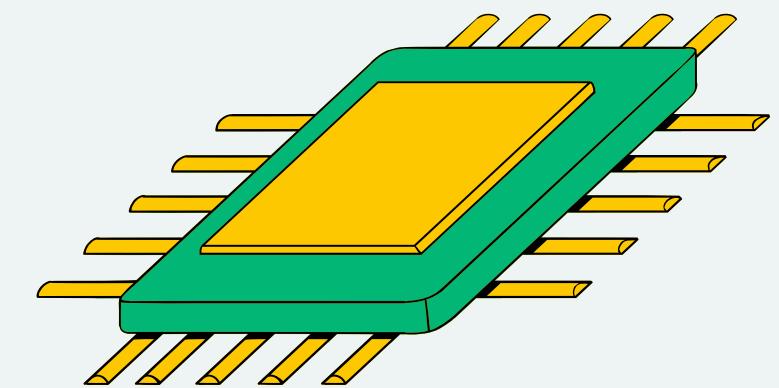


WATER QUALITY **ANALYSIS**



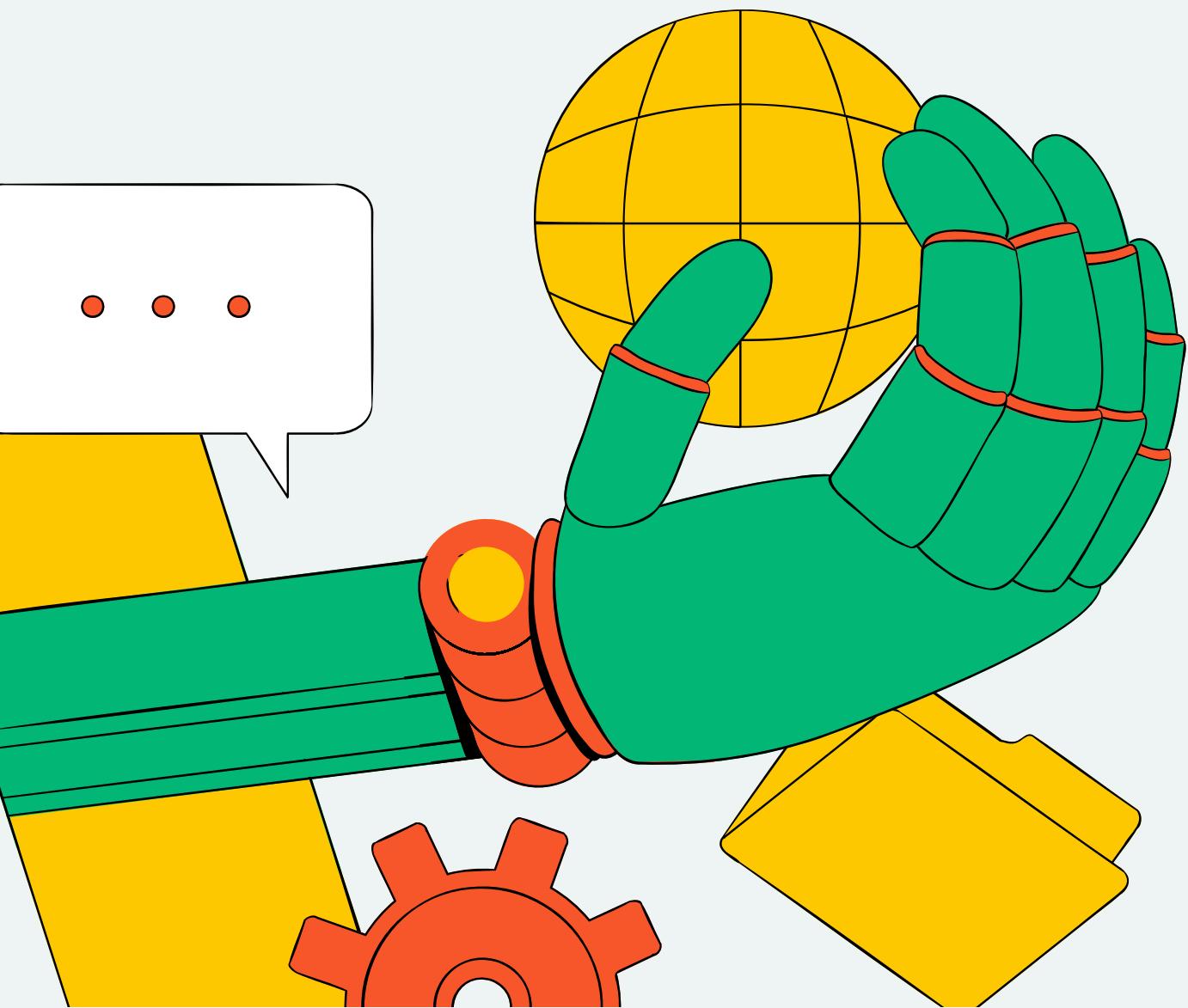
içindekiler



- Giriş
- Araştırma Soruları ve Bulgular
- Teknolojiler
- Motivasyon ve Arka Plan
- Kod Yapısı ve Veri Analizi Aşamaları
- Eksik Veri ve Metodoloji
- Aykırı Değerler
- Kategorikleştirme
- Machine Learning ve Sonuçlar



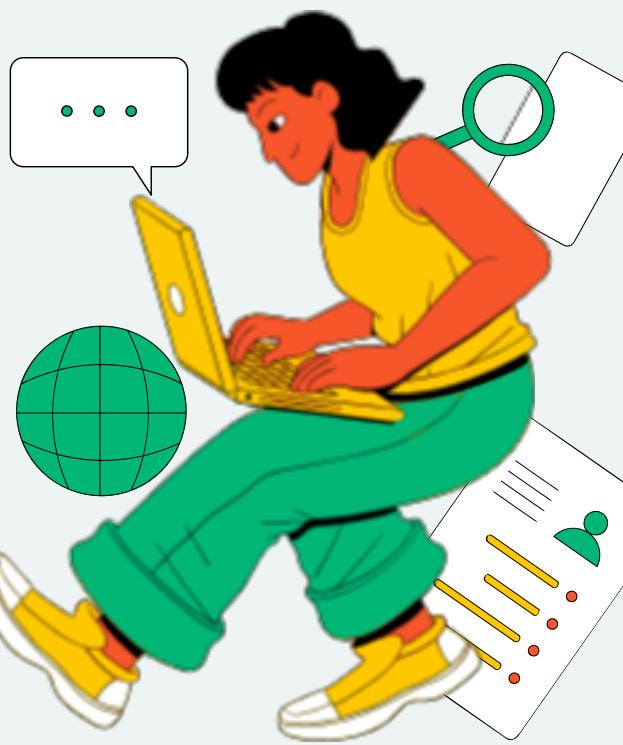
GİRİŞ



Veri Kümesi Bilgiler

- Veri seti , 3276 gözlem ve 10 değişkenden oluşmaktadır. Değişkenler arasında ph, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes, Turbidity, Potability bulunmaktadır. Potability değişken , suyun içilebilir (1) veya içilemez (0) olduğunu gösterir. Bazı değişkenlerde eksik veriler bulunmaktadır:
 - pH: 491 eksik
 - Sülfat: 781 eksik
 - Trihalometanlar: 162 eksik
 - Toplam eksik veri sayısı: 1434

ARAŞTIRMA SORULARI VE BULGULAR



1. Su Kalitesinin Potabilite Üzerindeki Etkisi:

- Su kalitesi özelliklerinin potabiliteye etkisini değerlendirmek için analiz yapıldı.
- Suyun belli özellikleri incelenerek suyun içilebilir olup olmadığı analiz edildi.

2. Su Kalitesi Özelliklerinin Birbirleriyle İlişkisi:

- Su özelliklerinin fiziksel ve kimyasal olarak birbirleriyle olan ilişkiler incelendi.
- Yapay zeka teknikleri kullanılarak bu ilişkiler analiz edildi.

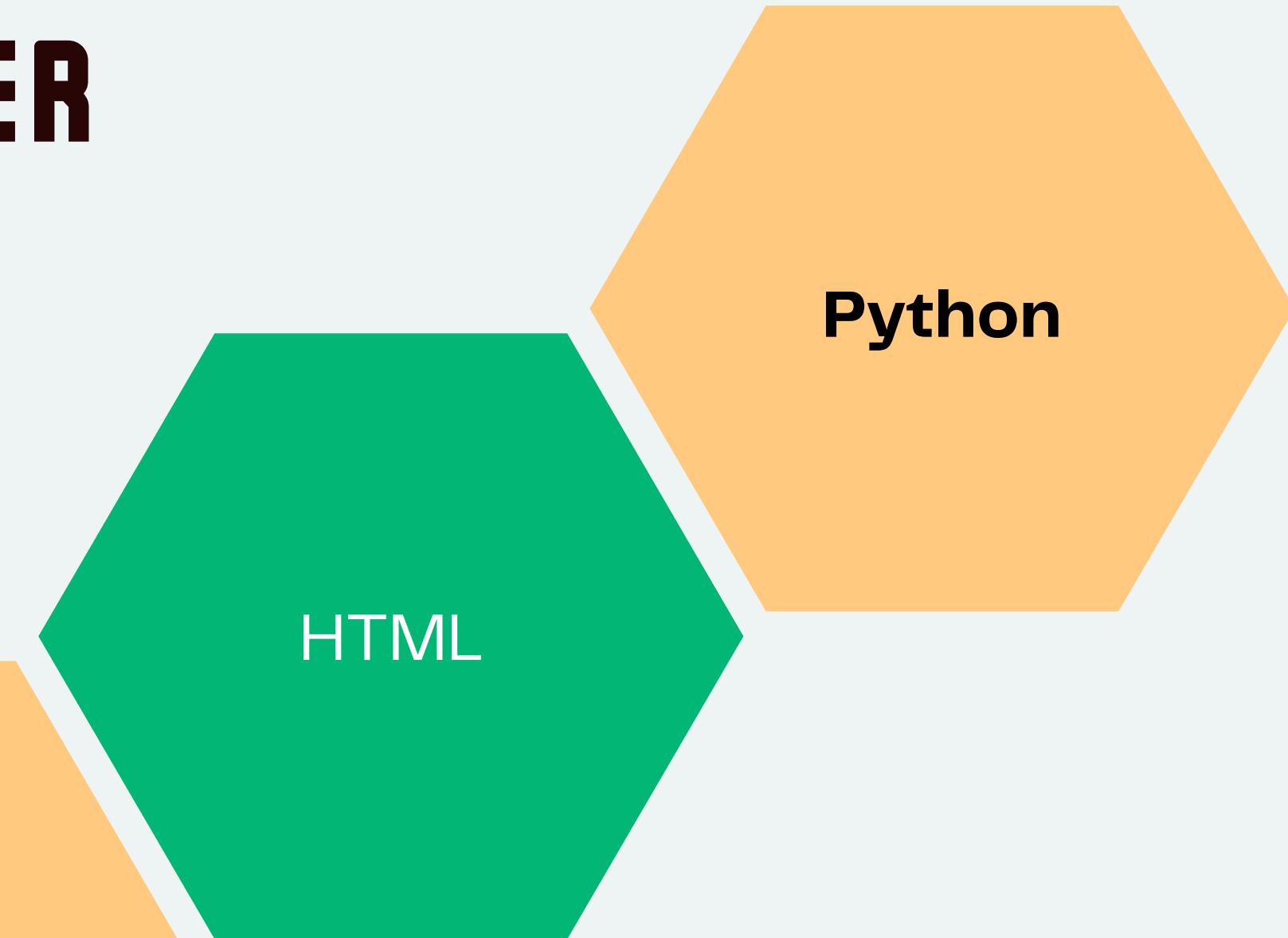
3. Potabilite Üzerindeki En Etkili Faktörlerin Belirlenmesi:

- Potabiliteyi etkileyen en önemli faktörleri belirlemek için özellikler arasındaki ilişkiler analiz edildi.
- Suyun içilebilirliğini etkileyen faktörler görselleştirildi.



TEKNOLOJİLER

Machine
Learning



Python



MOTİVASYON VE ARKA PLAN

- Su kalitesinin potabilite durumu üzerindeki etkisini değerlendirmek halkın sağlığı açısından hayatı öneme sahiptir.
- pH, sertlik, kimyasal bileşim gibi su özelliklerinin yönetimi ve güvenli içme suyu sağlanması açısından önemlidir.
- Bu analiz, sağlık kurumları, su temini ile ilgili kuruluşlar ve karar vericilere sağlık risklerini minimize etmek ve halkın sağlığını korumak için bilgi sağlar.



KOD YAPISI VE VERİ ANALİZİ AŞAMALARI

Kullanılan kütüphaneler arasında Flask, Pandas, NumPy, Matplotlib ve Plotly bulunmaktadır. Kod, Flask uygulamasını oluşturur ve veri analizi işlemlerini gerçekleştirir.

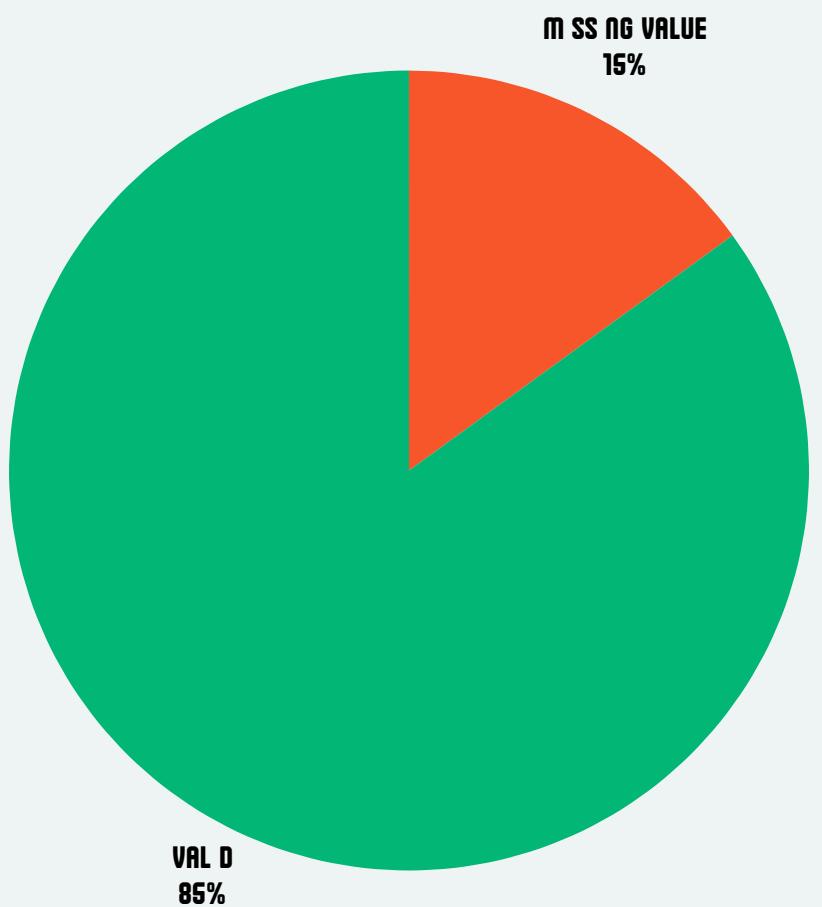
Veri analiz aşamaları şu adımları içerir:

- Veri yükleme
- Özeti istatistiklerin oluşturulması
- Eksik veri analizi
- Aykırı değer tespiti
- Veri görselleştirmesi

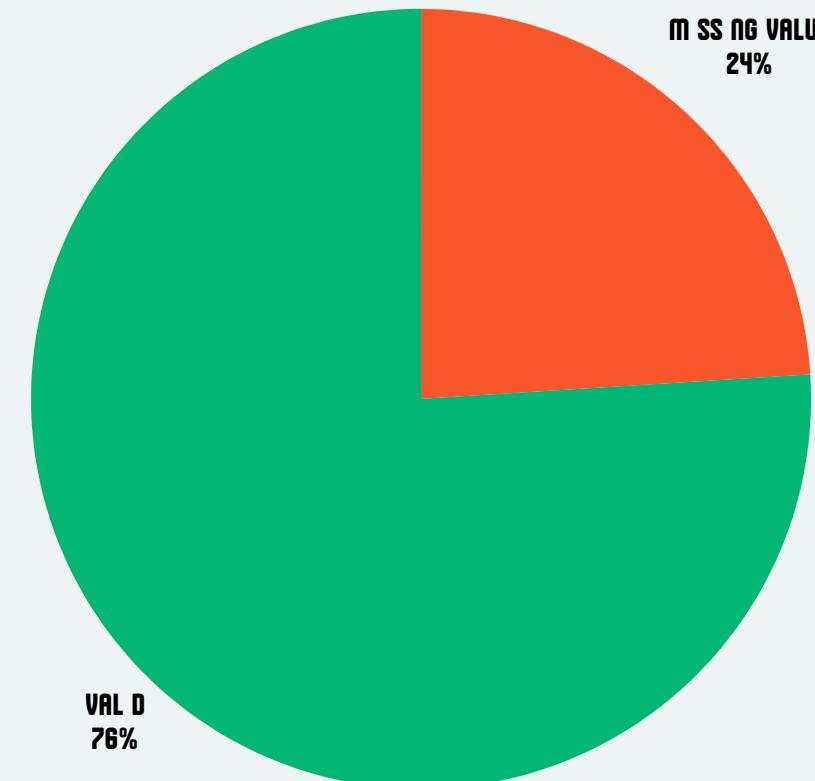


EKSİK VERİ

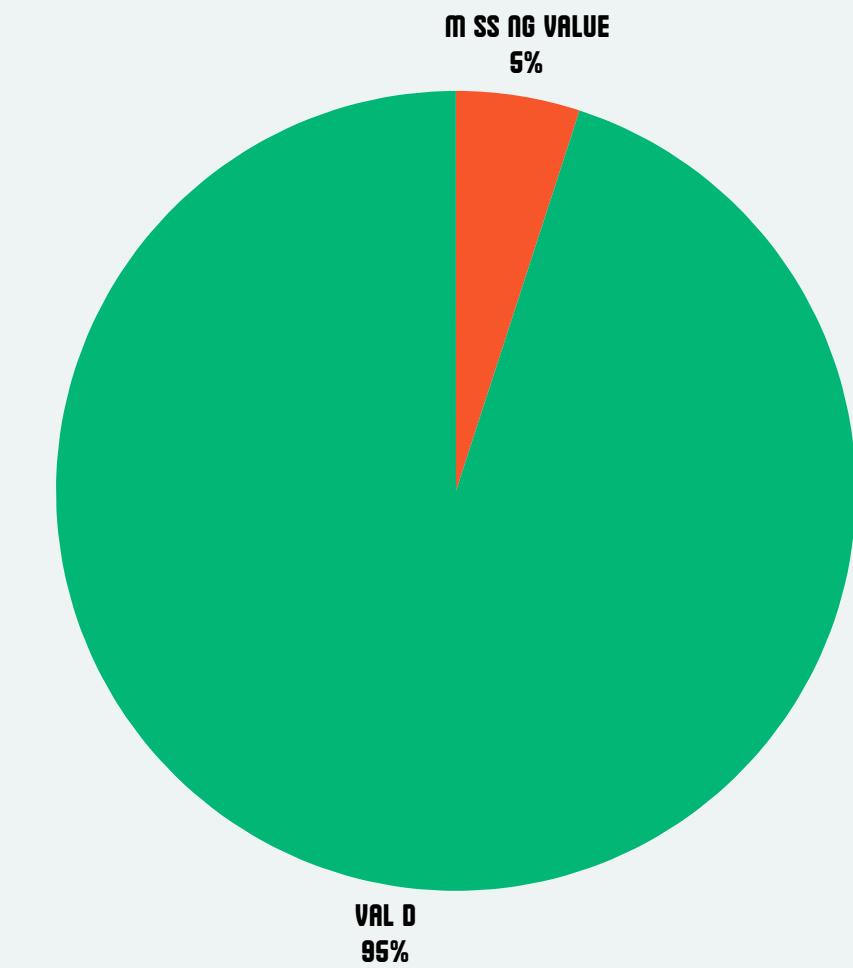
pH



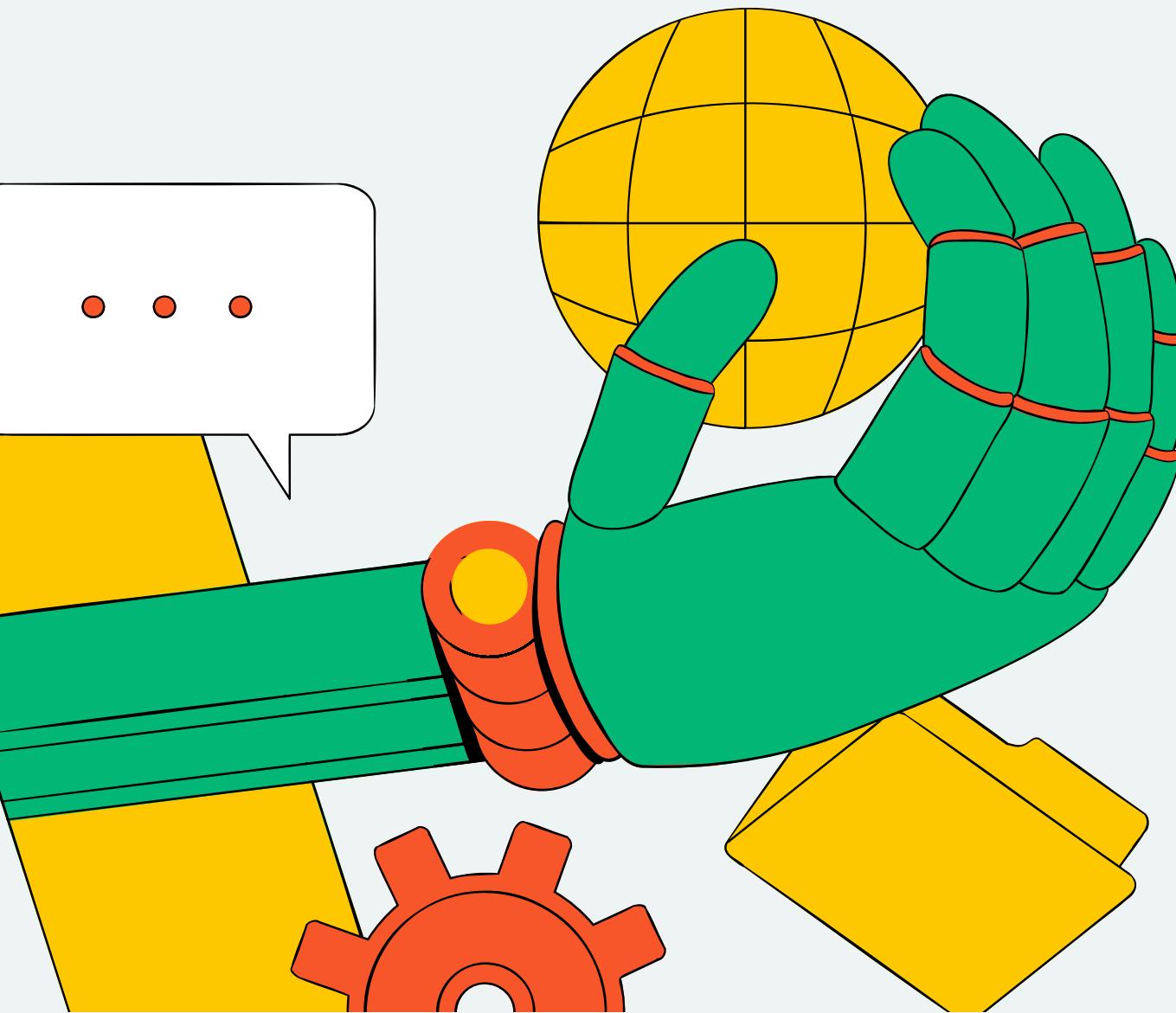
Sulfate



Tr halomethanes



METODOLOJİ



Kullanılan Algoritmalar:

- Logistic Regression
- Support vector machines (SVClassification)
- Decision Tree.

Veri Ön İşleme:

- pH ve Sülfattaki eksik veri bulunan satırlar drop edildi
- Trihalometanlar sınıf bazlı doldurma yöntemi kullanılarak mean ile dolduruldu ve daha doğru sonuçlar elde edildi.

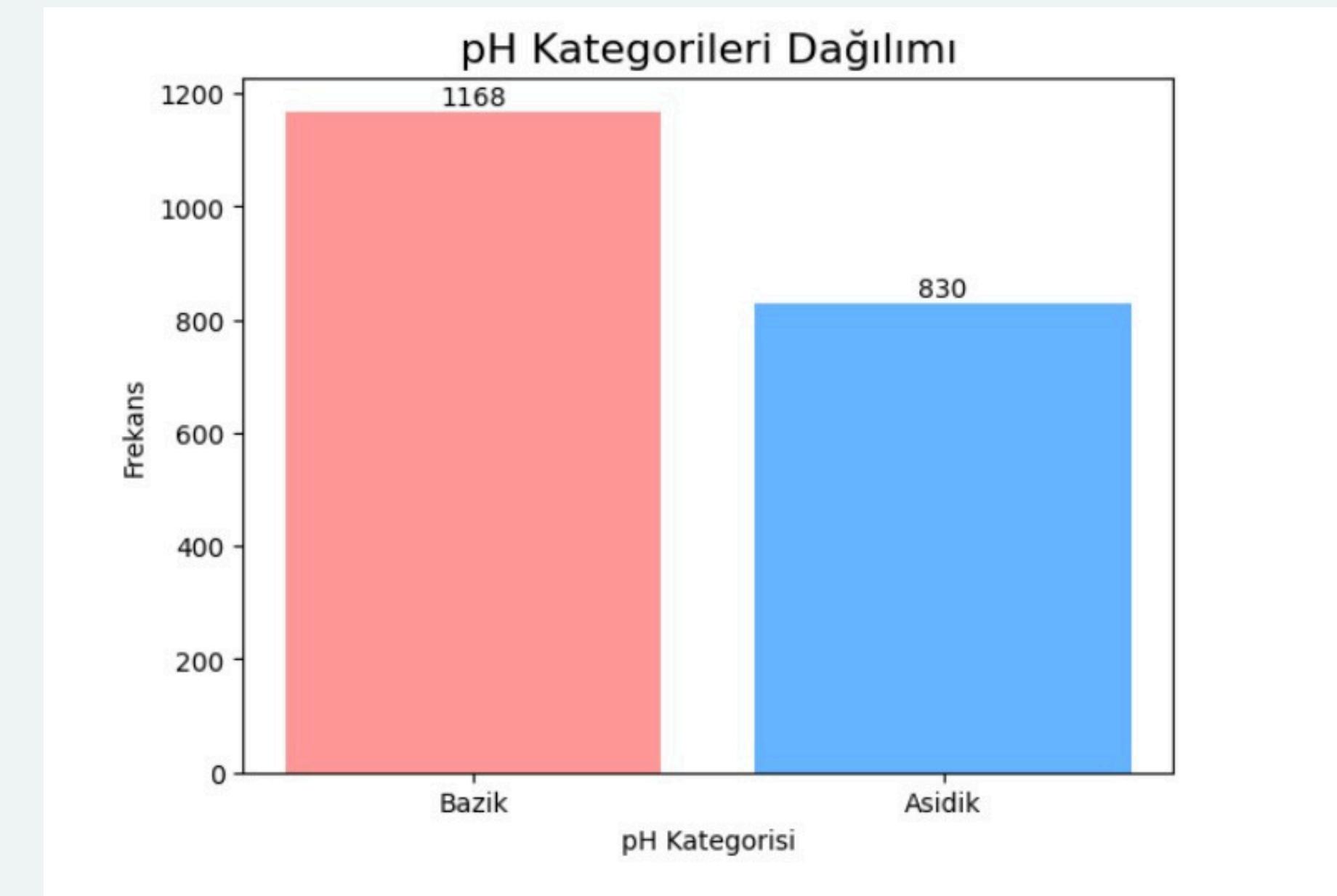
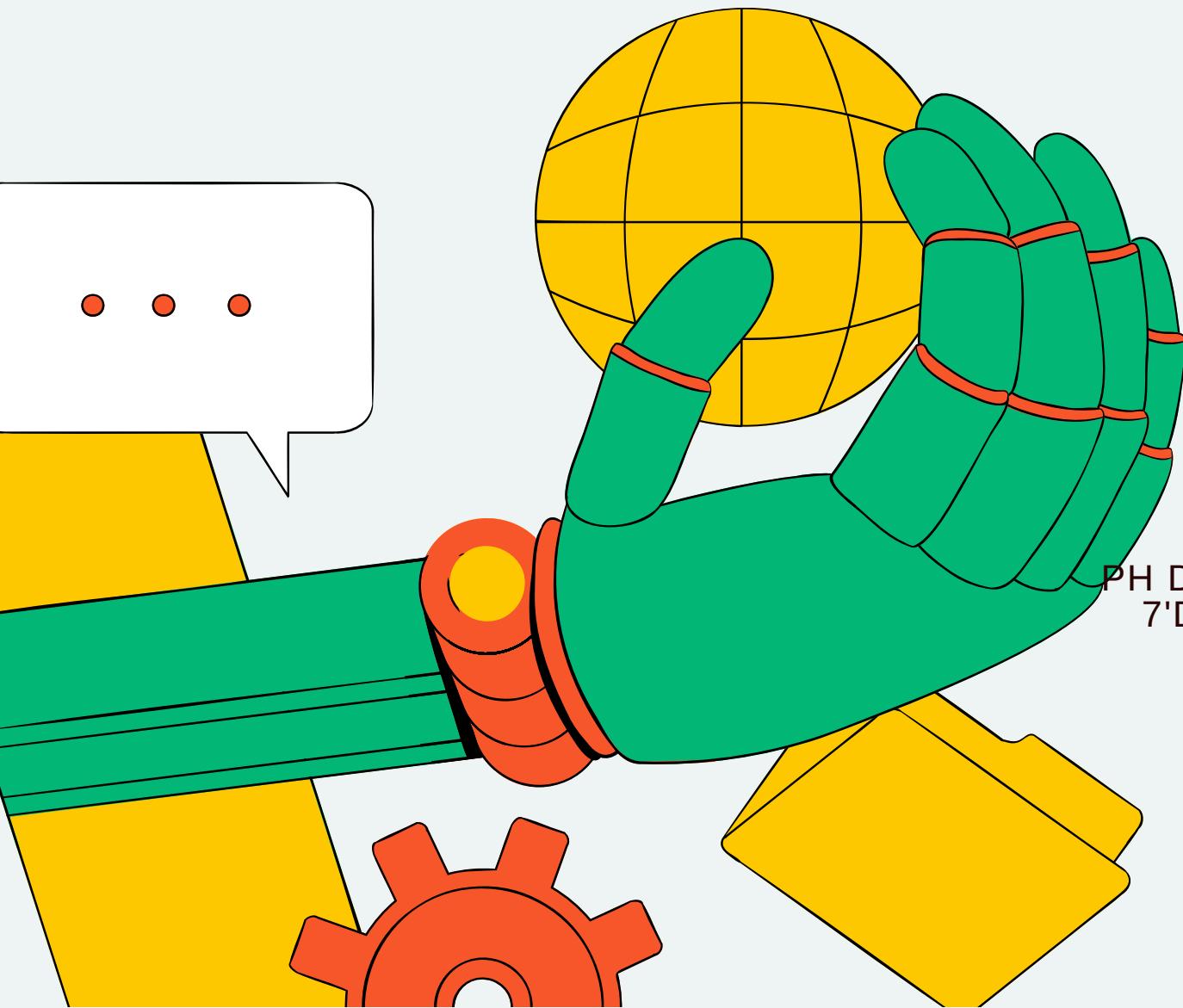
AYKIRI DEĞERLER

IQR yöntem ile aykırı değerleri tespit edilirken, Q1 ve Q3 çeyreklikler için 75 ve 25 yerine 95 ve 5 kullanılarak daha geniş bir aralık elde edildi. Bu sayede daha az sayıda veri noktası aykırı değer olarak tanımlandı.

Aykırı değerler tespit edildikten sonra, problemi çözmek için baskılama yöntemi kullanıldı.

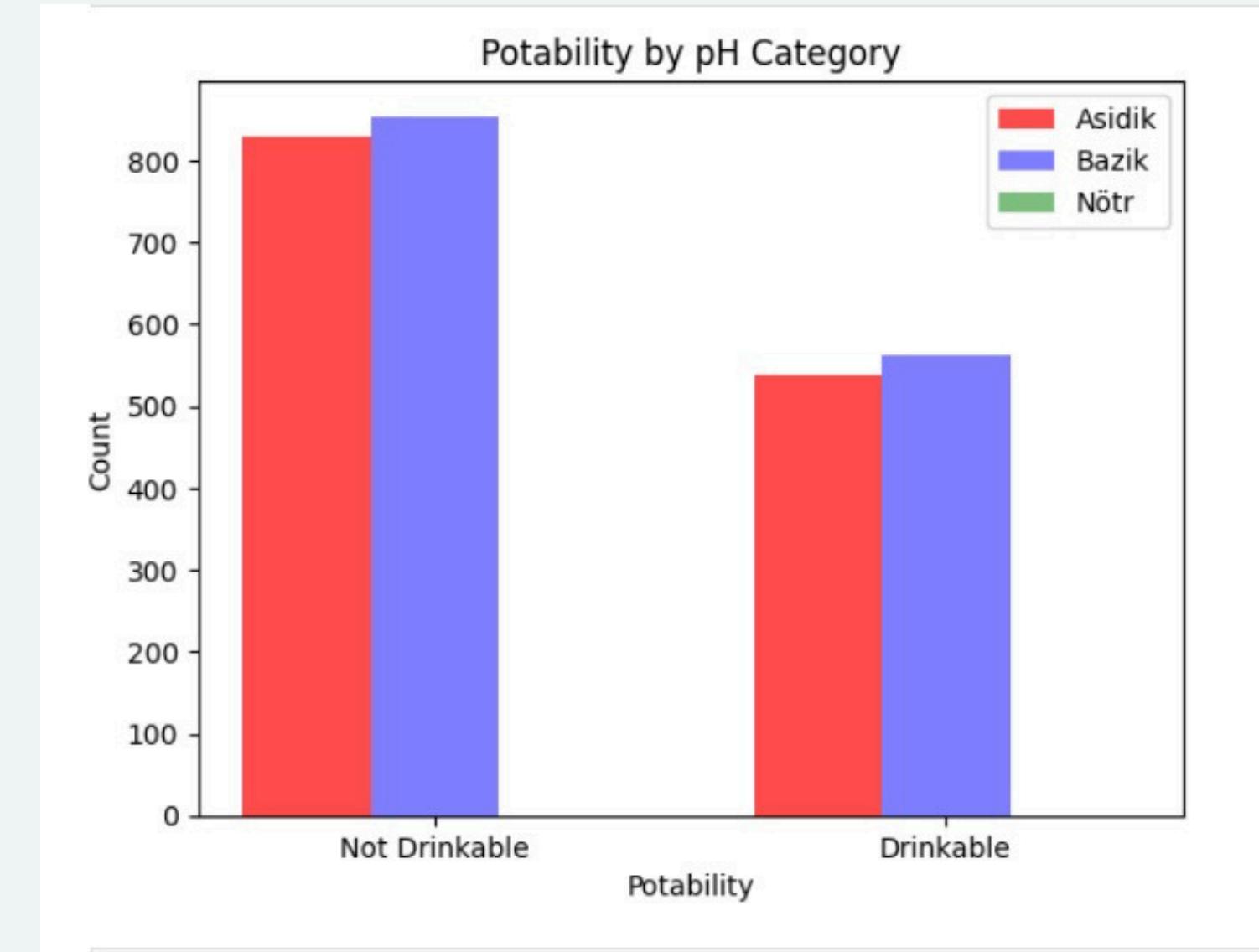
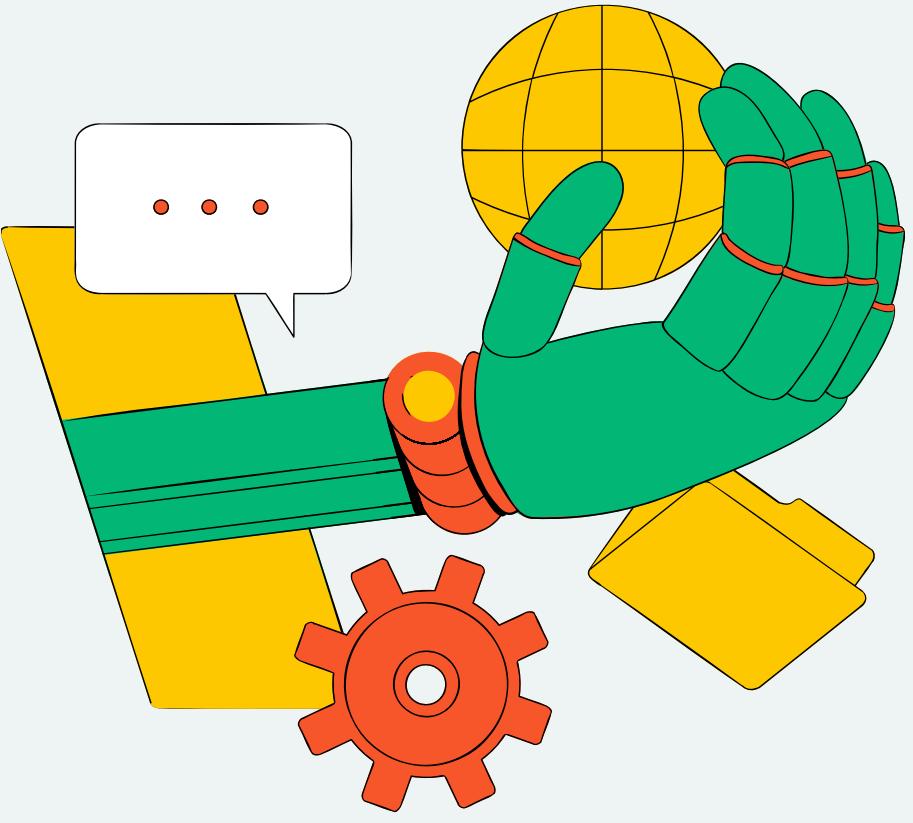


KATEGORİKLEŞTİRME



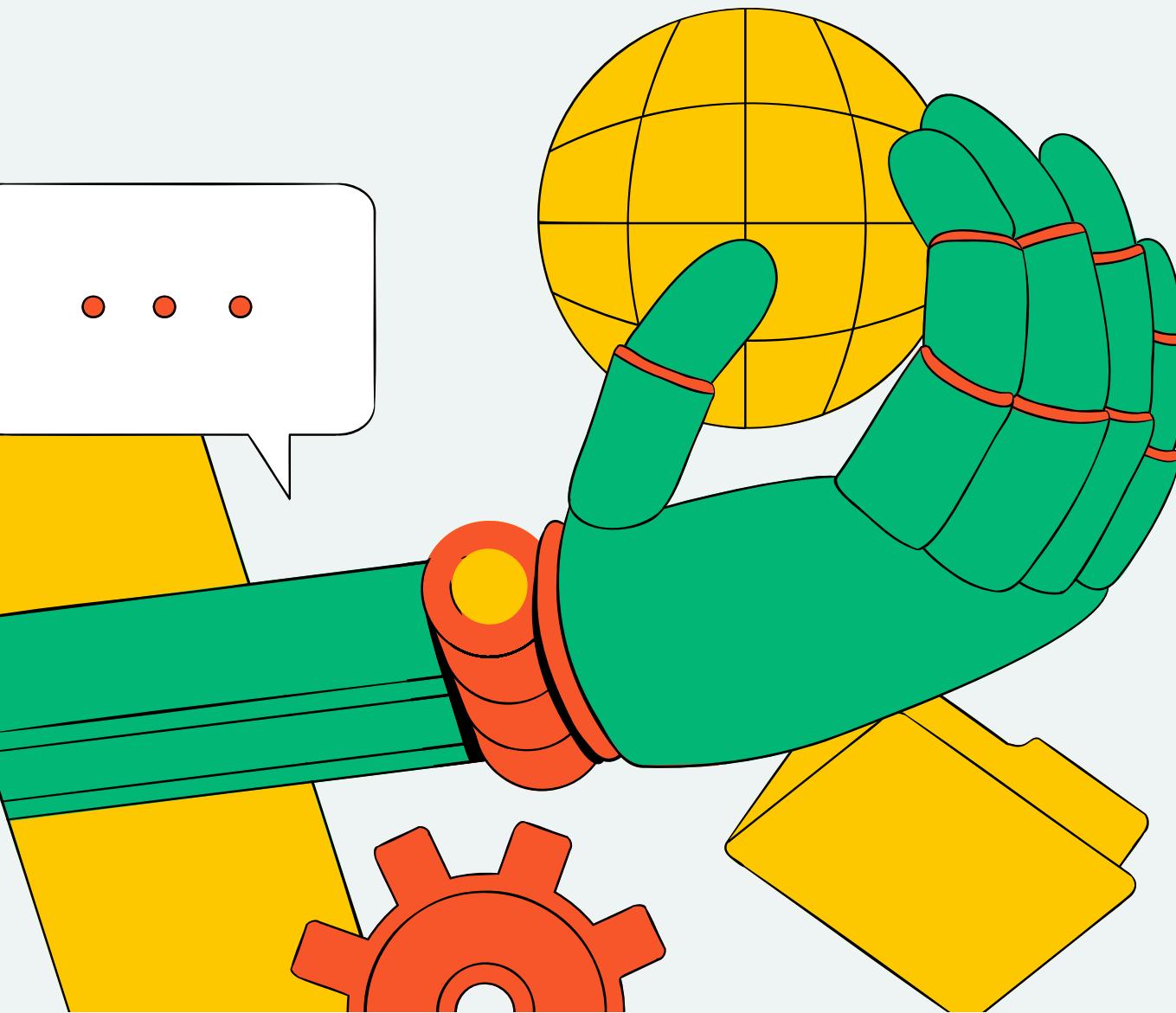
PH DEĞERLERİ BAZ ALINARAK SUYUN ASİDİK, BAZİK VEYA NÖTR OLUP OLMADIĞINI BELİRLİYORUZ. PH DEĞERİ 7'DEN KÜÇÜK OLAN SULARI "ASİDİK", 7'DEN BÜYÜK OLANLARI "BAZİK" VE TAM OLARAK 7 OLANLARI "NÖTR" OLARAK KATEGORİZE EDER.





her pH kategorisindeki suyun içilebilirlik durumunu incelemek için bir bar grafiği oluşturduk.

MACHINE LEARNING VE SONUÇLAR



Başlangıç Verisi:

- pH, sülfat ve trihalometanlarda eksik veriler vardı.
- Eksik veriler silinerek işlem yapıldı ancak bu algoritma performansını olumsuz etkiledi.

Algoritma Performansı:

- Logistic Regression: 57.7%
- SVC: 57.7%
- RandomForest: 69.5%
- XGBoost: 66.2%
- LightGBM: 69.1%
- CatBoost: 72.6%
- AdaBoost: 58.0%

Eksik verileri ortalama ile doldurulup, outliers üzerinde değişiklik yapıldığında sonuçlar:

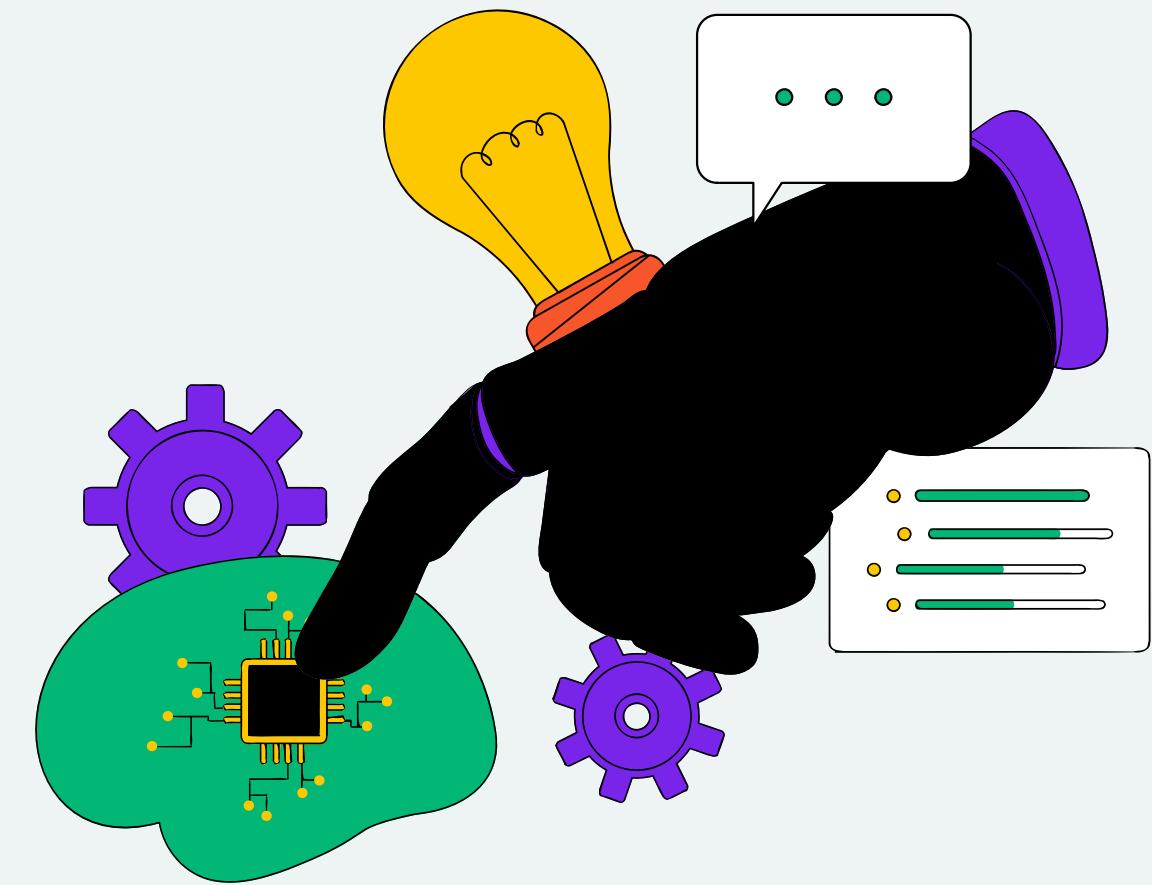


Algoritma Performansı:

- Logistic Regression: 63.1%
- SVC: 63.3%
- AdaBoost: 75.9%
- Random Forest: 76.3%
- XGBoost: 76.9%
- CatBoost: 77.7%
- LightGBM: 77.8%

Train ve Test:

- LightGBM Accuracy: 77.8%
- LightGBM Train Accuracy: 99.4%
- Random Forest Accuracy: 76.3%
- Random Forest Train Accuracy: 100%
- CatBoost Accuracy: 77.7%
- CatBoost Train Accuracy: 94.8%



Parametre Optimizasyonu Sonuçları:

- Random Forest: 80.4%
- XGBoost: Best Score: 79.4%
- LightGBM: Best Score: 79.5%
- CatBoost: Best Score: 80.1%



ÖZET

- pH, sertlik ve kimyasal bileşim gibi su kalitesi özellikleri potabiliteyi önemli ölçüde etkiler.

- CatBoost algoritması, suyun içilebilirliğini belirlemek için en doğru tahminler sağladı.

- Bu analiz ve bulgular, su kaynaklarının yönetimi ve güvenli içme suyu temini konusunda önemli bilgiler sunar.



dinLEDiĞiniz içiN TEŞEKKÜRLER

**BUSENUR GÖKLER
GÜLFEM ALBAYRAK
ALEYNA BARUT
DOĞA DURMAZ**

[HTTPS://GITHUB.COM/BUSE-NG/MIPOWERPROJECT](https://github.com/buse-nG/miPOWERPROJECT)