıklı müşteri işlem adetlerinde ise yine 0,56 gerçekleme ve 0,27 NRMSE değeriyle XgBoost algoritması en iyi sonucu üretmiştir [19].

Bir başka çalışmada, S.More ve arkadaşları, tesis konum problemlerinde makine öğrenmesinin etkinliğini araştırmışlardır. Makine öğrenmesi tekniklerinin tesis yeri

belirleme problemlerine çeşitli şekillerde uygulanabileceğini belirtmişlerdir. Örneğin, geçmiş verilere dayalı olarak farklı konumlardaki talebi tahmin etmek için denetimli öğrenme kullanılabileceğinden ve bu bilgilerin tesislerin nereye konumlandırılacağına ilişkin karar vermek için yararlı olacağından bahsetmişlerdir. Sonuç olarak, en düşük amaç değerine ulaşan ve dolayısıyla tesis yerleştirmek için en iyi yöntem Gradient Boosting algoritması olmuştur. Ancak, bu algoritma aynı zamanda en uzun çözüm süresine sahip algoritma olarak bulunmuştur. Dolayısıyla diğer yöntemlere göre daha yavaş olarak değerlendirilmektedir. İkinci olarak en iyi yöntem ise Rastgele Orman algoritması bulunmuştur. Destek Vektör Makinesi'nin ise en yüksek amaç değerine ve hızlı bir ortalama çözüm süresine sahip olduğunu belirtmişlerdir. En hızlı ortalama çözüm süresine sahip algoritma olarak ise K-Means kümeleme algoritması bulunmuştur [20].

Makine öğrenmesiyle yapılan bir başka çalışmada ise O. Yildirim, ATM yerleştirmek için iyi ve kötü mahalleleri skorlayan üç farklı algoritma denemiştir. Lojistik Regresyon ve Destek Vektör Makinesi algoritmaları iyi mahalleleri tahminlerken 0,87 ve 0,88 duyarlılık (recall) skoru üretmişlerdir. Karar Ağaçları ise iyi mahalleleri tahminlerken 0,7, kötü mahalleleri tahminlerken 0,92 duyarlılık sonucuna ulaşmıştır [21].

M. Zeydan ve S. Kayserili, ATM'lerin konumlarının belirlenmesine yönelik yaptıkları çalışmada, veri madenciliği tekniklerinden ve karar ağaçlarından faydalanmışlardır. Özel bir bankaya ait belirli bir şehirdeki ATM veri setini kullandıkları çalışmada, kural tabanlı konum tespitine dayalı bir karar destek sistemi oluşturmuşlardır. Karar ağacı algoritmaları arasında model tahmin doğruluğunu, STATISTICA 10.0 yazılımı kullanarak ortalama %92,96 olarak bulmuşlardır [22].

Makine öğrenmesi ile çalışılmış tesis yerleştirme problemine verilebilecek bir diğer örnekte de D.Wickramasuriya ve arkadaşları, baz istasyonu kurulumlarında bir sonraki baz istasyonunun nereye kurulacağı problemini çözmeye çalışmışlardır. Çalışmalarında, RNN (Tekrarlayan Sinir Ağı)'ye dayalı bir yöntem sunmuşlardır. Önerilen makine öğrenmesi yöntemi, en uygun sanal baz istasyonu kümesini tahmin etmede %98 başarı göstermiştir [23].

Genel yerleştirme ve ATM lokasyon problemlerinin çözümünde, Genetik Algoritma (GA) önemli bir yer tutmaktadır. Genetik Algoritmalar, biyolojik evrim süreçlerinden esinlenerek geliştirilen ve özellikle karmaşık optimizasyon problemlerinde etkili sonuçlar sağlayan yöntemlerdir. ATM lokasyon seçimlerinde Genetik Algoritmaların kullanımı, büyük veri setleri ve çok sayıda kısıt içeren sorunların çözümünde büyük bir esneklik sunar.

G. Zhou ve arkadaşları, elektrik istasyonlarının şarj istasyonu problemini çözmek için yaptıkları çalışmada, GA'dan, İrlanda bölgesini bir dikdörtgen boyutunda bölmek, böldükten sonra karelere ayırmak ve şarj talebini karşılamak için korelasyon katsayısını ayarlamak için faydalanmışlardır [24].

M. Kaveh ve arkadaşları, hastane yeri seçimi probleminde Genetik Algoritma'yı geliştirmeyi amaçlamışlardır. Geliştirilmiş Genetik Algoritma (IGA)'nın performansını değerlendirmek için GA ve Parçacık Sürü Optimizasyonu (PSO) yöntemleri kullanılmıştır. Sonuç olarak, yüksek uygunluk fonksiyonuna sahip kromozomların ve düşük uygunluk fonksiyonuna sahip kromozomların uygun kombinasyonlarının seçilmesinin, algoritmanın keşif yeteneğini artırdığını söylemişlerdir. Diğer bir deyişle, algoritma yerel minimumlarla uğraşmayarak, algoritmanın yakınsama sürecinin iyileşeceği ve algoritmanın farklı performanslarda yüksek düzeyde kararlılık göstereceği belirtilmiştir. Bulgulara göre, geliştirilmiş Genetik Algoritma diğer algoritmalara kıyasla daha iyi performans göstermiştir. IGA'nın yakınsama hızının GA ve PSO'dan bu problem özelinde daha yüksek olduğu tespit edilmiştir. Ayrıca, IGA'nın diğer algoritmalara göre kararlılık seviyesinin daha fazla olduğu bulunmuştur [25].

A. Alhaffa ve W. Abdulal yaptıkları çalışmada, ATM yer seçimlerini belirleme problemi için üç farklı algoritma tasarlayıp karşılaştırmışlardır: Sezgisel Yaklaşım, Evrişim Kullanan Sıralama Tabanlı Genetik Algoritma ve Evrişim Kullanan Benzetilmiş Tavlama. Bu çalışmada, en yüksek kapsama yüzdesine ulaşmayı ve amaçlanan çalışma alanında daha az ATM'ye gereksinim duymayı hedeflemişlerdir. Her bir algoritmanın performansını ölçtüklerinde, Sıralama Tabanlı Genetik Algoritma'nın, Sezgisel Yaklaşım'a göre minimum %2,2 ve maksimum %20,13 iyileşme kaydettiğini görmüşlerdir. Ayrıca, Benzetilmiş Tavlama'nın, Sezgisel Yaklaşım'a göre %26,32'ye varan oranda, Evrişim Kullanan

Genetik Algoritma'nın ise %2,28 oranında kapsama değeri açısından daha iyi performans gösterdiğini ortaya koymuşlardır [26].

M. Kulkarni, yine bir önceki çalışmada incelendiği gibi Sıralama Tabanlı Genetik Algoritma yöntemini kullanarak ATM lokasyon problemine çözüm aramıştır. Amaç olarak, ATM sayısını en aza indirmek ancak seçilen alandaki tüm müşteri taleplerini karşılamak olarak belirlemiştir. Deneysel sonuçlar, sıralama tabanlı Genetik Algoritma'nın, çalışma alanını maksimize ve ATM sayısını minimize ederek optimizasyon problemini çözdüğünü göstermiştir [27].

ATM ve diğer tesislerin yerlerinin belirlenmesinde çeşitli yöntemler kullanılmıştır. Veri analizi ve makine öğrenimi yöntemleri, eğilimleri ve örüntüleri analiz etmek için sıkça tercih edilmiş ve özellikle regresyon modelleri ile sınıflandırma algoritmaları öne çıkmıştır. Coğrafi Bilgi Sistemleri, coğrafi ve demografik verilerin görselleştirilmesi ve analizi için kullanılarak ATM yerlerinin coğrafi olarak en uygun noktalarını belirlemede önemli bir rol oynamıştır. Ayrıca, demografik ve müşteri davranış analizleri, ATM kullanımı ve müşteri davranışlarına dair kritik veriler sağlayarak yer seçimini optimize etmede etkili olmuştur. Çok kriterli karar analizi ise çeşitli kriterler ve ağırlıklar kullanarak en uygun ATM ve diğer tesislerin lokasyonlarını belirlemede önemli bir yöntem olarak benimsenmiştir.

1.2 Tezin Amacı

Tezin amacı, ATM yerleşim yeri seçiminde kullanılan geleneksel yöntemlerin sınırlılıklarını aşarak, makine öğrenmesi ve Genetik Algoritma tekniklerinin bu sürece nasıl entegre edilebileceğini araştırmaktır. Bu bağlamda, literatürde mevcut olan çalışmalar incelenecek, bu yöntemlerin avantajları ve dezavantajları tartışılacak ve ATM yerleşim yeri seçiminde uygulanabilirlikleri değerlendirilecektir. Bir diğer deyişle, makine öğrenmesi ve Genetik Algoritmaların birlikte kullanılarak daha etkili ve verimli bir yerleşim yeri seçim süreci oluşturup oluşturmayacağını araştırmaktır. Literatürdeki mevcut çalışmaların özetlenmesi, yeni yöntemlerin geliştirilmesi ve bu yöntemlerin gerçek dünya uygulamasında nasıl kullanılabileceğinin gösterilmesi hedeflenmektedir. Bu tez, bankaların ATM yerleşim yeri seçiminde daha bilimsel ve veri odaklı kararlar almasına katkı sağlamayı amaçlamaktadır.

1.3 Literatüre Katkı

Literatürde ATM lokasyon problemlerine odaklanan çok kısıtlı sayıda çalışma bulunmaktadır. Yapılan önceki çalışmalarda da ATM yer seçimi problemlerine daha çok geleneksel yöntemler olan GIS (Coğrafi Bilgi Sistemleri) ya da matematiksel modeller ile çözüm aranmıştır. Bu araştırma, gerçek dünya verilerinden beslenen, uygulamada özel bir banka tarafından denenme kapasitesine sahip, veri setinde kriterleri detaylı bir şekilde değerlendiren ve gereksiz özellikleri eleyen bir yaklaşım ile ele alınmıştır.

Model tarafından puanlanan lokasyonların güvenilirliği, yine aynı bankanın mevcutta sahada olan ATM'lerinin çoklu kriterlerinin hepsine bakılarak değerlendirme yapılmasıyla hesaplanabilir. Literatürde, algoritmaların veya matematiksel modellerin ürettiği sonuçların bu şekilde bir teyidinin yapıldığına rastlanılmamıştır.

Çalışma, makine öğrenmesi, Genetik Algoritma ve Weka aracındaki özellik seçim algoritmalarının kullanılması sayesinde, bu alanda daha önce yapılmış olan çalışmalardan farklılaşmaktadır. Bu çalışmanın, kullanılan yöntemler esas alınarak, daha farklı yaklaşımlarla, diğer lokasyon tespit çalışmalarına ışık tutması ve literatüre katkı sağlanması hedeflenmektedir.

1.4 Tez Organizasyonu

Bu bölümde, tezin bölümlerinin kısa bir özeti ve her bir bölümde ele alınan konular açıklanmıştır.

Veri Seti

Veri Seti bölümünde, verinin kaynağı, veri setinin özellikleri ve istatistiksel özetleri sunulmuş, ayrıca veri ön işleme adımları ve elenen özellikler açıklanmıştır. Önemli özelliklerin korelasyon haritası ve hedef değişken ile bağımsız değişken ilişkilerini gösteren grafikler de yer almaktadır.

Yöntemler

Yöntemler kısmında, veri setinden çıkarılan özellikler ve üç farklı özellik seçim yöntemi detaylandırılmış, Genetik Algoritma ve Weka ile yapılan özellik seçim süreçleri açıklanmıştır.

Deneysel Sonuçlar

Bu bölümde, farklı özellik sayılarıyla çalışan makine öğrenmesi algoritmalarının performansları karşılaştırılmış, hiper-parametre değişimlerinin etkileri incelenmiştir. Ayrıca, Genetik Algoritma ile seçilen özellikler ve parametrelerin etkileri değerlendirilmiş, bu özelliklerle Rastgele Orman algoritması tekrardan test edilmiştir.

Sonuç

Sonuç bölümünde, ATM lokasyon probleminin önemi vurgulanmış, makine öğrenmesi, Genetik Algoritma ve Weka ile elde edilen sonuçlar özetlenmiş ve yorumlanmıştır. Ayrıca, gelecekte yapılabilecek çalışmalar üzerinde durulmuştur.

1.5 Hipotez

Bu tez kapsamında üzerinde durulan hipotezler aşağıda maddeler halinde verilmiştir.

- Makine öğrenmesi modelleri, ATM yerleşim yerlerinin belirlenmesinde önemli ölçüde başarılı tahminler yapabilir.
- Genetik Algoritmalar, ATM yerleşim yeri problemlerinin çözümünde, veri setinden özellik seçimi yapabilir ve makine öğrenmesi modelleri bu özellikler ile daha başarılı bir performans sergileyebilir.
- Literatürde yer almayan yeni yaklaşımlar ile ATM lokasyon seçim çözümlerinin başarısı artırılabilir.
- ATM yerleşim yeri seçimlerinde makine öğrenmesi algoritmalarının kullanılması, geleneksel yöntemlere kıyasla daha etkili sonuçlar sağlayabilir.
- Weka ile özellik seçimi sonrası, makine öğrenmesi algoritmaları daha başarılı sonuçlar verebilir.

2
VERİ SETİ

Bu çalışmada kullanılan veri seti, bir bankanın dış bir firmadan elde ettiği, ATM yerleşim yeri kararlarını desteklemek amacıyla toplanmış kapsamlı bir veri setidir. Bu veri seti, bankanın stratejik kararlar almasına yardımcı olmak için gerekli olan çeşitli bilgileri içermekte ve ATM'lerin en uygun şekilde yerleştirilebileceği potansiyel konumları ve bu konumlarla ilgili detaylı verileri kapsamaktadır.

Veri setinde bulunan bilgiler arasında, ATM'lerin yerleştirilebileceği potansiyel konumlar, hedef değişken olan Ağırlıklı Puan, lokasyonlara yakın bölgelerde bulunan diğer ATM'lerin mesafesi, hangi bankanın kaç tane ATM'sinin bulunduğu gibi veriler yer almaktadır. Ayrıca, nüfus yoğunlukları, sosyo-ekonomik göstergeler ve bölgenin demografik özellikleri gibi faktörler de veri setinin içinde yer almaktadır. Bu detaylı bilgiler, ATM yerleşim yeri kararlarının daha bilinçli ve verimli bir şekilde alınmasına olanak tanımaktadır.

Veri setinde her bir potansiyel lokasyonun çevresindeki mevcut ATM'lerin sayısını ve hangi bankalara ait olduğunu bilmek, rekabet analizini ve pazar boşluklarının belirlenmesini kolaylaştırmaktadır. Nüfus yoğunlukları ve demografik bilgiler, belirli bir bölgedeki müşteri potansiyelini anlamak için kullanılırken, sosyo-ekonomik göstergeler, bölgedeki potansiyel müşterilerin harcama gücü ve alışkanlıkları hakkında önemli ipuçları sağlamaktadır.

Çalışmanın bu bölümünde, veri setinin genel özelliklerinden, veri setiyle alakalı bazı istatistiki bilgilerden ve özelliklerin veri tiplerinden bahsedilecektir.

2.1 Veri Seti Detayları

Veri seti, hedef değişken de dahil olmak üzere toplam 97 özellik ve 20.603 örnekten oluşmaktadır. Bu özelliklerin 89 tanesi nümerik, 8 tanesi ise kategorik değişkendir. Her bir özelliğin ismi Tablo 2.1'de verilmektedir.

Veri setindeki 20.603 gözlem üzerinde yapılan analize göre, bazı özellikler için oluşan genel istatistikler Tablo 2.1'de verilmiştir. Buna göre, incelenen bölgelerdeki toplam nüfus ortalaması 152,78, en yakın şube mesafe ortalaması 16.711,41 metre olarak bulunmuştur. Bu bölgelerdeki ortalama indeks değeri 28,16, XXX bankasının Türk müşteri indeksi 8,38, demografik indeks 0,93, ATM sayısı 1,15, muhafazakâr oran 50,17, muhafazakâr nüfus oranı 0,97, katılım ATM sayısı 0,02, ortalama hane geliri 6.338,61 TL, AB ses grubu nüfusu 28,91, C ses grubu nüfusu 90,89, DE ses grubu nüfusu ise 32,94 olarak hesaplanmıştır.

Tablo 2.1 Veri setinin bazı özellikler için genel istatistikleri

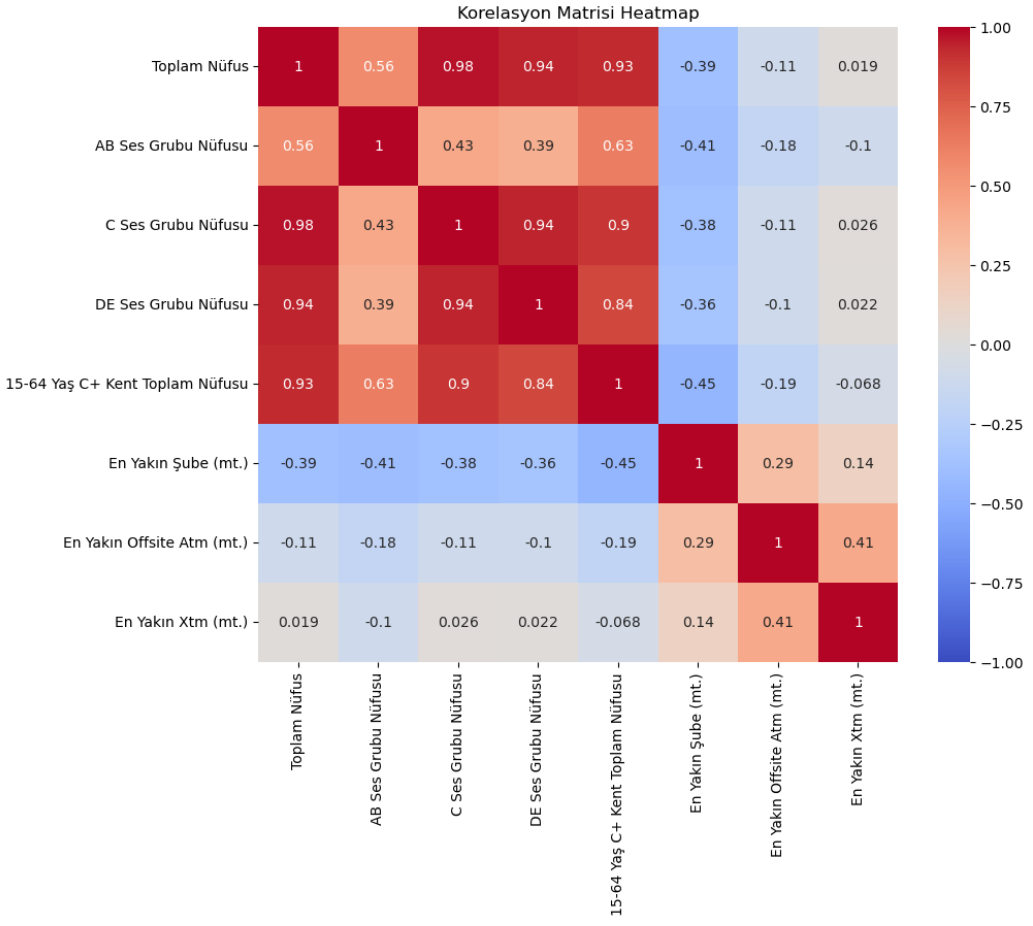
Özellikler	Adet	Ortalama	Standart Sapma	En Küçük	En Büyük
Toplam Nüfus	20.603	152,78	530,89	0	24.215
Foot Traffic	20.603	30,93	29,08	0	100
En Yakın Şube (mt.)	20.603	16.711,4	30.597,20	0	197.961,2
Ortalama Index	20.603	28,16	9,25	3	71,90
XXX Bankası Müşteri Indeksi	20.603	8,37	13,33	0	100
Muhafazakar Oran	20.603	50,17	20,79	0	99

Tablo 2.2'de ham veri setindeki birkaç özellik ve bu özelliklerin değerleri gösterilmekte ve hedef değişken olan Ağırlıklı Puan ve İl Adı, İlçe Adı, İlçe Türü, Alan Kategori, Alan alt Kategori, Potansiyel gibi bağımsız özellikler ve örnek değerler belirtilmektedir.

ALAN_KODU	ALAN_ADI	İl Adı	İlçe Adı	İlçe Türü	Alan Kategori	Alan alt Kategori	Potansiyel	Ağırlıklı Puan	Ortalama Index
ATM_ALAN_8321	Büyükdere Caddesi	İstanbul	Şişli	Metropol	Cadde	Lokal Segment	Hayır	100.0	67.84714285714286
ATM_ALAN_10377	Atatürk Bulvarı	Ankara	Çankaya	Metropol	Cadde	Lokal Segment	Hayır	100.0	56.69973319824384
ATM_ALAN_11004	Halaskargazi Caddesi	İstanbul	Şişli	Metropol	Cadde	Lokal Segment	Hayır	100.0	67.57045592705167
ATM_ALAN_17884	Kahramanmaraş Caddesi	Trabzon	Ortahisar	Merkez	Cadde	Lokal Segment	Hayır	92.94409465914025	52.048693009118544
ATM_ALAN_16279	Kosgeb Caddesi	Konya	Karatay	Merkez	Cadde	Lokal Segment	Hayır	81.83871580547113	45.82968085106383
ATM_ALAN_12610	100. Yıl Bulvarı	Ankara	Yenimahalle	Metropol	Cadde	Lokal Segment	Hayır	75.1686507936508	42.09444444444445
ATM_ALAN_14276	Fevzi Çakmak Caddesi	Bursa	Osmangazi	Metropol	Cadde	Lokal Segment	Hayır	100.0	71.76761904761905
ATM_ALAN_7825	Kısıklı Caddesi	İstanbul	Üsküdar	Metropol	Cadde	Lokal Segment	Hayır	93.48521855550732	52.351722391084095
ATM_ALAN_16486	Mustafa Kemal Paşa Caddesi	Manisa	Şehzadeler	Merkez	Cadde	Lokal Segment	Hayır	100.0	57.79431948665991
ATM_ALAN_13766	Hükümet Bulvarı	Aydın	Efeler	Merkez	Cadde	Lokal Segment	Hayır	90.2923240218073	50.56370145221209
ATM_ALAN_12896	Cumhuriyet Bulvarı	İzmir	Konak	Metropol	Cadde	Lokal Segment	Hayır	87.34351630723211	48.91236913204998
ATM_ALAN_14564	İnönü Caddesi	Çorum	Merkez	Merkez	Cadde	Lokal Segment	Hayır	81.90990917643653	45.86954913880446
ATM_ALAN_15155	Profesör Muammer Aksoy Bulvarı	Gaziantep	Şehitkamil	Metropol	Cadde	Lokal Segment	Hayır	93.73523966324119	52.491734211415064
ATM_ALAN_14099	İzzet Baysal Caddesi	Bolu	Merkez	Merkez	Cadde	Lokal Segment	Hayır	81.60981449317316	45.70149611617697
ATM_ALAN_17331	Kazım Paşa Caddesi	Samsun	İlkadım	Merkez	Cadde	Lokal Segment	Hayır	86.77824817870412	48.595818980074306

Şekil 2.1 Ham veri setini içeren .csv dosyası

Şekil 2.1'de, ham veri setinin içerdiği özelliklerin bir kısmı görülmektedir. Bu veri setindeki ALAN_KODU, ALAN_ADI gibi benzersiz özellikler veri setinden çıkarılmıştır. Ağırlıklı Puan ise hedef değişkeni belirtmektedir.



Şekil 2.2 Özellikler arası korelasyon matrisi

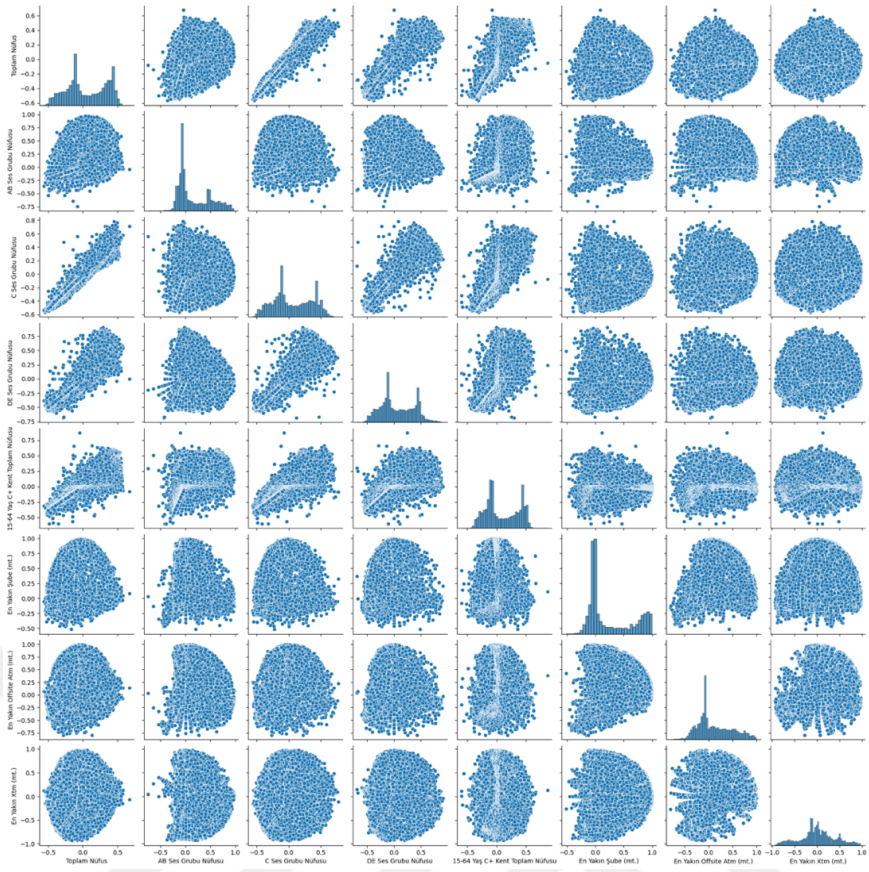
Bölüm 3.1'de, özellikler arası ilişkiler incelenmiştir. Bu bağlamda, özellikler arasındaki korelasyona, ısı haritalarına, histogram ve kutu grafiklerine bakılmıştır. Şekil 2.2'de örnek bir korelasyon haritası görülmektedir.

Korelasyon matrisindeki sonuçlar, çeşitli nüfus ve mesafe değişkenleri arasındaki ilişkileri ortaya koymaktadır. Toplam Nüfus ile AB Ses Grubu Nüfusu arasında 0,56'lık orta düzeyde pozitif bir korelasyon gözlemlenmektedir. Bu, toplam nüfusun arttığı bölgelerde AB Ses Grubu Nüfusu'nun da artma eğiliminde olduğunu göstermektedir. Toplam Nüfus ile C Ses Grubu Nüfusu arasındaki korelasyonun ise oldukça yüksek (0,98) olması, iki değişken arasındaki neredeyse tam doğrusal ilişkiyi ve C Ses Grubu Nüfusu'nun büyük ölçüde toplam nüfusa bağlı olduğunu gösterir. Ayrıca, Toplam Nüfus ile 15-64 Yaş Kent Toplam Nüfusu arasındaki yüksek korelasyon (0,93), yaş grubunun toplam nüfusla güçlü bir bağlantısı olduğunu göstermektedir. Diğer yandan, En Yakın Offsite ATM ile Toplam Nüfus arasındaki düşük negatif korelasyon (-0,11) ve En Yakın XTM ile Toplam Nüfus arasındaki neredeyse sıfır korelasyon (0,019), bu değişkenlerin toplam nüfus ile belirgin bir ilişkisi olmadığını işaret etmektedir.

Aynı zamanda, özellikler arasındaki ilişkileri gözlemlemek için Şekil 2.3'te görüleceği üzere özellikler arası pair plot grafiği çizdirilmiştir. Burada, özellikler arasında doğrusal bir ilişki gözlemlendiğinde bu ilişkiler, korelasyon matrisi yardımıyla da doğrulanmıştır.

Yapılan tüm analizler önce ham veri seti üzerinde özellikler arasındaki ilişkileri gözlemlemek amacıyla gerçekleştirilmiş, daha sonra Bölüm 3.1'de özellik seçimi kısmında anlatılan yöntemler gerçekleştirildikten sonra da aynı analizler ve veri görselleştirmeleri gerçekleştirilmiştir.

Bağımlı değişken ile diğer bağımsız değişkenler arasındaki gözlemler tek tek incelenmiştir. Şekil 2.1 ve Şekil 2.2'te görüleceği üzere, Toplam Nüfus ve C Ses Grubu Nüfusu özellikleri arasında doğrusal bir ilişki gözlemlenmiştir ve C Ses Grubu Nüfusu özelliğinin veri setinden çıkarılma kararı alınmıştır.



Şekil 2.3 Özellikler arası çift grafik (pair plot)

Pair plot grafiği, değişkenler arasındaki ilişkileri görselleştirerek korelasyon matrisindeki sayısal değerleri destekleyen bir analiz sağlar. Toplam Nüfus ile diğer değişkenler arasındaki ilişkiler incelendiğinde, Toplam Nüfus ile AB Ses Grubu Nüfusu arasında orta düzeyde pozitif bir ilişki görülmüştür. Pair plot'ta, bu iki özellik arasındaki ilişkide noktaların yukarı doğru eğilim gösterdiği ve aynı zamanda veri dağılması da gözlemlenmiştir. Toplam nüfus ile C Ses Grubu Nüfusu arasındaki ilişki oldukça güçlü gözükmektedir. Bu iki ilişki arasında pair plot'ta neredeyse doğrusal bir ilişki vardır ve noktalar sıralı bir şekilde hizalanmıştır. Benzer şekilde, Toplam Nüfus ile DE Ses Grubu Nüfusu arasındaki ilişki de yüksek bir pozitif korelasyon göstermiştir. Bu da pair plot'ta güçlü bir yukarı doğru eğilim olarak gözükmektedir.

3
YÖNTEMLER

Bu bölümde, özelliklerin belirlenmesiyle alakalı detaylar ve kullanılan yöntemler Bölüm 3.1'de verilmiştir. Bölüm 3.2'de ise makine öğrenmesi temelli yöntemler ile özelliklerin seçilmesi sağlanmıştır. Bölüm 3.3'te çeşitli makine öğrenmesi ve regresyon modelleriyle alakalı bilgilere yer verilmiştir. Model performansının hangi metriklerle değerlendirildiğine ilişkin bilgiler Bölüm 3.4'te özetlenmiştir. Makine öğrenmesinden sonra özellik seçimi Genetik Algoritma ile yapılmıştır. Bölüm 3.5'te Genetik Algoritma ile alakalı genel bilgiler aktarılmıştır. Bölüm 3.6'da Weka'da özellik seçimi ile alakalı genel bilgilerden, Bölüm 3.7'de Weka'da özellik seçimi için kullanılan yöntemlerden bahsedilmiştir. Weka ile özellik seçimi parametreleri ile alakalı genel detaylar Bölüm 3.8'de aktarılmıştır.

3.1 Özelliklerin Belirlenmesi

Veri hazırlığı bölümü kapsamında, ham veri seti analiz edilmiş olup, veri seti üzerinde çeşitli ön işleme adımları uygulanmıştır. Bu işlemler, veri setinin daha temiz, anlaşılır ve makine öğrenmesi modelleri için uygun hale getirilmesi amacıyla gerçekleştirilmiştir.

'Toplam İşlem Sayısı' ve 'Toplam Para Çekme TL' kolonları, tüm değerleri 0 olduğundan ve model performansına anlamlı bir etki yapmayacağından dolayı silinmiştir.

Şekil 2.1'de verilen korelasyon matrisine göre, 0,9 değerinin üzerinde yüksek korelasyon gösteren özelliklerin silinmesi kararlaştırılmıştır. Örneğin, Toplam Nüfus ile yüksek korale olan C Ses Grubu Nüfusu, DE Ses Grubu Nüfusu ve 15-64 Yaş C+ Kent Toplam Nüfusu özellikleri veri setinden çıkarılmıştır.

'XXX Xtm' ve 'Para Çekme/işlem' değerlerinin hepsi '1' olduğu için de veri setinden çıkarılmıştır.

'ALAN_KODU' ve 'ALAN_ADI' kolonlar benzersiz değerler olduğu için veri setinden çıkarılmıştır. Bu tür benzersiz veya kimlik belirleyici kolonlar, modelin

öğrenme sürecine katkı sağlamamakta ve modelin karmaşıklığını artırabilmektedir. Bu da modelin performansını olumsuz etkileyebilir.

Makine öğrenmesi algoritmalarının kategorik değerler yerine sayısal değerlerle çalışması göz önünde bulundurularak, 'Evet' ya da 'Hayır' şeklinde değerler alan 'Potansiyel' kolonu, Label Encoding (Etiket Kodlama) yöntemiyle '0' ve '1' olacak şekilde sayısal değerlere dönüştürülmüştür. Bu işlem için 'LabelEncoder' kullanılmıştır ve yeni oluşturulan 'Potansiyel_Encoded' kolonu veri setine eklenmiştir. Orijinal 'Potansiyel' kolonu ise veri setinden çıkarılmıştır.

'İlçe Türü', 'Alan Kategori' ve 'Alan alt Kategori' kolonları One-Hot Encoding yöntemiyle dönüştürülmüş ve yeni oluşturulan 'Metropol', 'Cadde', 'Askeri Alan', 'Üniversite', 'Havaalanı', 'Kurum', 'Sanayi' ve 'Ticari Bölge' sütunları veri setine eklenmiştir. Orijinal 'İlçe Türü', 'Alan Kategori' ve 'Alan Alt Kategori' kolonları veri setinden çıkarılmıştır.

Bazı değişkenler için ise aykırı değerler belirlenmiş ve üst limit ile değiştirilmiştir. Bu işlem için ilk olarak veriler 'RobustScaler' ile ölçeklendirilmiştir. Ardından, IQR (Çeyrekler Arası Aralık) yöntemiyle üst sınırın üzerindeki aykırı değerler üst sınır değerine çekilmiştir.

Bazı özellikler ise Standartlaştırma ve Normalizasyon işlemlerinden geçirilmiştir. Bu işlemle, verilerin aynı ölçeğe çekilerek modelleme sürecine hazırlanması sağlanmıştır. Tüm bu işlemlerin sonucunda 97 özellikten oluşan ham veri seti 81 özelliğe indirgenmiştir.

3.2 Özellik Seçimi

Veri seti ile alakalı kısımda ve Bölüm 3.1'de uygulanan özelliklerin belirleme işlemlerinin ardından, Python'da Scikit-learn kütüphanesinden üç farklı özellik seçim yöntemi denenmiştir. Bu yöntemler sırasıyla "GenericUnivariateSelect", "SelectPercentile" ve "RFE" sınıflarıdır. Bu yöntemler yardımıyla en iyi 12 özellik seçilmiştir. Seçilen özelliklerin hangi sırayla seçildiği Tablo 3.2'de verilmiştir.

Tablo 3.2 Özellik seçme yöntemleri ile elde edilen özellik sıralamaları

Özellik	**GenericUnivariate Select**	**SelectPercentile**	**RFE**
Ortalama Index	1	1	**1**
XXX Müşteri Index	2	2	**2**
Atm Index	3	4	**4**
Banka Index	4	5	**5**
Foot Traffic Index	5	6	**6**
En Yakın Şube (mt.)	6	8	**8**
Toplam Nüfus	7	9	**9**
Foot Traffic	8	11	**11**
Demog. Index	9	7	**7**
Rakip Atm	10	X	**X**
XXX Şube Yakın	11	X	**X**
XXX Şube Uzak	12	12	**12**
Segment Index	X	3	**3**
Ortalama Hane Geliri	X	10	**10**

3 farklı özellik seçim algoritması ile elde edilen özellikler karşılaştırıldığında, hepsinin hemen hemen aynı özellikleri tespit ettiği, sadece sıralama bazında bir farklılık olduğu tespit edilmiştir. GenericUnivariateSelect ile yapılan özellik seçimlerinde diğer sınıflardan farklı olarak Rakip ATM özelliğinin tespit edildiği gözlemlenmiştir. SelectPercentile ve RFE yöntemlerinde ortak olarak Ortalama Hane Geliri özelliği belirlenmiştir. Diğer özellikler; Ortalama Index, Atm Index, Foot Traffic Index vb. ortak özelliklerdir. Fonksiyonlar arasında benzer özellikler elde edildiğinden çalışmanın ilerleyen bölümlerinde SelectPercentile ile edilen özelliklerle devam edilecektir.

3.3 Makine Öğrenmesi ve Regresyon Modelleri

Makine öğrenmesi, veriden öğrenme ve gelecekteki olayları, durumları veya eğilimleri tahmin etme yeteneği sağlayan algoritma ve tekniklerin bütünüdür. Geleneksel programlama yaklaşımlarından farklı olarak, makine öğrenmesi modelleri, veri içindeki desenleri ve ilişkileri keşfederek otomatik olarak öğrenir ve genelleştirir. Bu öğrenme süreci, modelin veri setine dayalı olarak parametrelerini optimize etmesini sağlar ve böylece daha doğru tahminler yapılabilir.

Regresyon modelleri, sürekli ve nicel değişkenleri tahmin etmek amacıyla yaygın olarak kullanılan makine öğrenmesi tekniklerinden biridir. Makine öğrenmesinin önemli bir alt kümesini oluşturur ve bağımlı ya da hedef bir değişkenin bağımsız değişkenlerle olan ilişkisini modellemek için kullanılır. Bu modeller, genellikle sürekli ve nicel veri tahmini için çalışır. Regresyon analizindeki asıl amaç, bağımsız değişkenler ile bağımlı değişken arasındaki ilişkiyi matematiksel bir formül ile ifade etmek ve bu ilişkiye dayanarak gelecekteki değerleri tahmin etmektir.

Bu problem benzeri lokasyon tespit çalışmalarında da makine öğrenmesi tekniklerinin literatürde kullanıldığı görülmektedir. Yapılan bir çalışmada, lokasyon alanlarının tespitinde önemli kriterleri belirlemek için makine öğrenmesi algoritmalarından yararlanılmıştır [28].

Makine öğrenmesinde modellerin birbirleri arasındaki performans ya da başarı sonuçları problemin boyutuna ve veri setine göre farklılık gösterebilir. Bu tezde regresyon problemlerinde başarı gösteren Rastgele Orman, Karar Ağacı, KNN, Ridge, Lasso, Gradient Boosting, Destek Vektör Regresyonu (SVR), Elastic Net, Adaboost algoritmaları kullanılmış ve bu algoritmaların performans sonuçları karşılaştırmalı olarak verilmiştir. Bu bölümde, her bir algoritmanın teorik temelleri ve uygulama alanlarından bahsedilecek, makine öğrenmesi modellerinin performansını değerlendirmek için kullanılan metrikler açıklanacaktır. Aynı zamanda, model doğrulama ve hiper-parametre optimizasyonu gibi konulara değinilecektir.

3.3.1 Doğrusal Regresyon

Doğrusal Regresyon, makine öğrenmesinde en temel ve en yaygın kullanılan regresyon tekniklerinden biridir. Bu model, bağımlı (hedef) değişken ile bir veya daha fazla bağımsız değişken ya da özellik arasındaki doğrusal ilişkiyi

modellemeye çalışır. Doğrusal regresyon modeli, veri noktalarını en iyi temsil eden bir doğruyu bulmayı amaçlar. En temel anlamda, nicel bir yanıtı tahmin etmek için yararlı bir yöntemdir [29]. Doğrusal Regresyon'un sınırlaması olarak, bağımlı ve bağımsız değişkenler arasında doğrusal bir ilişki varsayılması gösterilebilir. Ancak gerçek dünyada, ilişkiler çoğu zaman doğrusal olmadığından, bu durum modelin tahmin performansını olumsuz etkileyebilir.

3.3.2 Rastgele Orman

Rastgele Orman, birden fazla karar ağacını bir araya getirerek daha güçlü ve dengeli tahminler yapan bir makine öğrenmesi algoritmasıdır. Bu yöntem, hem sınıflandırma hem de regresyon problemlerinde kullanılır ve özellikle doğrusal olmayan karmaşık ilişkileri modellemede etkilidir. Bu algoritma, bir topluluk yöntemidir ve temelinde karar ağaçları bulunur. Veri setinden birçok rastgele alt örnekleme yaparak farklı karar ağaçları oluşturur. Her bir karar ağacı, bu farklı alt örneklemlerle eğitilir. Ayrıca, her bir karar ağacı oluşturulurken, her düğümde rastgele bir alt özellik kümesi seçilir [30]. Böyle yapılması, modelin çeşitliliğini artırır ve aşırı öğrenmeyi ya da ezberlemeyi azaltır. Rastgele Orman modelinde genel tahmin yapılırken, regresyon için tüm ağaçların tahminlerinin ortalaması alınır. Sınıflandırma için ise çoğunluk oyu alınarak tahmin yapılır.

Rastgele Orman, çeşitli veri setlerinde iyi performans göstermesinin yanı sıra aşırı uyuma dayanıklı olduğundan genelleme yeteneği iyidir. Ayrıca, hangi özelliklerin hedef değişken üzerinden daha fazla etkili olduğunu belirleyebilir. Sınırlamalar olarak ise yüksek hesaplama gücü ve karmaşıklık sayılabilir.

3.3.3 Karar Ağaçları

Karar Ağaçları, hem sınıflandırma hem de regresyon problemlerinde kullanılan, veri setini dallara ayırarak tahminler yapan makine öğrenmesi algoritmasıdır. Bu modeller, veri içindeki karar kurallarını hiyerarşik bir yapıda temsil eder. Bu yapının görselleştirilmesi, karar ağaçlarının yorumlanması açısından faydalı olmaktadır.

Karar Ağaçları, kök düğümden başlayarak veri setini belirli kurallara göre iki veya daha fazla alt gruba ayırır. Her bir dal, bir karar kuralını temsil eder ve her yaprak düğümü bir tahmin sonucunu veya sınıf etiketini belirtir. Karar Ağaçları, kök düğüm, iç düğümler ve yaprak düğümlerden oluşur.

Karar Ağacı modeli, veri setinde üzerinde bölmeleri belirlemek için belirli ölçütler kullanır. Sınıflandırma problemlerinde, Gini ve Bilgi Kazancı; regresyon problemlerinde ise Ortalama Kare Hata metriği kullanılır. Karar Ağacı, her bölmede bu ölçütleri optimize ederek veri setini dallara ayırır.

3.3.4 Ridge Regresyon

Ridge Regresyon, doğrusal regresyonun bir varyantıdır ve modelin aşırı uyum yapmasını önlemek için kullanılır. Çoklu doğrusal bağlantı sorununu hafifletmek için çok kullanışlı bir algoritma olarak bilinmektedir [32]. Doğrusal Regresyon modeline bir ceza terimi ekleyerek katsayıların büyüklüğünü kontrol altına alır ve böylece modelin karmaşıklığını azaltmış olur. Bu yöntem, çoklu doğrusal bağlantı problemlerinin olduğu veri setlerinde etkili çalışır.

3.3.5 Lasso Regresyon

Lasso Regresyon, doğrusal regresyon modellerine L1 norm cezası ekleyerek özellik seçimi ve düzenleştirme sağlayan bir yöntemdir. Bu yaklaşımın bir faydası, üretilen modellerin daha yorumlanabilir olması olarak görülmektedir [33]. Lasso, özellikle birçok özellik içeren veri setlerinde, bazı katsayıları sıfıra indirerek, modelin sadeliğini ve yorumlanabilirliğini artırır. Ceza terimi sayesinde, modelin aşırı uyumu azaltılır ve genelleme yeteneği artırılır. Lasso, yüksek ilişkili özellikler arasında rastgele seçim yapabilir, bu da modelin kararlılığını etkileyebilir.

3.3.6 Gradient Boosting

Gradient Boosting, zayıf karar vericileri ya da öğrenicileri bir araya getirerek güçlü bir tahmin modeli oluşturan bir makine öğrenimi tekniğidir. Bu yöntem, ardışık olarak yeni modeller oluşturarak, önceki modellerin hatalarını düzelterek çalışır. Bir diğer deyişle, daha sağlam bir model oluşturmak için genellikle karar ağaçları olan zayıf tahmin topluluğunun güçlendirilmesini kullanarak tahminler yapar [34]. Gradient Boosting, birçok algoritmanın temelini oluşturan etkili bir yöntemdir ve XGBoost, LightGBM, CatBoost gibi çeşitli kütüphanelerde uygulanmıştır.

Gradient Boosting temelinde, kayıp fonksiyonunu minimize etmek için Gradient Descent (Gradyan İniş) yöntemini kullanır. Yüksek doğruluk ve genelleme yeteneği sağlar ve aynı zamanda özellik seçimi de yapabilir. Gradient Boosting, diğer modellere nazaran daha fazla hesaplama gücü gerektirebilir. Bir risk ise, parametreler doğru ayarlanmadığı zaman aşırı uyum ile karşı karşıya kalmasıdır.

3.3.7 Elastic Net Regresyonu

Elastic Net regresyonu, L1 ve L2 ceza terimlerini birleştirerek Ridge ve Lasso regresyonlarının avantajlarını birleştiren bir regresyon yöntemidir [35]. Bu yöntem, hem özellik seçimi yaparak hem de katsayıları küçülterek modelin performansını iyileştirmeyi amaçlar. Elastic Net, özellikle yüksek boyutlu veri setlerinde ve çoklu doğrusal bağlantı sorunu olan veri setlerinde etkilidir. Ceza parametresi ile L1 ve L2 ceza terimleri arasındaki oranı kontrol eden parametrenin iyi seçilmesi iyi bir performans için önemlidir.

3.3.8 SVR (Destek Vektör Regresyonu)

Destek Vektör Regresyonu (SVR), destek vektör makineleri prensibini kullanarak regresyon analizi yapmak için kullanılan bir makine öğrenmesi yöntemidir. Standart SVR yöntemi, küçük tutarsızlıklara karşı duyarsız olan 1-norm kayıp fonksiyonu kullanarak doğrusal bir regresyon modeli oluşturmaktır [36]. SVR, doğrusal olmayan ilişkileri modellemek için etkili bir biçimde kullanabilir ve aykırı değerlere karşı dirençlidir.

SVR'nin temel amacı, veri noktalarını bir hiper düzlem etrafında bir araya getirmektir. Ancak, tüm veri noktalarının hiper düzlem üzerinde olmasını beklemek gerçekçi değildir. Bu nedenle, bir hata marjı tanımlanır. Bu hatanın büyüklüğü, bir hiper düzlem ve veri noktaları arasındaki mesafeyi tanımlar. SVR, bu hatayı minimize ederek en iyi hiper düzlemi bulmaya çalışır.

SVR, doğrusal ve doğrusal olmayan ilişkileri modelleyebilir. Doğrusal olmayan ilişkileri modellemek için çekirdek fonksiyonları kullanılır. (Doğrusal çekirdek, Polinom çekirdeği, RBF çekirdeği vb.) C ve epsilon olmak üzere iki önemli hiper-parametreye sahiptir. C, marjın genişliğini kontrol eder. Küçük C değerleri, daha geniş bir marj oluştururken, büyük C değerleri daha dar bir marj oluşturur. Epsilon parametresi ise hata marjını tanımlar. Model, epsilon değeri içindeki hatalara odaklanır ve bu hataları minimize etmeye çalışır.

Marjlar sayesinde aykırı değerlere karşı dirençli olması, doğrusal ve doğrusal olmayan ilişkileri modelleyebilmesi ve karmaşık ilişkilere uygulanabilir olması avantajları arasındadır. Özellikle büyük veri setlerinde bellek ve hesaplama süresine paralel olarak çalışma süresi uzun olmaktadır.

3.3.9 K-En Yakın Komşu (KNN) Regresyonu

KNN Regresyonu, regresyon analizi için kullanılan basit ama etkili bir makine öğrenmesi yöntemidir. KNN regresyonu, bir veri noktasını tahmin etmek için en yakın k komşusunun ortalama değerini kullanır. Bu yöntem, veriler arasındaki ilişkileri yakalamak ve tahmin yapmak için kullanışlıdır.

K-komşudaki k değeri, tahmin yapmak için kullanılacak komşu sayısını temsil eder. K sayısı genellikle tek sayı seçilir, böylece çoğunluk oylaması yapılırken eşitlik durumunda bir sonuca varılabilir. KNN, genellikle öklid mesafesi gibi bir mesafe metriği kullanır. Öklid mesafesi, iki nokta arasındaki doğrusal mesafeyi ölçer.

KNN, özellikle veri seti küçük olduğunda iyi bir performans gösterebilir. Hem sınıflandırma hem de regresyon problemleri için kullanılabilir. KNN yöntemi, sürekli değerli fonksiyonların regresyonuna kolayca uyarlanabilir [37]. Dezavantajı ise büyük veri setlerinde yavaş çalışmasıdır. Ayrıca optimal k değerinin seçilmesi önemlidir. Küçük k değerleri aşırı uyuma, büyük k değerleri ise modelin genelleme yeteneğini azaltmaya neden olabilir.

3.4 Model Değerlendirme Metrikleri

Regresyon modellerinin doğruluğunu değerlendirmek için uygun metriklerin kullanılması önemlidir. Bu bölümde, regresyon modellerinin performansını değerlendirmek için yaygın olarak kullanılan yöntemlere yer verilmiştir.

3.4.1 Ortalama Kare Hata (MSE)

MSE, gerçek ve tahmin edilen değerler arasındaki farkın karesinin ortalamasıdır. Daha düşük MSE değeri, modelin daha iyi performans gösterdiğini gösterir. Ancak, MSE, hataların karesi olduğu için orijinal veriyle aynı ölçekte değildir. Şekil 3.1'de MSE'nin formülü görülmektedir.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Şekil 3.1 Ortalama kare hata (MSE) formülü

3.4.2 Ortalama Mutlak Hata (MAE)

MAE, gerçek ve tahmin edilen değerler arasındaki mutlak farkın ortalamasıdır. MSE'ye benzer şekilde, daha düşük bir MAE değeri, daha iyi bir model performansını gösterir. MAE, hataların orijinal veriyle aynı ölçekte olması nedeniyle yorumlanması MSE'ye göre daha kolaydır. MAE'nin formülü Şekil 3.2'de görülmektedir.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Şekil 3.2 Ortalama mutlak hata (MAE) formülü

3.4.3 R-Kare

R-Kare, bağımsız değişkenlerin bağımlı değişken üzerindeki varyansını açıklama yüzdesini gösterir. 1 değerine yaklaşan bir R-Kare değeri, modelin verileri ne kadar iyi açıkladığını gösterir. Ancak, R-Kare, modelin karmaşıklığını ve gereksiz değişkenlerin eklenmesini göz önünde bulundurmadığı için tek başına yeterli değildir. Şekil 3.3'te R-Kare'nin formülü gösterilmiştir.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

Şekil 3.3 R-Kare formülü

3.5 Genetik Algoritma ve Özellik Seçimi

Genetik Algoritma, biyolojik evrimin temel prensiplerini taklit eden ve problemlerin optimize edilmesi için kullanılan bir optimizasyon algoritmasıdır. Özellikle özellik seçimi gibi karmaşık ve büyük boyutlu problemlerde etkili bir şekilde kullanılabilir. Genetik Algoritma, doğal seçilim, çaprazlama, mutasyon gibi operatörler kullanarak popülasyonu evrimleştirir ve en iyi çözümü elde etmeye çalışır.

Özellik seçimi, makine öğrenimi ve veri analizi alanında önemli bir konudur. Veri setlerindeki gereksiz ve yetersiz özellikler, modelin karmaşıklığını artırabilir, aşırı uyuma yol açabilir ve genelleme yeteneğini azaltabilir. Bu nedenle, doğru özellikleri seçmek, modelin performansını artırmak ve daha iyi sonuçlar elde etmek için önemlidir.

GA, sezgisel optimizasyon yaklaşımlardır ve çok değişkenli regresyon modellerinde değişken seçimi için kullanılabilir [38]. Genetik Algoritma, özellik seçimi için kullanıldığında, veri setindeki özelliklerin bir alt kümesini seçmek için kullanılır. Başlangıçta rastgele oluşturulan bir popülasyon, çaprazlama ve mutasyon operatörleri ile evrimleştirilir ve her iterasyonda en iyi özelliklerin bulunduğu popülasyon güncellenir. Bu süreç, belirli bir uygunluk fonksiyonuna göre yapılarak, en iyi özelliklerin bulunduğu alt küme elde edilmeye çalışılır.

Bu bölümde, Genetik Algoritma operatörleri ele alınacaktır.

3.5.1 Popülasyon, Birey ve Genetik Operatörler

Genetik Algoritmalarda popülasyon, çözümlerin bir kümesini temsil eder. Başlangıçta rastgele oluşturulan çözümlerden oluşan bir popülasyon, Genetik Algoritma'nın temel çalışma birimidir. Popülasyonun büyüklüğü, Genetik Algoritma'nın başarımını ve çözüm kalitesini etkileyebilir. Genellikle, popülasyonun büyüklüğü arttıkça algoritmanın çözüm kalitesi artar ancak hesaplama maliyeti de artabilir.

Popülasyondaki her bir çözüm, bir bireyi temsil eder. Bir birey genellikle bir çözümün temsilidir ve genellikle bir dizi gen olarak temsil edilir. Örneğin, özellik seçimi probleminde, birey belirli bir özellik kümesini temsil edebilir. Bireyler, Genetik Algoritma'nın çaprazlama ve mutasyon operatörleri tarafından işlenir ve yeni bireyler oluşturulur.

Genetik Algoritmalarda kullanılan temel operatörler; Seçim, Çaprazlama, Mutasyon ve Elitizm'dir. Genetik operatörler, Genetik Algoritma'nın her iterasyonunda uygulanarak popülasyonun evrimini sağlar. Bu operatörlerin etkili bir şekilde kullanılması, GA'nın başarı oranını ve çözüm kalitesini artırabilir.

3.5.2 Uyum (Fitness) Fonksiyonu

Genetik Algoritmaların temelinde yer alan uyum fonksiyonu, her bir bireyin ne kadar iyi bir çözüm olduğunu ölçen ve bireylerin seçilmesinde kullanılan bir metriktir. Uyum fonksiyonu, genellikle bir bireyin hedeflenen probleme ne kadar iyi uyum sağladığını belirlemek için tasarlanır ve bireyin uygunluğunu bir sayısal değer olarak ifade eder.

Uyum fonksiyonu genellikle maksimize edilmesi gereken bir metriktir. Örneğin, bir optimizasyon problemi için uygunluk fonksiyonu, maksimize edilmesi gereken bir hedef fonksiyon olabilir. Ancak hata metrikleri gibi bazı durumlarda, uygunluk fonksiyonunun minimize edilmesi gerekmektedir.

Uyum fonksiyonu, GA'nın temel prensiplerinden biri olan doğal seçilim işlemi için kritik bir öneme sahiptir. Daha yüksek uygunluk değerine sahip bireyler, gelecek nesillere daha fazla olasılıkla aktarılır, böylece popülasyon daha iyi çözümlere doğru evrilir.

Uyum fonksiyonunun doğru bir şekilde tanımlanması ve hedeflenen probleme uygun olarak seçilmesi önemlidir. Yanlış bir uygunluk fonksiyonu seçimi, GA'nın yanlış yönlendirilmesine ve kötü çözümlerin bulunmasına neden olabilir.

Bu tez çalışmasında, GA özellik seçimi için uygulanmıştır. Özellik seçimi için GA uygularken, her birey bir özellik kümesini temsil etmektedir. Başlangıçta rastgele seçilen özellik kümesinden oluşan bir popülasyon oluşturulur. Her iterasyonda, uyum fonksiyonu kullanılarak her bireyin uygunluğu hesaplanır. Daha sonra, seçim, çaprazlama ve mutasyon operatörleri kullanılarak yeni nesil bireyler oluşturulur. Bu süreç, belirli bir durma kriteri sağlanana kadar veya belirlenen iterasyon sayısına ulaşılana kadar tekrarlanır. Bu şekilde, GA, veri setindeki özelliklerin bir alt kümesini seçerek model performansını artırabilir ve gereksiz özelliklerin etkisini azaltabilir.

3.5.3 Seçim Operatörü

Genetik Algoritmaların temel bileşenlerinden biri olan seçim operatörü, popülasyondaki bireyler arasından hangi bireylerin bir sonraki nesle geçeceğini belirleyen operatördür. Seçim operatörünün temel amacı, daha yüksek uyum değerine sahip bireylerin seçilme olasılığını artırarak, popülasyonun evrimsel süreçte daha iyi çözümler bulmasını sağlamaktır.

Seçim operatörünün temel işlevi, doğal seçilim ilkesini taklit ederek, daha uygun bireylerin hayatta kalma ve üreme şansını artırmaktır. Bu süreç GA'nın her iterasyonunda uygulanır ve popülasyonun genel uygunluğunu artırır. Rulet Tekerleği Seçimi (Roulette Wheel Selection), Turnuva Seçimi (Tournament Selection), Sıralama Seçimi (Rank Selection), Elitizm (Elitism) ve Orantılı Seçim (Proportional Selection) literatürde yaygın olarak kullanılan seçimlerdir.

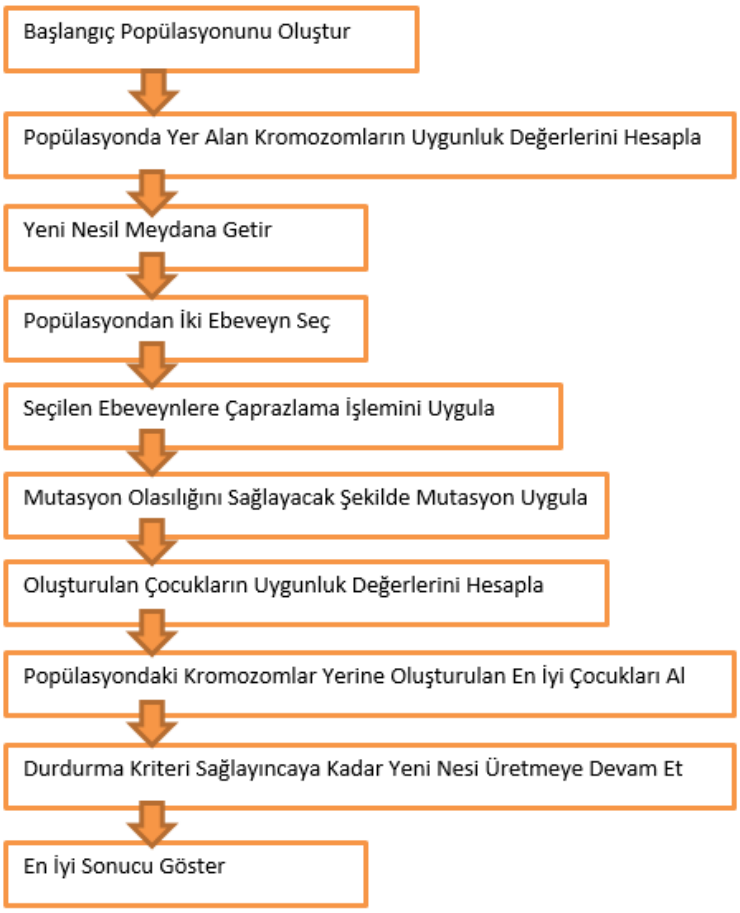
3.5.4 Mutasyon Operatörü

Mutasyon operatörü, GA'nın önemli bileşenlerinden olmakla beraber popülasyonun genetik çeşitliliğini artırmak için kullanılır. Mutasyon, bireylerin genetik yapısında rastgele değişiklikler yaparak, yeni çözümlerin keşfedilmesini sağlar [39]. Bu, GA'nın yerel minimumdan kaçınmasına ve daha iyi çözümler bulmasına yardımcı olur.

Mutasyon operatörünün temel amacı, popülasyonda varyasyonu korumak ve artırmaktır. Genetik Algoritmaların evrimsel süreci sırasında, popülasyonun çeşitliliği azalabilir ve bu da algoritmanın belirli bir çözüme sıkışmasına (yerel minimum) yol açabilir. Mutasyon, bu durumu engelleyerek, popülasyonda yeni genetik materyal oluşturur ve daha geniş bir arama alanı sağlar. Böylece, daha iyi çözümler bulunması mümkün hale gelir.

Mutasyon operatörünün etkinliği, mutasyon oranına bağlıdır. Mutasyon oranı, GA'nın her neslinde, bireylerin gen dizilerinde mutasyona uğrayan genlerin oranını belirler. Çok düşük bir mutasyon oranı, popülasyonun çeşitliliğini yeterince artırmaz ve algoritmanın yerel minimuma sıkışmasına neden olabilir. Çok yüksek bir mutasyon oranı ise, algoritmanın çözüm kalitesini düşürebilir ve GA'nın performansını olumsuz etkileyebilir. Bu nedenle, uygun bir mutasyon oranı seçmek, Genetik Algoritma'nın başarısı için önemlidir.

Mutasyon operatörünün etkinliği, yalnıza oranıyla değil, aynı zamanda mutasyon türüyle de ilişkilidir. Farklı mutasyon türleri, Genetik Algoritma'nın arama kapasitesini ve çözüm bulma yeteneğini etkileyebilir. Örneğin, bazı mutasyon türleri genetik materyali küçük adımlarla değiştirirken, diğerleri daha büyük değişiklikler yaparak daha geniş bir arama alanı sağlar. Bu tür farklılıklar, popülasyonun genetik çeşitliliğini artırmak ve yeni çözüm yolları keşfetmek için çeşitli stratejiler geliştirilmesini gerektirir.



Şekil 3.4 Genel genetik algoritma akışı

3.5.5 Çaprazlama Operatörü

Çaprazlama (Crossover) operatörü, GA süreçlerinde, popülasyonun yeni nesillerini oluştururken iki veya daha fazla ebeveyn bireyin genetik materyallerini birleştirerek yeni bireyler üretir [40]. Çaprazlama, Genetik Algoritmalarda bilgi değişimi ve çeşitliliğini sağlayarak, popülasyonun daha iyi çözümler bulmasına yardımcı olur.

Çaprazlama operatörünün temel amacı, mevcut çözümlerden yeni çözümler türetmek ve bu sayede arama alanında daha geniş bir keşif yapılmasını sağlamaktır. Çaprazlama, ebeveyn bireylerin genetik materyallerini birleştirerek, hem ebeveynlerin iyi özelliklerini yeni bireylere aktarır hem de yeni ve potansiyel olarak daha iyi çözüm yollarının bulunmasına olanak tanır. Böylece, GA'nın adaptasyon yeteneği ve optimizasyon performansı artar.

Yüksek bir çaprazlama oranı, popülasyonun çeşitliliğini artırır ve daha geniş bir arama alanı sağlar. Ancak, çok yüksek bir çaprazlama oranı, iyi çözümlerin korunmasını zorlaştırabilir.

3.6 Weka ile Özellik Seçimi

Weka, makine öğrenmesi ve veri madenciliği için yaygın kullanılan güçlü bir araçtır. Weka, özellik seçimi sürecini kolaylaştıran çeşitli algoritmalar ve yöntemler sunmaktadır. Özellik seçimi, makine öğrenmesi modellerinin performansını artırmak, modelin karmaşıklığını azaltmak ve daha anlamlı sonuçlar elde etmek için önemli bir adımdır. Weka, çeşitli özellik seçimi algoritmaları ve araçları sunarak, bu süreci kolaylaştırır ve otomatikleştirir.

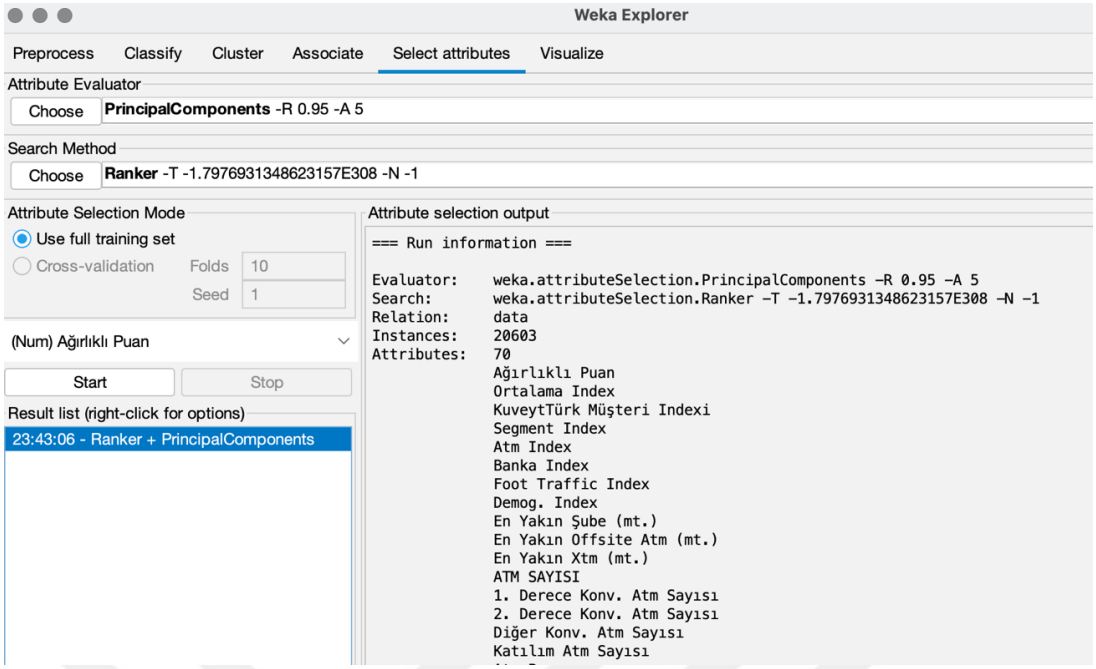
Weka'da özellik seçimi Attribute Evaluators (Özellik Değerlendiriciler) ve Search Methods (Arama Yöntemleri) olmak üzere iki ana bileşenle gerçekleştirilmektedir. Özellik değerlendiriciler, her bir özelliğin ne kadar önemli olduğunu belirlerken, arama yöntemleri ise bu değerlere dayanarak en iyi özellik kombinasyonunun bulunmasına yardımcı olurlar.

3.6.1 Weka'nın Özellik Seçimi Arayüzü

Makine öğrenmesi süreçlerinde yaygın olarak kullanılan Weka, kullanıcı dostu bir arayüze sahiptir. Şekil 3.2'de görüleceği üzere Weka Explorer'da, "Select Attributes" kısmının Özellik Değerlendirici ve Arama Yöntemi kısmından seçimler yapılabilmektedir.

Özellik değerlendiriciler, her bir özelliğin model performansına olan katkısını belirlemek için çeşitli değerlendirme kriterleri sunar. Bu kriterler, genellikle özelliklerin sınıflandırma ve regresyon modellerindeki önemini ölçen istatistiksel ve matematiksel metrikler içerir. Örneğin, bilgi kazancı, chi-squared testi, korelasyon tabanlı metrikler gibi çeşitli değerlendirme metrikleri kullanarak, hangi özelliklerin modelin başarısını artıracağını belirler.

Arama yöntemi ise, özelliklerin kombinasyonlarını belirlemek için kullanılan stratejileri içerir. Bu yöntemler arasında ileri seçim, geri eleme ve genetik algoritmalar gibi çeşitli teknikler bulunur.



Şekil 3.5 Weka'nın özellik seçim arayüzü

3.7 Weka'da Özellik Seçiminde Kullanılan Yöntemler

Weka'da yaygın olarak kullanılan bazı özellik seçimi yöntemleri şu şekildedir:

Filtre Yöntemleri (Filter Methods):

Veri setindeki özellikleri bağımsız olarak değerlendirir ve belirli bir kriter (korelasyon, bilgi kazancı vb.) kullanarak özellikleri sıralamaktadır.

Bu yöntem, modelin eğitim sürecinden bağımsız olarak özellik seçimi yapmaktadır.

Weka'da Chi-Kare (Chi-squared), Bilgi Kazancı ve Relief gibi çeşitli filtre yöntemleri bulunmaktadır.

Sarmalayıcı Yöntemler (Wrapper Methods):

Modelin performansını doğrudan optimize etmek için özellik alt kümelerini değerlendirmektedir.

Bu yöntem, belirli bir makine öğrenmesi algoritması kullanarak en uygun özellik kümesini belirlemektedir.

Weka'da, Greedy Stepwise ve Best First gibi çeşitli sarmalayıcı yöntemler bulunmaktadır.

3.7.1 InfoGainAttributeEval

InfoGainAttributeEval yöntemi, her bir özelliğin hedef değişkenle olan ilişkisini bilgi teorisi temelli bir ölçü olan Information Gain (Bilgi Kazancı) aracılığıyla değerlendirmektedir. Bilgi kazancı, özelliklerin hedef değişken üzerindeki etkisini ölçmek için kullanılır ve bu sayede en anlamlı özellikler belirlenmektedir.

InfoGainAttributeEval, her bir özelliğin bilgi kazancını hesaplamakta ve bu değerlere dayanarak özellikleri önem sırasına göre sıralamaktadır. İşlem adımları sırasıyla, entropi hesaplama, koşullu entropi, bilgi kazancı ve sıralamadır. İlk adımda, hedef değişkenin entropisi hesaplanmaktadır. Koşullu entropi kısmında, her bir özellik için koşullu entropi hesaplanmaktadır. Bilgi kazancı bölümünde ise, her bir özellik için bilgi kazancı, genel entropi ile koşullu entropi arasındaki fark olarak hesaplanmaktadır. Sıralama bölümünde, bilgi kazançları hesaplandıktan sonra, özellikler bilgi kazancı değerlerine göre sıralanmaktadır. En yüksek bilgi kazancına sahip özellikler, hedef değişkeni en iyi açıklayan özellikler olarak kabul edilmektedir.

InfoGainAttributeEval'in hesaplama gücü yüksektir ve büyük veri setlerinde etkili bir şekilde çalışmaktadır. Dezavantajı olarak, bilgi kazancı, özelliklerin bağımsız olduğunu varsayar, bu da bazı durumlarda yanıltıcı olabililmektedir.

3.7.2 GainRatioAttributeEval

GainRatioAttributeEval, bilgi kazancı yöntemine dayanan bir diğer özellik değerlendirici yöntemidir. Bilgi kazancı yüksek olan özellikler, çok farklı değere sahip olduklarında önyargılı olabilmektedir. Bu sebeple, bu yöntem, bilgi kazancı hesaplamasına ek olarak, bilgi kazancını normalize eden bir oranla çalışmaktadır. Bu sayede, bilgi kazancının doğrudan kullanımında karşılaşılan bazı problemleri azaltmakta ve daha dengeli özellik seçimini sağlamaktadır.

GainRatioAttributeEval, bilgi oranını hesaplamak için sırasıyla entropi, koşullu entropi, bilgi kazancı, ayrıklaştırma entropisi ve bilgi oranlarının hesaplanması, son olarak da özelliklerin sıralanması adımlarını izlemektedir. Avantajları olarak, bilgi kazancını normalize etmesi, özelliklerin hedef değişken üzerindeki etkisini daha adil bir şekilde değerlendirmesi ve dengeli özellik seçimi yapması sayılabilir. Ancak, bilgi oranının hesaplanması, bilgi kazancına göre daha karmaşıktır ve hesaplama süresi daha uzun olmaktadır.

3.7.3 ReliefFAttributeEval

ReliefFAttributeEval, her bir özelliğin hedef değişkenle olan ilişkisini değerlendirirken, örneklerin birbirine olan benzerliklerini de dikkate almaktadır. Özellikle, karmaşık ve yüksek boyutlu veri setlerinde etkili bir şekilde çalışmaktadır.

Orijinal Relief algoritmasının gelişmiş bir versiyonudur ve veri setindeki her bir özelliğin hedef değişkeni ne kadar iyi açıkladığını belirlemek için kullanılmaktadır. ReliefF, her bir veri örneği için en yakın komşularını ya da benzer örnekleri analiz ederek, özelliklerin önemini hesaplamaktadır. Bu yöntem, sınıflandırma ve regresyon problemlerinin ikisinde de kullanılmaktadır.

Veri setinden rastgele örnek seçimi, seçilen örneğe en yakın komşu seçimi, her bir özellik için ağırlıkların güncellenmesi, bu adımların belirli bir iterasyon boyunca tekrarlanması ve özelliklerin ağırlıklara göre sıralanması adımlarını izlemektedir.

ReliefF, veri setindeki karmaşık ve etkileşimli ilişkileri yakalayabilmektedir. Bu yöntem, yüksek boyutlu veri setlerinde etkili bir şekilde çalışabilmektedir. Aynı zamanda, gürültüye karşı dayanıklıdır ve etkili sonuçlar verebilmektedir. Ancak, hesaplama maliyeti yüksek olmaktadır.

3.7.4 CfsSubsetEval

CfsSubsetEval, veri seti içerisindeki özelliklerin alt kümelerini değerlendirmek için kullanılan bir özellik seçim yöntemidir. Bu yöntem, özelliklerin hedef değişkenle olan korelasyonlarını ve kendi aralarındaki korelasyonları dikkate alarak en uygun özellik alt kümelerini seçmeye çalışmaktadır.

CfsSubsetEval, makine öğrenmesi modellerinde kullanılacak en iyi özelliklerin belirlenmesine yardımcı olur. Özellikle büyük veri setlerinde, gereksiz ve tekrarlı özellikleri elemek, modelin performansını artırmak ve hesaplama maliyetlerini düşürmek için bu yöntem oldukça etkili olmaktadır.

3.7.5 WrapperSubsetEval

WrapperSubsetEval, özellik seçimi için kullanılan bir yöntemdir ve makine öğrenmesi modellerinde kullanılacak en iyi özellik alt kümelerini belirlemek amacıyla tasarlanmıştır. Bu yöntem, seçilen özelliklerin doğrudan model performansına olan etkisini değerlendirmekte ve özellik seçim sürecinde belirli bir

Makine öğrenmesi algoritmasını kullanarak en iyi alt kümeyi belirlemeye çalışmaktadır. WrapperSubsetEval yönteminin hesaplama maliyeti yüksek olsa da, sağladığı yüksek performans ve esneklik nedeniyle birçok uygulamada tercih edilen bir yöntemdir.

3.8 Weka ile Özellik Seçimi Parametreleri

Weka'da özellik seçimi yaparken Bölüm 3.7'de bahsedilen algoritmaların parametrelerini ayarlamak gerekmektedir. Bu parametreler, kullanılan algoritmaya ve veri setinin özelliklerine göre farklılık gösterebilmektedir. Aşağıda, çalışmada kullanılan özellik seçim algoritmalarının parametreleri gösterilmektedir:

CfsSubsetEval

- **Metrik Türü**: Kullanılacak korelasyon metrik türünü belirler. Örneğin: Pearson korelasyonu.
- **Gereken İstatistikler (NumRequiredStatistics):** İstatiktiksel değerlendirmede gereken minimum istatistik sayısını belirler.

InfoGainAttributeEval

- **Binning**: Sürekli özelliklerin sınıflandırılmasında kullanılan bin sayısını belirler.
- **Normalize**: Bilgi kazanımının normalleştirilip normalleştirilmeyeceğini belirler.

GainRatioAttributeEval

- **numFeatures**: Seçilecek maksimum özellik sayısını belirler.
- **evaluationGain**: Bilgi kazancını hesaplama ve değerlendirme sürecinde kullanılan parametredir.

Genetik Algoritma (GeneticSearch)

- **Popülasyon Boyutu (PopulationSize):** Genetik Algoritma popülasyonundaki birey sayısını belirler.
- **Maksimum Jenerasyon Sayısı (MaxGenerations):** Genetik Algoritma'nın çalıştırılacağı maksimum jenerasyon sayısını belirler.
- **Mutasyon Oranı (MutationRate):** Her jenerasyonda kaç bireyin mutasyona uğrayacağını belirler.

4
DENEYSEL SONUÇLAR

Bu bölümde makine öğrenmesi regresyon çalışmaları anlatılmıştır. Bölüm 4.1'de makine öğrenmesi modellerinin sonuçları verilmiş ve her birinin ayrı ayrı karşılaştırması yapılmıştır. Bölüm 4.2'de ise Genetik Algoritma ile yapılan çalışmalar anlatılmış, seçilen özellikler listelenmiştir. Daha sonra, GA ile seçilen özelliklerin makine öğrenmesi modeline verilmesiyle yapılan çalışmalardan bahsedilmiştir. Bölüm 4.3'te ise Weka'da farklı algoritmalar ile yapılan özellik seçme çalışmaları raporlanmıştır. Son olarak Bölüm 4.4'te ise tüm ham özelliklerin makine öğrenmesi sonuçları, Genetik Algoritma ve Weka ile seçilen özelliklerin tekrardan makine öğrenmesi modeline verilerek elde edilen sonuçların karşılaştırılması yapılmıştır.

4.1 Makine Öğrenmesi Modellerinin Sonuçları ve Karşılaştırması

Çalışmanın bu bölümünde, kullanılmış olan tüm regresyon algoritmalarının sonuçları verilmiştir. Her bir modelin performansı, R-Kare, Ortalama Mutlak Hata (MAE) ve Ortalama Kare Hatası (MSE) metrikleriyle ölçülmüştür. Performans değerlendirmeleri yapılırken, sadece tek bir metriğe bakılmamıştır. Sadece R-Kare sonucuna bakmak, modelin karmaşıklığı ve gereksiz değişkenlerin eklenmiş olmasını dikkate almadığı için yanıltıcı olabilir. Aynı zamanda, büyük hatalar ve aşırı uyum (overfitting) konusunda da tek başına çok iyi bilgi vermemektedir. MAE ve MSE, hataların doğası hakkında bilgi ortaya koymaktadır. Ortalama ve büyük hataların daha az olduğu modellere de bakılmıştır. Buradaki asıl amaç, hem R-kare'nin yüksek olduğu, hem de MAE ve MSE değerlerinin düşük olduğu bir denge bulmaktır. Tek bir metriğe bağımlı kalmak yerine, tüm metrikleri dikkate alarak en dengeli performansa sahip olan model seçilmeye çalışılmıştır. Tüm metrikler değerlendirilerek, en dengeli ve güvenilir performansı sunan modeller performans sonuçlarına göre Tablo 4.1'te sıralanmıştır. Araştırmada, hedef değişken olan Ağırlıklı Puan'ı tahminlemek için aşağıdaki makine öğrenmesi modelleri kullanılmıştır.

- Rastgele Orman
- Karar Ağaçları
- Ridge Regresyon
- Lasso Regresyon
- Gradient Boosting
- AdaBoost
- Elastic Net Regresyonu
- Destek Vektör Regresyonu (SVR)
- K-En Yakın Komşu (KNN) Regresyonu

İlk fazda, veri ön işleme adımları, uzmanlarla görüşüldükten sonra veri setinden özellik çıkarılması gibi adımlarla belirlenen 81 özellik üzerinden model sonuçları elde edilmiştir. İkinci aşamada ise, Bölüm 3.2'de anlatıldığı gibi SelectPercentile fonksiyonu ile elde edilen ve 10 özelliğin çalıştırıldığı modelin performansları verilmiştir. Daha sonra, SelectPercentile fonksiyonunun parametresi değiştirilerek farklı sayıda özellikler elde edilmiştir. En sonda da, farklı özellik sayılarıyla çalıştırılan modellerin performans karşılaştırmaları yapılmıştır. Böylece, özellik sayısındaki değişimin makine öğrenmesi algoritmalarının performanslarına olan etkisi ölçülmüştür.

Makine öğrenmesi modellerinin performansı, çapraz doğrulama yapılarak değerlendirilmiştir. Bu çalışmada, 10 katlı çapraz doğrulama ile modeller değerlendirilmiştir. Her bir kat için modellerin performansı ayrı ayrı hesaplanmış ve bu sonuçların ortalaması alınmıştır.

Rastgele Orman, Karar Ağaçları, Ridge Regresyon, Lasso Regresyon, Gradient Boosting, Adaboost, Elastic Net, SVR ve KNN Regresyon algoritmaları için 81 özellikten elde edilen performans değerleri Tablo 4.1'de verilmiştir. Tablo 4.2'de ise SelectPercentile fonksiyonuyla seçilen 10 özellikten elde edilen performans değerleri görülmektedir. R-Kare, MSE ve MAE sonuçlarına göre en başarılı algoritmalar Rastgele Orman ve Gradient Boosting olarak gözlemlenmiştir. KNN Regresyon ise %55 civarlarında oldukça düşük performans sergilemiştir.

Tablo 4.1 81 özellik için makine öğrenmesi algoritma başarı değerleri

Algoritma	R-Kare	MSE	MAE
Rastgele Orman	89,36	0,041	0,032
Karar Ağaçları	82,17	0,093	0,065
Ridge Regresyon	57,13	0,275	0,223
Lasso Regresyon	57,53	0,273	0,221
Gradient Boosting	87,16	0,088	0,06
AdaBoost	85,87	0,095	0,061
Elastic Net	60,12	0,23	0,189
Destek Vektör	64,26	0,212	0,181
KNN Regresyon	55,15	0,365	0,339

Tablo 4.2 10 özellik için makine öğrenmesi algoritma başarı değerleri

Algoritma	R-Kare	MSE	MAE
Rastgele Orman	91,27	0,027	0,019
Karar Ağaçları	88,17	0,093	0,066
Ridge Regresyon	60,01	0,24	0,191
Lasso Regresyon	60,12	0,24	0,191
Gradient Boosting	90,56	0,036	0,027
AdaBoost	87,89	0,087	0,059
Elastic Net	62,09	0,24	0,191
Destek Vektör	67,15	0,193	0,171
KNN Regresyon	59,23	0,286	0,236

Özellik seçimi ve veri ön işleme yapılmadan gerçekleştirilen modellerin performansı, özellik seçimi yapıldıktan sonraki süreçlere göre daha düşük performans sergilemiştir.

4.1.1 Makine Öğrenmesi Parametrelerinin Değiştirilmesiyle Oluşan Model Performansları

Bu bölümde, makine öğrenmesi modellerinin performanslarını iyileştirmek amacıyla kullanılan hiper-parametre optimizasyonu ve bu süreçte elde edilen

sonuçlar ele alınmıştır. Hiper-parametre optimizasyonu, modelin öğrenme sürecini ve tahmin gücünü etkileyen parametrelerin ayarlanmasıdır. Bu çalışmada, farklı makine öğrenmesi algoritmaları üzerinde hiper-parametre optimizasyonu yapılarak, en iyi performans gösteren modeller belirlenmiştir.

Hiper-parametre optimizasyonu, modelin performansını maksimize etmek amacıyla kullanılan çeşitli arama stratejileri ile gerçekleştirilmiştir. Izgara Araması (Grid Search) ve Rastgele Arama (Random Search) en çok kullanılan yöntemlerdendir. Her bir model için farklı parametre kombinasyonları denenmiş ve çapraz doğrulama kullanılarak en iyi performans gösteren kombinasyonlar seçilmiştir.

Rastgele Orman algoritması için "n_estimators", "max_depth", "min_samples_split" ve "min_samples_leaf" parametreleri farklı değerlerde denenmiş ve sonuçları gösterilmiştir. Gradient Boosting için "n_estimators", "learning_rate", "max_depth" ve "subsample" parametrelerinden oluşan en iyi kombinasyon bulunmaya çalışılmıştır. SVR için ise "C", "epsilon" ve "kernel" parametreleri, Ridge ve Lasso Regresyon modelleri için ise "alpha" değeriyle farklı sonuçlar üretilmiştir.

Her bir model için hiper-parametre optimizasyonu sonucunda elde edilen en iyi parametre kombinasyonları ve bu kombinasyonlarla elde edilen performans metrikleri Tablo 4.3, Tablo 4.4, Tablo 4.5 ve Tablo 4.6'da gösterilmiştir.

Tablo 4.3 Rastgele Orman algoritmasının hiper-parametrelere göre başarıları

n_estimators	**max_ depth**	**min_samples _split**	**min_samples _leaf**	**R-Kare**	**MSE**	**MAE**
100	8	5	2	91,22	0,017	0,013
150	10	4	1	91,31	0,015	0,012
200	12	6	3	91,07	0,019	0,015
120	9	5	2	91,37	0,014	0,011
180	11	4	1	91,12	0,018	0,016

Tablo 4.4 Karar Ağacı algoritmasının hiper-parametrelere göre başarıları

criterion	splitter	max_ depth	min_samp les_split	min_samp les_leaf	R-Kare	MSE	MAE
gini	best	10	2	1	88,32	0,08	0,06
entropy	random	15	4	2	88,15	0,09	0,06
gini	best	8	3	1	90,07	0,07	0,05
entropy	best	12	5	3	89,43	0,82	0,05
gini	random	None	2	1	89,01	0,08	0,05

Tablo 4.5 SVR algoritmasının hiper-parametrelere göre başarıları

C	epsilon	kernel	R-Kare	MSE	MAE
1	0,1	rbf	66,32	0,184	0,168
10	0,2	linear	67,15	0,177	0,155
0,1	0,05	poly	68,53	0,163	0,147
1	0,1	sigmoid	69,17	0,152	0,133
100	0,1	rbf	65,17	0,192	0,179

Tablo 4.6 Gradient Boosting algoritmasının hiper-parametrelere göre başarıları

max_depth	n_estimator	learning_rate	R-Kare	MSE	MAE
5	100	0,1	90,15	0,022	0,017
10	200	0,05	90,91	0,02	0,015
15	300	0,01	90,07	0,023	0,018
20	400	0,1	89,87	0,025	0,019
25	500	0,05	88,91	0,027	0,021

Hiper-parametre optimizasyonu, her model için performansı önemli ölçüde iyileştirmiştir. Özellikle Rastgele Orman modeli, optimize edilmiş parametrelerle en yüksek R-Kare değerine ulaşmış ve diğer performans metriklerinde de üstünlük sağlamıştır. Bu sonuçlar, hiper-parametre ayarlarının model performansını artırmada önemli bir rol oynadığını göstermektedir.

4.1.2 Özellik Sayısı Değişimin Model Performanslarına Etkisi

Bu bölümde, model performanslarının kullanılan özellik sayısının değişimi ile nasıl etkilendiği incelenmiştir. Özellik sayısının azaltılması, modelin genel performansını artırabilir ve aşırı öğrenmeyi önleyebilir. Bu çalışmada, SelectPercentile fonksiyonu kullanılarak, her bir model için farklı sayıda özelliklerle eğitim yapılmış ve sonuçlar değerlendirilmiştir.

SelectPercentile yöntemi, belirli bir yüzdelik dilime göre en iyi özellikleri seçmektedir. Seçim işlemi, özelliklerin hedef değişken ile olan ilişkisine göre yapılmaktadır. Bu aşamada, modelin performansını belirlemek için, farklı yüzdelik dilimlerde (%10, %15, %20, %25, %30, %35) seçilen özelliklerle modeller eğitilmiştir. Özellik sayısının artması, her model için performansın artacağı anlamına gelmediği görülmüştür. Özelliklerin değişmesi, farklı modeller için performansları farklı ölçüde etkilemiştir. Genel olarak tüm modellerde, özellik sayısının artmasıyla model performansları artmıştır. Ancak, tüm özellik artımlarında model performansının artışı söz konusu değildir. Ayrıca, fazla sayıda azaltılan özellik sayısı, modelin performansında belirgin bir düşüşe yol açabilmektedir. Algoritmaların genelinde, özellik sayısının düşmesiyle, model performansında gözle görülür bir düşüş gerçekleşmiştir. Bu nedenle, en uygun özellik setini belirlemek için dikkatli bir denge kurulmalıdır.

Tablo 4.7 5 özellik için makine öğrenmesi algoritma başarı değerleri

Algoritma	R-Kare (%)	MSE	MAE
Rastgele Orman	90,94	0,041	0,021
Karar Ağaçları	87,97	0,107	0,073
Ridge Regresyon	59,24	0,252	0,214
Lasso Regresyon	59,56	0,251	0,202
Gradient Boosting	90,41	0,042	0,034
AdaBoost	87,35	0,091	0,061
Elastic Net	61,64	0,253	0,207
Destek Vektör	66,45	0,207	0,188
KNN Regresyon	58,85	0,301	0,256

Tablo 4.8 10 özellik için makine öğrenmesi algoritma başarı değerleri

Algoritma	R-Kare	MSE	MAE
Rastgele Orman	91,27	0,027	0,019
Karar Ağaçları	88,17	0,093	0,066
Ridge Regresyon	60,01	0,242	0,192
Lasso Regresyon	60,12	0,24	0,191
Gradient Boosting	90,56	0,036	0,027
AdaBoost	87,89	0,087	0,059
Elastic Net	62,09	0,24	0,191
Destek Vektör	67,15	0,193	0,171
KNN Regresyon	59,23	0,286	0,236

Tablo 4.9 15 özellik için makine öğrenmesi algoritma başarı değerleri

Algoritma	R-Kare	MSE	MAE
Rastgele Orman	91,21	0,025	0,018
Karar Ağaçları	88,43	0,086	0,063
Ridge Regresyon	60,31	0,232	0,184
Lasso Regresyon	60,48	0,23	0,183
Gradient Boosting	90,69	0,032	0,024
AdaBoost	88,08	0,083	0,057
Elastic Net	62,91	0,23	0,183
Destek Vektör	67,87	0,186	0,163
KNN Regresyon	59,85	0,273	0,222

Tablo 4.10 20 özellik için makine öğrenmesi algoritma başarı değerleri

Algoritma	R-Kare	MSE	MAE
Rastgele Orman	91,23	0,022	0,016
Karar Ağaçları	88,65	0,079	0,059
Ridge Regresyon	60,55	0,222	0,175
Lasso Regresyon	60,59	0,22	0,174
Gradient Boosting	90,78	0,029	0,022
AdaBoost	88,18	0,079	0,055
Elastic Net	63,24	0,22	0,174
Destek Vektör	68,18	0,179	0,154
KNN Regresyon	60,32	0,258	0,208

Tablo 4.11 25 özellik için makine öğrenmesi algoritma başarı değerleri

Algoritma	R-Kare	MSE	MAE
Rastgele Orman	91,19	0,02	0,015
Karar Ağaçları	88,77	0,071	0,056
Ridge Regresyon	60,88	0,212	0,168
Lasso Regresyon	60,95	0,21	0,166
Gradient Boosting	90,84	0,025	0,02
AdaBoost	88,28	0,075	0,053
Elastic Net	64,11	0,204	0,159
Destek Vektör	68,86	0,171	0,146
KNN Regresyon	61,03	0,243	0,194

Tablo 4.12 30 özellik için makine öğrenmesi algoritma başarı değerleri

Algoritma	R-Kare	MSE	MAE
Rastgele Orman	91,22	0,017	0,013
Karar Ağaçları	88,84	0,064	0,053
Ridge Regresyon	61,04	0,208	0,162
Lasso Regresyon	61,09	0,207	0,161
Gradient Boosting	90,91	0,022	0,018
AdaBoost	88,87	0,071	0,051
Elastic Net	64,23	0,201	0,157
Destek Vektör	69,17	0,164	0,138
KNN Regresyon	61,12	0,229	0,18

Tablo 4.13 35 özellik için makine öğrenmesi algoritma başarı değerleri

Algoritma	R-Kare	MSE	MAE
Rastgele Orman	91,25	0,018	0,014
Karar Ağaçları	88,76	0,066	0,055
Ridge Regresyon	60,91	0,211	0,167
Lasso Regresyon	60,94	0,21	0,166
Gradient Boosting	90,89	0,023	0,019
AdaBoost	88,75	0,072	0,052
Elastic Net	64,77	0,199	0,152
Destek Vektör	69,98	0,161	0,135
KNN Regresyon	60,76	0,223	0,176

Sonuç olarak, bu çalışma, özellik seçimi sürecinin, makine öğrenmesi modellerinin performansını önemli ölçüde etkileyebileceğini göstermektedir. Doğru özelliklerin seçilmesiyle, hem modelin doğruluğunun arttığı hem de hesaplama maliyetinin düştüğü gözlemlenmiştir.

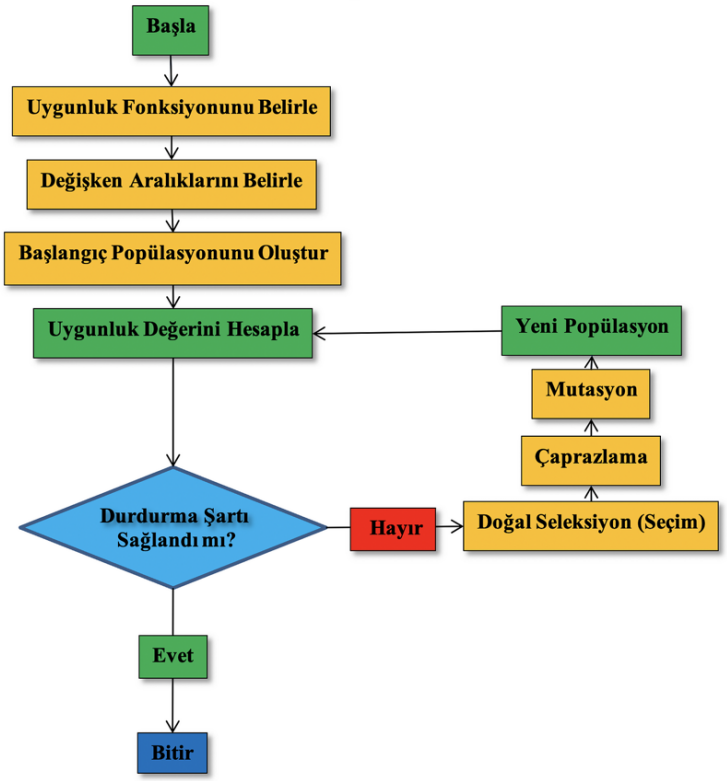
Her özellik kombinasyonu için modellerin performans ölçütleri Tablo 4.7'den 4.13'e kadar gösterilmiştir. Tüm özellik sayıları için Rastgele Orman ve ardından Gradient Boosting algoritmaları en iyi sonuçları vermiştir. Ridge Regresyon, Lasso

Regresyon, Destek Vektör (SVR) ve KNN Regresyon algoritmaları tüm özellik sayıları için %70 R-Kare skoru üzerine çıkamamıştır.

4.2 Genetik Algoritma ile Özellik Seçim Sonuçları

Bu bölümde Genetik Algoritma ile yapılmış olan çalışmalar ve özellik seçimi işlemlerinin sonuçları sunulmuştur. Genetik Algoritma, doğal seçilim ve genetik prensiplerine dayanan bir optimizasyon tekniğidir ve bu çalışmada özellik seçimi amacıyla kullanılmıştır.

Başlangıçta, rastgele belirlenen özellik kombinasyonlarından oluşan bir popülasyon oluşturulmuştur. Bu popülasyon, yüz bireyden (özellik kombinasyonları) oluşmaktadır. Her birey, özelliklerin seçilip seçilmediğini belirten ikili bir dizidir (0 veya 1). Eğer bir bireyin pozisyonunda 1 değeri varsa, o pozisyondaki özellik seçilir. Her bireyin uygunluğunu değerlendirmek için uygunluk (fitness) fonksiyonu kullanılmıştır. Bu çalışmada, fitness fonksiyonu olarak modelin R-kare değeri kullanılmıştır. Seçilen özellikler kullanılarak model çapraz doğrulama ile değerlendirilmiştir. "cross_val_score" fonksiyonu, modeli alt küme veri seti üzerinde on katlı çapraz doğrulama ile çalıştırmış ve R-Kare skorunu hesaplamıştır. Bu sayede, yüksek R-kare değerine sahip özellik kombinasyonları daha yüksek uygunluk değeri ile değerlendirilmiştir. Fitness değeri, seçilen özelliklerle oluşturulan alt veri seti kullanılarak, bir Rastgele Orman modelinin çapraz doğrulama sonuçlarının ortalaması olarak hesaplanmıştır. Yüksek uygunluk değerine sahip bireylerin bir sonraki nesilde daha fazla yer alması sağlanmıştır. Bu işlem, en uygun bireylerin seçilmesi ve yeni bireylerin oluşturulması sürecini içermektedir. Seçilen bireyler üzerinde çaprazlama ve mutasyon işlemleri gerçekleştirilmiştir. Çaprazlama, iki ebeveyn bireyden yeni bireyler oluşturma sürecidir ve çeşitliliği artırır. Mutasyon ise, belirli bir olasılıkla bireylerin genlerinde rastgele değişiklikler yaparak yeni özellik kombinasyonları oluşturur. Mutasyon işlemi sonucunda, eğer bir gen mutasyona uğrarsa, "1" değerleri "0", "0" değerleri ise "1" değerine dönüşmektedir. Mutasyon ve çaprazlama fonksiyonlarında parametreler mutasyon için 0,1 ve çaprazlama için ilk aşamada 0,8 seçilmiştir.



Şekil 4.1 Genetik algoritma akışı

Mutasyon ve çaprazlama oranlarının değişmesi, GA'nın hesaplama maliyetini değiştirdiği görülmüştür. Çaprazlama ve mutasyon işlemleri sonucunda yeni nesiller oluşturulmuş ve bu süreç belirli bir iterasyon sayısına (50 jenerasyon) ulaşana kadar tekrarlanmıştır. Yeni nesiller, popülasyonun en uygun bireyleri kullanılarak oluşturulmuş ve her jenerasyon sonunda popülasyon sıralanarak en iyi bireyler seçilmiştir. Genetik Algoritma'nın akışı Şekil 4.1'de gösterilmiştir.

Genetik Algoritma ile özellik seçim sonuçlarında, en uygun özellik sayısı olarak toplamda 11 özellik seçilmiştir. Bu özellikler, en iyi uyum değerine sahip özellik setidir. Seçilen özellikler Tablo 4.14'te gösterilmektedir. Makine öğrenmesinde SelectPercentile ile seçilen özelliklerden farklı olarak 5, 6, 7, 8, 9 ve 10 no'lu özellikler olan Muhafazakar Oran, Tüzel Müşteri Sayısı, Katılım ATM Sayısı, YYY Bankası ATM Sayısı, XXX Bankası ATM Sayısı ve XTM Sayısı özellikleri belirlenmiştir.

Tablo 4.15'te ise GA ile seçilen bu özelliklerin, ilk çalışmadaki makine öğrenmesinde en başarılı performansı gösteren Rastgele Orman algoritmasına verilmesiyle oluşan ölçüm sonuçları gösterilmiştir. GA'nın bulduğu özelliklerle

değerlendirilen modelin, ilk çalışmaya göre daha başarılı sonuçlar ürettiği gözlemlenmiştir.

Tablo 4.14 Genetik Algoritma ile seçilen özellikler

No	Özellik	No	Özellik
1	En Yakın Şube (mt.)	**7**	Katılım ATM Sayısı
2	En Yakın Offsite ATM (mt.)	**8**	YYY Bankası ATM Sayısı
3	Toplam Nüfus	**9**	XXX Bankası ATM Sayısı
4	Foot Traffix Index	**10**	XTM Sayısı
5	Muhafazakar Oran	**11**	Demog. Index
6	Tüzel Müşteri Sayısı		

Tablo 4.15 Genetik Algoritma ile seçilen özelliklerin Rastgele Orman algoritmasına verilmesiyle oluşan performans sonuçları

Algoritma	R-Kare	MSE	MAE
Rastgele Orman	0,97	0,04	0,02

4.3 Weka ile Özellik Seçimi Sonuçları

Bu bölümde, regresyon problemi için Weka uygulaması kullanılarak gerçekleştirilen özellik seçimi sonuçları özetlenmiştir. Bu çalışmada, Weka platformu üzerinden farklı özellik seçimi yöntemleri uygulanmış ve bir özellik seti elde edilmiştir. Sonrasında, seçilen özelliklerin daha önceki çalışmalarda en başarılı performansı gösteren Rastgele Orman algoritmasıyla değerlendirilmesi sonucu çıkan performans sonucu verilmiştir.

İlk aşamada, korelasyon bazlı özellik seçme yöntemi olan CfsSubsetEval kullanılmıştır. Bu yöntemdeki amaç, hedef değişkenle yüksek korelasyona sahip ve birbirleriyle düşük korelasyona sahip özellikleri seçmektir. İkinci aşamada, Wrapper Subset Evaluation yöntemi kullanarak, bir makine öğrenmesi algoritması olan Rastgele Orman ile en iyi performans gösteren özellik alt kümelerini seçmek amaçlanmıştır. Üçüncü özellik seçme yöntemi olarak, InfoGainAttributeEval kullanılmıştır. Bilgi kazancı ile hedef değişken hakkında en fazla bilgi sağlayan

özelliklerin seçilmesi amaçlanmıştır. Dördüncü ve beşinci aşamalarda ReliefFAttributeEval ve GainRatioAttributeEval yöntemleri kullanılmıştır.

Yukarıda belirtilen beş farklı yöntem kullanılarak, en iyi özellikler seçilmiş ve bu özellikler Tablo 4.16'dan 4.20'ye kadar listelenmiştir. Özellik seçiminde kullanılan yöntemlerin her biri farklı özellikler önermiş, GA çalışmasında tespit edilen özellik sayısı kadar; yani on birer özellik seçilmiştir. Bu özellikler Rastgele Orman modeline verilmiş ve R-Kare, MSE ve MAE her bir model için hesaplanmıştır. Seçilen özellikler ile oluşturulan modellerin performans değerleri Tablo 4.21'de gösterilmiştir. Ayrıca, Tablo 4.22'de bu beş farklı yöntemde kullanılan parametre değerleri verilmiştir.

Sonuç olarak en başarılı performansı veren özelliklerin InfoGain yöntemi ile seçilmiş olduğu tespit edilmiştir.

Tablo 4.16 CfsSubsetEval ile seçilen özellikler

No	Özellik	No	Özellik
1	Atm Index	**7**	Denizbank
2	Ortalama Index	**8**	1. Derece Konv. Atm
3	En Yakın Offsite Atm (mt.)	**9**	XXX Müşteri Indexi
4	Segment Index	**10**	XTM Sayısı
5	Muhafazakar Oran	**11**	Ortalama Hane Geliri
6	Demog Index		

Tablo 4.17 WrapperSubsetEvaluation ile seçilen özellikler

No	Özellik	No	Özellik
1	En Yakın Şube (mt.)	**7**	Ortalama Index
2	Banka Index	**8**	Ortalama Hane Geliri
3	Demog Index	**9**	XXX Şube Uzak
4	Atm Index	**10**	Foot Traffic
5	Foot Traffic Index	**11**	XXX Müşteri Indexi
6	Toplam Nüfus		

Tablo 4.18 InfoGainAttributeEval ile seçilen özellikler

No	Özellik	No	Özellik
1	XXX Xtm	**7**	Katılım ATM Sayısı
2	En Yakın Offsite ATM (mt.)	**8**	YYY Bankası ATM Sayısı
3	Toplam Nüfus	**9**	Ortalama Hane Geliri
4	Ortalama Index	**10**	Banka Index
5	Muhafazakar Oran	**11**	XXX Müşteri Indexi
6	Tüzel Müşteri Sayısı		

Tablo 4.19 ReliefFAttributeEval ile seçilen özellikler

No	Özellik	No	Özellik
1	Atm Index	**7**	Banka Index
2	En Yakın Offsite ATM (mt.)	**8**	YYY Bankası ATM Sayısı
3	Muhafazakar Oran	**9**	Foot Traffic
4	Foot Traffix Index	**10**	Toplam Nüfus
5	Katılım ATM Sayısı	**11**	Demografik Endeks
6	En Yakın Xtm (mt.)		

Tablo 4.20 GainRatioAttributeEval ile seçilen özellikler

No	Özellik	No	Özellik
1	Katılım Atm Sayısı	**7**	XXX Xtm
2	Ortalama Index	**8**	Toplam Nüfus
3	En Yakın Offsite Atm (mt.)	**9**	En Yakın Şube (mt.)
4	Foot Traffix Index	**10**	XTM Sayısı
5	Muhafazakar Oran	**11**	Demog. Index
6	Banka Index		

Tablo 4.21 Weka'da seçilen özelliklerin Rastgele Orman algoritması ile modellenmesi sonucu oluşan performans sonuçları

Model	R-Kare	MSE	MAE
CfsSubsetEval	0,93	**0,09**	0,06
WrapperSubsetEvaluation	0,91	**0,12**	0,09
InfoGainAttributeEval	0,94	**0,07**	0,05
ReliefFAttributeEval	0,92	**0,1**	0,08
GainRatioAttributeEval	0,94	**0,08**	0,05

Tablo 4.22 Özellik seçim yöntemlerinde kullanılan parametreler

Yöntem	Parametreler	Parametre Değerleri
CfsSubsetEval	NumRequiredStatistics	20
	evaluationMeasure	Correlation
WrapperSubsetEval	classifier	RandomForest
	searchMethod	BestFirst
	evaluationMeasure	Cross-Validation (10
InfoGainAttributeEval	numFeatures	11
	evaluationMeasure	InformationGain
ReliefFAttributeEval	numNeighbours	10
	sampleSize	100
	evaluationMeasure	reliefFscore
GainRatioAttributeEva	numFeatures	11
	evaluationMeasure	Gain Ratio

Tüm özellik seçimi sonucunda çalıştırılan modeller, ilk fazda makine öğrenmesi model bölümünde gerçekleştirilen özellik seçim sonucunda çalıştırılan modelden daha başarılı performans sergilemiştir. Özellikle InfoGain ve GainRatio yöntemleri %94,04 R-Kare skoruyla en yüksek performansları sergileyen modeller için en uygun özellikleri seçmiştir. İlk çalışmadaki performansa en yakın ve en düşük performansı sergileyen yöntem WrapperSubset olmuştur. Bu yöntemde, parametre

değeri olarak Rastgele Orman algoritmasının kullanılmasın, benzer sonuçlar üretmiş olma ihtimalini artırması muhtemeldir.

4.4 Tüm Sonuçların Karşılaştırılması

Bu bölümde, makine öğrenmesi modelleri ile yapılan performans ölçümleri sonucu en yüksek R-Kare skorunu veren Rastgele Orman algoritmasının sonuçlarıyla, GA ve Weka ile elde edilen özelliklerin tekrardan Rastgele Orman algoritmasına verilerek değerlendirilmesi sonucu oluşan performans sonuçlarının karşılaştırılması yapılmıştır. Tablo 4.22'de üç farklı çalışmaya ait R-Kare, MSE ve MAE skorları görülmektedir. Bu sonuçlara göre, GA'nın bulduğu özelliklerin Rastgele Orman algoritmasıyla çalıştırılması sonucu en yüksek performans değerlerine ulaşılmıştır. Tablo 4.22'de de görüldüğü üzere, GA ile seçilen özellikler kullanılarak Rastgele Orman algoritması ile elde edilen R-Kare skoru, diğer yöntemlerle elde edilen skorlardan daha yüksektir. MSE ve MAE değerleri de benzer şekilde daha düşük çıkmıştır. Genetik Algoritma, özelliklerin etkileşimlerini daha iyi yakalayarak, modelin genel performansını artırmıştır.

Tablo 4.23 Makine öğrenmesi, Genetik Algoritma ve Weka'nın performans karşılaştırması

Model	R-Kare	MSE	MAE
Genetik Algoritma + Rastgele Orman (GA_RO)	0,97	**0,05**	0,03
Weka + Rastgele Orman (Weka_RO)	0,93	**0,12**	0,22
Rastgele Orman	0,91	**0,31**	0,43

5
SONUÇ

ATM'ler (bankamatik), bankacılık hizmetlerinin önemli bir parçası olup, müşterilere 7/24 erişim imkanı sunarak nakit para çekme, nakit yatırma, bakiye sorgulama gibi işlemleri kolay ve hızlı bir şekilde gerçekleştirme imkanı sağlar. ATM'lerin doğru konumlandırılması, bankaların hizmet kalitesini artırırken, müşteri memnuniyeti ve işlem hacmini de olumlu yönde etkilemektedir. Bu nedenle ATM lokasyonlarının hassas ve özenli bir şekilde seçilmesi, bankalar için stratejik bir önem taşımaktadır. Bu tez çalışmasında, makine öğrenmesi, Genetik Algoritmalar ve Weka uygulaması kullanılarak ATM yerleşim yerlerinin tahmin edilmesi sağlanmıştır.

İlk aşamada, literatür taraması detaylı bir şekilde yapılmıştır. ATM lokasyon problem özeline inmeden önce, literatürde tüm tesisleri kapsayan lokasyon problemleri ele alınmıştır. Gerçek dünya problemi olan ATM lokasyon problemi için, gerçek bir banka veri seti kullanılmıştır. Ham veri seti 89 nümerik ve 8 kategorik olmak üzere toplam 97 özellikten oluşmaktadır. İlk analizler, veri setinin genel özelliklerini ve istatistiksel dağılımlarını ortaya koyarak, verilerin temizlenmesi ve işlenmesi sürecini içermektedir. Makine öğrenmesi çalışmaları kapsamında, Python dilinde, üç farklı makine öğrenmesi fonksiyonuyla belirli sayıda özellikler seçilmiştir. Çalışmanın devamında, makine öğrenmesi regresyon algoritmaları ile oluşturulan modeller sayesinde, bu modellerin regresyon performansları ölçülmüştür. Bu algoritmalar arasında en başarılısı Rastgele Orman olarak belirlenmiştir. Rastgele Orman algoritmasının R-Kare skoru %91,37 olarak elde edilmiştir. Bu algoritmayı Karar Ağacı %90,07 ile takip etmiştir.

İkinci aşamada, GA ile fonksiyonlar yazılarak ve mutasyon, çaprazlama, seçim operatörleri kullanılarak en iyi özellikler seçilmiştir. Bu özellikler, birinci aşamada olduğu gibi, en iyi sonucu veren Rastgele Orman algoritması ile modellenmiştir. Bu model sonucunda %97,12 model başarısı ölçülmüştür.

Üçüncü aşamada ise Weka uygulaması kullanılarak, özellik seçimi çalışmaları yapılmıştır. Beş farklı yöntemle elde edilen özellikler, ikinci aşamada olduğu gibi

Rastgele Orman algoritmasıyla modellenerek performans değerleri elde edilmiştir. Weka'dan seçilen özelliklerin girdi olarak verildiği modellerin başarısı, GA ile yapılan çalışmanın gerisinde kalmasına rağmen, ilk aşamadaki performans değerlerini geçmiştir.

Çalışmanın bulguları, GA'nın karmaşık veri setlerinde etkili bir özellik seçimi yöntemi olduğunu göstermektedir. Genetik Algoritma'nın popülasyon tabanlı arama stratejisi, özellikler arasındaki etkileşimleri daha iyi yakalayarak model performansını artırmaktadır.

Bu çalışmanın sonuçları, makine öğrenmesi ve Genetik Algoritma tekniklerinin, veri analizinde ve karar destek sistemlerinde etkin bir şekilde kullanılabileceğini göstermektedir. Genetik Algoritma ile elde edilen yüksek performanslı modeller, bankaların stratejik karar verme süreçlerine önemli katkılar sağlayabilir.

Gelecek çalışmalarda, farklı sektörlerdeki yer seçimi ve kaynak tahsisi problemleri için Genetik Algoritma ve makine öğrenmesi yöntemlerinin uygulanabilirliği araştırılabilir. Daha büyük ve karmaşık veri setleri ile çalışarak, modelin genelleme yeteneği ve performansının değerlendirmesi yapılabilir. GA ile birlikte farklı optimizasyon yöntemleri kullanılarak hibrit özellik seçim yöntemleri geliştirilebilir ve uygulanabilir.

KAYNAKÇA

[1] F.Burelli, F.Gorelikov, M.Labianca, "Benchmarking Study 2014 and Industry Report, 2014. http://www.valuepartners.com/downloads/PDF-Communicati/2014.02-ATMIA-and-Value-Partners-ATM-Benchmarking-and-Industry-report-report.pdf

[2] Proverbio, M., Gera, P., & Burelli, F. (2016). ATM Benchmarking Study 2016 and Industry Report, 2016. https://www.accenture.com/_acnmedia/PDF-10/Accenture-Banking-ATM-Benchmarking-2016.pdf

[3] Esther M. Arkin, P. Carmi, J. Katz, S.B. Joseph, M. Segal, "Locating battery charging stations to facilitate almost shortest paths", Discrete Applied Mathematics, vol. 254, pp. 10-16, 2019.

[4] M.Basset, A.Gamal, R.K. Chakrabortty, M. Ryan, "A new hybrid multi-criteria decision-making approach for location selection of sustainable offshore wind energy stations: A case study", Journal of Cleaner Production, vol. 280, 2021.

[5] N. Cinar, S.Ahiska, "A decision support model for bank branch location selection." International Journal of Mechanical, Industrial Science and Engineering, vol. 3 (12), pp. 26-31, 2009.

[6] B. Sennaroglu, G.V. Celebi, "A military airport location selection by AHP integrated PROMETHEE and VIKOR methods." Transportation Research Part D: Transport and Environment, vol. 59, pp. 160-173, 2018.

[7] A. Arabani, R. Farahani, "Facility location dynamics: An overview of classifications and applications", Computers & Industrial Engineering, vol. 62, pp. 408-420, 2012.

[8] S. Awaghade, P. Dandekar, P. Ranade, "Site Selection and Closest Facility Analysis for Automated Teller Machine (ATM) Centers: Case Study for Aundh (Pune), India." International Journal of Advancement in Remote Sensing, GIS and Geography. vol 2, pp. 19-29, 2012.

[9] J. Zhou, Y. Wu, C. Wu, F. He, B. Zhang, F. Liu, "A geographical information system based multi-criteria decision-making approach for location analysis and evaluation of urban photovoltaic charging station: A case study in Beijing", Energy Conversion and Management, vol. 205, p.112340, 2020.

[10] M.S. Khan and N.Yesmin, "A GIS based approach to manage spatial distribution and location of financial services: a case study of ATM services" J Bangladesh Inst Plan, vol. 9, pp. 125-140, 2016.

[11] S. Lozano, J.M Garcia-Cascales, M.S. and M.T Lamata, "GIS-based onshore wind farm site selection using Fuzzy Multi-Criteria Decision Making methods. Evaluating the case of Southeastern Spain" Applied Energy, vol. 171, pp. 86-102, 2016.

[12] Kisore, N.Raghu and CH B. Koteswaraiah, "Improving ATM coverage area using density based clustering algorithm and voronoi diagrams" Information Sciences, vol. 376, pp.1-20, 2017.

[13] P.Kolesar, “Stalking the endangered CAT: A queueing analysis of congestion at automatic teller machines” Interfaces, vol. 14, pp. 16-26, 1984.

[14] O. Berman “The maximizing market size discretionary facility location problem with congestion” Socio-Economic Planning Sciences, vol. 29, pp. 39-46, 1995.

[15] A. Mansour. Aldajani, K. Alfares, “Location of banking automatic teller machines based on convolution”, Computers & Industrial Engineering, vol. 57, pp. 1194-1201, 2009.

[16] M. Erol Genevois, D. Celik, H.Z. Ulukan, “ATM Location Problem and Cash Management in Automated Teller Machines”, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering, vol. 9(7), pp. 2543-2548, 2015.

[17] Y.Li, H.Sun, C.Zhang and G.Li, “Sites Selection of ATMs Based on Particle Swarm Optimization,” 2009 International Conference on Information Technology and Computer Science, pp. 526-530, 2009, Kiev, Ukraine.

[18] K. Bilginol, H.H.Denli, D.Z.Şeker, “Ordinary Least Squares Regression Method Approach for Site Selection of Automated Teller Machines (ATMs).” Procedia Environmental Sciences, vol. 26, pp. 66-69, 2015.

[19] H.H.Yurdakul et.al, “ATM Allocation Using Decision Tree-Based Algorithms,” 2021 29th Signal Processing and Communications Applications Conference (SIU), pp. 1-4, 2021, Istanbul, Turkey.

[20] S.K. More, L.R. Gupta et.al, “Exploring the Effectiveness of Machine Learning in Facility Location Problems.” 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) , IEEE, pp. 154-158.

[21] O. Yildirim, “Scoring Neighborhoods for Locating ATM Using Machine Learning”, MEF University, pp. 24, 2018, Istanbul.

[22] M. Zeydan and S. Kayserili, “A rule-based decision support approach for site selection of Automated Teller Machines (ATMs)”, Intelligent Decision Technologies, vol. 13, pp. 161-175, 2019.

[23] D.S. Wickramasuriya, C.A. Perumalla, K. Davaslioglu and R.D. Gitlin, “Base station prediction and proactive mobility management in virtual cells using recurrent neural networks” 2017 IEEE 18th Wireless and Microwave Technology Conference (WAMICON), pp. 1-6, 2017, Florida, USA.

[24] G. Zhou, Z. Zhu and S. Luo, “Location optimization of electric vehicle charging stations: Based on cost model and genetic algorithm”, Energy, vol. 247, (123437), 2022.

[25] M. Kaveh, M. Kaveh, M. Mesgari and R.S. Paland, “Multiple Criteria Decision-Making for Hospital Location-Allocation Based on Improved Genetic Algorithm”, Appl Geomat, vol. 12, pp. 291-306, 2020.

[26] A. Alhaffa and W. Abdulal, “A Market-Based Study of Optimal ATM’s Deployment Strategy”, International Journal of Machine Learning and Computing, vol. 1, no. 1, April, 2011.

[27] M. Kulkarni, “ATM Deployment Using Rank Based Genetic Algorithm”, International Journal of Engineering Research & Technology (IJERT), 2015, Bangalore, India.

[28] Y. Yang, J. Tang, H. Luo and R. Law, “Hotel Location Evaluation: A Combination of Machine Learning Tools and Web GIS”, International of Hospitality Management, vol. 47, pp. 14-24, 2015.

[29] G. James, D. Witten, T. Hastie, R. Tibshirani and J. Taylor, “Linear Regression”, An Introduction to Statistical Learning, pp. 69-134, July, 2023.

[30] A. Callens, D. Morichon, S. Abadie, M. Delpey and B. Liquet, “Using Random Forest and Gradient Boosting Trees to Improve Wave Forecast at a Specific Location”, Applied Ocean Research, vol. 104, 2020.

[31] H.H. Yurdakul et al. “ATM Allocation Using Decision Tree-Based Algorithms,” 2021 29th Signal Processing and Communications Applications Conference (SIU), 2021, Istanbul, Turkey.

[32] J.J Ahn, H.W. Byun, K.J. Oh and T.Y. Kim, “Using Ridge Regression with Genetic Algorithm to Enhance Real Estate Appraisal Forecasting”, Expert Systems with Applications, vol. 39, pp. 8369-8379, 2012.

[33] B. Zhang, J. Geng and L. Lai, “Multiple Change-Points Estimation in Linear Regression Models via Sparse Group Lasso,” IEEE Transactions on Signal Processing, vol. 63, pp. 2209-2224, May, 2015.

[34] H. Rao, X. Shi, A.K. Rodrigue, J. Feng, Y. Xia, M. Elhonesy, X. Yuan and L. Gu, “Feature Selection Based on Artificial Bee Colony and Gradient Boosting Decision Tree”, Applied Soft Computing, vol. 74, pp. 634-642, 2019.

[35] K. Krol, T. Rymarczyk, E. Kozlowski and K. Niderla, “Using Principal Component Analysis and Elastic Net in Logistic Regression to Identify the Location of Objects in EIT”, Journal of Physics: Conference Series, vol. 2408, 2021, Jastarnia, Poland.

[36] Z. Wu, C. Li, J. Ng and K. Leung, “Location Estimation via Support Vector Regression,” IEEE Transactions on Mobile Computing, vol. 6, pp. 311-321, March, 2007.

[37] G. Morales-Espana, J. Mora-Florez and G. Carrillo-Caicedo, “A Complete Fault Location Formulation for Distribution Systems Using the k-Nearest Neighbors for Regression and Classification,” 2010 IEEE/PES Transmission and Distribution Conference and Exposition: Latin America (T&D-LA), pp. 810-815, 2010, Sao Paulo, Brazil.

[38] Z. Zhang et al. “Variable Selection in Logistic Regression Model with Genetic Algoritm”, vol. 6, Feb, 2018.

[39] A. Alhaffa, O.A. Jadaan, W. Abdulal and A. Jabas, “Rank Based Genetic Algorithm for Solving the Banking ATM’s Location Problem using Convolution”, 2011 IEEE Symposium on Computers & Informatics, pp. 6-11, 2011, Kuala Lumpur, Malaysia.

[40] M.H. Moradi and M. Abedini, “A Combination of Genetic Algorithm and Particle Swarm Optimization for Optimal DG Location and Sizing in Distribution Systems”, Internation Journal of Electrical Power & Energy Systems, vol. 34, pp. 66-74, 2012.

A
ÖZELLİKLER

Tablo A.1 Ham veri seti

Özellik	Özellik	Özellik	Özellik
ALAN_KODU	ALAN_ADI	İl Adı	İlçe Adı
İlçe Türü	Alan Kategori	Alan alt Kategori	Potansiyel
Ağırlıklı Puan	Ortalama Index	XXX Müşteri Ind.	Segment Index
Atm Index	Banka Index	Foot Traffic Ind.	Demog. Index
İlçe Türü	Toplam İşlem Sayısı	Toplam Para Çekme TL	En Yakın Şube (mt.)
En Yakın Offsite Atm (mt.)	En Yakın Xtm (mt.)	ATM SAYISI	1.Derece Konv. Atm Sayısı
2.Derece Konv. Atm Sayısı	Diğer Konv. Atm Sayısı	Katılım Atm Sayısı	Atm Puanı
Akbank	Garanti Bankası	Halk Bankası	İş Bankası
Vakıflar Bankası	YYY Bankası	Ziraat Bankası	Denizbank
Finansbank	Hsbc Bank	Ing Bank	Şekerbank
Teb	Alternatif Bank	Anadolu Bank	Eurobank Tekfen
Fortis	Odea Bank	Provus	TekstilBank
Turkish Bank	Albaraka Türk	Ziraat Katılım	XXX Bankası
XXX Xtm	Türkiye Finans	Vakıf Katılım	1000 mt. XXX Tüzel Müşteri
Toplam Nüfus	AB Ses Grubu Nüfusu	C Ses Grubu Nüfusu	DE Ses Grubu Nüfusu
Ortalama Hane Geliri	15-64 Yaş C+ Kent Nüfusu	Muhafazakâr Oran	Toplam İşyeri Sayısı
Foot Traffic	Banka Puan	Akbank.1	Garanti Bankası.1
Halk Bankası.1	İş Bankası.1	Vakıflar Bank.1	Yapı Kredi.1
Ziraat Bankası.1	Denizbank.1	Finansbank.1	Hsbc Bank.1
Ing Bank.1	Şekerbank.1	Teb.1	Alternatif Bank.1
Anadolu Bank.1	Odea Bank.1	TekstilBank.1	Turkish Bank.1
Albaraka Türk.1	Ziraat Katılım	XXX.1	Türkiye Finans.1
Vakıf Katılım.1	Puan	Rakip Atm	XXX Atm
XXX Şube Yakın	XXX Şube Uzak	XXX Atm Mesafe	XXX Xtm
Para Çekme/İşlem	Muhafazakâr		

Tablo A.2 Özelliklerin açıklamaları

Özellikler	Açıklama
Potansiyel	Veri setindeki lokasyonun, ATM kurulumuna uygunluğunu belirtir.
Ağırlıklı Puan	ATM lokasyonlarının kuruluma uygunluğunu belirten puandır. 0-100 arasında değer alır.
Ortalama Index	Özelliklerdeki diğer index değerlerinin ortalama değeridir.
XXX Müşteri Indexi	XXX bankasının müşterilerinin, veri setinde yer alan lokasyondaki yoğunluğunu belirtir.
Katılım ATM Sayısı	Lokasyona belirli bir mesafede bulunan katılım bankalarının toplam ATM sayısını belirtir.
Muhafazakar	Veri setindeki lokasyonun toplumsal olarak muhafazarlık oranını gösterir.
Foot Traffic Index	Lokasyonun yaya trafiği yoğunluğunu gösterir.
En Yakın Offsite ATM (mt.)	Lokasyona en yakın konumda bulunan şube dışı ATM'nin mesafelerini belirtir.
En Yakın XTM (mt.)	XXX bankasının XTM cihazlarının, lokasyonlara olan mesafesini belirtir.
Alan Kategori	ATM'lerin bulunduğu lokasyonların caddede mi yoksa kapalı alanda mı olduğunu belirtir.
Alan alt Kategori	ATM'lerin bulunduğu lokasyonların hastane, havaalanı, sanayi gibi yerlerde olduğunu belirtir.
1000 mt. XXX Tüzel Müşteri	XXX bankasının, lokasyona 1000 mt. mesafe içindeki tüzel müşteri yoğunluğunu gösterir.
Diğer Konv. ATM Sayısı	Lokasyon civarında, katılım bankası haricindeki konvensiyonel bankaların ATM sayısını belirtir.

TEZDEN ÜRETİLMİŞ YAYINLAR

Konferans Bildirileri

1. I.Birden and B.Diri, “Determination of ATM Locations Based on Machine Learning and Genetic Algorithm” 17th International Istanbul Scientific Research Congress on Life, Engineering, Architecture and Mathematical Sciences Proceedings Book, 28-30 April, 2024, Istanbul.