28.07.2025

Hava Kalitesi Veri Kümesinin Makine Öğrenmesi Yöntemleriyle Analizi

1. Giriş

Bu proje, hava kalitesinin sınıflandırılması ve mevcut veriler üzerinden gizli yapılarının analiz edilmesini amaçlamaktadır. Özellikle çevresel faktörler ve kirletici maddelerle ilişkili hava kalitesi sınıflarını (örneğin, İyi, Orta, Kötü, Tehlikeli) makine öğrenmesi algoritmaları yardımıyla tahmin etmek hedeflenmektedir. Ayrıca, verinin yapısal örüntülerini daha iyi anlamak için çeşitli kümeleme teknikleri de uygulanacaktır.

Proje kapsamında gözetimli öğrenme algoritmalarına (sınıflandırma), gözetimsiz öğrenme algoritmalarına (kümeleme), boyut indirgeme tekniklerine ve model optimizasyonuna odaklanılacaktır. Model performanslarını değerlendirmek için çapraz doğrulama (cross-validation) uygulanacaktır.

İlgili veri kümesine aşağıdaki bağlantıdan erişebilirsiniz:

https://www.kaggle.com/datasets/mujtabamatin/air-quality-and-pollution-assessment?resource=download

2. Veri Kümesi Açıklaması

Veri seti, güncellenmiş çevresel ölçümleri içermekte ve hava kirliliğiyle ilgili çeşitli faktörleri kapsamaktadır. Sütunlar aşağıdaki gibidir:

- Temperature: Derece Celsius cinsinden sıcaklık.
- Humidity: Yüzde cinsinden nem oranı.
- PM2.5: 2.5 mikron ve altı partikül madde konsantrasyonu.
- PM10: 10 mikron ve altı partikül madde konsantrasyonu.
- NO2, SO2, CO: Azot dioksit, kükürt dioksit ve karbon monoksit konsantrasyonları.
- Proximity_to_Industrial_Areas: Sanayi bölgelerine yakınlık ölçüsü.
- Population_Density: Nüfus yoğunluğu.
- Air Quality: Kategorik hedef değişken (Good, Moderate, Poor, Hazardous).

3. Projenin Amaçları

- Hava kalitesi sınıflarını tahmin edebilecek bir sınıflandırma modeli geliştirmek.
- Veri seti içerisinde doğal kümeleri ve desenleri kümeleme ile keşfetmek.
- Boyut indirgeme tekniklerini kullanarak veriyi daha anlaşılır hale getirmek.
- Hiperparametre ayarlaması ve düzenlileştirme ile modelleri optimize etmek.
- Model performanslarını karşılaştırmalı olarak analiz etmek.
- Elde edilen sonuçları uygun grafik ve tablolarla görselleştirmek.

4. Metodoloji

Bu projede aşağıdaki adımlar sistematik bir şekilde izlenecektir:

1. Veri Hazırlığı ve Keşifçi Veri Analizi (EDA)
2. Özellik Mühendisliği ve Ön İşleme
3. Boyut İndirgeme Teknikleri (PCA, LDA, t-SNE)
4. Sınıflandırma Modelleri Eğitimi ve Değerlendirme
5. Kümeleme Teknikleri ile Veri Keşfi
6. Sonuçların Karşılaştırılması ve Yorumlanması

4.1. Veri Hazırlığı ve EDA (Keşifsel Veri Analizi)

- Eksik değerlerin kontrolü ve işlenmesi.
- Kategorik verilerin benzersiz değer analizleri.
- Sayısal değişkenlerin betimleyici istatistikleri (ortalama, medyan, std).
- Korelasyon analizi ile değişkenler arası ilişkilerin incelenmesi.
- Aykırı değerlerin kutu grafikleri ile tespiti.
- Hedef değişkenin dağılımı ve dengesizlik analizi.

4.2. Özellik Mühendisliği ve Ön İşleme

- Kategorik değişkenlerin etiketlenmesi (Label Encoding / One-Hot Encoding).

- Sayısal verilerin standardizasyonu (StandardScaler).

- Gerekirse özellik seçimi veya çıkarımı yapılması.

4.3. Boyut İndirgeme Teknikleri

- PCA: Özelliklerin varyansını koruyarak boyut indirgeme.

- LDA: Sınıf ayrımını maksimize ederek indirgeme.

- t-SNE: Özellikle görselleştirme için yüksekten düşük boyuta geçiş.

4.4. Sınıflandırma Modelleri ve Değerlendirme

Aşağıdaki sınıflandırıcılar eğitilecek ve 5-katlı çapraz doğrulama (cross-validation) ile değerlendirilecektir:

- Lojistik Regresyon (L1/L2 regularization)

- Karar Ağaçları

- Rastgele Orman

- Destek Vektör Makineleri (SVM)

- K-En Yakın Komşular (KNN)

Her model için hiperparametre ayarlamaları GridSearchCV veya RandomizedSearchCV ile yapılacaktır. Düzenlileştirme (regularization) etkisi, özellikle Lojistik Regresyon ve SVM üzerinde karşılaştırmalı analizle incelenecektir.

Performans metrikleri:

- Accuracy, Precision, Recall, F1-Skoru

- ROC AUC (multi-class için One-vs-Rest)

- Confusion Matrix, Classification Report

4.5. Kümeleme Analizi

- K-Means (elbow yöntemi ile en uygun k değeri)
- DBSCAN

Her kümeleme yöntemi, PCA veya t-SNE ile boyut indirgenmiş uzayda görselleştirilecektir.

5. Kod Yapısı ve Uygulama

Kod parçaları aşağıdaki modüllere bölünecektir:

- Veri Yükleme ve Ön İşleme
- Boyut İndirgeme ve Görselleştirme
- Sınıflandırma Modelleri Eğitimi ve Değerlendirme
- Kümeleme Analizi ve Görselleştirme
- Sonuçların Grafiklerle Sunumu

6. Beklenen Sonuçlar

- Hangi modelin en iyi sınıflandırma başarımını verdiği belirlenecek.
- Düzenlileştirme etkileri analiz edilecek.
- Boyut indirgeme tekniklerinin veriyi ayırma gücü görselleştirilecek.
- Kümeleme teknikleri ile verideki olası gruplar keşfedilecek.
- Tüm analizler bilimsel görseller ve tablolarla desteklenecektir.

Önemli: Bulgular 5 dk'lık bir sunum eşliğinde aktarılmalıdır. Kod ve Sunum paylaşılmalıdır.