LOTUS AI

YAPAY ZEKA VE BİLİŞİM ÇÖZÜMLERİ A.Ş.

ANOMALİ TESPİTİ RAPORU

HAZIRLAYANIN;

ADI: SENANUR
SOYADI: BAYRAM

09/12/2024

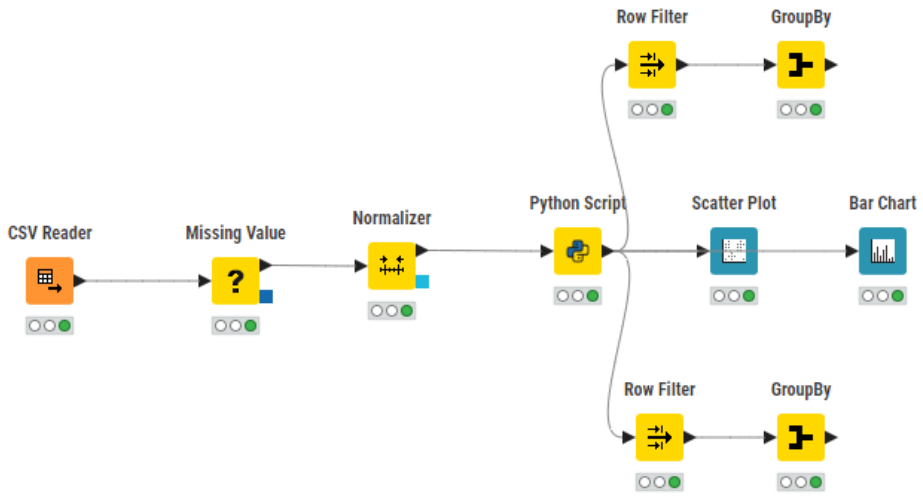
GİRİŞ

Sağlık sektörü, büyük miktarda verinin üretilmesiyle birlikte veri analitiği ve makine öğrenimi yaklaşımlarını benimseyerek önemli ilerlemeler kaydetmektedir. Bu çalışmada kullanılan "Healthcare Providers Data" adlı veri seti, sağlık hizmeti sağlayıcılarına ilişkin çeşitli finansal ve operasyonel bilgileri içermektedir. Veri seti, sağlık kurumlarının hizmet süreçlerini optimize etmelerine, maliyet etkinliği sağlamalarına ve hizmet kalitesini artırmalarına yardımcı olabilecek değerli içgörüler sunmaktadır.

Bu bağlamda, çalışmanın amacı, veri setinde yer alan potansiyel anomali durumlarını tespit etmektir. Anomaliler, genellikle veri setinde nadiren görülen ve diğer veri noktalarından önemli ölçüde farklılık gösteren gözlemler olarak tanımlanır. Bu tür anomalilerin tespiti, finansal hilelerin önlenmesi, operasyonel hataların belirlenmesi ve karar destek süreçlerinin iyileştirilmesi gibi kritik alanlarda değerli bilgiler sağlayabilir.

Anomali tespiti için yaygın olarak kullanılan yöntemlerden biri olan **Isolation Forest (IF)** algoritması, bu çalışmada tercih edilmiştir. Isolation Forest, yüksek boyutlu veri kümelerinde etkili bir şekilde çalışabilmesi ve açıklanabilirlik avantajı sunması nedeniyle seçilmiştir. Bu çalışmada, hem KNIME hem de Python platformları üzerinde IF algoritmasının uygulanması ele alınacaktır.

1. KNIME



Bu çalışma, Kaggle platformunda sunulan sağlık hizmeti sağlayıcıları verisini kullanarak bir veri ön işleme ve analiz süreci gerçekleştirmek amacıyla KNIME Analytics Platformu'nda geliştirilmiş bir iş akışını tanımlamaktadır. İş akışının temel amacı, veri setinde eksik değerlerin yönetilmesi, verilerin normalize edilmesi, anlamlı gruplamalar yapılması ve bu gruplar üzerinden görselleştirme yapılarak içgörülerin elde edilmesidir. Aşağıda iş akışındaki adımlar ve kullanılan düğümlerin işlevleri detaylı bir şekilde açıklanmıştır.

İş Akışı Düğümleri ve Fonksiyonları

1. **CSV Reader (CSV Okuyucu)**
   Bu düğüm, veri setini iş akışına dahil etmek için kullanılmıştır. Sağlık hizmeti sağlayıcılarına ilişkin bilgiler içeren CSV dosyası yüklenmiş ve veri seti işlenebilir hale getirilmiştir.

2. **Missing Value (Eksik Değer İşleme)**
   Eksik veri problemlerini gidermek için kullanılan bu düğüm, veri setindeki eksik gözlemleri tespit ederek uygun bir stratejiyle (örneğin ortalama, medyan veya mod gibi) doldurmuştur. Bu adım, eksik verilerin analitik sonuçlara olumsuz etkilerini önlemek için kritik öneme sahiptir.

3. **Normalizer (Normalleştirici)**
   Verilerin ölçeklenmesini sağlayarak değişkenlerin farklı birimlerden kaynaklı etkilerini ortadan kaldırır. Bu adım, değişkenlerin eşit öneme sahip olmasını sağlamak ve analiz süreçlerinin güvenilirliğini artırmak için uygulanmıştır.

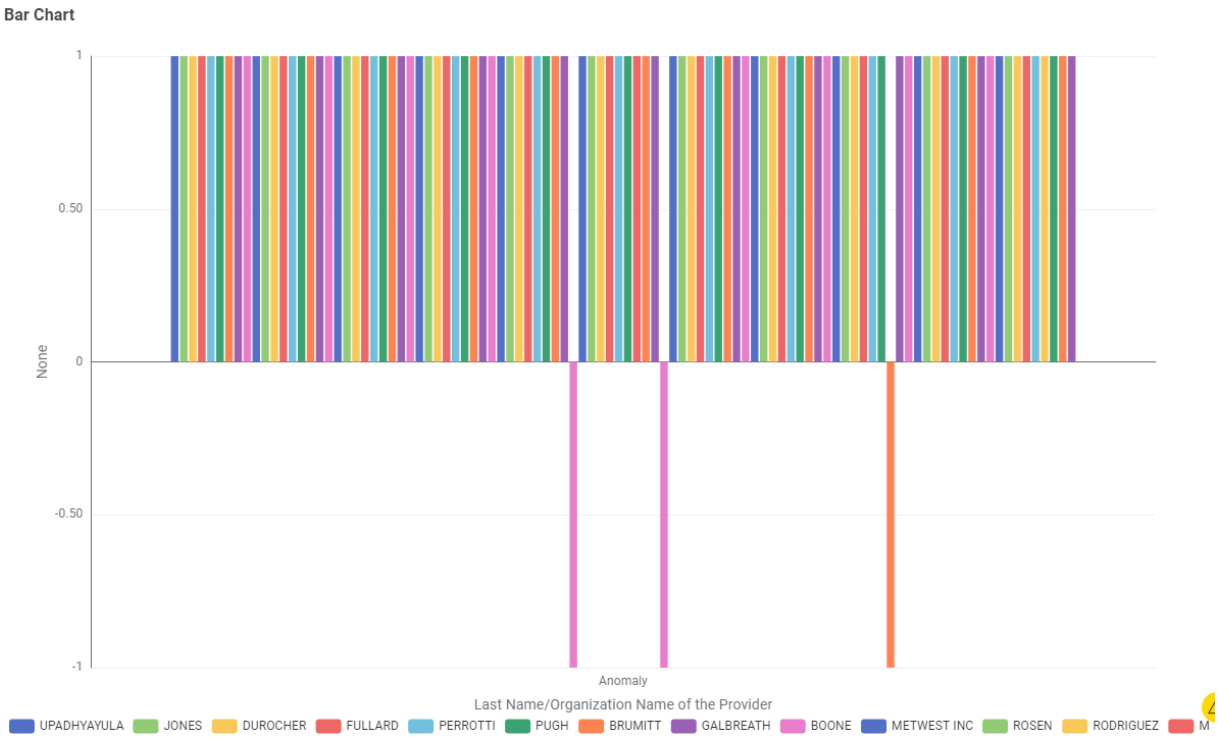
4. **Python Script (Python Betiği)**
   Python programlama diliyle özelleştirilmiş veri manipülasyonu gerçekleştirilmiştir. Bu adım, veri özelliklerinden türetilmiş yeni değişkenlerin eklenmesi, istatistiksel hesaplamalar veya veri temizliği gibi işlemleri içerebilir.

5. **Scatter Plot (Dağılım Grafiği)**
   Veriler arasındaki ilişkilerin görselleştirilmesi amacıyla dağılım grafiği oluşturulmuştur. Bu adım, iki değişken arasındaki korelasyonları değerlendirme olanağı sağlamıştır.

6. **Row Filter (Satır Filtresi)**
   Veriler, belirli kriterlere göre alt kümelere ayrılmıştır. Örneğin, yalnızca belirli bir coğrafi bölgede yer alan sağlık hizmeti sağlayıcıları analiz edilmek üzere seçilmiştir.

7. **GroupBy (Gruplama)**
   Bu düğüm, verilerin belirli değişkenlere göre gruplandırılmasını sağlamıştır. Örneğin, sağlık hizmeti sağlayıcıları türlerine, coğrafi bölgelerine veya diğer kategorik özelliklerine göre gruplandırılarak özet istatistikler oluşturulmuştur.

8. **Bar Chart (Çubuk Grafik)**
   Gruplanmış verilerin görselleştirilmesi için kullanılan bu düğüm, sağlık hizmeti sağlayıcılarının kategorik dağılımlarını çubuk grafiklerle sunmuştur. Bu görselleştirme, gruplar arasındaki farkları açıkça ortaya koymaktadır.



Bu KNIME iş akışı, sağlık hizmeti sağlayıcıları verisinin temizlenmesi, dönüştürülmesi ve görselleştirilmesi süreçlerini içermektedir. İş akışı, veri setinin eksik değer problemlerinin çözülmesi, ölçeklendirilmesi ve anlamlı gruplar oluşturularak analiz edilmesi için tasarlanmıştır. Elde edilen sonuçlar, sağlık hizmeti sağlayıcılarının coğrafi dağılımını, türlerini ve potansiyel ilişkilerini görselleştirme ve değerlendirme olanağı sunmaktadır.

2. PYTHON

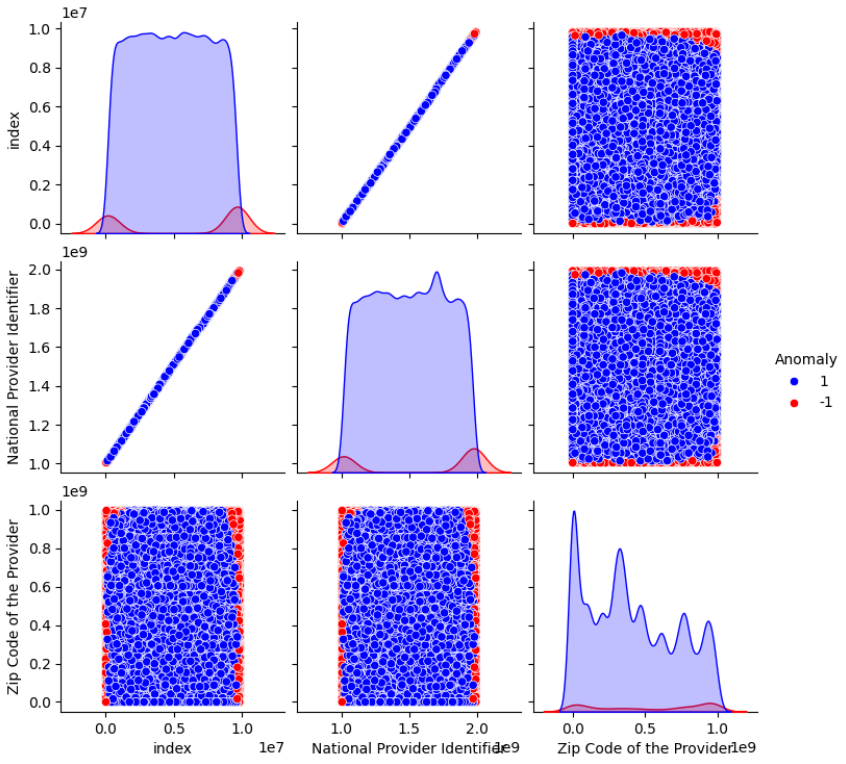
Bu çalışma, bir veri setinde anomali tespiti yapmak amacıyla **Isolation Forest** algoritmasını kullanarak bir analiz süreci gerçekleştirmektedir. Anomali tespiti, verilerde alışılmadık veya beklenmedik gözlemleri belirlemek için kullanılan bir yöntemdir ve veri analitiği, siber güvenlik, sağlık ve finans gibi pek çok alanda önemli uygulamalara sahiptir. Bu çalışmada kullanılan yöntem, bir veri setindeki sayısal değişkenlere odaklanmış ve anomali olarak belirlenen verileri görselleştirerek sonuçları analiz etmiştir.

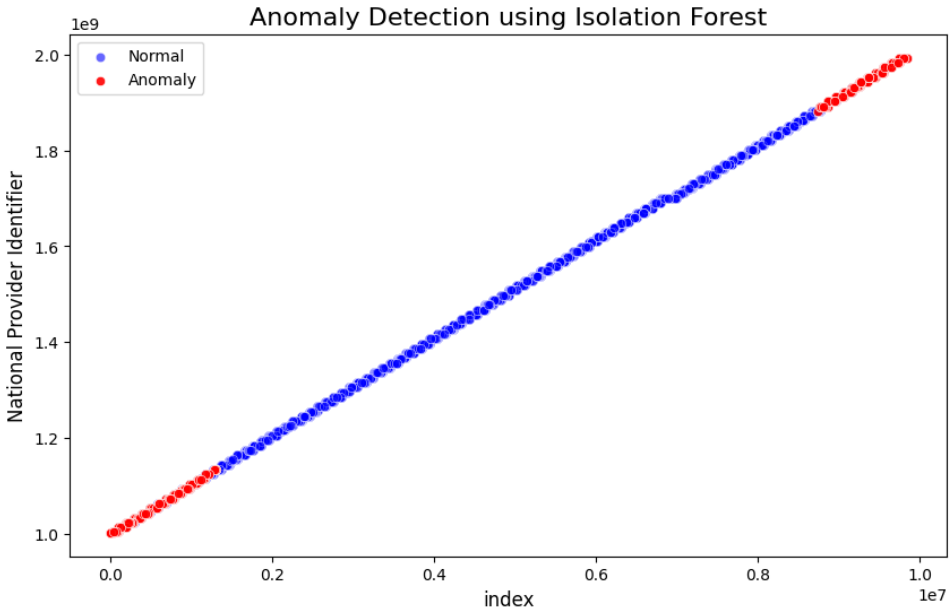
İlk olarak, veri seti üzerinde sadece sayısal sütunlar seçilerek işlem yapılmıştır. Eksik değer içeren gözlemler, modelin performansını etkileyebileceğinden, analizden çıkarılmıştır. Daha sonra, **Isolation Forest** algoritması uygulanmıştır. Bu algoritma, anomali tespitinde sıkça kullanılan bir makine öğrenimi yöntemidir ve verileri ağaç yapıları kullanarak izole etmeye çalışır. Normal veriler, genellikle diğer verilerle benzer özelliklere sahipken, anomaliler daha kolay izole edilir. Çalışmada, anomali oranı veri setinin %5'i olarak belirlenmiş ve model bu oran doğrultusunda veri setini sınıflandırmıştır. Modelin çıktısında, her bir gözlem "normal" veya "anomalik" olarak etiketlenmiştir.

Elde edilen sonuçlar, öncelikle toplam anomali ve normal gözlem sayısının hesaplanmasıyla değerlendirilmiştir. Daha sonra, bu sınıflandırma sonuçları bir **pairplot** grafiği aracılığıyla görselleştirilmiştir. Bu grafik, çoklu değişkenler arasındaki ilişkileri ve anomalilerin genel dağılımını renkli bir şekilde sunarak görsel bir analiz sağlamıştır. Normal veriler mavi renk ile, anomalik veriler ise kırmızı renk ile gösterilmiştir.

Sonraki adımda, normal ve anomalik veriler ayrıştırılmış ve iki temel değişken seçilerek bir dağılım grafiği oluşturulmuştur. Bu grafik, normal ve anomalik verilerin iki boyutlu bir düzlemdeki dağılımını daha ayrıntılı bir şekilde inceleme fırsatı sunmuştur. Normal verilerin genel olarak homojen bir dağılıma sahip olduğu gözlemlenirken, anomalik veriler genellikle veri yoğunluğunun az olduğu bölgelerde konumlanmıştır.

Bu çalışma, anomali tespiti süreçlerinde görselleştirmenin önemini vurgulamaktadır. Özellikle büyük ve karmaşık veri setlerinde anomalilerin tespiti, veri kalitesinin artırılması, potansiyel hataların belirlenmesi veya olağan dışı durumların araştırılması için kritik bir adımdır. Isolation Forest algoritmasının güçlü performansı ve görselleştirme teknikleri, veri setindeki anomalilerin etkili bir şekilde analiz edilmesini sağlamıştır. Bu yaklaşım, benzer yöntemlerin diğer veri setlerinde uygulanması için bir rehber niteliğindedir.





1.grafikte veri setindeki değişkenler arasındaki çift değişkenli ilişkiler ve değişkenlerin tek değişkenli dağılımları yer almaktadır. Bu grafik, normal ve anomalik gözlemler arasındaki farklılıkları daha ayrıntılı inceleme fırsatı sunmaktadır. Mavi renkle gösterilen normal veriler, veri setinde daha geniş bir dağılıma sahiptir ve yoğun olarak ortalama değerlere yakın konumlanmıştır. Kırmızı ile işaretlenen anomalik gözlemler ise veri dağılımının uç bölgelerinde yoğunlaşmış ve veri setinin genel yapısından farklılık göstermiştir. Özellikle "Zip Code of the Provider" değişkeninde, anomali gözlemleri daha dikkat çekici şekilde belirginleşmiştir.

2.grafikte veri setindeki gözlemler iki kategoride görselleştirilmiştir: mavi renkli noktalar normal verileri temsil ederken, kırmızı renkli noktalar anomali olarak belirlenen gözlemleri göstermektedir. Yatay eksen "index" (veri dizisinin sırası), dikey eksen ise "National Provider Identifier" (sağlık hizmeti sağlayıcılarının kimlik numarası) değerlerini göstermektedir. Görüldüğü üzere, anomali noktaları veri setinin başında ve sonunda yoğunlaşmış durumdadır. Bu, özellikle veri setinde sıradışı değerlerin belirli sınır bölgelerde yoğunlaştığını göstermektedir.