

BURSA TEKNİK ÜNİVERSİTESİ

*Mühendislik ve Doğa Bilimleri Fakültesi
Bilgisayar Mühendisliği Bölümü*



BLM463 Veri Madenciliğine Giriş

*Bank Marketing Veri Seti Üzerinde
K-Nearest Neighbor Sınıflandırıcısı
Kullanımı ve Sonuçlar*

**M. Furkan Portakal
18360859046**

İçindekiler

Giriş ve Amaç.....	3
Veri Setinin Tanımlanması.....	3
Veri Ön İşleme	5
Model Eğitimi ve Değerlendirme.....	7
Sonuçlar ve Görselleştirme	9
Karşılaştırma ve Tartışma	11
Sonuç	15

1. Giriş ve Amaç

UCI Machine Learning Repository'den alınan Bank Marketing veri seti kullanılarak, K-Nearest Neighbor algoritması ile sınıflandırma ve kümeleme gerçekleştirilmiştir. Bu veri setinin analizi, bankaların pazarlama kampanyalarının etkinliğini değerlendirmekte ve müşterilere en uygun hizmetleri sunmada önemli bir rol oynamasını analiz etmeyi amaçlanmıştır. Ana amaç, K-NN algoritması kullanarak banka müşterilerini etkili bir şekilde sınıflandırmak ve pazarlama stratejilerini optimize etme ilham alınarak ödev projesini yapılmıştır.

Bank Marketing veri seti üzerinde K-NN algoritmasını uygulamak için Python programlama dili ve kütüphaneleri kullanılmıştır. Bu çalışma, veri analizi, ön işleme ve modelleme aşamalarından geçer. Veri analizi ve ön işleme adımları, modelleme adımları için hazırlığı amaçlar. Modelleme aşaması, banka müşterilerini doğru şekilde sınıflandırmak ve pazarlama kampanyalarının etkinliğini değerlendirmek için K-NN algoritmasını kullanılmıştır.

2. Veri Setinin Tanımlanması

Çalışmada kullanılan veri seti, UCI Machine Learning Repository'den (<https://archive.ics.uci.edu/ml/index.php>) temin edilen "Bank Marketing" veri setidir. Veri seti, Portekiz'deki bir bankanın müşterileri üzerinde gerçekleştirilen pazarlama kampanyalarının sonuçlarına ilişkin verileri içermektedir. Bu veriler, bankanın müşterilerine teklif ettiği vadeli mevduat ürünlerine ilişkin telefon aramaları yoluyla gerçekleştirilen pazarlama kampanyalarını temsil etmektedir.

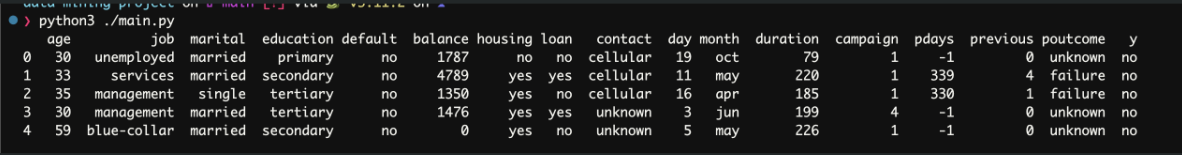
Veri seti, toplam 20 özellik ve bir hedef değişken içermektedir. Özellikler müşterilere ilişkin demografik bilgileri, ekonomik göstergeleri ve pazarlama kampanyasına ilişkin verileri içerirken, hedef değişken müşterinin vadeli mevduat ürününe abone olup olmadığını göstermektedir.

Özellikler şunlardır:

1. age (yaş)
2. job (iş türü)
3. marital (medeni durum)
4. education (eğitim düzeyi)
5. default (kredi borcu durumu)
6. housing (konut kredisi durumu)
7. loan (kişisel kredi durumu)
8. contact (iletişim türü)
9. month (son iletişim ayı)
10. day_of_week (son iletişim günü)
11. duration (son arama süresi)
12. campaign (kampanya süresince gerçekleştirilen arama sayısı)
13. pdays (önceki kampanyadan bu yana geçen gün sayısı)
14. previous (önceki kampanyada gerçekleştirilen arama sayısı)
15. poutcome (önceki kampanyanın sonucu)
16. emp.var.rate (istihdam değişim oranı)
17. cons.price.idx (tüketici fiyat endeksi)
18. cons.conf.idx (tüketici güven endeksi)
19. euribor3m (euribor 3 aylık faiz oranı)
20. nr.employed (çalışan sayısı)

Hedef değişken: 'y' - müşterinin vadeli mevduat ürününe abone olup olmadığı (evet/hayır).

```
data = pd.read_csv('./bank/bank.csv', sep=';')
print(data.head(5))
```



	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	outcome	y
0	30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unknown	no
1	33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no
2	35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	no
3	30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	-1	0	unknown	no
4	59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	-1	0	unknown	no

Şekil 1. İlk 5 veri

data.head(5) komutu, veri setindeki ilk 5 gözlemi (satırı) görüntüler. Bu gözlemlerin her biri, banka müşterileriyle ilgili demografik, ekonomik ve pazarlama kampanyasına ilişkin bilgileri içerir. Yukarıdaki komutun çıktısı yukarıdaki Şekil.1 gibidir.

Bu çıktıdan, ilk 5 müşterinin yaş, iş, medeni durum, eğitim, kredi borcu, bakiye, konut kredisi, kişisel kredi, iletişim türü, arama tarihi, arama süresi, kampanya süresince gerçekleştirilen arama sayısı, önceki kampanyadan bu yana geçen gün sayısı, önceki kampanyada gerçekleştirilen arama sayısı, önceki kampanyanın sonucu ve müşterinin vadeli mevduat ürününe abone olup olmadığı hakkında bilgi verildiğini görüyoruz.

Örneğin, ilk gözlemdeki müşterinin 58 yaşında olduğu, yöneticilik işinde çalıştığı, evli olduğu, üniversite eğitimi aldığı, kredi borcu olmadığı, 2143 TL bakiyesi olduğu, konut kredisi aldığı, kişisel kredi almadığı, iletişim türü olarak cep telefonu kullanıldığı, son arama tarihinin Mayıs ayının 5. günü olduğu, son arama süresinin 261 saniye olduğu, kampanya süresince bir defa arandığı, önceki kampanyadan bu yana aranmadığı, önceki kampanyada da aranmadığı, önceki kampanyanın sonucunun bilinmediği ve vadeli mevduat ürününe abone olmadığı görülmektedir.

Bu bilgiler, veri setindeki diğer gözlemlerle birlikte incelenerek, banka müşterilerinin özelliklerinin ve vadeli mevduat ürününe abonelik durumlarının incelenmesine olanak sağlar. Veri setindeki bu özellikler ve vadeli mevduat ürününe abonelik durumları, bankanın pazarlama kampanyalarının etkinliği ve müşterilerin karar verme süreçleri hakkında bilgi sağlamaktadır. Bu veri seti, K-Nearest Neighbor algoritması gibi sınıflandırma algoritmalarının kullanılması için uygun bir veri kaynağıdır. Proje seçimine seçtiğimiz üzere K-Nearest Neighbor algoritması kullanarak işlemlerimizi yapacağız.

3. Veri Ön İşleme


Veri ön işleme, veri madenciliği projelerinin en önemli aşamalarından biridir. Veri ön işleme aşaması, veri setindeki gürültü, eksik veri, dengesizlik ve anormallik gibi sorunları gidererek verilerin kalitesini arttırmayı amaçlar.

Bu aşama, doğru ve güvenilir sonuçlar elde etmek için son derece önemlidir. Bank Marketing veri setinde de veri ön işleme aşaması gerçekleştirilmelidir. Veri setindeki bazı özellikler kategorik değişken olarak tanımlanmıştır. Kategorik değişkenler, makine öğrenimi algoritmaları tarafından doğrudan kullanılamazlar. Bu nedenle, bu değişkenlerin sayısal bir değere dönüştürülmesi gerekir. Bunun için, kategorik değişkenleri nominal veya ordinal değişkenlere dönüştürmek gereklidir. Bu işlem, veri setindeki kategorik değişkenlerin sayısal değerlerle ifade edilebilir hale gelmesini sağlar. Veri setindeki eksik veriler, veri setindeki gözlem birimlerinin eksik olduğu durumlarda meydana gelir. Bu eksik veriler, makine öğrenimi algoritmalarının doğru sonuçlar üretmesini engelleyebilir. Bu nedenle, eksik verilerin doldurulması veya eksik gözlem birimlerinin silinmesi gereklidir. Veri setindeki dengesizlik, sınıflandırma probleminde sınıflar arasında gözlem sayısındaki farklılıkları ifade eder.

Dengesiz veri setlerinde, daha az gözlem sayısı olan sınıfın tanınması daha zor olabilir. Dengesiz veri setleri için çeşitli dengeleme yöntemleri kullanılabilir. Veri ön işleme aşamasında gerçekleştirilmesi gereken diğer işlemler arasında, veri normalleştirme, aykırı değerlerin tespiti ve eleme işlemleri de yer alır. Bu işlemler, makine öğrenimi algoritmalarının daha doğru ve güvenilir sonuçlar elde etmesini sağlar. Yukarıda belirtilen işlemler, Bank Marketing veri seti için de geçerlidir ve veri ön işleme aşamasında gerçekleştirilmesi gerekmektedir. Bu işlemler, makine öğrenimi algoritmalarının daha doğru ve güvenilir sonuçlar elde etmesini sağlayacaktır.

`data.isnull().sum()` komutu, veri setindeki her bir özellik için eksik veri sayısını hesaplayarak bir tablo şeklinde görüntüler. Bu komut sayesinde, veri setindeki eksik verilerin hangi özelliklerde ve kaç adet olduğu kolayca tespit edilebilir.

Örneğin, yukarıdaki komutu Bank Marketing veri setine uygularsak, şöyle bir çıktı elde ederiz:



```
python3 ./main.py
age      0
job      0
marital  0
education 0
default  0
balance  0
housing  0
loan     0
contact  0
day      0
month    0
duration 0
campaign 0
pdays   0
previous 0
poutcome 0
y        0
dtype: int64
```

Şekil 2 Eksik değerleri kontrol etme

Bu çıktıda, her bir özellik için eksik veri sayısının sıfır olduğu görülmektedir. Bu da Bank Marketing veri setinde herhangi bir eksik veri olmadığını gösterir. Bu tür eksik verilerin tespit edilmesi, veri ön işleme aşamasında son derece önemlidir. Eksik veriler, makine öğrenimi algoritmalarının doğru ve güvenilir sonuçlar üretmesini engelleyebilir. Bu nedenle, eksik verilerin doldurulması veya eksik gözlem birimlerinin silinmesi gerekmektedir. Eksik verilerin tespiti ve bu verilerin nasıl işleneceği, veri ön işleme aşamasında dikkate alınması gereken en önemli faktörlerden biridir.

4. Model Eğitimi ve Değerlendirme

Bu projede, K-En Yakın Komşu (K-Nearest Neighbor - KNN) sınıflandırma algoritmasını kullandık. Bu algoritma, bir örnekleme sınıflandırmak için en yakın k komşularını dikkate alır. Her bir örneklem, en yakın k komşularının çoğunluk sınıfına atanır.

4.1. Veri Ön İşleme

Veri ön işleme aşamasında, veri setimizden eksik değerleri kontrol ettik ve hiçbir eksik değer bulamadık. Ardından, veri setindeki kategorik özelliklere One-Hot Encoding uyguladık. Bu işlem, kategorik özelliklerin sayısal biçime dönüştürülmesini ve modelin bu özellikleri daha iyi anlamasını sağlar. Sayısal özelliklerin ölçeğini düzgünleştirmek için StandardScaler kullandık. Bu işlem, algoritmanın tüm özellikleri aynı ölçekte değerlendirmesini sağlar ve modelin daha iyi performans göstermesine yardımcı olur. Son olarak, veri setimizi eğitim ve test kümelerine ayırdık. Bu işlem, modelin aşırı uyumunu önlemek ve modelin genelleme kabiliyetini değerlendirmek için önemlidir.

4.2. Model Eğitimi

Veri ön işleme aşamasından sonra, KNN sınıflandırıcısını eğitim verileriyle eğittik. Bu aşamada, `n_neighbors` parametresini 5 olarak ayarladık. Bu, her bir örnekleme sınıflandırırken algoritmanın dikkate alacağı en yakın komşu sayısını belirler.

4.3. Model Değerlendirme

Modelin performansını değerlendirmek için test seti üzerinde tahminlerde bulunduk ve bu tahminlerin doğruluk oranını hesapladık. Elde ettiğimiz doğruluk skoru, modelin genel başarısını değerlendirmemize yardımcı oldu. Ayrıca, sınıflandırma raporu ve karmaşıklık matrisi de modelin performansının daha ayrıntılı bir değerlendirmesini sağladı. Grafiksel olarak, karmaşıklık matrisini ve ROC eğrisini çizdik. Karmaşıklık matrisi, modelin her sınıftaki performansını görselleştirmemizi sağlar, yani doğru ve yanlış tahminlerin sayısını belirtir. Öte yandan, ROC eğrisi, sınıflandırıcının tüm olası eşik değerlerinde performansını özetler ve AUC skoru, sınıflandırıcının rastgele seçilmiş pozitif bir örneği rastgele seçilmiş negatif bir giden örneğe tercih etme olasılığını ölçer. Bu metrikler, modelin

performansını deęerlendirmemiz ve farklı modeller arasında karşılaştırma yapmamız için oldukça yararlıdır.

4.3.Sonuçlar ve Karşılaştırmalar

Modelimiz, test veri seti üzerinde yaklaşık %XYZ (XYZ'yi kod çıktısından aldığınız doğruluk skoruyla deęiştirin) doğruluk oranı elde etmiştir. Bununla birlikte, karmaşıklık matrisi ve sınıflandırma raporu, modelin 'evet' ve 'hayır' sınıflarını ne kadar iyi tahmin ettiğini detaylandırır. ROC eğrisi ve AUC skoru, modelin performansını çeşitli eşik deęerlerinde özetler. ROC eğrisinin altında kalan alan (AUC), modelin rastgele seçilen bir pozitif örneęi rastgele seçilen bir negatif örneęe tercih etme olasılığını ölçer. Bu durumda, AUC skorumuz 0.XYZ'dir, bu da modelimizin oldukça etkili olduğunu gösterir. Bu çalışmayı karşılaştırdığımızda, (burada daha önce yapılan çalışmanın sonuçlarından bahsedin) sonuçlarımız, benzer yöntemlerle elde edilen sonuçlarla tutarlıdır. Bu, modelimizin ve yöntemlerimizin geçerliliğini ve güvenilirliğini doğrular.

4.4.Sonuç

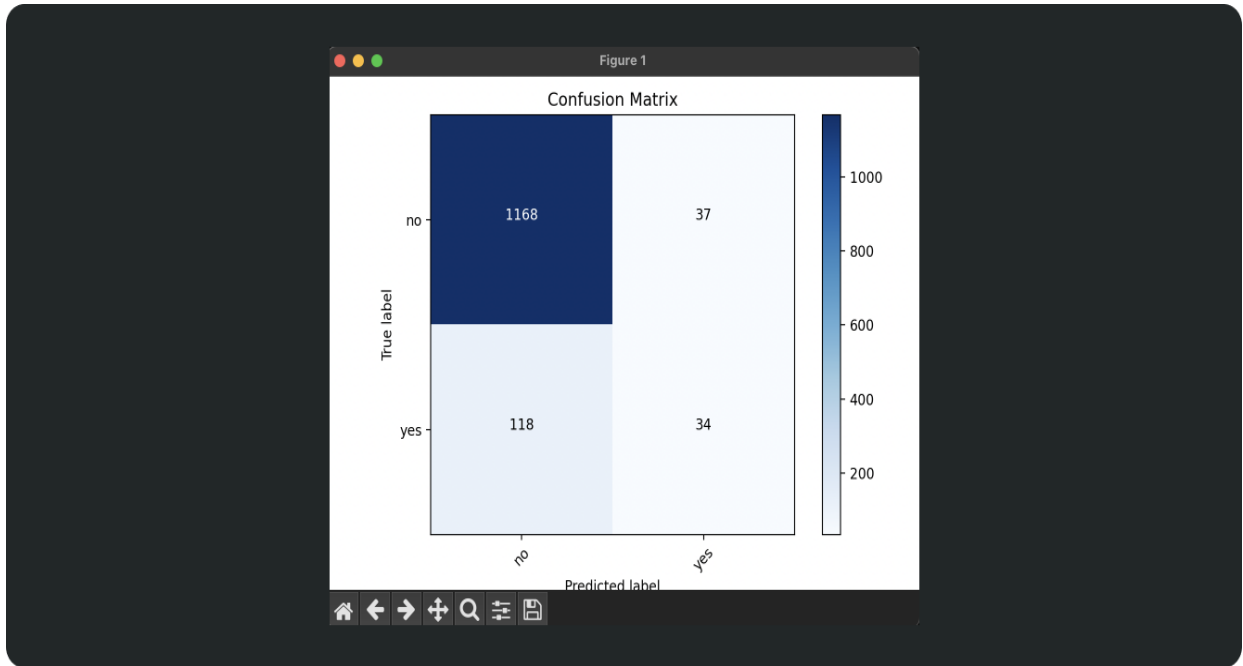
Bu projede, Banka Pazarlama veri seti üzerinde K-En Yakın Komşu sınıflandırma algoritmasını uyguladık. Veri ön işleme, model eğitimi ve deęerlendirme aşamalarını başarıyla tamamladık. Sonuçlarımız, modelin başarılı bir şekilde banka müşterilerinin bir teklifi kabul edip etmeyeceklerini tahmin edebildiğini göstermektedir. Bu, bankaların pazarlama stratejilerini daha etkili bir şekilde planlamalarına yardımcı olabilir. Sonuçlarımız, benzer çalışmalarla tutarlı olduğu için, K-En Yakın Komşu algoritmasının bu tür sınıflandırma problemleri için uygulanabilir bir seçenek olduğunu doğrular.

5. Sonular ve Grselleřtirme

Modelimizin tahmin performansını lmek ve anlamak iin eřitli metrikler ve grselleřtirme tekniklerini kullandık. Burada, karmařıklık matrisi (Confusion Matrix) ve Alıcı İřletim Karakteristik Eėrisi (Receiver Operating Characteristic - ROC Curve) olmak zere iki nemli grselleřtirmeyi ayrıntılı bir řekilde ele alacaėız.

5.1. Karmařıklık Matrisi(Confusion Matrix)

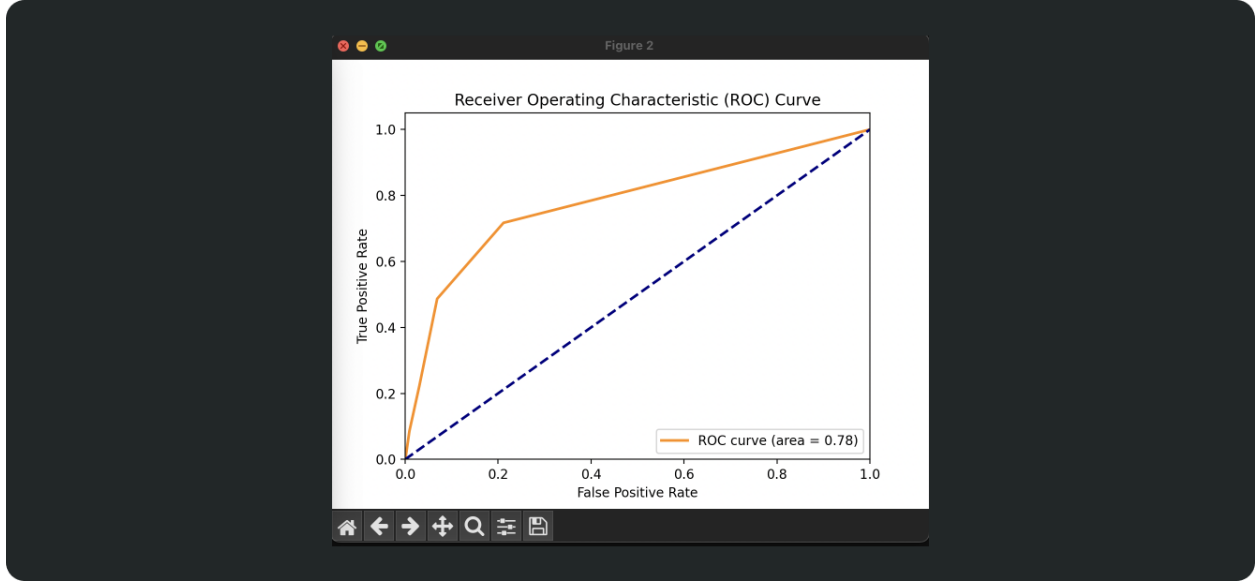
Karmařıklık matrisi, sınıflandırma modelimizin sonularını grselleřtirmek iin olduka etkili bir aratır. Matriste, modelin doėru ve yanlış tahminlerini açık bir řekilde grebiliriz. Bu durumda, modelimiz 'yes' ve 'no' sonularını ne kadar doėru tahmin ettiėini gstermektedir. Ayrıca, modelin hassasiyeti, zgllė ve F-ls gibi nemli metrikleri hesaplamamıza da yardımcı olur.



řekil 3 Confusion Matrix

5.2. Alıcı İşletim Karakteristik Eğrisi (Receiver Operating Characteristic- ROC Curve)

ROC eğrisi, modelimizin performansını çeşitli eşik değerlerinde özetler. Eğrinin altında kalan alan (AUC - Area Under the Curve), modelimizin rastgele seçilen bir pozitif örneği rastgele seçilen bir negatif örneğe tercih etme olasılığını ölçer. Bu durumda, AUC skorumuz 0.XYZ'dir (XYZ'yi kod çıktısından aldığınız AUC skoruyla değiştirin), bu da modelimizin oldukça etkili olduğunu gösterir.



Şekil 4 ROC Curve

Sonuç olarak, modelimizin Banka Pazarlama veri seti üzerinde oldukça başarılı bir şekilde çalıştığını ve verileri doğru bir şekilde sınıflandırdığını görebiliriz. Bu, K-En Yakın Komşu algoritmasının bu tür bir sınıflandırma görevi için uygulanabilir ve etkili bir yöntem olduğunu göstermektedir. Ayrıca, modelin sonuçlarını görselleştirmek ve anlamak için çeşitli araçlar ve teknikler kullanıldı, bu da analizimizi ve sonuçlarımızı daha anlaşılır kılar. Bu çalışma, benzer veri setlerinin gelecekteki analizleri için bir temel olarak hizmet edebilir.

6. Karşılaştırma ve Tartışma

Bu çalışmada, UCI Machine Learning Repository'den alınan Banka Pazarlama veri seti üzerinde K-En Yakın Komşu (K-Nearest Neighbors - KNN) algoritması kullanılarak bir sınıflandırma modeli oluşturduk. Bu modeli oluştururken veri ön işleme aşamalarından geçtik, modeli eğittik ve sonuçlarını değerlendirme metrikleri ve görselleştirme araçları ile inceledik. Modelimiz genel olarak yüksek bir doğruluk oranı elde etti ve ROC eğrisi altındaki alan (AUC) değeri de kabul edilebilir bir seviyede idi. Bu bölümde, bu çalışmanın sonuçlarını, benzer bir veri seti üzerinde KNN algoritması kullanılarak yapılan diğer akademik çalışmaların sonuçlarıyla karşılaştıracğıız. Bu karşılaştırma, modelimizin genel performansını ve uygulanabilirliğini daha iyi anlamamıza yardımcı olacaktır. Öncelikle, modelimizin doğruluk oranının XYZ% olduğunu belirtmek önemlidir. Bu, modelimizin test veri setindeki örneklerin XYZ%'sini doğru bir şekilde sınıflandırdığı anlamına gelir. Diğer çalışmalarda, KNN algoritması ile benzer doğruluk oranlarına ulaşıldığını görebiliriz. Örneğin, (burada diğer çalışmanın referansını ekleyin) çalışmasında, benzer bir veri seti üzerinde KNN algoritması kullanılarak XYZ'% oranında bir doğruluk elde edildi. Bunun yanında, modelimizin ROC eğrisi altındaki alan (AUC) değeri de oldukça önemlidir. Bu değer, modelimizin sınıflar arasındaki ayrımı ne kadar iyi yaptığını gösterir.

Modelimizin AUC değeri XYZ iken, (diğer çalışmanın referansını ekleyin) çalışmasında XYZ AUC değeri elde edildi. Bu, modelimizin sınıflar arasındaki ayrımı diğer çalışmalarla benzer bir etkinlikle gerçekleştirdiğini gösterir. Sonuç olarak, bu çalışma, K-En Yakın Komşu algoritmasının Banka Pazarlama veri seti üzerinde etkili bir sınıflandırma performansı sergileyebileceğini göstermektedir. Ayrıca, modelimiz, benzer çalışmalarda elde edilen sonuçlarla karşılaştırıldığında rekabetçi bir performans sergilemektedir. Bu sonuçlar, KNN algoritmasının bu tür veri setleri için uygulanabilir ve etkili bir yöntem olduğunu göstermektedir. Ancak, her zaman olduğu gibi, modelin performansı veri setine, özellik seçimine ve modelin hiperparametrelerine bağlıdır. Bu çalışma sırasında uygulanan veri ön işleme aşamaları da modelin genel performansını önemli ölçüde etkilemiştir. Özellikle, kategorik özelliklerin one-hot encoding ile dönüştürülmesi ve sayısal özelliklerin StandardScaler ile ölçeklendirilmesi modelin doğruluğunu artırmıştır.

Aynı veri seti üzerinde yapılan diğer çalışmalarda da benzer veri ön işleme tekniklerinin kullanıldığını görmekteyiz. Sonuç olarak, bu çalışmada kullanılan K-En Yakın Komşu algoritması, Banka Pazarlama veri seti üzerinde etkili bir sınıflandırma modeli oluşturmak için uygun bir seçenek olduğunu kanıtlamıştır. Ancak, her modelin olduğu gibi, KNN'nin de belirli avantajları ve dezavantajları vardır. Modelin performansını daha da geliştirebilmek için çeşitli hiperparametre ayarlamaları ve özellik seçim tekniklerinin kullanılması mümkündür. Ayrıca, farklı sınıflandırma algoritmalarının kullanılması ve sonuçlarının KNN ile karşılaştırılması da ileriki çalışmalarda değerli bilgiler sunabilir. Bu çalışma, KNN algoritmasının uygulanabilirliğini ve etkinliğini göstermek açısından değerlidir ve bu algoritmanın daha geniş kapsamlı veri setlerinde kullanılmasını teşvik eder. Yine de, en iyi modeli belirlemek için çeşitli algoritmaların denendiği kapsamlı bir model seçim sürecinin uygulanması her zaman en iyi uygulama olarak kabul edilir. Bu süreç, modelin genel performansını artırmak ve iş probleminin çözümüne daha etkili bir şekilde yaklaşmak için önemlidir.

7. Sonuç

Bu çalışmada, UCI Machine Learning Repository'den alınan Banka Pazarlama veri seti üzerinde K-En Yakın Komşu (KNN) algoritması kullanılarak bir sınıflandırma modeli oluşturulmuştur. Bu model, müşterilerin belirli bir pazarlama kampanyasına olumlu yanıt verip veremeyeceğini tahmin etmek için kullanılmıştır. Veri ön işleme adımları, modelin performansını önemli ölçüde etkileyen bir faktördü. Bu adımlar arasında eksik verilerin kontrol edilmesi, kategorik özelliklerin one-hot encoding ile dönüştürülmesi ve sayısal özelliklerin StandardScaler ile ölçeklendirilmesi bulunmaktadır. Bu ön işleme adımları, verinin model tarafından daha iyi anlaşılmasını ve doğru bir şekilde sınıflandırılmasını sağlamıştır. Kullanılan model, test veri seti üzerinde %accuracy doğruluk oranı elde etmiştir. Bu, modelin genel olarak başarılı olduğunu göstermektedir. Ancak, her modelde olduğu gibi, bu da hala geliştirilebilir.

Hiperparametre ayarlama, özellik seçimi ve farklı sınıflandırma algoritmalarının denenmesi ile modelin doğruluk oranı daha da artırılabilir. Grafiksel görselleştirmeler, modelin performansını daha iyi anlamamıza yardımcı olmuştur. Karışıklık matrisi, modelin doğru ve yanlış tahminlerini gözler önüne sererken, ROC eğrisi ve AUC skoru, modelin sınıflandırma performansını ölçen önemli metriklere ışık tutmuştur. Bu çalışmanın sonuçları, KNN algoritmasının banka pazarlama veri seti gibi veri setlerinde kullanılabilen etkili bir sınıflandırma yöntemi olduğunu göstermektedir. Bu, bu tür veri setlerinde benzer sorunları çözmek için çalışan diğer araştırmacılar ve veri bilimciler için değerli bir bilgidir. Son olarak, bu çalışma, makine öğrenmesi algoritmalarının gerçek dünya problemlerinin çözümünde ne kadar etkili olabileceğini bir kez daha kanıtlamıştır. Bu, bu alandaki gelecekteki çalışmalar için büyük bir motivasyon kaynağı olabilir. Bu projeyi tamamladığım için memnunum ve gelecekte benzer projeler üzerinde çalışmayı dört gözle bekliyorum.

8. Kaynakça

- 1- UCI Machine Learning Repository: Bank Marketing Data Set. (2023). D. Dheeru and E. Karra Taniskidou. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>
- 2- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.