



Alınış tarihi (Received): 06.02.2023
Kabul tarihi (Accepted): 12.07.2023

Makine Öğrenmesi Teknikleri Kullanılarak Bankalardaki Potansiyel Müşterilerin Sınıflandırılması

Semih OKTAY¹, Halit BAKIR^{2,*}, Timuçin Emre TABARU³

¹Sivas Bilim ve Teknoloji Üniversitesi, Savunma Teknolojileri Bölümü, Sivas, Türkiye,

²Sivas Bilim ve Teknoloji Üniversitesi, Bilgisayar Mühendisliği Bölümü, Sivas, Türkiye,

³Sivas Bilim ve Teknoloji Üniversitesi, Elektrik-Elektronik Mühendisliği Bölümü, Sivas, Türkiye,

Sorumlu yazar: halit.bakir@sivas.edu.tr

ÖZET: Günümüzdeki bankalar kendi kampanya ve fırsatları müşterilerine bildirmek için birçok yol kullanır ve bunlardan biri de telefon görüşmeleridir. Ancak her müşteri için her kampanya ve fırsat cazip ve avantajlı olmayabilir. Bu tür bilgilendirme aramalarından dolayı belirli bir süre sonra rahatsız olan müşteriler artık bankaların aramalarına karşı duyarsızlaşmaktadır ve ilgilerini çekebilecek kampanyalara bile gereken ilgiyi göstermemektedir. Bu problemin çözümü ise her kampanya için müşterilerin verileri kullanılarak bir potansiyel müşteri grubu oluşturulmasıdır. Bu sayede hem ilgili müşterilere ulaşılacak hem de banka için daha kısa zamanda bir başarı sağlanmış olacaktır. Daha önceki çalışmalarda bu problem için yapay zekâ teknolojisi kullanılmıştır ancak kullanılan algoritmaların performansı daha önce yapılan hiçbir çalışmada optimize edilmemiştir. Özellikle, bu modellerin en iyi performansta çalışmasını sağlayabilecek hiperparametrelerin bulunması için Grid Search, Random Search ve Bayesian Search gibi model optimizasyon yöntemleri kullanılmamıştır. Bu çalışmada ise bu modellerin kıyaslaması yapılmış ve en iyi iki model olan XGBoost ve Random Forest modellerine yukarıda belirtilen model optimizasyon yöntemleri uygulanmıştır. Kullanılan veri kümesi bir bankanın verilerinden alınmıştır ve modellerin eğitiminde kullanılmadan önce temizlenip düzenlenmiştir. Aynı zamanda homojen bir veri kümesi olmadığı için de SMOTE-Tomek yöntemi kullanılarak homojen hâle getirilmiştir. Bu veri seti üzerindeki en yüksek sonuç ise %95,1 doğruluk, %94,8 kesinlik, %95,3 duyarlılık, %95,1 f1 skoru ve %99,2 AUC değeri ile XGBoost modeli olmuştur. Ardından %94,6, %92,3 kesinlik, %97,2 duyarlılık, %94,7 f1 skoru ve %99,1 AUC değeri doğruluk ile Random Forest modeli olmuştur. Daha sonrasında da optimizasyon yöntemleri uygulanmış ve sadece Random Forest modelinde %0,4'lük bir doğruluk artışı gözlemlenmiştir.

Anahtar Kelimeler – Makine öğrenmesi, banka mevduat hesabı, sınıflandırma, hiper parametre optimizasyonu

Classification of Potential Customers in Banks Using Machine Learning Techniques

ABSTRACT: Today's banks use many ways to notify their customers about their campaigns and opportunities, and one of them is phone calls. However, not every campaign and opportunity may be attractive and advantageous for every customer. Customers, who are disturbed by such information calls after a certain period of time, become insensitive to the calls of banks and do not show the necessary attention even to campaigns that may interest them. The solution to this problem is to create a potential customer group for each campaign using customer data. In this way, both relevant customers will be reached and success will be achieved in a shorter time for the bank. Artificial intelligence technology has been used for this problem in previous studies, but the performance of the used algorithms has not been optimized in any previous study. Particularly, model optimization methods such as Grid Search, Random Search, and Bayesian Search were not used to find hyperparameters that could make these models run at the best performance. In this study, these models were compared and the above-mentioned model optimization methods were applied to the two best models, XGBoost and Random Forest. The dataset used was taken from a bank's data and was cleaned and organized before being used in training models. At the same time, since there is no homogeneous data set, it has been homogenized using the SMOTE-Tomek method. The highest result on this data set was the XGBoost model with 95.1% accuracy, 94.8% precision, 95.3% recall, 95.1% f1 score and 99.2% AUC value. Then, Random Forest model became 94.6%, 92.3% precision, 97.2% recall, 94.7% f1 score and 99.1% AUC value accuracy. Afterwards, optimization methods were applied and an accuracy increase of 0.4% was observed only in the Random Forest model.

Keywords – Machine learning, bank deposit account, classification, hyperparameter optimization

1. Giriş

Son birkaç yılda bankalar vadeli mevduat ürünlerini yeni ürünlerini paylaşmakta ve onlara cazip şekilde pazarlama konusunda güçlük çekmektedir. Bunun en büyük sebeplerinden birisi bankaların ürünleriyle gerçekten ilgileyecek ve potansiyel müşteri olacak hedef kitlesini belirleyememesidir (Alexandra et al., 2021.; Ruangthong & Jaiyen, 2015). Ayrıca birçok faktörün bir müşterinin vadeli mevduat hesabı açmasında etkili olmaktadır (Raiter, 2021). Bununla birlikte daha öncesinde sürekli olarak ürün tanıtımı için aranan ve ilgisini çekmeyen ürünlere mâruz kalan müşteri, banka tarafından gelen aramaları reddetmekte ya da arandığı için şikâyet etmektedir (Elsalamony, 2014). Bu yüzden bankalar daha sonrasında oluşturacakları yeni vadeli mevduat ürünlerini potansiyel müşterilerine ulaştırmakta zorluk çekmektedirler (Das, 2015). Bankalar, bu tarz kârlarını ve şirket geleceklerini etkileyecek problemlere profesyonel çözümler bulmak, müşteri davranışları ile satın alma tercihleri hakkında bir istatistik elde etmek ve bunlar sayesinde bir satış stratejisi oluşturarak kârlarını artırmak için çok büyük müşteri verileri oluşturmuş ve yararlanmaya başlamışlardır. Bu stratejik iş girişimi için ise yapay zekâ ve büyük veri teknolojileri kullanılmaktadır (Moro et al., 2011).

Yapay zekâ, normalde insanların yapmasını gerektiren görevleri gerçekleştirmek için bilgisayar sistemlerini kullanmayı ifade eder ve genellikle birbirinin yerine kullanılsa da makine öğreniminden daha geniş bir kavram olarak kabul edilebilir. Günümüzde sayıları gittikçe artan birçok kuruluş, trend olan bu teknolojiyi pazardaki diğer rakip kuruluşlara üstünlük sağlamak için kullanmaktadır (Hou et al., n.d.). Müşterilerin bir vadeli mevduat hesabı açıp açmayacağını tahmin etmek için kullanılan çeşitli makine öğrenimi tabanlı tahmin teknikleri vardır (Moro et al., 2014). Birçok araştırmacı, makine öğrenimini, verileri akıllı bir şekilde işlemek ve geçmiş veri bilgilerinden problemlerin çözümü için işe yarayacak gizli ayrıntıların anlaşılabilmesi için en efektif araştırma metodolojilerinden biri olarak görmektedir (Elsalamony, 2014).

Bu makalede, müşterilerin banka vadeli mevduat hesabı açıp açmayacağını tahmin etmeye yönelik tahmin modelleri geliştirmek için makine öğrenimi teknikleri kullanılmıştır. Bunun için bu zamana kadar en gelişmiş ve en iyi sonuçları veren teknikleri deneyip sonuçları birbirleriyle kıyaslanmıştır. Bunun sonucunda ise bu problem üzerinde en iyi sonucu veren teknik tespit edilmiştir. Kullanılacak olan veri seti 45211 veriden oluşmuş ve 16 değişkene bağlıdır.

2. Literatür Taraması

Daha iyi bir çözüm elde etmeye çalıştığımız problemle alakalı literatürde birçok çalışma bulunmaktadır. Bu çalışmaların literatüre katkıları sayesinde gün geçtikçe daha iyi sonuçlar veren çalışmalar ortaya koyulmuştur. Bu bölümde, bizim çalışmamızdaki gibi potansiyel müşterilerin sınıflandırılması problemi hakkında en güncel çözümlerin kullanıldığı çalışmalara yer verilmiştir. Alexandra ve arkadaşları (Alexandra et al., 2021.) bir Portekiz pazarlama kampanyasından toplanan gerçek verileri kullanmıştır. Bu veri seti üzerinde Decision Tree (DT), Naive Bayes (NB) ve Random Forest (RF) gibi sınıflandırma modelleri kullanmışlardır. Bunlara ek olarak K-means, K-medoids, and DBSCAN gibi kümeleme teknikleri de uygulanmıştır. Sonuçlarda ise sınıflandırma modelleri içinde %91,13 doğruluk değeri ile (RF) modeli en iyi sonucu verirken kümeleme teknikleri arasında en iyi sonucu veren %96,03 doğruluk değeriyle K-medoids tekniği olmuştur. Gupta ve arkadaşları (Gupta, Raghav, et al.,

2021.) çalışmalarında, bir Portekiz bankasının vadeli mevduat hesabına yatırım yapacak müşterilerin verilerini ele almıştır. Veri seti üzerinde Gaussian Naïve Bayes (GNB), RF, Light Gradient Boosting (GBM) ve Extreme Gradient Boosting (EGB) modelleri kullanılmıştır. Çalışmanın sonucunda ise GNB, RF ve GBM modelleri yaklaşık %91 gibi bir doğruluk oranı ile en iyi sonuçları veren modeller olmuştur. Ancak GBM modeli diğer iyi sonuç veren modellerden işlem süresi olarak daha avantajlı olduğu görülmüştür. Borugadda ve arkadaşlarının (Borugadda et al., 2021.) çalışması uzun vadeli mevduat hesabı açabilecek olan potansiyel müşterileri tespit etmeye yönelik bir çalışmadır. Bu tespit ve analizi yapabilmek için bir Portekiz bankacılık kurumunun doğrudan pazarlama kampanyalarına (telefon görüşmeleri) ilişkin veri seti kullanılmıştır. Çalışmada, RF, Support Vector Machine (SVM), GNB, DT ve Logistic Regression (LR) gibi çeşitli makine öğrenimi algoritmaları kullanılmıştır. Çalışmanın sonunda ise LR modelinin %92,48'lik bir doğruluk oranı ile en iyi sonucu vermiştir. Tekouabou ve arkadaşları (Stéphane et al., 2019) bir Portekiz perakende bankasının 2008'den 2013'e kadar müşterilerinin bazı özellikleri hakkında oluşturulan bir veri setini ele almıştır. Her özellik türünü ayrı ayrı önceden işleyen ve tahmin performansını optimize etmek için bunları normalleştiren yeni bir modelleme yaklaşımını tanıtmaktadır. Önerilen yaklaşımı değerlendirebilmek için en iyi bilinen 5 yaygın model ile karşılaştırılmıştır. Kullanılan modeller ise NB, LR, DT, Artificial Neural Network (ANN) ve SVM modelleridir. Çalışma sonucunda elde edilen sonuçlar, ön işleme yaklaşımlarının, bu algoritmaların performansını, önceki yazarlar tarafından aynı veri tabanı üzerinde yapılan önceki çalışmalarda elde edilen sonuçlara kıyasla önemli ölçüde iyileştirdiğini göstermektedir. Ayrıca daha da ilginç olan işleme yaklaşımlarının DT haricindeki diğer modellerde performansın artışına sebep olurken DT modelinde performans düşüklüğüne sebep olmasıdır. Ancak yine de en iyi sonucu veren normalleştirme işlemi yapılmadan kullanılan DT modelidir. Hung ve arkadaşlarının (Hung et al., 2019) yaptıkları çalışmada banka pazarlama problemine yönelik çalışmışlardır. Bunun için problemin çözümüne uygun bir veri seti bulmuşlardır ve bu veri seti üzerinde RF, DT ve Gradient Boosting (GB) kullanılmış ve sonuçlarda en yüksek doğruluk oranı veren model normalize edilmiş bir şekilde olan RF modeli olmuştur (%86). Hou ve arkadaşları (Hou et al., n.d.) yaptıkları çalışmada, bir banka müşterisinin vadeli mevduat hesabı açıp açmayacağını tahmin eden en iyi modeli oluşturmak istemişlerdir. Bunun için NB, DT, RF, SVM ve Neural Network (NN) algoritmalarını kullanmışlardır. Ancak çalışma sonucunda tüm bu modellerin iyi bir performans verdikleri ve aralarında en iyi olan tek bir model olmadığı sonucuna varmışlardır. Ghosh ve arkadaşlarının (Ghosh et al., 2018) çalışmasında, Multilayer Perception Neural Network (MLPNN), DT ve SVM kullanılmaktadır. Amaç, MLPNN, DT ve SVM tekniklerinin gerçek dünya banka mevduatı aboneliği verileri üzerindeki performansını incelemektir. Çalışma sonucundaki deneysel sonuçlar, DT modelinin MLPNN ve SVM modellerine kıyasla bu problem için daha etkili bir çözüm ortaya koyduğu görülmüştür. Frempong ve arkadaşları (Asare-Frempong et al., 2017) çalışmasında iki amaç belirlemişlerdir. Bunlardan birincisi vadeli mevduat hesabı açması potansiyel müşterileri tahmin etmek iken ikincisi ise hangi temel özelliklere sahip müşteri profilinin daha yüksek ihtimâle vadeli mevduat hesabı açtığını ortaya koyan bir kümeleme işlemidir. İlk amaç için MLPNN, DT (C4.5), LR ve RF gibi makine öğrenmesi algoritmaları denenmiş ve çalışmanın sonucunda en iyi modelin %86,80'lik bir oran ile RF modeli olduğu belirlenmiştir. İkinci amaç için ise K-means kümeleme işlemi veri seti üzerine uygulanmış ve kümeleme analizi sonucunda en olası müşterilerin, evli ve minimum eğitim seviyesi ortaöğretim olan yönetici pozisyonlarındaki kişiler olduğunu göstermiştir. Koumétio ve arkadaşlarının (Koumétio et al., 2018.) yaptığı

çalışmada, bankanın uzun vadeli mevduatlarının satışı için telefonla pazarlama hedef çağrılarının tahminini optimize etmek için yeni bir sınıflandırma tekniği sunmaktadır. Ayrıca, bu yeni oluşturulan sınıflandırma tekniği diğer iyi bilinen sınıflandırma teknikleri ile karşılaştırılmıştır (NB, DT), ANN ve SVM. Deneysel çalışmalar sonucunda ise ortaya konulan yeni teknik diğer iyi bilinen teknikler ile yaklaşık olarak aynı doğruluk oranını yakalamıştır. Ghatasheh ve arkadaşlarının (Ghatasheh et al., 2020.) yaptığı çalışmada, dengesiz verilerin etkisini hafifletmek ve hangi müşterilerin vadeli mevduat hesabı açıp açmayacağını öngörmesi için çok katmanlı bir algılayıcı (MLP) sınıflandırıcı ile derin öğrenme kullanılarak gelişmiş bir sinir ağı uygulanmıştır. Aralarında en yüksek doğruluğun, %89,98'in RF tarafından sunulduğu birkaç geleneksel makine öğrenimi sınıflandırıcısı karşılaştırılmıştır. Jin ve arkadaşları (Shao et al., n.d.) yaptıkları çalışmada, banka tarafından telefon ile çeşitli teklifleri pazarlamanın başarı oranını nasıl daha fazla yükseltilebileceğini incelenmiştir. Çalışmada, veri seti DT modeline göre işlenmiş ve daha sonra işlenmiş veriler DT modeli ile sınıflandırılmıştır. Çalışmada ayrıca Vector Machine ve NN modelleri de denenmiş ancak en iyi sonucu DT modeli vermiştir. Abu-Srhan ve arkadaşları (Abu-Srhan et al., 2019) yaptıkları çalışmada, büyük ölçüde dengesiz verilerin yüksek hızda örneklenmesinin etkisini tartışmıştır ve birleştirici hiyerarşik kümelemenin (Agglomerative Hierarchical Clustering) yüksek örneklemeden daha iyi olduğu sonucuna varmışlardır. Ayrıca NB'nin böyle bir veri kümesi için en iyi sınıflandırıcı olduğu sonucuna varmışlardır. SMOTE için farklı yüzde değerleri karşılaştırıldığında, en iyi sonuçlar yüzde 400 olarak kabul edilmiş ve bunun için en iyi doğruluk RF modelinden alınan %88,8'lik doğruluk oranıydı. Gupta ve arkadaşları (Gupta, of Artificial Intelligence Techniques in, et al., n.d.) yaptıkları çalışmada Portekiz Bankacılık Kurumundan aldıkları veri seti üzerinden potansiyel müşterilere ulaşmak için bir model oluşturmayı amaçlamışlardır. Bu problem için ise NN ve RF tekniklerini karşılaştırmışlardır. Her iki tekniğin de belirli parametreleri değiştirilerek farklı sonuçlar elde edilmiş ve deney sonuçlarında, bu problem için RF modelinin daha iyi sonuçlar verdiği görülmüştür. İlham ve arkadaşlarının (Fetah et al., 2019.) yapmış oldukları çalışmada, bir banka müşterilerinin çeşitli bilgilerinin olduğu veri seti üzerinde çeşitli sınıflandırma algoritma uygulamaları denenmiş ve karşılaştırılmıştır. Herhangi bir ön işleme işlemi yapılmayan veri setine uygulanan DT, NB, RF, K Nearest Neighbour (K-NN), SVM, NN, ve LR sınıflandırma algoritmalarından en iyi sonucu veren SVM olmuştur. Rahman ve arkadaşları (Rahman et al., 2020.) yapay zekanın bir alt dalı olan makine öğrenmesi tekniklerini kullanarak bir bankadaki müşteri kaybını tahmin etmek için bir yöntem önermişlerdir. Çalışma, müşteri davranışını analiz ederek müşteri kaybı olasılığının araştırılmasını sağlamıştır. Kullanılan KNN, SVM, DT ve RF sınıflandırıcı algoritmalarından bu sorun için en iyi sonucu veren RF modeli olmuştur. Aşağıdaki Tablo 1'de de bu bölümde benzer veri kümesine sahip çalışmaların makine öğrenme teknikleri kullanılarak aldıkları sonuçların tablo hâlinde karşılaştırılması yapılmıştır.

Yaptığımız literatür taramasına göre bu problemle alakalı olarak birçok sınıflandırma, kümeleme ve yapay sinir ağı modelleri kullanılmıştır. Ancak SMOTE-Tomek yöntemi uygulanarak iyi bilinen makine öğrenmesi metotları uygulanmamıştır. Aynı zamanda uygulanan modellerin daha efektif sonuçlar vermesi için hiperparametre değerleri ile alakalı bir çalışma bulunmamaktadır.

Bu çalışmada diğer çalışmalardan farklı olarak literatüre aşağıdaki katkıları vermiştir;

- RF, DT, NB, XGBoost ve Bagging makine öğrenimi modelleri uygulanmış ve karşılaştırılmıştır.
- Heterojen olan veri kümesini homojen hâle getirmek için SMOTE-Tomek yöntemi kullanılmıştır.
- En iyi iki modelin performansını etkileyen hiperparametrelerin probleme ve modele göre en iyi değerleri bulunmuştur
- Son olarak en iyi iki modele Grid Search, Random Search ve Bayesian Search model optimizasyon yöntemleri uygulanmıştır ve sonuçları karşılaştırılmıştır.

Tablo 1. Literatürdeki çalışmaların performans karşılaştırma tablosu
Table 1. Performance comparison table of studies in the literature

Çalışma	Kullanılan Metot	H.P. Tuning	Sonuçlar
Alexandra ve arkadaşları (Alexandra et al., 2021)	RF	-	%91,13
Gupta ve arkadaşları (Gupta, Raghav, et al., 2019)	GNB, RF ve GBM	-	~%91
Borugadda ve arkadaşları (Borugadda et al., 2021)	LR	-	%92,48
Tekouabou ve arkadaşları (Stéphane et al., 2019)	DT	-	%100
Hung ve arkadaşları (Hung et al., 2019)	RF	-	%86
Hou ve arkadaşları (Hou et al., 2022.)	RF	-	%91,89
Ghosh ve arkadaşları (Ghosh et al., 2018)	DT	-	%92,48
Frempong ve arkadaşları (Asare-Frempong et al., 2017)	RF	-	%86,80
Ghatasheh ve arkadaşları (Ghatasheh et al., 2020)	RF	-	%89,98
Abu-Srhan ve arkadaşları (Abu-Srhan et al., 2019)	RF	-	%88,8

3. Metaryal ve Metot

Günümüzde birçok banka firması doğrudan müşterilerini telefonla arayarak ulaşıyor ve onlara çeşitli kampanyalarını tanıtıyor. Bu yol, firmaların müşterileri bilgilendirmek, onları kazanmak ve vadeli mevduat hesabı açtırmak için kullanılan etkin bir yoldur. Bu sayede müşteri sahip olabileceği fırsatlardan haberdar olur ve ekonomik anlamda daha profesyonel ve kazançlı bir hizmet almak isteyebilir. Ancak aranacak her müşteri profilindeki kişi için aynı kampanya ve fırsatlar ilgi çekici olmayabiliyor. Bu yüzden hedef müşteri profilini belirleyebilmek çok önemlidir. Bizim de kullandığımız makine öğrenmesi teknikleri bu amaca hizmet etmek için kullanılmıştır.

3.1. Materyal

Yapılan çalışmada çok fazla veri setinin bulunduğu Kaggle platformundaki “Bank Marketing Campaign” veri seti kullanılmış (*Bank Marketing Campaign / Kaggle*, n.d.) ve model eğitimi bu veri seti içindeki müşteri bilgileri ile yapılmıştır. Bu veri kümesinde 45211 veri bulunmakla birlikte bu verilerin 16 farklı parametresi ve iki sınıflı bir çıktısı bulunmaktadır. Veri setinin %80’i model eğitimi, %20’si test için kullanılmıştır. Şekil 1’de kullanılan veri setinden örnek olarak birkaç veri gösterilmektedir.

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no

Şekil 1. Veri seti örneği.
Figure 1. Examples from the dataset.

3.2. Metot

3.2.1. Veri Artırma Yöntemleri

Günümüzdeki her veri seti homojen olarak oluşturulamamış ve bunun sonucunda da dengesiz veri setleri ortaya çıkmıştır (Sundarkumar et al., 2015; X. Wu & Liu, 2022). Ancak bu durum o veri setlerinin kullanılamayacağı anlamına gelmemektedir. Bu problem için veri artırma ya da veri azaltma metotları bulunmaktadır. Ancak bunların arasında en etkili olanlarından biri “SMOTE-Tomek” metodudur.

SMOTE-Tomek Veri Artırma Metodu

Veri artırmanın da veri azaltmanın da amacı veri setini dengelemektir. Ancak veri azaltılarak veri seti dengelenirse model eğitiminde çok büyük öneme sahip verilerin de kullanılmama ihtimâli de vardır. Aynı zamanda çok fazla veri artırmanın yapılması da çok fazla tekrarlanan veri olacağından modelin öğrenme değil de aşırı şekilde ezberlemesine neden olur. SMOTE-Tomek veri dengeleme yöntemi ise her iki tekniğinde pozitif yönlerini alan hibrit bir modeldir. SMOTE-Tomek tekniği, imbalanced_learn kütüphanesi kullanılarak uygulanabilen ve SMOTE veri artırma metodu ile Tomek veri azaltma metodu hibritleyen bir tekniktir. Bu tekniğin çalışma şekli, örneğin dengesiz bir D veri kümesinde birçok yeni veri oluşturabilmek için SMOTE metodu ile D’ artırılmış kümesi oluşturulur. Daha sonrasında, TomekLink metodu ile D’ veri kümesindeki veri çiftleri kaldırılır (Wang et al., 2019).

3.2.2. Sınıflandırma Algoritmaları

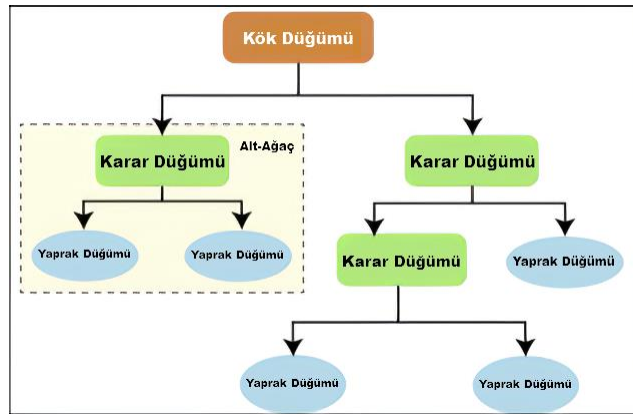
Makine öğrenmesi çok fazla problem için kullanılmaya başlayınca birçok algoritma ortaya çıkmış ve sürekli gelişmiştir. Bunun sonucunda ise birçok gelişmiş algoritma modelleri ortaya çıkmıştır. Bunlardan bazıları DT, RF, NB, XGBoost, Bagging ve Boosting modelleridir. modelleridir.

Karar Ağacı (Decision Tree - DT)

DT, Şekil 2’de de görüldüğü gibi her farklı ihtimâlin olabileceği durumda sayısal değeri olan bir özelliğin spesifik bir eşik değerine göre nasıl bir yol izleyeceğini bir dizi karar mekanizmalarını efektif ve uyumlu bir şekilde birleştiren ardışık bir modeldir. Çalışma prensibinden kaynaklı olarak gruplama amacıyla kullanılır. Bir ağacın genelden özele doğru ilerleyen bölgelerinde olduğu gibi düğümler ve dallardan oluşur. Her düğüm, sınıflandırılacak bir kategorideki özellikleri temsil eder ve her alt küme, düğüm tarafından alınabilecek bir değeri tanımlar (Charbuty et al., 2021). Yüksek hassasiyeti sebebiyle birçok uygulama alanı mevcuttur.

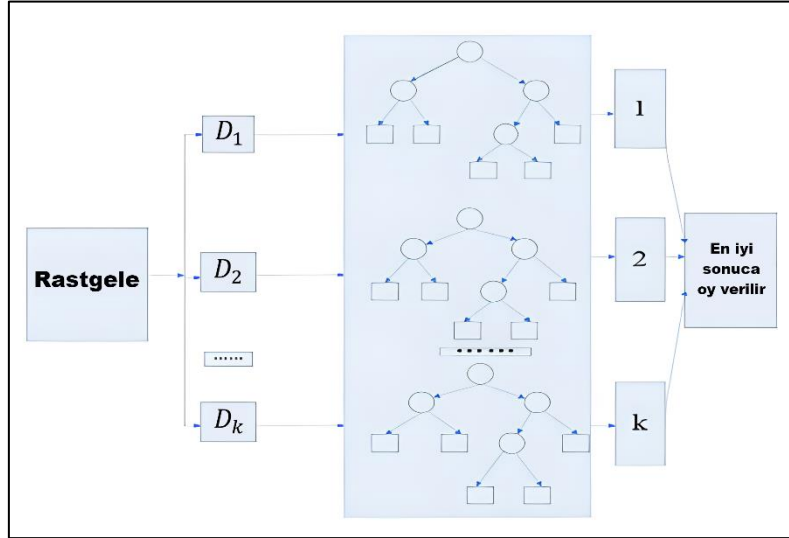
Rastgele Orman (Random Forest – RF)

RF, denetimli öğrenme tekniğine ait popüler bir makine öğrenimi algoritmasıdır. Makine öğreniminde hem sınıflandırma hem de regresyon problemleri için kullanılabilir. Komplike bir problemi sonuca kavuşturmak ve modelin performansını yükseltmek için birden fazla sınıflandırıcıyı birleştirme süreci olan topluluk öğrenmesi kavramına dayanır. Rastgele orman sınıflandırıcısı, eğitim setinin rastgele seçilen bir alt kümesinden Şekil 3’te de görüldüğü gibi birden fazla karar ağacı (DT) oluşturur ve daha sonra farklı karar ağaçlarından gelen oyları toplayarak oylamanın sonucuna göre finalde bir karar verir.



Şekil 2. DT şematiği (Charbuty et al., 2021).

Figure 2. DT schematic.



Şekil 3. RF şeması (Chen et al., n.d.).
Figure 3. RF schematic.

Naive Bayes (NB)

Bayes teoremi sınıflandırma amacıyla kullanılır ve sınıflandırmanın tahmin edici tarafından bağımsız bir şekilde işlem yaptığı varsayılır. Yani bir sınıftaki belirli bir özelliğin NB sınıflandırıcısının diğer tüm özelliklerden bağımsız olduğunu varsayarak işlem yapar. NB modeli, çok büyük veri kümeleri oluşturabilme ve veri seti hakkında daha iyi bir analiz yapılabilmesi için kullanılır. Aynı zamanda çok basit ve özel bir modeldir. Basit bir çalışma prensibi olmasından dolayı komplike problemlerde bile yüksek performans göstermektedir (Jackins et al., 2021). Bayes teoremi matematiksel olarak aşağıdaki denklemle ifade edilir:

$$P\left(\frac{a}{y}\right) = \frac{P\left(\frac{y}{a}\right)P(a)}{P(y)} \quad (1)$$

Burada a ve y birer olaydır ve $y \neq 0$ 'dır. Kısacası burada y olayının doğru olması durumunda a olayının olma olasılığı bulunmaya çalışılmaktadır.

XGBoost

XGBoost, topluluk ağaçlarına dayalı bir denetimli öğrenme algoritmasıdır. Bir kayıp fonksiyonu (d) ve bir düzenleme teriminden (β) oluşan bir maliyet amaç fonksiyonunu optimize etmeyi amaçlar:

$$\Omega(\theta) = \sum_{i=1}^n d(y_i, \hat{y}_i) + \sum_{k=1}^K \beta(f_k) \quad (2)$$

Burada, \hat{y}_i tahmin değeri, n eğitim setindeki örnek sayısı, K oluşturulacak ağaç sayısı ve f_k topluluk ağaçlarından bir ağaçtır. Düzenlilik terimi şu şekilde tanımlanır:

$$\beta(f_t) = \gamma T + \frac{1}{2} \left[\alpha \sum_{j=1}^T |c_j| + \lambda \sum_{j=1}^T c_j^2 \right] \quad (3)$$

Burada γ minimum bölünme kaybı azaltımı, λ ağırlık üzerinde bir düzenleme terimi ve c her bir yaprakla ilişkili ağırlıktır. T ise yaprak sayısıdır (Cherif et al., n.d.).

Torbalama (Bagging)

Boosting'in temel çalışma prensibi, her veri örneğine aynı başlangıç katsayısını vermek ve daha sonra ki yinelemelerde modelin o örneği doğru tahmin edip etmemesine bakılarak devamlı bir şekilde katsayıları ayarlamaktır. Gelecekteki yinelemelerde yanlış tahmin edilen eğitim örneklerinin önemini artırarak önceki yinelemelerin hatalarından öğrenir. Ancak Bagging'de ise temel öğrencilerini birbirinden bağımsız olarak eğitir ve modelin tahminlerinde çeşitliliği teşvik etmek için veri dönüşümlerini kullanırlar (González et al., 2020). Bagging, tahmin modellerinin kararlılığını ve doğruluğunu artırmak için kullanılan bir topluluk öğrenme tekniğidir. Aşağıdaki adımları içerir:

- **Örnekleme:** Orijinal veri kümesinden rastgele örnekleme yapılarak birden fazla eğitim veri kümesi oluşturulur.
- **Model Eğitimi:** Her bir veri örneği, genellikle zayıf sınıflandırıcı veya temel model olarak adlandırılan bireysel bir modeli eğitmek için kullanılır.
- **Model Bağımsızlığı:** Bireysel modeller, bağımsızlıklarını sağlamak için farklı rastgeleleştirme teknikleri kullanılarak bağımsız olarak eğitilir.
- **Paralel Eğitim:** Bireysel modeller birbirlerine bağlı olmadıkları için paralel olarak eğitilirler.
- **Tahmin Birleştirme:** Bireysel modellerin tahminleri, çoğunluk oylaması veya ortalama alma gibi toplama yöntemleri kullanılarak birleştirilir.
- **Nihai Tahmin:** Toplanan tahminler, torbalama topluluk modelinin nihai çıktısını oluşturur.

3.2.3. Model Optimizasyonu için Hiperparametreleri Fine-Tuning Etme Yöntemleri

Bir makine öğrenimi modeli, faydalı bilgileri ortaya çıkarmak ve tahminler oluşturabilmek için kullanılır. Aynı zamanda verilerden öğrenilmesi gereken bir dizi parametreye sahip olan matematiksel bir model olarak tanımlanır. Makine öğrenimi modellerinde iki farklı türde parametre bulunmaktadır. İlk parametre çeşidi model parametreleridir. Bu parametreler, oluşturulan model tarafından eğitim sırasında kullanılacak verilerden tahmin edilen parametrelerdir. İkinci parametre çeşidi ise model hiperparametreleridir. Model hiperparametreleri, modelin normal eğitim sürecinde doğrudan öğrenilen parametreler değildir ve tahmin edilemezler. Model, eğitime başlamadan önce belirlenir, sabitlenir ve sonrasında da model eğitimine başlanılır. Bu parametreler model parametrelerini tahmin etmek için kullanılırlar ve model performansında hayati öneme sahiptirler (Bakır et al., 2023; BAKIR et al., 2023; Bakır & Bakır, 2023; Bakır & Elmabruk, 2023; Bayırbağ & Bakır, 2023; Demircioğlu et al., 2023; Demircioğlu & Bakır, 2023; Doğan & BAKIR, 2023; DURAN et al., 2023; SUNGUR & BAKIR, 2023). Bu yüzden hiperparametreleri ayarlamak için birden fazla Python kütüphanesi ve yöntem

bulunmaktadır. Bunlardan bazıları Grid Search, Random Search ve Bayesian hiperparametreleri ayarlama algoritmalarıdır.

Izgara Arama (Grid Search)

Grid Search, en temel hiperparametre ayarlama yöntemidir. Eğitilecek modelin hiperparametrelerinden, istenilen parameterlere ve değerlere göre oluşabilecek tüm kombinasyonlar denenerek modelin en iyi performans vereceği hiperparametreleri bulunur. En büyük dezavantajı birçok kombinasyon oluşacağı için modelde en iyi performans gösteren hiperparametreleri bulmak uzun zaman alabilir.

Rastgele Arama (Random Search)

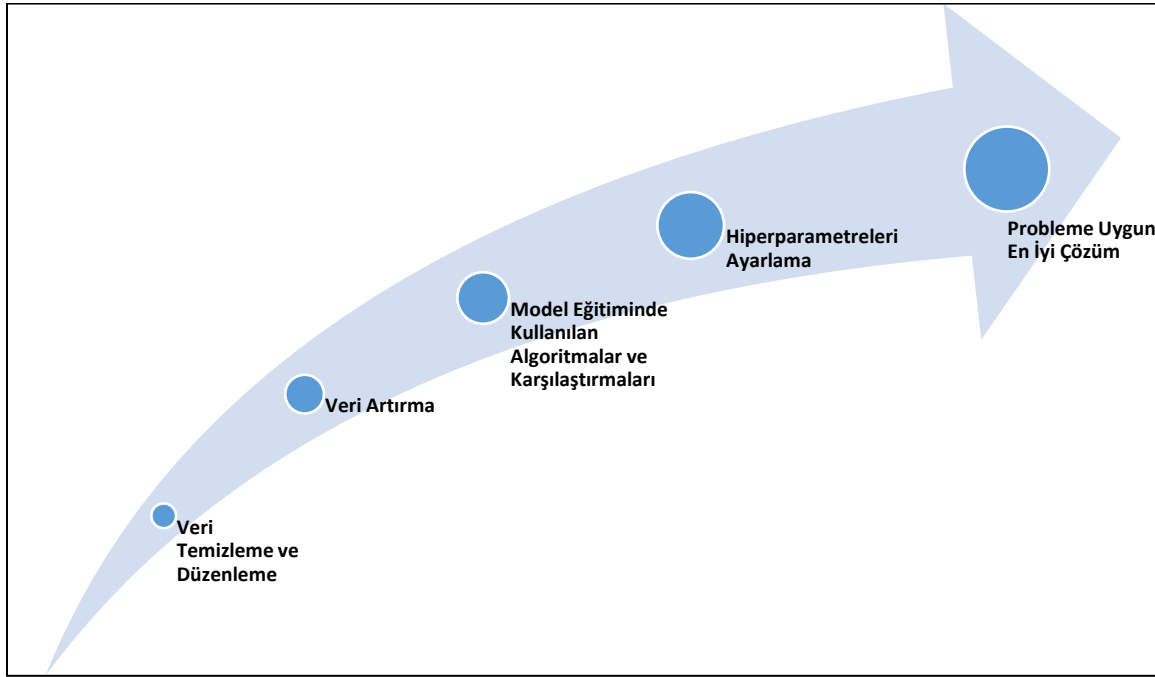
Random Search, Grid Search'e benzerdir ancak aralarındaki fark Random Search'te bütün kombinasyonların denemesi yerine adından da anlaşılacağı gibi rastgele seçilen bir alt kümedeki kombinasyonlar denenerek bu sonuçların arasındaki en iyi sonucu veren hiperparametreler bulunur. Rastgele seçilen alt küme ne kadar küçükse işlem o kadar hızlı ancak daha az doğru olur. Alt küme ne kadar büyük olursa işlem o kadar yavaş ancak daha doğru olur.

Bayes Araması (Bayesian Search)

Bayesian optimizasyon algoritması Grid Search ve Random Search algoritmalarının aksine, hiperparametrelerden Bayesian teoremi yaklaşımına göre olasılıksal bir model oluşturur. Bu model yapılan geçmiş değerlendirme sonuçlarını takip eder. Bu yüzden bir sonra ki parametre setini rastgele seçmek yerine, o zamana kadar en yüksek sonucu veren nokta neresi ise o noktaya yakın noktalarda deneme yapar. Yani geçmişteki yapılan değerlendirmelere göre bir daha ki potansiyel maksimum performans verecek noktayı bulmaya çalışır. Bu da hem süre hem de doğruluk açısından çok efektif bir metottur (J. Wu et al.,2019).

4. Deneysel Çalışmalar

Bu bölümde deneysel çalışmalar yapılırken uygulanan teknik ve metotların nasıl kullanıldığı ayrıntılı bir şekilde açıklanmış ve kullanıldıktan sonra çıkan sonuçlar gösterilmiştir. Şekil 4'te de deneysel çalışmalar boyunca yapılan işlemler temel başlıklar hâlinde gösterilmiştir.



Şekil 4. Çalışmada Yapılanları Özetleyen Akış Diyagramı.
Figure 4. Flow Chart Summarizing the conducted study.

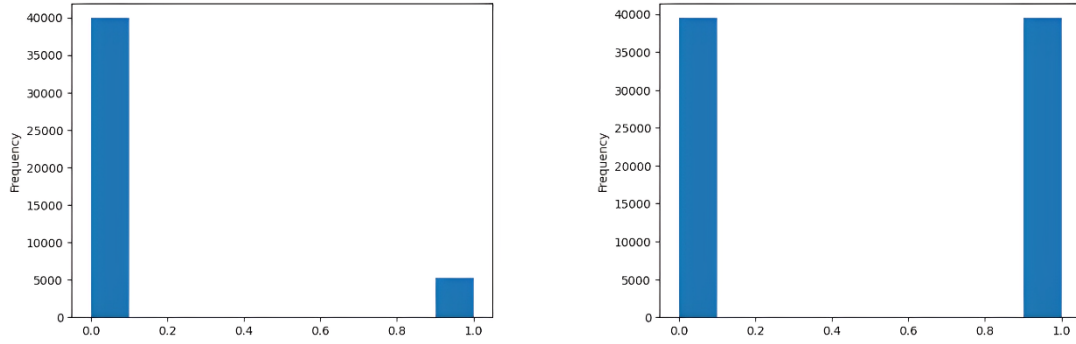
4.1 Veri Temizleme ve Düzenleme

Çalışmada öncelikle modellerin düzgün ve daha iyi bir şekilde çalışabilmesi için veri seti düzenlenmiştir. İlk olarak veri setindeki eksik ya da boş verilerin var olup olmadığına bakılmıştır ancak herhangi bir eksik ya da boş bir veri bulunmamıştır. Sonrasında tekrarlanan verilerin olup olmadığı kontrol edilmiş ancak tekrar edilen veri de bulunmamıştır. Bu kontrol de gerçekleştirildikten sonra veri setindeki özelliklerin “object” veri tipinde olanları tespit edilmiş ve bunlar “int” veri tipine dönüştürülmüştür. Bununla birlikte veri setindeki bütün değerler normalize edilmiştir.

4.2 Veri Artırma

Başlangıçtaki veri seti 45211 veriden oluşmaktadır ve bu verilerin iki sınıflı bir sonuç çıktısı bulunmaktadır. Ancak veri setine bakıldığında bu çıktılarının sayısı dengesizdir. Şekil 6’nın sol tarafta görüldüğü gibi bir sınıf 39922 tane veriden oluşurken diğer sınıf ise 5289 tane veriden oluşmaktadır.

Veri setinin dengesiz olması kullanılacak modellerin sağlıklı ve doğru bir şekilde eğitilememesine sebep olmaktadır. Bu yüzden veri setinin belirli yöntemlerle dengelenmesi gerekir. Bu çalışmada “SMOTE-Tomek” metodu kullanılarak veri dengeleme işlemi gerçekleştirilmiştir. Bunun sonucunda ise veri setimiz 78934 tane veriden oluşan bir veri seti hâline gelmiştir. Şekil 5’in sağ tarafında görüldüğü gibi yarısı birinci sınıfı, diğer yarısı ise ikinci sınıfı oluşturacak şekilde dağılmış ve veri seti dengeli bir hâle getirilmiştir. Grafiklerin x eksenini verilerin hangi sınıfa ait olduğunu gösterirken y eksenini ise belirli sınıflardaki veri sayısını göstermektedir.



Şekil 5. Veri dengeleme işleminde önceki ve sonraki veri seti dağılımı.

Figure 5. Distribution of the data set before and after the data balancing process.

4.3. Model Eğitiminde Kullanılan Algoritmalar ve Karşılaştırmaları

4.3.1. Değerlendirme metrikleri

Makine öğrenmesi teknikleri kullanılarak bir model oluşturulduğunda bunu bazı metriklere göre değerlendirmek, çalışmanın önemli bir parçasıdır. Bu metriklerin bazıları önerdiğimiz modellerin performansını incelerken çalışmamızda kullanılmıştır ve aşağıda da ne anlama geldiği açıklanmıştır.

Sınıflandırma Doğruluğu (Accuracy – Acc)

Genellikle bir modelin performansını değerlendirmek için sınıflandırma doğruluğu kullanılır ancak model hakkında detaylı bilgi edinmek istenilirse diğer metriklerin sonuçlarına da bakılması gereklidir. Bir modelin sınıflandırma doğruluğunun bulunabilmesi için, o modelin doğru tahmin sayısının toplam tahmin sayısına bölünmesi yeterlidir.

$$Acc = \frac{\text{Doğru Tahmin Sayısı}}{\text{Toplam Tahmin Sayısı}}$$

Karmaşıklık Matrisi (Confusion Matrix)

Karmaşıklık matrisi, çıktı olarak modelin performansı ile alakalı net bir bilgi sahibi olunabilecek verileri içinde bulunduran bir matristir.

Tablo 2. Karmaşıklık matrisi.

Table 2. Confusion Matrix.

	Tahmin Edilen Pozitif	Tahmin Edilen Negatif
Gerçek Pozitif	Doğru Pozitif (DP)	Yanlış Negatif (YN)
Gerçek Negatif	Yanlış Pozitif (YP)	Doğru Negatif (DN)

Tablo 2’de de görüldüğü gibi 4 tane önemli terim bulunmaktadır:

- **Doğru Pozitif (DP):** Veri setinin pozitif sınıfında olan bir verinin eğitilen model tarafından da pozitif olarak tahmin edildiği durumdur.

- **Yanlış Pozitif (YP):** Veri setinin negatif sınıfında olan bir verinin eğitilen model tarafından da pozitif olarak tahmin edildiği durumdur.
- **Doğru Negatif (DN):** Veri setinin negatif sınıfında olan bir verinin eğitilen model tarafından da negatif olarak tahmin edildiği durumdur.
- **Yanlış Negatif (YN):** Veri setinin negatif sınıfında olan bir verinin eğitilen model tarafından da pozitif olarak tahmin edildiği durumdur.

Bu tanımlardan yola çıkarak karmaşıklık matrisinin köşegeni üzerindeki sayılar ne kadar büyükse model o kadar başarılıdır çıkarımının yapılması doğru olur.

Eğri Altındaki Alan (Area Under Curve – AUC)

AUC bir modelin değerlendirilmesinde en çok kullanılan metriklerden birtanesidir. Genellikle ikili sınıflandırma yapılan modellerin değerlendirilmesinde kullanılır. AUC, modelin sınıfları ne kadar başarılı ayırt edebildiğini anlatır. x eksenini gerçek pozitif oranı (hassasiyet), y eksenini ise yanlış pozitif oranını temsil eder. ROC eğrisinin altındaki alan ne kadar büyürse, modelin sınıflar arasındaki ayırt etme performansı da o kadar artmaktadır.

$$\text{Hassasiyet} = \frac{DP}{YN + DP}$$

$$\text{Yanlış Pozitif Oranı} = \frac{YP}{DN + YP}$$

Kesinlik (Precision)

Kesinlik, model tarafından doğru pozitif sayısının toplam tahmin edilen pozitif sayısına bölümü ile bulunur.

$$\text{Kesinlik} = \frac{DP}{DP + YP}$$

Duyarlılık (Recall)

Duyarlılık, sadece doğru pozitiflerin tüm doğrulara bölümüyle bulunur.

$$\text{Duyarlılık} = \frac{DP}{DP + DN}$$

F1 Puanı (F1-Score)

F1 puanı, kesinlik ve hatırlama arasındaki dengeyi bulmaya çalışır. Bu yüzden de hesaplanırken ikisi arasındaki harmonik ortalama alınır. Ayrıca F1 puanı 0 ile 1 aralığındadır ve ne kadar büyükse model o kadar başarılıdır.

$$F1 \text{ Puanı} = 2 * \frac{\text{kesinlik} * \text{duyarlılık}}{\text{kesinlik} + \text{duyarlılık}}$$

4.3.2 Araştırma Bulguları

Veri seti, temizleme, düzenleme ve dengeleme evrelerinden geçtikten sonra artık model eğitimi için hazır hâle getirilmiş ve çeşitli algoritmalarla işlenmiştir. Problemin çözümü için en iyi doğruluk oranını ve çözümü verecek olan modeli bulmak için DT, RF, NB, XGBoost ve Bagging algoritmaları kullanılmıştır. Yapılan deneyler sonucunda farklı doğruluk oranları ve değerler ortaya çıkmıştır. Tablo 1’de de görüldüğü gibi çıkan sonuçlarda bu problem için en iyi sonuç veren modeller RF ile XGBoost modelleri olmuştur.

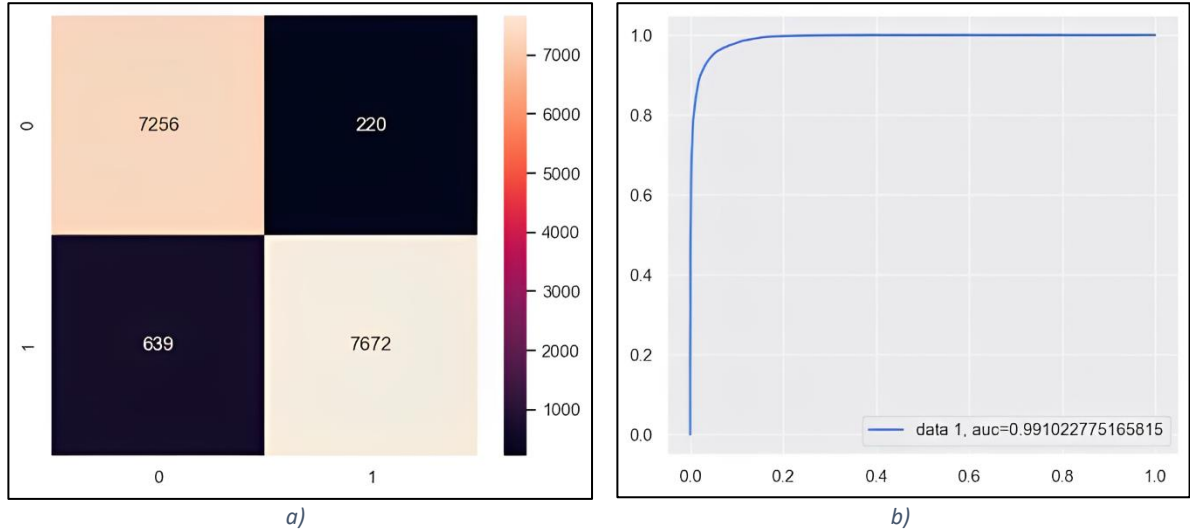
En iyi sonuç veren modelleri daha iyi analiz etmek gerekmektedir. Bunun için modellerin karmaşıklık matrisine ve ROC grafiğine bakmak gerekir. Aşağıda en iyi sonuçları veren iki modelin karmaşıklık matrisi ve ROC grafiği verilmiştir.

Tablo 3. Model Karşılaştırma metriklerinin değerleri.

Table 3. The performance of testing models.

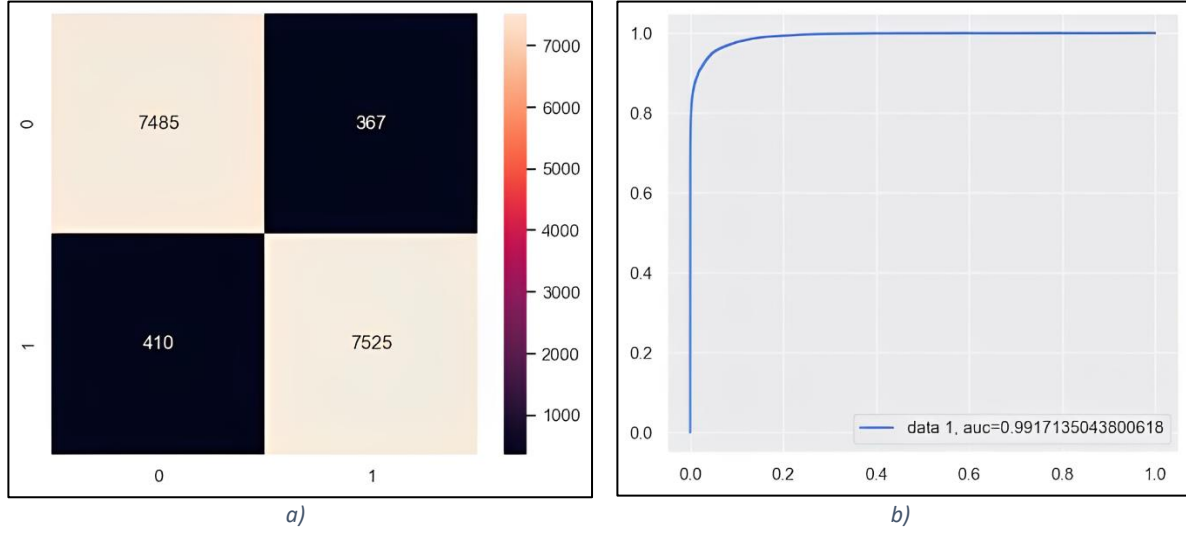
	Accuracy	Precision	Recall	F1-Score	AUC
DT	%90,5	%89,8	%91,3	%90,6	%90,5
RF	%94,6	%92,3	%97,2	%94,7	%99,1
NB	%75,9	%78,2	%71,9	%74,9	%82,5
XGBoost	%95,1	%94,8	%95,3	%95,1	%99,2
Bagging	%92,7	%92,1	%93,5	%92,8	%97,7

Şekil 6 ve Şekil 7’ye baktığımızda 0 numaralı sınıf banka teklifini kabul etmeyen müşterilerin bulunduğu sınıftan 1 numaralı sınıf ise banka teklifini kabul eden müşterilerin bulunduğu sınıftır. XGBoost modelinin 1 numaralı sınıfı, Random Forest modelinin ise 0 numaralı sınıfı daha başarılı bir şekilde sınıflandırdığı görülmektedir. Ancak toplam doğru sınıflandırmaya bakıldığında XGBoost modeli biraz daha yüksektir.



Şekil 6. Random Forest modelinin a) karmaşıklık matrisi b) ROC grafiği.

Figure 6. Random Forest model's a) complexity matrix b) ROC curve.



Şekil 7. XGBoost modelinin a) karmaşıklık matrisi b) ROC grafiği.
Figure 7. (a) Confusion matrix b) ROC curve of XGBoost model.

4.4 Hiperparametreleri Ayarlama

Çalışmada bütün algoritmaların hiperparametreleri default olarak eğitime sokulmuştur. Ancak bu parametreler değiştirilerek eğitimden alınacak sonuçlar negatif ya da pozitif yönde etkilemektedir. Bu yüzden çalışmada farklı hiperparametre ayarlama metotları en iyi sonuç elde eden iki model üzerinde denenmiş ve farklı sonuçlar alınmıştır. Kullanılan metotlar, Grid Search, Random Search ve Bayesian Optimization teknikleri kullanılmıştır. Bu tekniklerle ayarlama yapılan parametreler, bu parametrelere verilen değerler ve en iyi sonucu veren parametreler Tablo 4’de gösterilmektedir. Tablo 4’de gösterilen değerler sonucuna göre RF ve XGBoost algoritmaları tekrar eğitip test edilmiştir. Hiperparametre ayarlama işleminden sonra alınan doğruluk oranları Tablo 5’te gösterilmektedir.

Tablo 4. Hiperparametre ayarlaması yapıldığında en iyi alınan sonuçlar.

Table 4. Best results after conducting hyperparameter tuning.

Model	Kullanılan Optimizasyon Teknikleri	Ayarlama Yapılan Parametreler ve Kullanılan Değerler	En İyi Sonucu Veren Değerler
RF	Random Search	min_samples_leaf: 1, 2, 3, 4, 5 min_samples_split: 2, 3, 4, 5	n_estimators: 100, min_samples_split: 3, min_samples_leaf: 1
	Grid Search	n_estimators: 80, 100, 120, 140, 160	n_estimators: 140, min_samples_split: 2, min_samples_leaf: 1
	Bayesian		n_estimators: 200, min_samples_split: 2, min_samples_leaf: 1
XGBoost	Random Search	max_depth: 3, 4, 5, 7 booster: gbtrees, gblinear reg_alpha: 0, 0.5, 1 base_score: 0.2, 0.5, 1	subsample: 0.8 scale_pos_weight: 5 reg_lambda: 10 reg_alpha: 0.5 max_depth: 5 learning_rate: 0.1 gamma: 0

		learning_rate: 0.1, 0.01, 0.05	colsample_bytree: 0.5 booster: gbtree base_score: 0.2
	Grid Search	gamma: 0, 0.25, 1 reg_lambda: 0, 1, 10 scale_pos_weight: 1, 3, 5 subsample: 0.8 colsample_bytree: 0.5	subsample: 0.8 scale_pos_weight: 1 reg_lambda: 0 reg_alpha: 0 max_depth: 7 learning_rate: 0.1 gamma: 0 colsample_bytree: 0.5 booster: gbtree base_score: 0.2
	Bayesian		subsample: 0.8 scale_pos_weight: 1 reg_lambda: 0 reg_alpha: 0 max_depth: 7 learning_rate: 0.1 gamma: 0 colsample_bytree: 0.5 booster: gbtree base_score: 0.5

Tablo 5. Hiperparametre ayarlama teknikleri kullanıldığında modelden alınan doğruluk değerleri.

Table 5. Accuracy values taken from the model after using hyperparameter tuning techniques.

	Random Search	Grid Search	Bayesian
RF Test Score	%94,6	%94,7	%95
XGBoost Test Score	%88,6	%93,4	%94

5.Tartışma

Bu bölümde deneysel çalışmalardan elde edilen sonuçlara ilişkin tablo üzerinde tartışması yapılmıştır. Tabloda da görüldüğü gibi bu veri seti ve problem üzerinde yapılan çalışmalarda daha önce hiperparametre optimizasyonu yapılarak performansta iyileştirme çalışması yapılmamıştır. Ayrıca SMOTE-Tomek yöntemi ile dengesiz olan veriyi dengeli hâle getirilerek model eğitimi yapılmamıştır. Bu çalışmada ise bu yöntemler kullanılarak literatürdeki diğer çalışmalardan daha iyi bir performans gösterilmiştir. Sadece Tekouabou ve arkadaşları (Stéphane et al., 2019) bu çalışmadaki sonuçlardan daha iyi sonuçlar almışlardır. Ancak onlarda her bir özellik türünü ayrı ayrı ön işleme tabi tuttıkları yeni bir modelleme yaklaşımı oluşturdukları için süreç çok uzamaktadır. Ayrıca şunu belirtmekte fayda vardır ki, gerçekleştirilen çalışmalarda hiperparametre optimizasyonu her model için pozitif yönde etki sağlayamamıştır. Örneğin yapılan hiperparametre optimizasyonu işlemi RF için pozitif bir etki oluştururken XGBoost modeli için negatif bir etki oluşturmuştur. Bu yüzden bu çalışmadaki performans optimizasyon yöntemleri gelecekte daha ayrıntılı deneyler yapılarak modeller üzerindeki etkisi incelenirse daha performanslı sonuçların ortaya çıkacağını göstermektedir.

Tablo 6. Önerilen modellerin literatürlerle kıyaslanması.**Table 6.** Comparison of the proposed models with the literature.

Çalışma	Kullanılan Metot	H.P. Tuning	Sonuçlar
Alexandra ve arkadaşları (Alexandra et al.,2021)	RF	-	%91,13
Gupta ve arkadaşları (Gupta, Raghav, et al.,2021)	GNB, RF ve GBM	-	~%91
Borugadda ve arkadaşları (Borugadda et al., n.d.)	LR	-	%92,48
Tekouabou ve arkadaşları (Stéphane et al., 2019)	DT	-	%100
Hung ve arkadaşları (Hung et al., 2019)	RF	-	%86
Hou ve arkadaşları (Hou et al., 2022)	RF	-	%91,89
Ghosh ve arkadaşları (Ghosh et al., 2018)	DT	-	%92,48
Frempong ve arkadaşları (Asare-Frempong et al., 2017)	RF	-	%86,80
Ghatasheh ve arkadaşları (Ghatasheh et al., 2020)	RF	-	%89,98
Abu-Srhan ve arkadaşları (Abu-Srhan, et al., 2019)	RF	-	%88,8
Önerilen Model-RF	RF	-	%94,6
Önerilen Model-XGBoost	XGBoost	-	%95,1
Önerilen Model-RF	RF	Bayesian	%95
Önerilen Model-XGBoost	XGBoost	Bayesian	%94

6. Sonuç

Çalışmada, ilk önce problem tanımı yapılmış daha sonrasında bu problemin çözümüne yönelik makine öğrenme tekniklerinin faydalı olacağı düşünülmüştür. Bunun için öncelikle çözüme uygun bir veri seti bulunmuş ve model eğitimi için üzerinde çeşitli işlemler uygulanmıştır. Sonrasında probleme en uygun algoritmanın hangisi olacağını öğrenmek için en iyi bilinen algoritmalar işlenmiş veri ile test edilmiştir. Ortaya çıkan sonuçlarda en iyi sonucu veren modeller %95,1 ile XGBoost modeli ve %94,6 ile RF modeli olmuştur. Son olarak daha iyi bir doğruluk oranı alabilmek için hiperparametre ayarlaması yapılmıştır. Ancak bu işlem XGBoost algoritmasında negatif bir etkiye sebep olmuş, Random Forest algoritmasında ise çok düşük bir iyileştirmeye sebep olmuştur.

Gelecek çalışmalarda modellerin hiperparametreleri daha ayrıntılı şekilde taranıp daha fazla test yapılabilir. Bunun sonucunda ise modelin doğruluk oranının yükseleceği ve probleme daha iyi çözüm getirileceği düşünülmektedir. Bu yüzden derin öğrenme teknikleri de kullanılacaktır.

7. Kaynaklar

- Abu-Srhan, A., of, R. A.-S.-I. J., & undefined 2019. (2019). Visualization and Analysis in Bank Direct Marketing Prediction. *Pdfs.Semanticscholar.Org*, 10(7). <https://pdfs.semanticscholar.org/8274/71809c42b759b6433bc2b7875f8e36ec74b8.pdf>
- Alexandra, J., on, K. P. S.-2021 3rd I. C., & undefined 2021. (n.d.). Machine learning approaches for marketing campaign in portuguese banks. *Ieeexplore.Ieee.Org*. https://ieeexplore.ieee.org/abstract/document/9649623/?casa_token=nLvFZRk4bW8AAAAA:I_Q1mAB_GLY-R2huGPAT_D5Vg1u5K1xJIwxHjVWBhwMSwP3C0awAhvejWlouvoszjXoaKaB_eR-_Q
- Asare-Frempong, J., on ..., M. J.-I. C., & undefined 2017. (2017). Predicting customer response to bank direct telemarketing campaign. *Ieeexplore.Ieee.Org*. <https://doi.org/10.1109/ICE2T.2017.8215961>
- Bakır, H., & Bakır, R. (2023). DroidEncoder: Malware detection using auto-encoder based feature extractor and machine learning algorithms. *Computers and Electrical Engineering*, 110, 108804.
- Bakır, H., Çayır, A. N., & Navruz, T. S. (2023). A comprehensive experimental study for analyzing the effects of data augmentation techniques on voice classification. *Multimedia Tools and Applications*, 1–28.
- Bakır, H., & Elmabruk, K. (2023). Deep learning-based approach for detection of turbulence-induced distortions in free-space optical communication links. *Physica Scripta*, 98(6), 065521.
- Bakır, H., Oktay, S., & Tabaru, E. (2023). Detection Of Pneumonia From X-Ray Images Using Deep Learning Techniques. *Journal of Scientific Reports-A*, 052, 419–440.
- Bank Marketing Campaign | Kaggle*. (n.d.). <https://www.kaggle.com/datasets/edith2021/bank-marketing-campaign?resource=download>
- Bayırbağ, V., & Bakır, H. (2023). Çalışan Yıpranması Tahmin Etmek için Hiper Parametresi Ayarlanmış Makine Öğrenme Algoritmalarının Kullanılması. *International Conference on Scientific and Academic Research*, 1, 466–471.
- Borugadda, P., Nandru, P., Madhavaiah, C., & Theresa, S. (n.d.). Predicting the success of bank telemarketing for selling long-term deposits: An application of machine learning algorithms. *Journal.Stic.Ac.Th*, 7(1). <https://journal.stic.ac.th/index.php/sjhs/article/view/296>
- Charbuty, B., of Applied Science, A. A.-J., Technology, & undefined 2021. (2021). Classification based on decision tree algorithm for machine learning. *Jastt.Org*, 02(01), 20–28. <https://doi.org/10.38094/jastt20165>
- Chen, Y., Zheng, W., Li, W., Letters, Y. H.-P. R., & undefined 2021. (n.d.). Large group activity security risk assessment and risk early warning based on random forest algorithm. *Elsevier*. https://www.sciencedirect.com/science/article/pii/S0167865521000192?casa_token=T6mSmj0Hrk8AAAAA:FG-isbvmLJD7WG42vXbqSXkIFz9Cshc13u37qQq6cux5X5_cKk3tU8bsboGGwiDI40J2Nrb1w9U
- Cherif, I. L., (WD), A. K.-2019 W. D., & undefined 2019. (n.d.). On using extreme gradient boosting (XGBoost) machine learning algorithm for home network traffic classification. *Ieeexplore.Ieee.Org*. https://ieeexplore.ieee.org/abstract/document/8734193/?casa_token=BfU2LT1NO0sAAAAA:cbZ6gZQs5N63xxLXyEsnR4AHBpBgO4vb_mx_FOOTrakN_7Ej4i2MwvfQDeN3rRo9JWPPVmiR_DkTgg
- Das, T. K. (2015). A customer classification prediction model based on machine learning techniques. *2015 International Conference on Applied and Theoretical Computing and Communication Technology (ICATccT)*, 321–326.
- Demircioğlu, U., & Bakır, H. (2023). Deep learning-based prediction of delamination growth in composite structures: Bayesian optimization and hyperparameter refinement. *Physica Scripta*.
- Demircioğlu, U., Sayil, A., & Bakır, H. (2023). Detecting Cutout Shape and Predicting Its Location in Sandwich Structures Using Free Vibration Analysis and Tuned Machine-Learning Algorithms. *Arabian Journal for Science and Engineering*, 1–14.

- Doğan, E., & BAKIR, H. (2023). Hiperparametreleri Ayarlanmış Makine Öğrenmesi Yöntemleri Kullanılarak Ağdaki Saldırıların Tespiti. *International Conference on Pioneer and Innovative Studies, 1*, 274–286.
- Duran, A., Bakır, H., & Guzey, H. M. (2023). Bilinen CNN mimarilerinin görsel Captcha sınıflandırması açısından değerlendirilmesi. *International Conference on Pioneer and Innovative Studies, 1*, 379–385.
- Elsalamony, H. A. (2014). Bank direct marketing analysis of data mining techniques. *International Journal of Computer Applications, 85*(7), 12–22.
- Fetah, K., Tebon, P., Ilham, A., Khikmah, L., & Iswara, I. B. A. I. (n.d.). Long-term deposits prediction: a comparative framework of classification model for predict the success of bank telemarketing. *Iopscience.Iop.Org*. <https://doi.org/10.1088/1742-6596/1175/1/012035>
- Ghatasheh, N., Faris, H., AlTaharwa, I., Harb, Y., Sciences, A. H.-A., & undefined 2020. (n.d.). Business analytics in telemarketing: cost-sensitive analysis of bank campaigns using artificial neural networks. *Mdpi.Com*. <https://doi.org/10.3390/app10072581>
- Ghosh, S., Hazra, A., Choudhury, B., on Machine ..., P. B.-... C., & undefined 2018. (2018). A comparative study to the bank market prediction. *Springer*. https://link.springer.com/chapter/10.1007/978-3-319-96136-1_21
- González, S., García, S., Del Ser, J., Rokach, L., & Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion, 64*, 205–237.
- Gupta, A., of Artificial Intelligence Techniques in, G. G.-A., & undefined 2019. (n.d.). Comparative study of random forest and neural network for prediction in direct marketing. *Springer*. https://link.springer.com/chapter/10.1007/978-981-13-1822-1_37
- Gupta, A., Raghav, A., on ..., S. S.-I. C., & undefined 2021. (n.d.). Comparative study of machine learning algorithms for Portuguese bank data. *Ieeexplore.Ieee.Org*. https://ieeexplore.ieee.org/abstract/document/9397083/?casa_token=EZlp423zAcgAAAAA:E-OgdyrA5eiAtSMUp55MtpvYBHPnhtPAielgcyclGZiDVB5tPEYnxiWqHI1XqEs_D8pwq1kwzSNw
- Hou, S., Cai, Z., Wu, J., Du, H., of Business, P. X.-I. J., & undefined 2022. (n.d.). Applying Machine Learning to the Development of Prediction Models for Bank Deposit Subscription. *Igi-Global.Com, 9*(1). <https://doi.org/10.4018/IJBAN.288514>
- Hung, P. D., Hanh, T. D., & Tung, T. D. (2019). Term deposit subscription prediction using spark MLlib and ML packages. *ACM International Conference Proceeding Series, 88–93*. <https://doi.org/10.1145/3317614.3317618>
- Jackins, V., Vimal, S., Kaliappan, M., & Lee, M. Y. (2021). AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *Journal of Supercomputing, 77*(5), 5198–5219. <https://doi.org/10.1007/S11227-020-03481-X>
- Koumético, C. S. T., ... W. C.-2018 6th I., & undefined 2018. (n.d.). Optimizing the prediction of telemarketing target calls by a classification technique. *Ieeexplore.Ieee.Org*. https://ieeexplore.ieee.org/abstract/document/8629675/?casa_token=wz6gYTmH5-YAAAAA:iLYTvjFGCXGf2QM0xM_kz8w3uB3o99DNBq39Sjk1Ae7DyJgssxfbJqLrAQvIqi9o hc7FoVHhqzVjsA
- Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems, 62*, 22–31.
- Moro, S., Laureano, R., & Cortez, P. (2011). *Using data mining for bank direct marketing: An application of the crisp-dm methodology*.
- Rahman, M., on, V. K.-2020 4th I. C., & undefined 2020. (n.d.). Machine learning based customer churn prediction in banking. *Ieeexplore.Ieee.Org*. https://ieeexplore.ieee.org/abstract/document/9297529/?casa_token=x7HcTf2ZbyQAAAAA:VyWmUgpWMF8Q_zfxNAxZt9F0ki9ZFalkGvvsFQHAcEXpZTRZlqmvEzAJrKnYsm4-u0kYSeAyBFDEgA
- Raiter, O. (2021). Segmentation of bank consumers for artificial intelligence marketing. *International Journal of Contemporary Financial Issues, 1*(1), 39–54.

- Ruangthong, P., & Jaiyen, S. (2015). Bank direct marketing analysis of asymmetric information based on machine learning. *2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 93–96.
- Shao, M., Bin, G., Wu, S., Jin, W., & He, Y. (n.d.). Three data mining models to predict bank telemarketing. *Iopscience.Iop.Org*. <https://doi.org/10.1088/1757-899X/490/6/062075>
- Stéphane, C., Tekouabou, K., Hassan, S., Koumetio, S. C., Cherif, W., & Silkan, H. (2019). A data modeling approach for classification problems: application to bank telemarketing prediction. *DL.Acm.Org, Part F148154*. <https://doi.org/10.1145/3320326.3320389>
- Sundarkumar, G. G., Ravi, V., & Siddeshwar, V. (2015). One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection. *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICIC)*, 1–7.
- Sungur, F., & Bakır, H. (2023). Yapay Zekâ Teknikleri Kullanılarak IOT Cihazlarda DDos Saldırı Tespiti. *International Journal of Advanced Natural Sciences and Engineering Researches*, 7(7), 275–280.
- Wang, Z. H. E., Wu, C., Zheng, K., Niu, X., Access, X. W.-I., & undefined 2019. (n.d.). SMOTETomek-based resampling for personality recognition. *Ieeexplore.Ieee.Org*. <https://ieeexplore.ieee.org/abstract/document/8827490/>
- Wu, J., Chen, X. Y., Zhang, H., Xiong, L. D., of Electronic ..., H. L.-J., & undefined 2019. (n.d.). Hyperparameter optimization for machine learning models based on Bayesian optimization. *Elsevier*. <https://www.sciencedirect.com/science/article/pii/S1674862X19300047>
- Wu, X., & Liu, H. (2022). Application of Big Data Unbalanced Classification Algorithm in Credit Risk Analysis of Insurance Companies. *Journal of Mathematics*, 2022.