# KNIME Anomali Tespiti Proje Raporu

## 1. Business Understanding (İş Problemini Anlama)

Bu çalışmanın amacı, süt bileşenlerine ait veriler üzerinde normal dağılımdan sapma gösteren örnekleri (anomali gözlemlerini) tespit etmektir.  
Anomali olarak tanımlanan bu gözlemler, ölçüm hatası, olağandışı biyolojik durumlar veya kalite problemlerini temsil edebilir.

Bu bağlamda hedef:

Süt örnekleri arasında normalden farklı davranan örnekleri otomatik olarak belirlemek ve görselleştirmektir.

## 2. Data Understanding (Veriyi Anlama)

### 1. Box Plot Analizi (Outliers)

Box Plot grafikleri incelendiğinde, birçok değişkende aykırı değerlerin (outliers) bulunduğu görülmüştür.

![A screenshot of a graph

AI-generated content may be incorrect.]()  
Özellikle:

* Protein\_content
* beta\_casein
* amino asit içerikleri (Cysteic Acid, Serine, Glycine, Valine vb.)
* pH ve Total\_Solids

değişkenlerinde medyan etrafında yoğunlaşmış bir dağılım olmasına rağmen, üst ve alt sınırların dışında kalan çok sayıda gözlem mevcuttur.

Bu durum:

* Ölçüm hatalarına,
* Biyolojik farklılıklara,
* Numuneler arasındaki doğal varyasyona

işaret edebilir.

Aykırı değerlerin varlığı, mesafe tabanlı yöntemlerde (kNN, Euclidean distance) doğrudan sonucu etkileyebileceği için, veri ön işleme aşamasında bu değerlerin dikkate alınması gerektiğini göstermektedir.

### 2. Eksik Veri (Missing Value) Problemi

Veri setinde yaklaşık 580 sütun bulunmasına rağmen, bazı değişkenlerde ciddi oranda eksik değer tespit edilmiştir.  
Örneğin:

* Phenylalanine (418 kayıt eksik)
* Gamma\_Aminobutyric\_acid (420 kayıt eksik)

Bu değişkenlerde eksik değer oranının oldukça yüksek olduğu görülmektedir.

Bu durum:

* Bu değişkenlerin tüm gözlemler için ölçülmediğini,
* Bazı analizlerde güvenilir bilgi sağlamadığını göstermektedir.

Bu nedenle:

Yüksek oranda eksik veri içeren sütunlar analiz dışı bırakılmış veya uygun yöntemlerle doldurulmuştur.

### 3. Skewness (Çarpıklık) Analizi

Statistics node çıktıları incelendiğinde, bazı değişkenler için:

Skewness > 1 olduğu gözlemlenmiştir.

Bu durum:

* Dağılımların simetrik olmadığını,
* Sağ ya da sol kuyrukta yoğunlaşma olduğunu,
* Verinin normal dağılıma uymadığını göstermektedir.

Özellikle amino asit içerikleri ve bazı kimyasal bileşenlerde yüksek skewness değerleri bulunmuştur.

Bu tür çarpık dağılımlar:

* Ortalama yerine medyanın daha temsil edici olabileceğini,
* Mesafe hesaplamalarında uç değerlerin etkisinin artabileceğini göstermektedir.

Bu nedenle veri seti normalize edilerek değişkenlerin etkisi dengelenmiştir.

## 3. Data Preparation (Veri Hazırlama)

Veri seti, modelleme aşamasına geçmeden önce çeşitli ön işleme adımlarından geçirilmiştir. Bu adımların amacı, veri kalitesini artırmak, gürültüyü azaltmak ve mesafe tabanlı anomali tespit yöntemine uygun bir yapı oluşturmaktır.

Öncelikle, analiz açısından anlamlı bilgi taşımayan ve birbirleriyle yüksek korelasyona sahip olan bazı değişkenler veri setinden çıkarılmıştır. Yüksek korelasyonlu değişkenlerin birlikte kullanılması, modelin belirli özelliklere aşırı ağırlık vermesine neden olabileceğinden, bu değişkenler veri setinden elenmiştir.

Ayrıca, çok yüksek oranda eksik değer içeren sütunlar analiz dışı bırakılmıştır. Bu sütunların güvenilir bilgi sağlamadığı ve model performansını olumsuz etkileyebileceği değerlendirilmiştir.

Eksik değerlerin bulunduğu sütunlar için veri tipine göre farklı doldurma stratejileri uygulanmıştır:

* **String (metinsel) sütunlarda:** eksik değerler bir önceki gözlemin değeri ile doldurulmuştur.
* **Integer (tam sayı) sütunlarda:** eksik değerler bir önceki gözlemin değeri kullanılarak giderilmiştir.
* **Float (ondalıklı) sütunlarda:** eksik değerler lineer interpolasyon (linear imputation) yöntemi ile doldurulmuştur.
* **Date (tarih) sütunlarında:** eksik değerler bir önceki tarih değeri ile tamamlanmıştır.

Bu yöntemler sayesinde veri setindeki boş değerler tutarlı bir biçimde giderilmiş ve veri bütünlüğü korunmuştur.

Eksik değerlerin giderilmesinin ardından, sayısal değişkenler arasında ölçek farklılıklarını ortadan kaldırmak amacıyla **Min-Max Normalizasyon (MinMax Scaler)** uygulanmıştır. Bu işlem ile tüm sayısal değişkenler 0 ile 1 aralığına dönüştürülmüştür. Böylece mesafe hesaplamalarında her değişkenin eşit ağırlıkta katkı sağlaması hedeflenmiştir.

Sonuç olarak, veri seti:

* gereksiz ve yüksek korelasyonlu değişkenlerden arındırılmış,
* eksik değerleri giderilmiş,
* normalize edilmiş

ve modelleme aşaması için uygun hale getirilmiştir.

## 4. Modeling (Modelleme)

Bu çalışmada **k-En Yakın Komşu (kNN) tabanlı anomali tespiti yöntemi** kullanılmıştır.

Uygulanan adımlar:

1. Distance Matrix Calculate:  
   Normalize edilmiş sayısal değişkenler kullanılarak tüm gözlemler arasındaki Öklidyen mesafeler hesaplanmıştır.
2. Distance Matrix Pair Extractor:  
   Mesafe matrisi, her gözlem çifti için mesafe değerlerini içeren tabloya dönüştürülmüştür.
3. Rule-based Row Filter:  
   Gözlemin kendisiyle olan mesafeleri (Object1 = Object2) veri setinden çıkarılmıştır.
4. Sorter:  
   Her gözlem için mesafeler küçükten büyüğe sıralanmıştır.
5. Group Loop Start – Row Filter – GroupBy – Loop End:  
   Her gözlem için en yakın 10 komşu seçilmiş ve bu komşulara ait mesafelerin ortalaması hesaplanmıştır.  
   Bu değer her gözlem için bir **anomaly score (Mean Distance)** olarak tanımlanmıştır.
6. Rule Engine:  
   Anomaly score değerleri için bir eşik belirlenmiş ve bu eşik değerin üzerinde kalan gözlemler **Anomaly**, diğerleri **Normal** olarak etiketlenmiştir.

## 5. Evaluation (Değerlendirme)

### 5.1 Scatter Plot Yorumu (Anomaly – Feature ilişkisi)

Cysteic\_Acid ve Protein\_content değişkenleri kullanılarak oluşturulan scatter plot grafiği incelendiğinde, normal gözlemlerin belirli bir yoğunluk bölgesinde kümelendiği görülmektedir. Bu yoğun bölge, veri setindeki baskın ve tipik süt örneklerini temsil etmektedir.

Anomali olarak etiketlenen gözlemler ise bu yoğun kümenin dışında veya sınır bölgelerinde konumlanmaktadır. Özellikle:

* Düşük Protein\_content ve düşük Cysteic\_Acid değerlerine sahip bazı gözlemler,
* Yüksek Cysteic\_Acid değerlerine sahip ancak genel dağılımdan ayrışan örnekler,

anomaliler olarak belirlenmiştir.

Bu durum, mesafe tabanlı anomali tespit yönteminin, veri dağılımının genel yapısından sapma gösteren noktaları başarılı bir şekilde ayırt edebildiğini göstermektedir.

![A graph with blue and red dots

AI-generated content may be incorrect.]()

Scatter plot üzerindeki renkli ayrım, modelin yalnızca tek bir değişkene değil, çok boyutlu özellik uzayındaki uzaklığa dayalı olarak anomali belirlediğini ve bu anomalilerin görsel olarak da doğrulanabilir nitelikte olduğunu ortaya koymaktadır.

### 5.2 Histogram Yorumu (Mean Distance Dağılımı)

Mean(Distance) değerlerine ait histogram incelendiğinde, anomali skorlarının belirli bir aralıkta yoğunlaştığı ve sağa çarpık (right-skewed) bir dağılım gösterdiği görülmektedir.

Histogramda:

* Düşük Mean(Distance) değerleri normal gözlemleri temsil ederken,
* Yüksek Mean(Distance) değerlerine sahip az sayıda gözlemin dağılımın sağ kuyruğunda yer aldığı görülmektedir.

Bu sağ kuyrukta yer alan gözlemler, model tarafından anomali olarak etiketlenen örnekleri temsil etmektedir. Bu yapı, veri setinde çoğunluğun benzer özelliklere sahip olduğunu, ancak küçük bir kısmın bu genel yapıdan anlamlı ölçüde uzaklaştığını göstermektedir.

![A graph with a bar graph

AI-generated content may be incorrect.]()

Histogram üzerinde belirlenen eşik değerin üzerinde kalan gözlemlerin anomali olarak sınıflandırılması, istatistiksel olarak tutarlı bir ayırım sağlamıştır.

### 5.3 Genel Değerlendirme

Scatter plot ve histogram birlikte değerlendirildiğinde:

* Normal gözlemlerin yoğun bir küme oluşturduğu,
* Anomali gözlemlerin ise bu kümelerin dışında veya sınır bölgelerinde yer aldığı,
* Anomali skorlarının dağılımının sağ kuyrukta toplandığı,

gözlemlenmiştir.

Bu sonuçlar, uygulanan kNN tabanlı anomali tespit yönteminin veri setindeki olağandışı örnekleri başarılı bir şekilde belirlediğini göstermektedir.

Modelin ürettiği anomalilerin rastgele değil, veri dağılımına göre anlamlı ve yorumlanabilir örnekler olduğu görülmektedir.

## 6. Deployment (Sonuçların Kullanımı)

Bu çalışma sonucunda:

* Her süt örneği için bir **anomaly score** ve bir **etiket (Normal / Anomaly)** elde edilmiştir.
* Anomali olarak tespit edilen gözlemler kalite kontrol süreçlerinde öncelikli olarak incelenebilir.
* Model, yeni gelen süt verileri üzerinde de çalıştırılarak otomatik anomali tespiti yapılabilir.

KNIME iş akışı sayesinde süreç görsel ve tekrar edilebilir bir yapı kazanmıştır.

## 7. Genel Sonuç ve Kazanımlar

Bu çalışmada CRISP-DM metodolojisi izlenerek:

* Veri ön işleme,
* Mesafe tabanlı anomali tespiti,
* Görselleştirme ve yorumlama adımları başarıyla uygulanmıştır.

Elde edilen sonuçlar, süt verisi içinde normalden sapma gösteren örneklerin etkili bir şekilde belirlenebileceğini göstermiştir.  
Bu yöntem, etiketli veri gerektirmeden (unsupervised) çalışması nedeniyle gerçek hayat senaryolarında kullanılabilir niteliktedir.

A computer screen shot of a diagram

AI-generated content may be incorrect.