📄 Teknik Rapor: Global Food Wastage Dataset

Amaç

📜Senaryo

Bu projenin amacı, küresel ölçekte gıda israfına ilişkin verileri analiz ederek israfın hangi gıda kategorilerinde ve ülkelerde daha yoğun olduğunu ortaya koymak ve çeşitli makine öğrenmesi modelleri kullanarak ekonomik kaybın tahmin edilmesini sağlamaktır. Bu amaçla, Kaggle'dan alınan Global Wastage veri seti incelenmektedir. Proje, görselleştirme ve modelleme teknikleriyle karar vericilere içgörü sunmayı hedeflemektedir.

Veri Seti: Global Food Wastage Dataset

Gözlem Sayısı: 5000

Değişken Sayısı: 8

Hedef Değişken: Economic Loss (Million $)

Temel Davranışsal Değişkenler: Country, Year, Food Category, Total Waste(Tons), Household Waste(%), Economic Loss (Millions $)

🔑Değişken Tanımları (Variable Definitions)

Country

Verinin ait olduğu ülke. Hangi ülkeye ait israf ve ekonomik kayıp verisinin gözlemlendiğini belirtir.

Year

Gözlemin ait olduğu yıl. Yıllık bazda israf ve ekonomik kayıpların takibi için kullanılır.

Food Category

Gıda türünü belirtir. Örneğin: Sebze, Tahıl, Et, Süt Ürünleri, vb. Bu değişken, israfın hangi kategorilerde yoğunlaştığını analiz etmek için kullanılır.

Total Waste (Tons)

Belirtilen ülke, yıl ve gıda kategorisine ait toplam gıda israfı miktarı (ton cinsinden). Ana sayısal analiz değişkenlerinden biridir.

Household Waste (%)

Hane halkı kaynaklı israf oranı. Toplam israfın ne kadarının evsel tüketimden kaynaklandığını yüzde (%) olarak gösterir.

Economic Loss (Million $)

İlgili israfın neden olduğu ekonomik kayıp, milyon Amerikan doları ($) cinsinden ifade edilir. Bu değişken, projede **hedef değişken (target variable)** olarak kullanılır.

Region (opsiyonel)

Eğer varsa, ülkenin ait olduğu coğrafi bölgeyi belirtir (Avrupa, Asya vb.). Segmentasyon analizleri için kullanılabilir.

GDP (opsiyonel)	Ülkenin yıllık gayrisafi yurt içi hasılası. Ekonomik göstergelerle israfın ilişkisini ölçmek için kullanılabilir.
Population (opsiyonel)	Ülke nüfusu. Kişi başı israf analizi yapılacaksa faydalı bir değişkendir.
Waste per Capita (kg)	Kişi başına düşen yıllık gıda israfı miktarı (kilogram cinsinden). Toplam israfı nüfusa oranlamak için kullanılabilir.

---

2. Yöntem

2.1. Veri Yükleme

Veri kümesi Google Drive üzerinden `.csv` formatında yüklenmiş ve `pandas` kütüphanesiyle okunmuştur.

2.2. Veri Ön İşleme

- Verideki genel bilgiler `.info()`, `.describe()` ve `.isnull().sum()` komutları ile incelenmiştir.
- Eksik veri içeren sütunlar tespit edilmiş ve analizde dikkate alınmıştır.

2.3. Keşifsel Veri Analizi (EDA)

- Gıda kategorilerine göre toplam israf `barplot` ile görselleştirilmiştir.
- Ülkelere göre ortalama ekonomik kayıplar analiz edilmiştir.
- Ülke – Gıda Kategorisi bazında kişi başına ortalama gıda israfı görselleştirilmiştir.
- Sayısal Değişkenler arası korelasyon yapılarak, görselleştirilmiştir.

2.4. Modelleme

Bağımlı değişken: **Economic Loss (Million $)**
Bağımsız değişkenler: Diğer sayısal sütunlar.

Kullanılan modeller:

- Linear Regression
- Random Forest
- XGBoost
- KNN
- Ridge Regression

Veri, eğitim ve test olarak ayrılmış ve modeller `GridSearchCV` kullanılarak hiperparametre optimizasyonu ile eğitilmiştir. `StandardScaler` ile ölçekleme uygulanmıştır (özellikle KNN ve LR için).

---

3. Veri

Kullanılan veri seti: `global_food_wastage_dataset.csv`
Sütunlar:

- `Country`
- `Year`
- `Food Category`
- `Total Waste (Tons)`
- `Household Waste (%)`
- `Economic Loss (Million $)`
- ve diğer türev sütunlar.

Veri, dünya çapında ülkelerden yıllar boyunca toplanan gıda israfı ve ekonomik kayıp bilgilerini içermektedir.

---

4. Model

Her modelin eğitimi sonrası test verisi üzerinde performans metrikleri hesaplanmıştır:

- **R^2 (Determination Coefficient)**
- **RMSE (Root Mean Squared Error)**
- **MAE (Mean Absolute Error)**

Ayrıca hata karşılaştırmaları için grafik çizilmiştir (bar chart).

---

5. Sonuçlar

Model	R^2 Skoru	RMSE	MAE
Linear Regression	0.XX	3336.50	2512.92
Random Forest	0.XX	3501.70	2614.13
XGBoost	0.XX	3539.21	2623.69
KNN	0.XX	5022.63	3885.23

En başarılı model, test seti üzerinde en düşük RMSE ve en yüksek R^2 skoruna sahip olan **Linear Regression** olarak görülmektedir.

---

6. Yorumlar

- Gıda israfı, özellikle bazı kategorilerde (örneğin sebzeler, tahıllar) oldukça fazladır.
- Hane halkı israf oranları ülkeden ülkeye değişiklik göstermektedir.
- Regresyon modelleri genel olarak iyi sonuçlar vermektedir, ancak model performansları veri kalitesi ve feature engineering ile daha da geliştirilebilir.
- Random Forest ve XGBoost gibi ensemble yöntemleri daha stabil sonuçlar sunmaktadır.
- Ridge Regression, overfitting'i kontrol altına almak için alternatif bir yöntem olarak denenmiştir.