T.C. YALOVA ÜNİVERSİTESİ

MÜHENDİSLİK FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

MAKİNE ÖĞRENMESİ DERSİ

PROJE RAPORU

Hazırlayan

180101039- Muhammed Furkan Uysal

Danışman

Prof. Dr. Murat Gök

Bank Marketing Analysis

Yalova-2022

İçindekiler

1. Giriş ve Proje Hakkında

Bu proje Makine Öğrenmesi dersi için ilk olarak makine öğrenmesinin temellerini öğrenme amacıyla yapılmaktadır. Projemde kullanacağım veri seti bir sonraki başlık altında olup, veri setinden bahsetmeden önce projenin amacından bahsedeceğim.

Pazarlama satış kampanyaları, business dediğimiz işi geliştirmek için tipik bir strateji kullanır. Şirketler, belirli bir hedefe ulaşmak için doğrudan pazarlama denilen yöntem ile hedeflenen müşteri bölümleriyle iletişim kurar. Müşterilerle uzaktan etkileşimleri bir iletişim merkezinde merkezileştirmek, kampanyaların operasyonel yönetimini kolaylaştırır. Müşterilerle birçok farklı uzaktan iletişim kurma yöntemi varken, en çok kullanılanlardan biri de telefonlardır.

Bu projedeki veriler ise bir Portekiz bankacılık kurumunun doğrudan pazarlama kampanyaları (telefon görüşmeleri) ile ilgilidir. Sınıflandırmanın amacı müşterinin vadeli bir mevduata abone olup olmayacağını tahmin etmektir.

2. Veri Seti (Data Set)

Projemde veri seti olarak California Irvine Üniversitesi websitesi arşivinde bulunan 'Bank Marketing Data Set' i kullandım. Veri setinde 17 adet özellik, 45211 adet örnek vardır. Veri setinin özellikleri *'Tablo 1'* de görülebilir.

ÖZELLİK	AÇIKLAMA
Age	Yaş değeri
Job	Yaptığı iş
Marital	Medeni hâli
Education	Aldığı en yüksek eğitim
Default	Varsayılan olarak kredisi var mı?
Balance	Hesabında bulunan para
Housing	Konut kredisi var mı?
Loan	Kişisel kredisi var mı?
Contact	İletişime geçme türü
Day	En son iletişime geçilen gün
Month	En son iletişime geçilen ay
Duration	Son iletişime geçildiğindeki iletişim süresi
Campaign	Bu kampanyada kaç kere iletişime geçildiğinin sayısı
Pdays	En son iletişim kurulduktan sonra geçen gün sayısı
Previous	Bu kampanyadan önce kaç kere iletişime geçildiğinin sayısı
Poutcome	Bir önceki pazarlama kampanyasının sonucu
Y	Müşteri vadeli bir mevduat abonesi oldu mu? (Hedeflenen)

Tablo 1

3. Yöntemler

Sınıflandırma, toplanan veriden tahmin modeli oluşturmak için öğrenme teknikleridir. Bunlara ek olarak bir de performans metrikleri vardır. Performans metrikleri ise oluşturulan modelin ya da modellerin başarısını ölçen değerlerdir.

3.1. Sınıflandırma Algoritmaları

Bu projede ben 5 farklı sınıflandırma algoritması kullandım. Bunlar sırasıyla Logistic Regression, kNN, Naive Bayes, Random Forest ve SVM'dir.

3.1.1. Logistic Regression

Logistic Regression, ikili bir sonucu tahmin etmek için kullanılan bir algoritmadır. Bir şeyin cevabı ya evet ya da hayır gibi bir ikili olabilir. İki kategoriden birine düşen sonuçlarla ikili sonucu belirlemek için bağımsız değişkenler analiz edilir.

3.1.2. k-Nearest Neighbor (kNN)

kNN adından da anlaşılacağı üzere en yakın komşuya göre sınıflandırma yapar. Verilen 'k' değerine en fazla hangi veri çok yakınsa, bütün veriler o sınıfa aitmiş gibi bir değerlendirme yapılır.

3.1.3. Naive Bayes

Naive Bayes, bir veri noktasının belirli bir kategoriye ait olup olmama olasılığını hesaplar. Metin analizinde, önceden ayarlanmış bir etikete ait olan veya olmayan cümleleri kategorize etmek için kullanılabilir.

3.1.4. Random Forest

Random Forest, birden fazla karar ağacının çıktısını tek bir sonuca ulaşmak için birleştiren bir algoritmadır.

3.1.5. SVM (Support Vector Machines)

Bir SVM, verileri polarite dereceleri içinde eğitmek ve sınıflandırmak için algoritmalar kullanır ve onu X/Y tahmininin ötesine götürür. Makine öğrenimini en üst düzeye çıkarmak için en iyi düzlem, etiketle arasında en büyük mesafeye sahip olandır. SVM, çok boyutlu olduğu için daha doğru makine öğrenimine izin verir.

3.2. Performans Metrikleri

Bu projede 5 farklı performans metriği kullanıldı. Bunlar sırasıyla Accuracy, Precision, Recall, F-Score ve AUC-ROC eğrisidir.

3.2.1. Accuracy

Türkçe'ye Doğruluk olarak geçen Accuracy, modelde doğru etiketlenmiş verilerin, toplam veri havuzuna oranıdır.

$$\frac{TP + TN}{TP + FP + TN + FN}$$

3.2.2. Precision

Precision ya da Kesinlik, modelde doğru pozitif etiklenmiş verilerin, tüm pozitif verilere oranıdır.

$$\frac{TP}{TP + FP}$$

3.2.3. Recall

Duyarlılık ya da Recall, pozitif olarak tahmin edilmesi gereken işlemlerin ne kadarının pozitif olarak tahmin edildiğini gösteren orandır.

$$\frac{TP}{TP + FN}$$

3.2.4. F-Score

F-Score, precision ve recall değerlerinin harmonik ortalamasını gösterir.

$$2 \times \frac{precision \times recall}{precision + recall}$$

3.2.5. AUC-ROC Eğrisi

ROC eğrisi, tüm sınıflandırma eşiklerinde bir sınıflandırma modelinin performansını gösteren grafiktir. Bu eğri iki parametreyi (TPR, FPR) çizer. AUC ise ROC eğrisi altındaki 'alan' anlamına gelir.

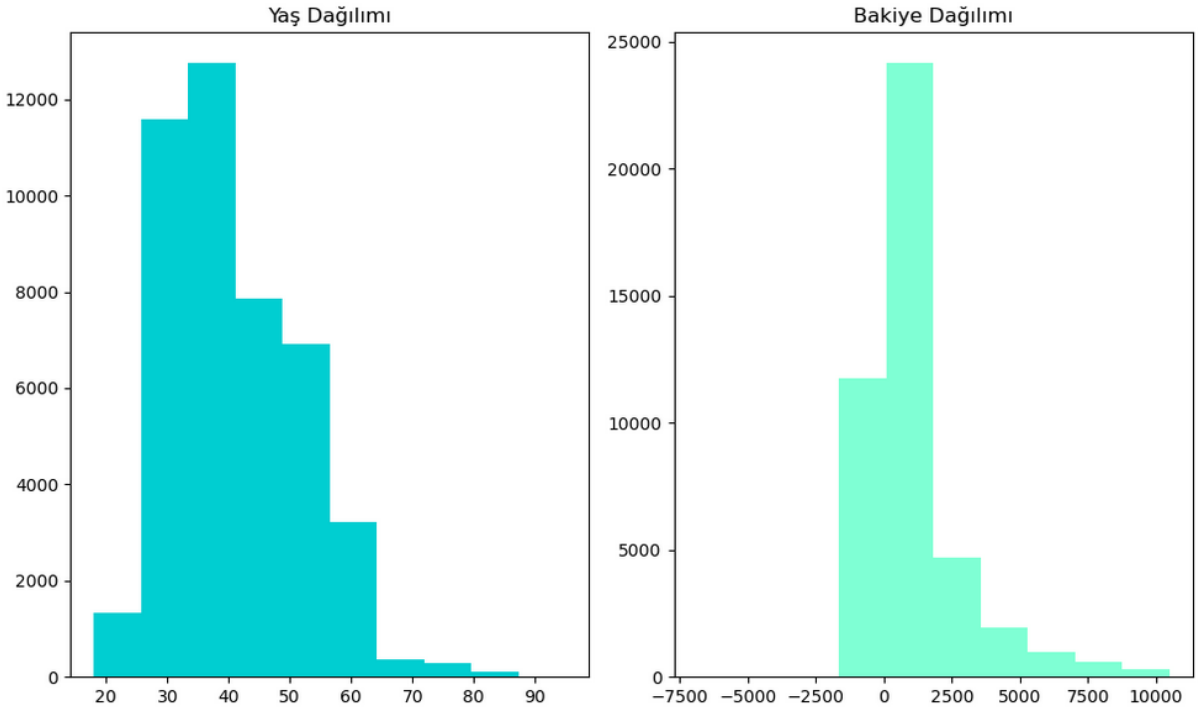
3.3. k-Fold Cross Validation

K-Fold cross validation, makine öğrenimi modellerinin becerisini tahmin etmek için kullanılan istatiksel bir yöntemdir. İşlem, belirli bir veri örneğinin bölüneceği grupların sayısını ifade eden bir 'k' adlı parametreye sahiptir. Bu nedenle prosedür k-katlı çapraz doğrulama olarak adlandırılır.

4. Veri Görselleştirme

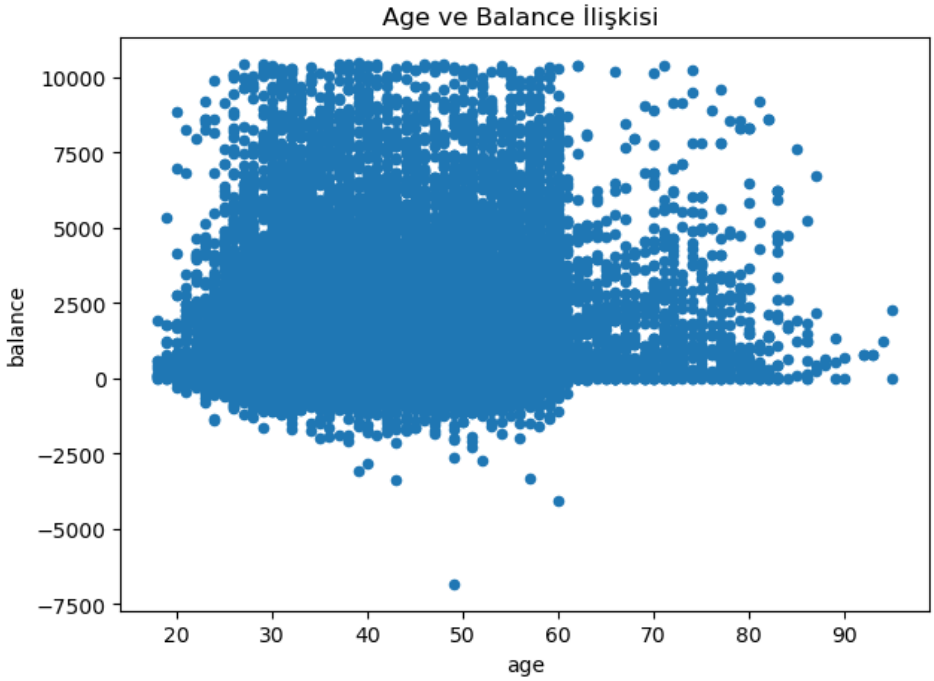
Veri üzerinde model oluşturmaya geçmeden önce birkaç adet görselleştirme yaparak veri hakkında bilgi edinelim.

- Yaş ve Bakiye Dağılımları



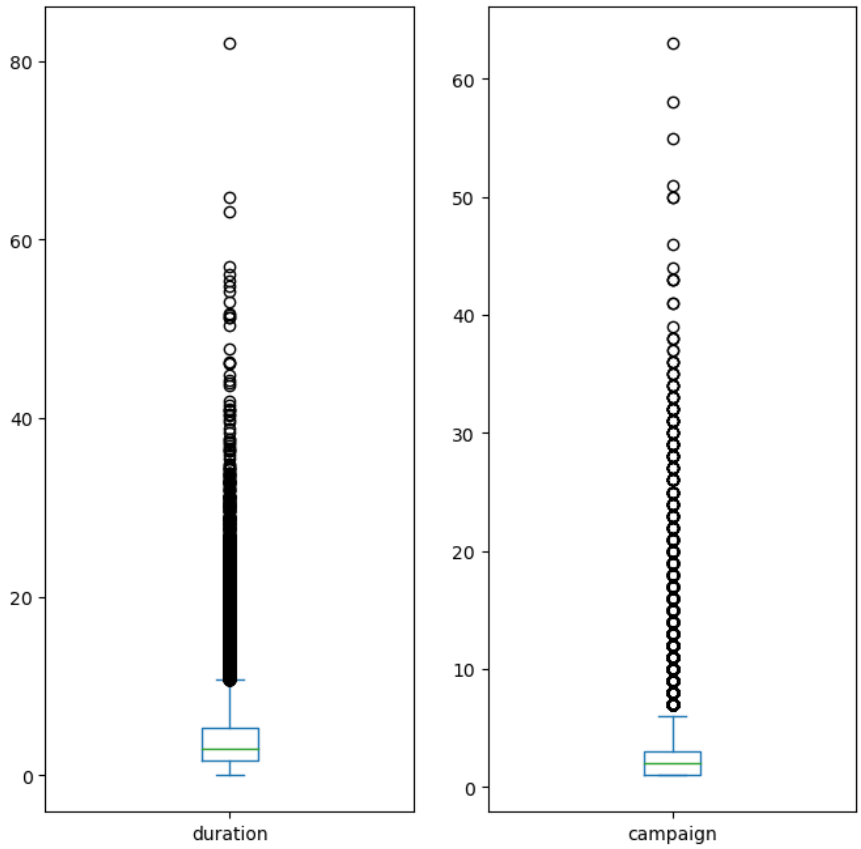
Grafik 1

- Yaş ve Bakiye İlişkisi Dağılımı



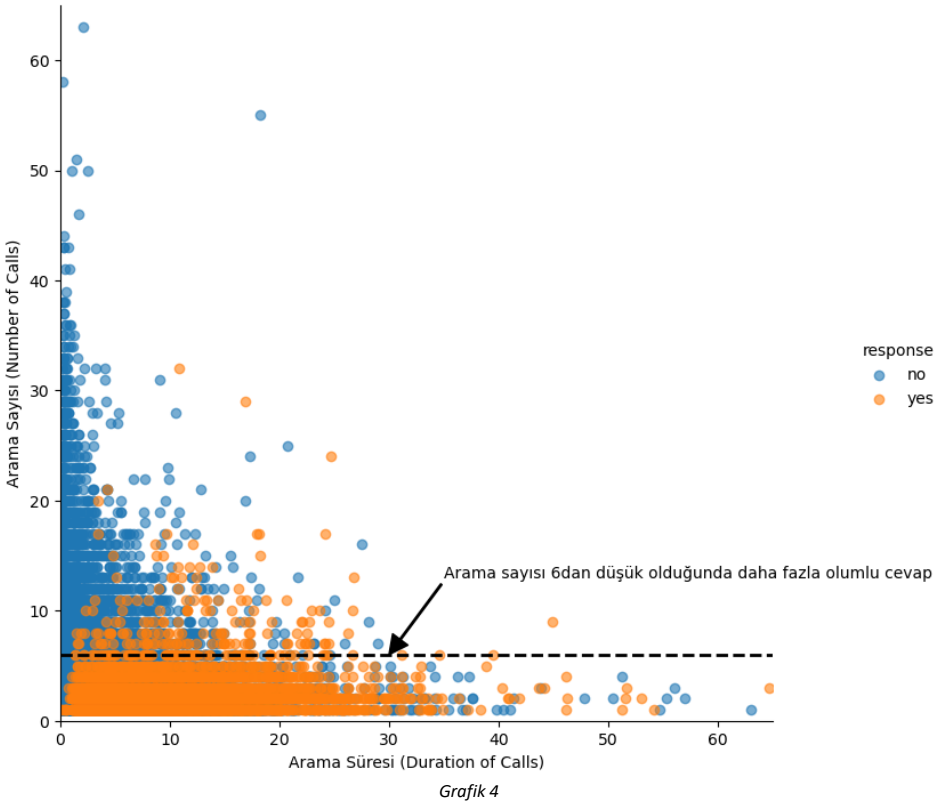
Grafik 2

- Duration ve Campaign Dağılımı



Grafik 3

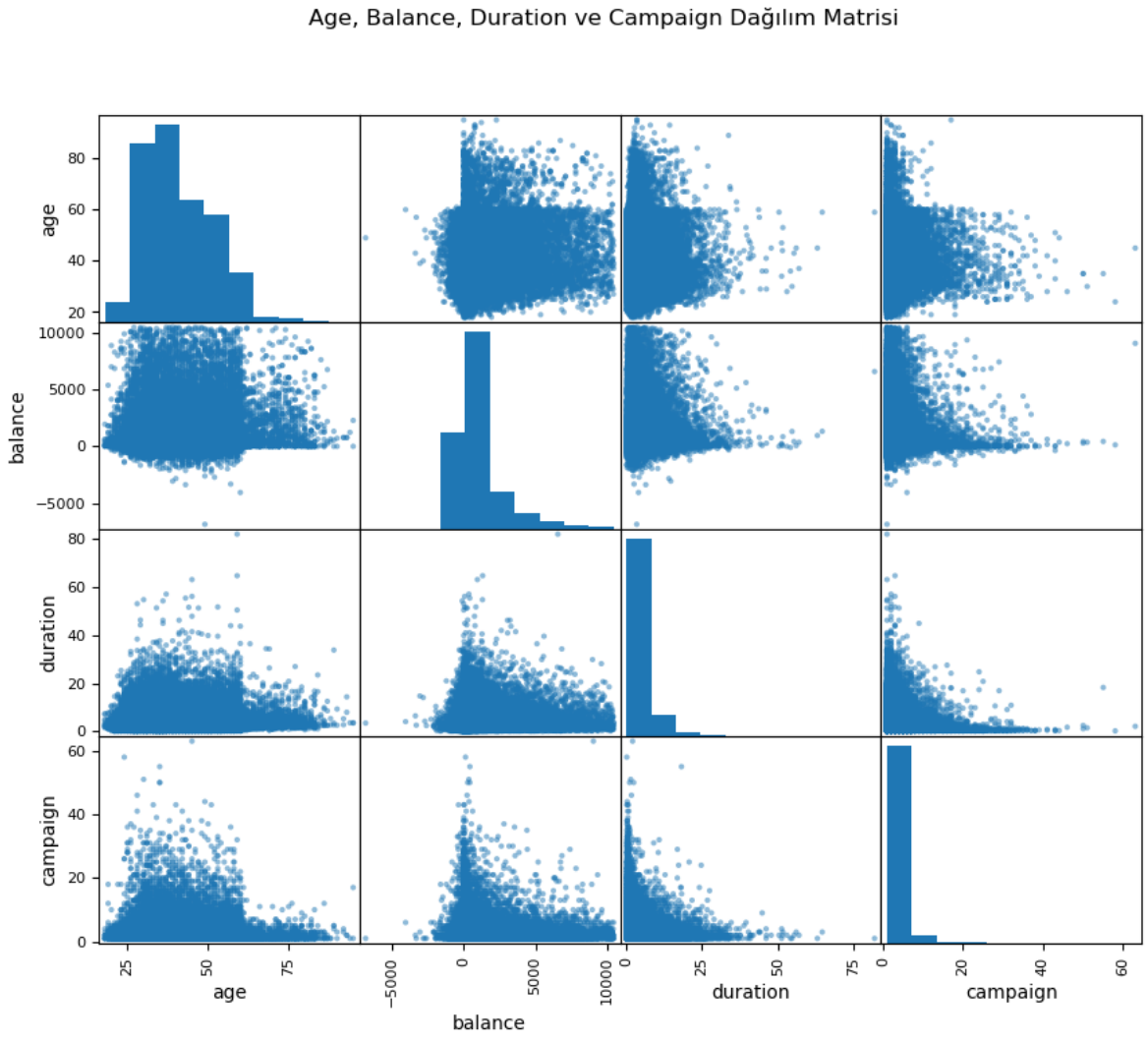
- Arama Sayısı ve Süresi Arasındaki İlişki (Müşterinin Cevabına Göre)



Grafik 4

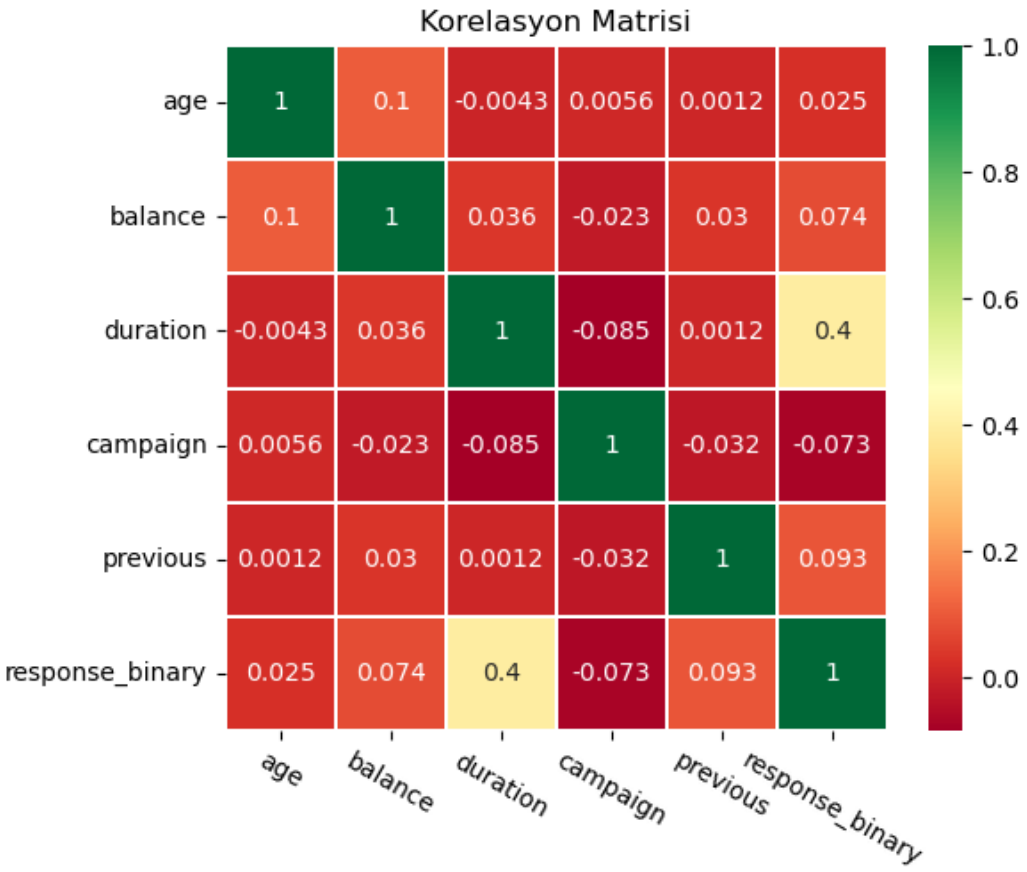
Grafik 4’te görüleceği üzere olumlu dönüşler arama sayısının 6’dan az olduğu yerlerde birikmiştir. Yani bir müşteri eğer olumlu bir cevap veriyorsa bunu genellikle ilk 6 aramada yapmaktadır. 6 aramadan sonra müşterilerin olumlu cevap verme olasılığı azalmaktadır.

- Age, Balance, Duration ve Campaign Dağılım Matrisi



Grafik 5

- Korelasyon Matrisi



Grafik 6

Korelasyon matrisine bakıldığında ise müşterinin cevabını etkileyen en yüksek oranlı değerin arama süresi olan ‘duration’ değeri olduğunu görüyoruz.

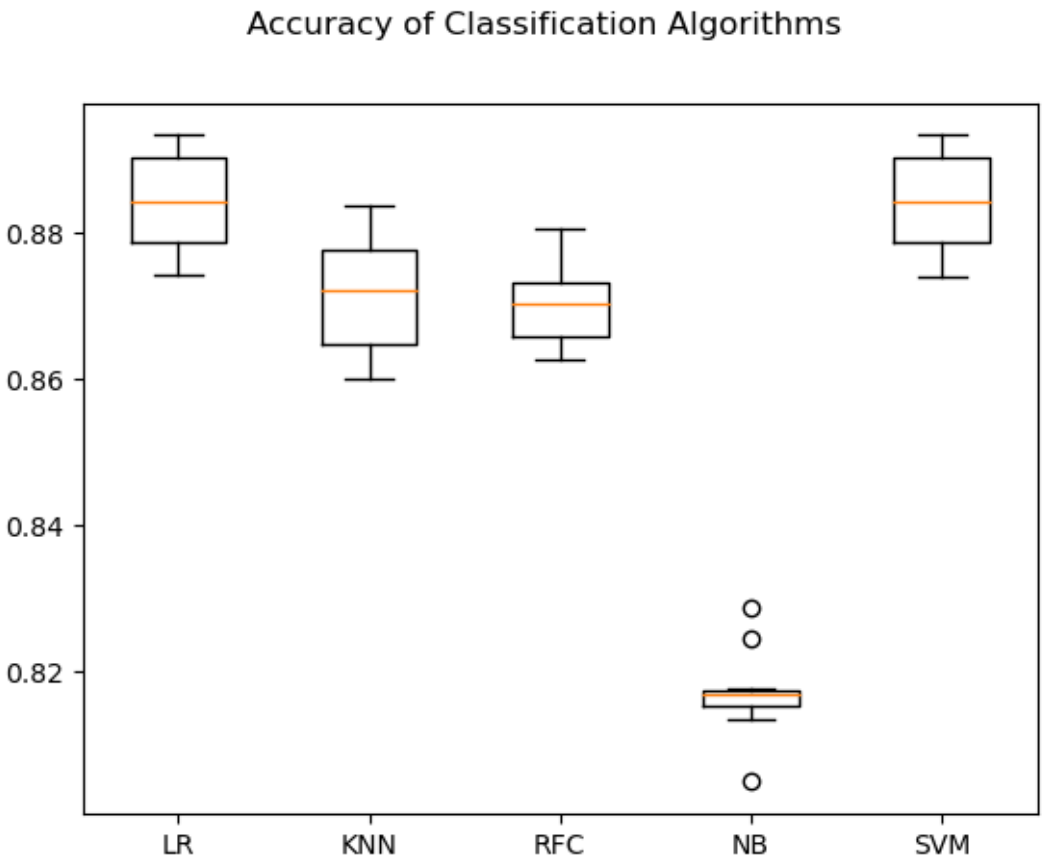
5. Sonuç

Sklearn kütüphanesinden k-Fold ve cross_val_score özellikleri kullanılarak modeller oluşturduk. Oluşturulan modellerin accuracy değerleri '*Tablo 2*'de bulunmaktadır.

Sınıflandırma Algoritması	**Accuracy Değeri**
Logistic Regression	0.884206
k-Nearest Neighbor	0.871949
Random Forest Classifier	0.869925
Naive Bayes	0.817018
Support Vector Machines (SVM)	0.884178

Tablo 2

k-Fold kullanılarak oluşturulan modellerde ortaya çıkan accuracy değerlerine göre en başarılı algoritma 'Logistic Regression' olmuştur.



Grafik 7

Diğer performans metriklerinin sonuçlarını da görmek istediğimiz için farklı bir yola başvurduk. Bir önceki modellerimizi oluştururken k-Fold ve cross_val_score kullandığımızda bize sadece istenilen bir metrik sonucunu veriyor. Diğer metrikler için tekrar bu yöntemi kullanırsak zaman olarak büyük bir sıkıntı yaşadığımızdan dolayı klasik model oluşturma yöntemini kullanıyoruz. Diğer metriklerin sonuçlarını almak için ise classification_report özelliğini kullanıyoruz.

- Logistic Regression Classification Report

	precision	recall	f1-score	support
0	0.88	1.00	0.94	7846
1	0.00	0.00	0.00	1048
accuracy			0.88	8894
macro avg	0.44	0.50	0.47	8894
weighted avg	0.78	0.88	0.83	8894

AUC: 0.500

Şekil 1

- K-Nearest Neighbor (kNN) Classification Report

	precision	recall	f1-score	support
0	0.89	0.97	0.93	7846
1	0.32	0.10	0.15	1048
accuracy			0.87	8894
macro avg	0.61	0.54	0.54	8894
weighted avg	0.82	0.87	0.84	8894

AUC: 0.535

Şekil 2

- Random Forest Classification Report

	precision	recall	f1-score	support
0	0.90	0.95	0.93	7846
1	0.40	0.23	0.29	1048
accuracy			0.87	8894
macro avg	0.65	0.59	0.61	8894
weighted avg	0.84	0.87	0.85	8894

AUC: 0.592

Şekil 3

- Naive Bayes Classification Report

	precision	recall	f1-score	support
0	0.90	0.90	0.90	7846
1	0.23	0.23	0.23	1048
accuracy			0.82	8894
macro avg	0.56	0.56	0.56	8894
weighted avg	0.82	0.82	0.82	8894

AUC: 0.562

Şekil 4

- Support Vector Machines (SVM) Classification Report

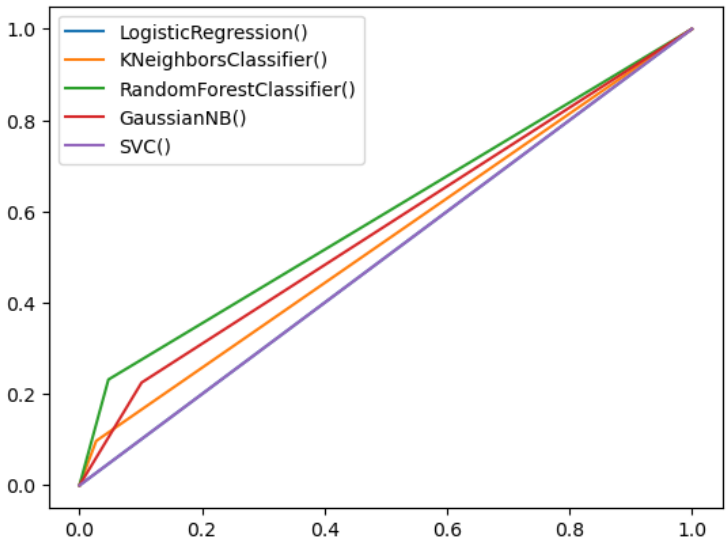
	precision	recall	f1-score	support
0	0.88	1.00	0.94	7846
1	0.00	0.00	0.00	1048
accuracy			0.88	8894
macro avg	0.44	0.50	0.47	8894
weighted avg	0.78	0.88	0.83	8894

AUC: 0.500

Şekil 5

Metriklerimize ulaştıktan sonra sırada ROC eğrilerini çizdirmek kalıyor.

'*Grafik 8*'den ya da classification raporlarına baktığımızda AUC-ROC değeri en yüksek olan algoritma 0.592 ile RFC oluyor.



Grafik 8

Kaynakça

1. https://monkeylearn.com/blog/classification-algorithms/
2. https://www.ibm.com/topics/random-forest
3. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
4. https://machinelearningmastery.com/k-fold-cross-validation/
5. https://github.com/sukanta-27/Predicting-Success-of-Bank-Telemarketing/blob/master/PDF/Bank_Marketing_Case_Study.pdf
6. https://github.com/furkanuysal/bank-marketing-ml/blob/main/PDFs/ml_homework_1.pdf
7. https://www.kaggle.com/code/yufengsui/ml-project-bank-telemarketing-analysis/
8. https://archive.ics.uci.edu/ml/datasets/Bank+Marketing