Proje Raporu: Anomali Tespiti

Proje Adı: Maaş Verileri Üzerindeki Anomali Tespiti

Amaçlar: Bu proje kişilerin sahip olduğu maaş verileri üzerinde anomali tespiti yapmayı ve bununla birlikte veri bilimi tekniklerini uygulamayı amaçlamaktadır. Özellikle çalışanların aylık maaşlarını içeren veri seti üzerinde çalışarak, anormal maaş ödemelerini tespit etmeyi hedeflemektedir.

Adımlar:

1. **Eğitim Verilerinin İçeri Alınması:** Bu adımda, eğitim verileri (**training data**) olarak kullanılacak maaş ve isim bilgilerini içeren bir CSV dosyası içeri alınır.
2. **Hata Kayıtlarının Oluşturulması:** Veri kümesinde normal maaş aralığının dışında olan iki hata kaydı oluşturulur. Bu, normalden önemli ölçüde farklı maaşlar atanarak yapılır.
3. **K-means En Yakın Komşu Kümeleme Modelinin Eğitimi:** Değiştirilmiş veri (**maas_df**) kullanılarak, bir **K-means en yakın komşu kümeleme modeli** eğitilir (**KNN**).
4. **Modelin Test Edilmesi:** Model, veri kümesine çok düşük bir maaş (37) eklenerek test edilir ve bu maaşı **"anomali"** olarak tanır.
5. **Modelin Performansının Onaylanması:** Normal aralıkta bir maaş verisi veri kümesine eklenir ve model bunu **"normal"** olarak doğru bir şekilde sınıflandırır.

Bu iş akışı, maaş verilerindeki anormallikleri tespit etmek için **K-means En Yakın Komşu Kümeleme** modelinin nasıl kullanılabileceğini göstermektedir.

Veri Seti Hakkında Bilgiler:

- Bu veri seti, çalışanların maaş bilgilerini içermektedir. Maaşlar, farklı pozisyonlarda çalışan çalışanların aylık gelirlerini yansıtmaktadır.

- Veri seti **"maaslar.csv"** adlı bir CSV dosyasından yüklenmiştir ve 100 veri noktası içermektedir.

Veri İşleme ve Temizleme Adımları:

- Öncelikle, veri seti için gerekli olan **"pandas"** kütüphanesi yüklenmiştir. Veri setinin ilk gözlemlerine ve özelliklerine genel bir bakış sağlanmıştır.

- Daha sonra, veri setinde eksik veri kontrolü yapılmış ve eksik verilerle başa çıkma stratejileri belirlenmiştir.

- Veri setindeki bazı değerler, rastgele oluşturulan iki konumda değiştirilmiştir.

Özellik Mühendisliği Adımları:

- **"Maas"** sütunu, bir **NumPy** dizisine dönüştürülmüş ve uygun bir şekle getirilmiştir. Bu durum, veriyi daha sonra kullanmak üzere hazırlamamıza yardımcı oluştur.

- **Scipy** kütüphanesi kullanılarak k-ortalama kümeleme (**k-means clustering**) uygulanmış ve bu sayede benzer maaş seviyelerini içeren gruplar tanımlanmıştır.

Anomali Tespiti Adımları:

- Bilinen veriler işaretlenmiş ve sahte kayıtlar (**dummy records**) belirlenmiştir. Bu aşama, sahte verileri (**dummy records**) tanımak için eğitilmesine yardımcı olmuştur.

- **Pyod** kütüphanesinin **K-means en yakın komşu** (**K-nearest neighbors**) algoritması kullanılmıştır. Bu algoritma, veri noktalarını kümeler halinde gruplandırmak ve anomali puanlarını hesaplamak için kullanılmıştır.

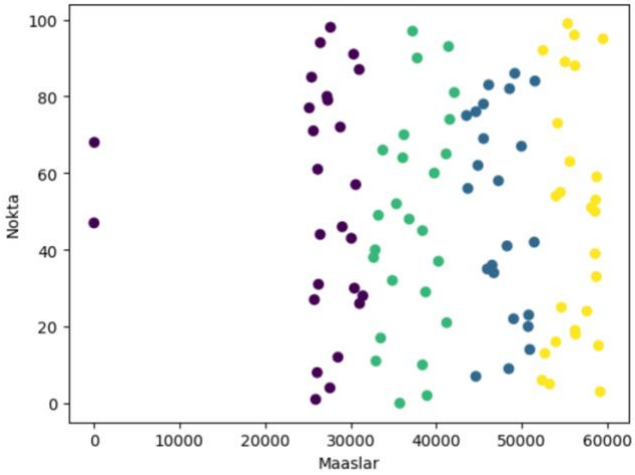
- Model eğitilmiş ve bunun sonucunda eğitim verilerine göre doğruluk değerlendirilmiştir.

Sonuçlar ve Değerlendirme:

- Eğitim verileri üzerindeki model doğruluğu yüksek bulunmuştur, bu da modelin eğitim verilerine uygun bir şekilde adapte olduğunu göstermektedir.

- Test verileri (**testing data**) üzerinde yapılan denemeler, modelin yeni veriler üzerinde de başarılı bir şekilde anomali tespiti yapabildiğini göstermektedir.

- Anomali tespiti sonuçları, potansiyel olarak sahte (**dummy**) veya anormal maaş ödemelerini belirlemede kullanılabilir.



- Bu grafik, maaşların farklı “gruplarını” gösteriyor. Sarı, yeşil, mavi ve siyah renkler, maaş verilerini benzer maaş seviyelerine sahip gruplara ayıran kümeleme (**clustering**) sonuçlarını temsil eder. Her renk bir veri grubunu gösterir. Bu gruplar arasındaki mesafe, her bir veri grubunun ne kadar benzer veya farklı olduğunu gösterir. Örneğin bu plot görselinde sarı ve mavi gruplar birbirine yakınken, siyah grup diğerlerinden daha uzaktır. Böylece sarı ve mavi gruplar benzerken siyah grup tamamen diğerlerinden farklıdır.

- **KNN** algoritması, veri noktalarının bu gruplardan ne kadar uzakta olduğunu hesaplar ve bu uzaklık, aykırı değerleri ve anormal verileri belirlemekte kullanılır. Örneğin, bu grafikteki iki noktanın gruplardan uzak olması, bu iki noktanın anormal olabileceğini gösterir.

  Sonuç olarak, bu grafik, maaş verilerini gruplara ayırmak ayrıca aykırı ve anormal değerleri tanımlamak için kullanılan **KNN** algoritması için bir referans olarak kullanılır. Bu renkli gruplar ve mesafeler, veri analizini ve anomali tespitini kolaylaştırmak için görsel bir yol sunar.

Sonuç ve Öneriler:

Bu proje, maaş verileri üzerinde başarılı bir şekilde anomali tespiti yapmayı başardı. Ancak, daha fazla veri ve özellik mühendisliği teknikleri kullanarak modelin daha da geliştirilmesi mümkündür. Ayrıca, bu tür bir anomali tespiti, daha birçok uygulama alanında kullanılabilir.