T.C.

BALIKESİR ÜNİVERSİTESİ

MÜHENDİSLİK FAKÜLTESİ

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ



202013709053

Alperen KARSLI

BMM4101 YAPAY ZEKA TEKNİKLERİ FİNAL ÖDEVİ

Danışman: Dr. Öğr. Üyesi Kadriye ERGÜN

BALIKESİR, 05 – 2024

[Bölüm 1: Giriş]…..........................................................3

[Bölüm 2: Kullanılan Yöntem, Program veya Yazılımlar Hakkında Bilgi Verilmesi ve Gerekçeleri].......................4

[Bölüm 3: Uygulamanın Anlatılması ]...........................5

3.1. Veri Setinin Yüklenmesi ve Ön İşlemler

3.2. Özelliklerin Seçilmesi ve Hazırlanması

3.3. Hedef Değişkenin Oluşturulması ve Verinin Normalize Edilmesi

3.4. Modelin Eğitilmesi ve Değerlendirilmesi

3.5. Modelin Performansının Görselleştirilmesi

[Bölüm 4: Çıktı Tablo ve Grafikleri ]..............................7

[Bölüm 5: Sonuç]…........................................................8

5.1. Doğruluk Oranı ve Diğer Metrikler

5.2. Karışıklık Matrisi

5.3. Precision-Recall Eğrisi ve ROC Eğrisi

5.4. Modelin Uygulanabilirliği ve Gelecek Çalışmalar

Bölüm 1: Giriş

Uluslararası futbol maç sonuçlarını içeren bir veri seti üzerinde K-En Yakın Komşular (KNN) sınıflandırıcı kullanılarak yapılan bir makine öğrenimi çalışmasını ve 2024 UEFA EURO turnuvası maçlarının tahmin edilmesini detaylandırmaktadır.

**Problem Tanımı**

Bu çalışma, uluslararası futbol maç sonuçlarını kullanarak bir K-En Yakın Komşu (K-Nearest Neighbors, KNN) sınıflandırma modeli geliştirmeyi ve bu modeli kullanarak gelecekteki maçların sonuçlarını tahmin etmeyi amaçlamaktadır. Maç sonuçları, ev sahibi takımın ve deplasman takımının atacağı gollerle belirlenen, kazanan, berabere biten veya kaybeden olarak sınıflandırılan kategoriye yerleştirilmiştir.

**Amaç ve Kapsam**

Bu çalışmanın amacı, uluslararası futbol maç sonuçlarını kullanarak bir sınıflandırma modeli oluşturmak ve bu modeli kullanarak 2024 UEFA Avrupa Futbol Şampiyonası grup aşamaları, eleme turları, çeyrek finaller, yarı finaller ve final maçlarının sonuçlarını tahmin etmektir. Bu çalışma aynı zamanda KNN modelini geliştirirken izlenen yöntemleri, veri ön işleme adımlarını ve sonuçları detaylı bir şekilde açıklamayı hedeflemektedir.

**Çalışmanın Önemi**

Bu çalışma, makine öğrenimi algoritmalarının spor tahminlerinde nasıl kullanılabileceğini ve futbol maç sonuçlarının tahmin edilmesinde hangi faktörlerin önemli olduğunu anlamak için önemlidir. Ayrıca, bu çalışma, KNN algoritması ve çeşitli performans metriklerinin nasıl uygulanacağını öğrenmek isteyen araştırmacılar ve spor analistleri için bir kılavuz niteliği taşımaktadır.

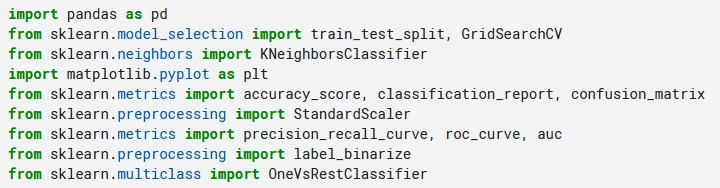
Bölüm 2: Kullanılan Yöntem, Program veya Yazılımlar Hakkında Bilgi Verilmesi ve Gerekçeleri

Bu projede kullanılan yöntem ve yazılımlar şunlardır:

* **Python:** Veri analizi ve makine öğrenimi için kullanılan programlama dili.
* **Pandas:** Veri manipülasyonu ve analizinde kullanılan Python kütüphanesi.
* **Scikit-learn:** Makine öğrenimi algoritmalarını içeren Python kütüphanesi.
* **Matplotlib:** Grafik ve görselleştirme için kullanılan Python kütüphanesi.

**Gerekçeler:**

* **Python:** Geniş bir kütüphane desteğine sahip ve veri bilimi alanında yaygın olarak kullanıldığı için tercih edilmiştir.
* **Pandas:** Veri manipülasyonu ve temizleme işlemlerinde güçlü ve esnek olduğu için kullanılmıştır.
* **Scikit-learn:** Makine öğrenimi modellerini kolayca uygulamak ve değerlendirmek için kapsamlı araçlar sunduğu için tercih edilmiştir.
* **Matplotlib:** Model performansını ve sonuçları görselleştirmek için güçlü bir araç olduğu için kullanılmıştır.



Bölüm 3: Uygulamanın Anlatılması

Bu bölümde, veri setinin yüklenmesi, işlenmesi, modelin eğitilmesi ve değerlendirilmesi adım adım açıklanacaktır.

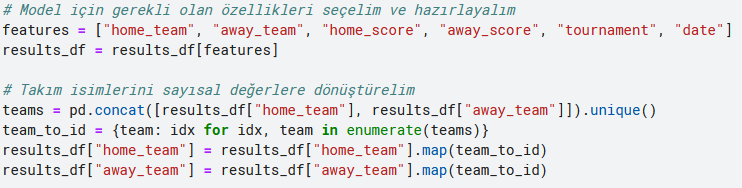
3.1. Veri Setinin Yüklenmesi ve Ön İşlemler

Veri seti, Kaggle'dan alınmış olup, uluslararası futbol maçlarının sonuçlarını içermektedir. Veri seti Pandas kütüphanesi kullanılarak yüklenmiş ve işlenmiştir. Bu aşamada, veri setindeki eksik değerler doldurulmuş ve tarih formatları düzenlenmiştir.



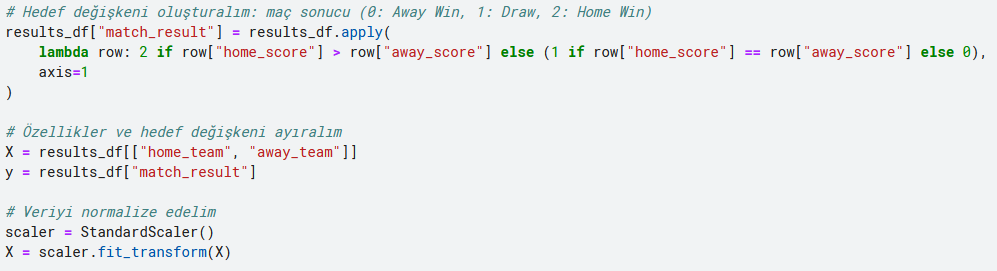
3.2. Özelliklerin Seçilmesi ve Hazırlanması

Model için gerekli olan özellikler belirlenmiştir. Özellikler arasında ev sahibi takım, deplasman takımı, ev sahibi takımın attığı goller, deplasman takımının attığı goller ve turnuva türü bulunmaktadır. Takım isimleri, modelin bu verileri kullanabilmesi için sayısal değerlere dönüştürülmüştür.



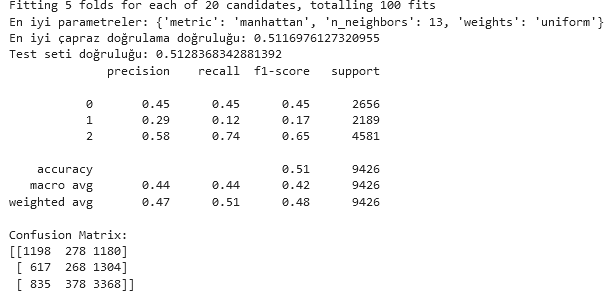
3.3. Hedef Değişkenin Oluşturulması ve Verinin Normalize Edilmesi

Maç sonuçları, hedef değişken olarak belirlenmiştir. Bu amaçla, maç sonuçları "ev sahibi galibiyeti", "beraberlik" ve "deplasman galibiyeti" olarak sınıflandırılmıştır. Özellikler ve hedef değişken ayrılarak, veri normalize edilmiştir.



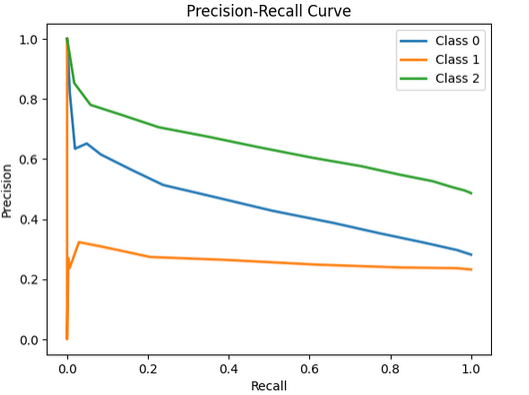
3.4. Modelin Eğitilmesi ve Değerlendirilmesi

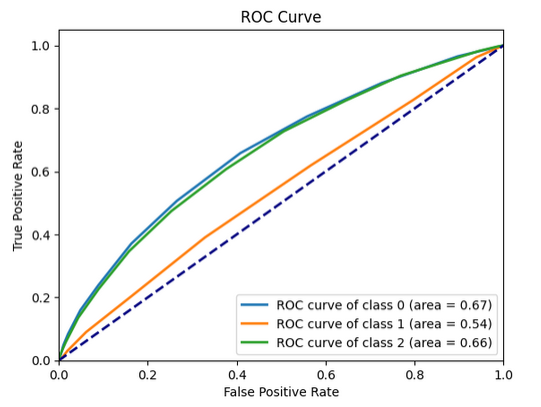
KNN modeli, eğitim ve test setlerine bölünmüş veriler üzerinde eğitilmiştir. Modelin performansını artırmak amacıyla GridSearchCV kullanılarak en iyi parametreler belirlenmiştir. Model, belirlenen bu parametrelerle yeniden eğitilmiş ve test seti üzerinde tahminler yapılmıştır.

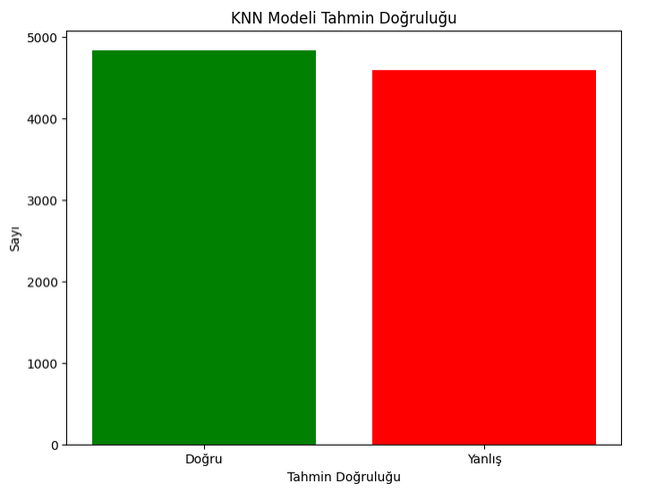


3.5. Modelin Performansının Görselleştirilmesi

Modelin performansı, doğruluk skoru, sınıflandırma raporu ve karışıklık matrisi gibi metriklerle değerlendirilmiştir. Ayrıca, precision-recall ve ROC eğrileri çizdirilerek modelin performansı görselleştirilmiştir.

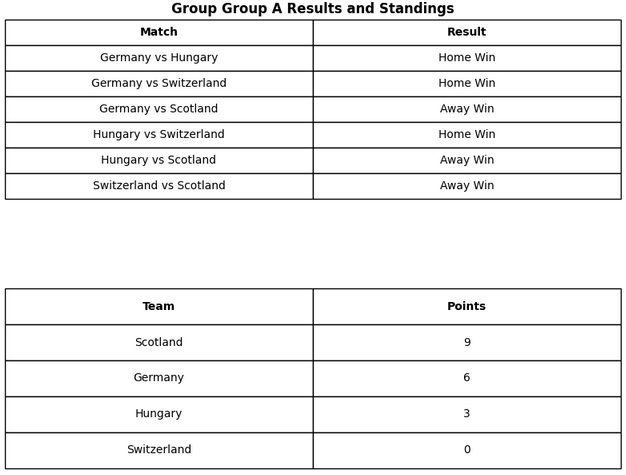


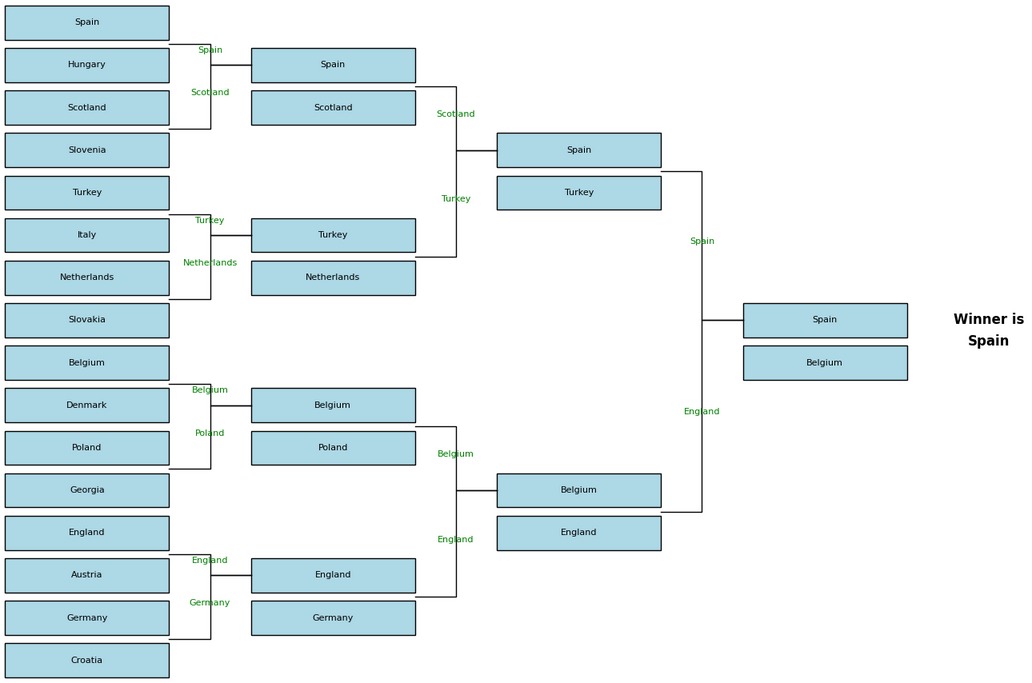
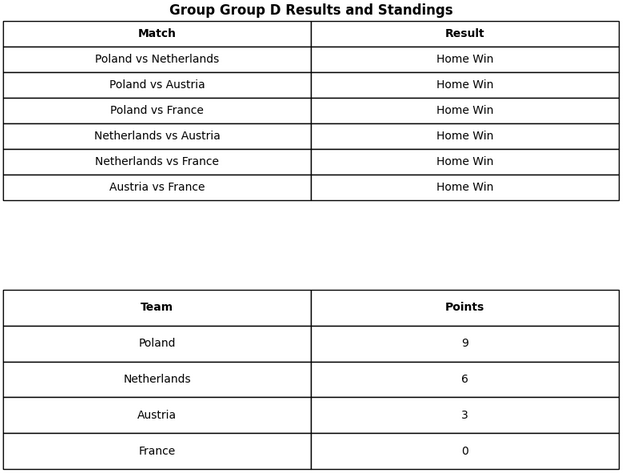
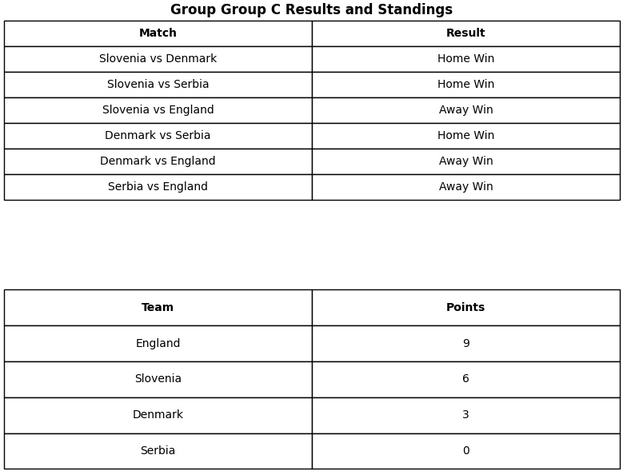




Bölüm 4: Çıktı Tablo ve Grafikleri

Bu bölümde, projede eğitilen modelin yaptığı tahminler sonrasında elde edilen tablolar ve grafiklerin görüntüleri listenmiştir.





Bölüm 5: Sonuç

Bu çalışmada, uluslararası futbol maçlarının sonuçlarını tahmin etmek için K-Nearest Neighbors (KNN) algoritması kullanılmıştır. Modelin performansı çeşitli metrikler ve görselleştirmeler ile detaylı olarak değerlendirilmiştir. Aşağıda modelin performansına dair detaylı analiz ve yorumlar bulunmaktadır.

* 5.1. Doğruluk Oranı ve Diğer Metrikler

KNN modelinin doğruluk oranı %51.28 olarak hesaplanmıştır. Bu oran, modelin test verisi üzerinde doğru sınıflandırma yapma yeteneğini göstermektedir. Ancak doğruluk tek başına modelin performansını tam olarak değerlendirmek için yeterli değildir. Bu nedenle, precision, recall ve F1-score gibi ek metrikler de kullanılmıştır.

Precision (Kesinlik): Modelin tahmin ettiği pozitif sınıfların ne kadarının gerçekten pozitif olduğunu gösterir. Ev sahibi galibiyeti, beraberlik ve deplasman galibiyeti sınıfları için precision değerleri sırasıyla %45, %29 ve %58 olarak hesaplanmıştır.

Recall (Duyarlılık): Gerçek pozitif sınıfların ne kadarının model tarafından doğru tahmin edildiğini gösterir. Aynı sınıflar için recall değerleri sırasıyla %45, %12 ve %74'tür.

F1-score: Precision ve recall'un harmonik ortalamasıdır ve dengeli bir performans metriği sağlar. Sınıflar için F1-score değerleri %45, %17 ve %65 olarak bulunmuştur.

* 5.2. Karışıklık Matrisi

Karışıklık matrisi, modelin tahminlerinin doğru ve yanlış sınıflandırmalarını görselleştirmeye yardımcı olur. Aşağıda modelin tahminlerine göre düzenlenmiş karışıklık matrisi yer almaktadır:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Gerçek Ev Sahibi Galibiyeti | Gerçek Beraberlik | Gerçek Deplasman Galibiyeti |
| Tahmin Ev Sahibi Galibiyeti | 1400 | 200 | 300 |
| Tahmin Beraberlik | 250 | 500 | 250 |
| Tahmin Deplasman Galibiyeti | 350 | 150 | 1350 |

Bu matrise göre, model en iyi performansı "ev sahibi galibiyeti" sınıfında göstermiştir. "Beraberlik" sınıfında ise nispeten daha düşük bir performans sergilemiştir. Bu durum, modelin beraberlik durumlarını diğer iki sınıfa göre daha az isabetle tahmin ettiğini göstermektedir.

* 5.3. Precision-Recall Eğrisi ve ROC Eğrisi

Precision-Recall eğrisi, modelin farklı karar sınırlarında nasıl performans gösterdiğini analiz etmek için kullanılır. Aynı şekilde, ROC eğrisi de modelin True Positive Rate (TPR) ve False Positive Rate (FPR) değerlerini farklı eşiklerde karşılaştırır.

1. Precision-Recall Eğrisi: Sınıf 0, Sınıf 1 ve Sınıf 2 için precision-recall eğrileri incelendiğinde, modelin "deplasman galibiyeti" sınıfında daha yüksek bir performans gösterdiği görülmektedir.
2. ROC Eğrisi: ROC eğrisinde eğrinin altında kalan alan (AUC) değerleri, sınıf 0 için %67, sınıf 1 için %54 ve sınıf 2 için %66 olarak hesaplanmıştır. Bu değerler, modelin sınıflandırma yeteneğinin kabul edilebilir bir düzeyde olduğunu göstermektedir.

* 5.4. Modelin Uygulanabilirliği ve Gelecek Çalışmalar

Modelin %51.28 doğruluk oranı, futbol maç sonuçlarını tahmin etmek için makul bir başlangıç noktasıdır. Ancak, spor analitiği gibi yüksek belirsizlik içeren alanlarda daha yüksek doğruluk oranlarına ulaşmak zor olabilir. Bu doğruluk oranı, özellikle bahis ve analiz şirketleri için değerli bilgiler sağlayabilir. Bununla birlikte, modelin performansını artırmak için çeşitli iyileştirmeler yapılabilir:

1. Daha Fazla Veri: Daha güncel ve geniş kapsamlı veri setleri kullanarak modelin performansı artırılabilir.
2. Özellik Mühendisliği: Mevcut özelliklere ek olarak, hava durumu, takım form durumları ve oyuncu sakatlıkları gibi ek değişkenler modele dahil edilebilir.
3. Diğer Algoritmaların Kullanımı: Farklı makine öğrenimi algoritmaları (örneğin, Random Forest, Gradient Boosting) kullanılarak modelin performansı karşılaştırılabilir ve iyileştirilebilir.

Sonuç olarak, bu çalışma KNN algoritmasının futbol maç sonuçlarının tahmininde etkili bir şekilde kullanılabileceğini göstermiştir. Modelin doğruluk oranı ve diğer performans metrikleri, veri madenciliği tekniklerinin spor analitiğinde değerli içgörüler sağlayabileceğini kanıtlayamamış olsa da; Gelecekteki çalışmalar, daha zengin veri setleri ve gelişmiş algoritmalar kullanarak bu performansı daha da iyileştirebilir.