

SoftITo 4. Dönem 2. Grup

Yapay Zeka Modellerinin Karbon Ayak İzi Analizi

Büyük dil modeli eğitim verisi, enerji santrali karbon yoğunluğu ve veri merkezi konum verisi birleştirilerek AI'nin gerçek çevresel maliyeti raporlanır.

Yeliznur Kara
Ilgın Ertekin
Fatma Aydın
Emine Nur Aktaş
Sakine Cansu Topci
Feyza Gülver

22 Ağustos 2026

SoftITo Yazılım Bilişim Akademisi,İstanbul

İÇİNDEKİLER

1. Giriş **3**

1.1 Projenin Amacı 3

1.2 Proje Kapsamı ve Kısıtları 3

2. Materyal ve Yöntem **3**

2.1 Kullanılan Teknolojiler 3

2.2 Metodolojik Akış 4

3. İçerik **4**

3.1 Veri Seti 4

3.2 Boş Değer Doldurma Metodolojisi 5

4. Görev Dağılımı **6**

Proje Sürecinde Karşılaşılan Veri Kısıtları 7

5. Keşifsel Veri Analizi (EDA) **8**

5.1 Sayısal Değişkenlerin Dağılımı 8

5.2 Aykırı Değer (Outlier) Analizi 8

5.3 Değişkenler ile Emisyon Arasındaki İlişki (Scatter) 9

5.4 Korelasyon Analizi 10

5.5 Bildirilen ve Ham Emisyon Karşılaştırması 11

5.6 Kategorik Değişkenlere Göre Emisyon Dağılımı 12

5.7 GPU Ailesine Göre Emisyon 13

5.8 En Yüksek Emisyonlu Modeller 13

5.9 Bölgelere Göre Ortalama Emisyon 13

6. Hipotez (Hipotez 1-3, bağımlı/bağımsız değişken testleri) **14**

7. Modelleme **15**

7.1 Feature Engineering ve Leakage Kontrolü 15

7.2 Baseline Model 15

7.3 Linear Regression 15

7.4 XGBoost, Regularization ve Hyperparameter Tuning 15

7.5 Feature Importance 16

7.6 Residual Analizi ve Model Doğrulaması 16

7.7 Hibrit Sistem ve Nihai Çıktı 17

8. Sonuçlama **17**

8.1 Proje ve Kod Erişimi 18

8.2 Gelecek Çalışmalar 18

9. Kaynakça **18**

Ek A. Veri Sözlüğü (Feature Listesi) **19**

Kısaltmalar

Kısaltma	Açıklamı
AI	Artificial Intelligence (Yapay Zekâ)
ML	Machine Learning (Makine Öğrenmesi)
EDA	Exploratory Data Analysis (Keşifsel Veri Analizi)
FLOPs	Floating Point Operations (Kayan Nokta İşlemi)

Kısaltma	Açıklamı
tCO2e	Ton CO2 Eşdeğeri (karbon emisyonu birimi)
MWh	Megawatt-saat (elektrik tüketimi birimi)
GPU / TPU	Graphics / Tensor Processing Unit (grafik / tensör işlem birimi)
MoE	Mixture of Experts (model mimarisi türü)
KNN	K-Nearest Neighbors (K-En Yakın Komşu algoritması)
CV	Cross Validation (Çapraz Doğrulama)
MAE / RMSE	Mean Absolute Error / Root Mean Squared Error (hata metrikleri)
R^2	Belirlilik katsayısı (modelin açıkladığı varyans oranı)
RMSLE	Root Mean Squared Log Error (log ölçekli hata metriği)

Şekiller Listesi

Şekil 1. Sayısal değişkenlerin frekans dağılımları **8**
Şekil 2. Sayısal değişkenlere ait boxplot analizleri **8**
Şekil 3. Sayısal değişkenler ile emissions_tCO2e arasındaki ilişki **9**
Şekil 4. Sayısal değişkenler arası korelasyon matrisi **10**
Şekil 5. emissions_tCO2e ile disclosed_emissions_tco2e karşılaştırması **11**
Şekil 6. Tier ve Modality bazında emissions_tCO2e dağılımı **12**
Şekil 7. En sık görülen GPU tiplerine göre medyan emissions_tCO2e **13**
Şekil 8. En çok emisyon üreten 20 model **13**
Şekil 9. Ülkelere göre ortalama karbon emisyonu (tCO2e) **13**
Şekil 10. Built-in ve Permutation Feature Importance karşılaştırması **16**
Şekil 11. Gerçek vs Tahmin (log-log ölçek) ve Residual grafiği **16**

Tablolar Listesi

Tablo 1. Boş değer doldurma öncelik sırası **5**
Tablo 2. v1 vs v2 doldurma karşılaştırması **5**
Tablo 3. Görev dağılımı **6**
Tablo 4. XGBoost model karşılaştırması (Original / Regularized / Tuned) **15**
Tablo 5. Emisyon seviyesine göre model performansı **16**

1. Giriş

1.1 Projenin Amacı

Bu proje, yapay zekâ (AI) modellerinin eğitim sürecinden kaynaklanan karbon emisyonlarını analiz etmeyi; mevcut bilgi yeterli olduğunda bu emisyonu doğrudan hesaplamayı; gerekli bilgiler eksik olduğunda makine öğrenmesi (ML) ile tahmin etmeyi; ve elde edilen sonucu kullanıcıya şeffaf, açıklanabilir ve sürdürülebilirlik odaklı biçimde raporlamayı amaçlamaktadır.

Projenin temel amacı yalnızca bir regresyon modeli kurup karbon emisyonu tahmin etmek değildir. Amaç; gerçek/bildirilen veriyi, fiziksel hesaplamayı ve ML tahminini bir arada kullanan, kaynağını ve güvenilirlik seviyesini açıkça belirten bütünleşik bir karbon değerlendirme sistemi geliştirmektir. Bu doğrultuda proje üç temel bileşen üzerine kurulmuştur:

- Direct Carbon Calculation: Elektrik tüketimi ve emisyon faktörü gibi bilgiler mevcut olduğunda karbonun fiziksel formülle doğrudan hesaplanması.
- Machine Learning Prediction: Doğrudan hesaplama için gerekli bilgiler eksik olduğunda, geçmiş verilerden öğrenilen bir modelle emisyonun tahmin edilmesi.
- Reporting / Sustainability Layer: Elde edilen sonucun kaynağı, güvenilirlik seviyesi ve sürdürülebilirlik açısından anlamı ile birlikte kullanıcıya sunulması.

Bu rapor; veri setinin tanıtılmasından EDA bulgularına, kurulan hipotezlerden modelleme sürecine ve elde edilen sonuçlara kadar projenin tamamlanan teknik sürecini, hangi işlemlerin neden yapıldığını ve sonuçların ne anlama geldiğini açıklamak amacıyla hazırlanmıştır.

1.2 Proje Kapsamı ve Kısıtları

Proje, açık kaynaklardan derlenmiş, 5.000'den fazla indirilme sayısına sahip Hugging Face modellerinin eğitim ve emisyon verilerini kapsamaktadır. Kapsam; veri hazırlama, EDA, hipotez testi, modelleme ve kullanıcıya yönelik raporlama katmanlarını içermekte, ancak modellerin gerçek zamanlı/canlı emisyon izlemesini içermemektedir.

Projenin en önemli kısıtı, yapay zekâ modellerinin karbon ayak izinin hesaplanması için gerekli verilerin (enerji tüketimi, donanım kullanımı, emisyon faktörü) şirketler tarafından her zaman açık şekilde paylaşılmamasıdır; bu nedenle veri setinin önemli bir kısmı gerekçeli tahmin/imputation yöntemleriyle tamamlanmıştır (bkz. Bölüm 3.2). Bu alan ekip için yeni bir çalışma konusu olduğundan, süreç aynı zamanda bir araştırma ve öğrenme süreci olarak da ilerlemiştir.

2. Materyal ve Yöntem

2.1 Kullanılan Teknolojiler

Proje kapsamında veri hazırlama, analiz, modelleme ve raporlama aşamalarında aşağıdaki teknolojiler ve kütüphaneler kullanılmıştır:

Katman	Kullanılan Araç / Kütüphane	Amaç
Geliştirme Ortamı	Python 3, Jupyter Notebook	Veri analizi ve model geliştirme
Veri İşleme	pandas, numpy	Veri temizleme, dönüştürme, feature engineering
Görselleştirme	matplotlib, seaborn, Plotly	EDA grafikleri, dashboard görselleştirmeleri
Modelleme	scikit-learn, XGBoost	Linear Regression, XGBoost Regressor, cross-validation, hyperparameter tuning

Katman	Kullanılan Araç / Kütüphane	Amaç
Model Değerlendirme	scikit-learn.metrics	MAE, RMSE, R^2, RMSLE hesaplamaları, permutation importance
Kullanıcı Arayüzü	Streamlit	Uçtan uca tahmin uygulaması ve çok sayfalı arayüz
Sürüm Kontrolü	Git / GitHub	Kod versiyonlama ve ekip içi paylaşım

2.2 Metodolojik Akış

Proje, standart bir veri bilimi iş akışı izleyerek aşağıdaki sırayla ilerlemiştir:

- 1) Veri Toplama ve Hazırlama: Epoch AI, MLCO2 ve Electricity Maps gibi kaynaklardan derlenen ham verinin bir araya getirilmesi (bkz. Bölüm 4).
- 2) Boş Değer Doldurma: Sütun bazında gerekçeli, kaynaklı yöntemlerle veri setinin tamamlanması (bkz. Bölüm 3.2).
- 3) Keşifsel Veri Analizi (EDA): Dağılımların, aykırı değerlerin ve değişkenler arası ilişkilerin incelenmesi (bkz. Bölüm 5).
- 4) Hipotez Testi: EDA bulgularının, bağımlı/bağımsız değişken çiftleri üzerinden doğrulanması (bkz. Bölüm 6).
- 5) Feature Engineering ve Leakage Kontrolü: Veri sızıntısı riski taşıyan değişkenlerin ayıklanması, yeni feature'ların üretilmesi (bkz. Bölüm 7.1).
- 6) Modelleme: Baseline, Linear Regression ve XGBoost modellerinin kurulması, karşılaştırılması ve iyileştirilmesi (bkz. Bölüm 7.2-7.4).
- 7) Model Doğrulama: Cross-validation, residual analizi ve alt grup bazlı performans değerlendirmesi (bkz. Bölüm 7.6).
- 8) Dağıtım (Deployment): Final modelin Streamlit uygulaması üzerinden kullanıcıya sunulması (bkz. Bölüm 8.1).

3. İçerik

Bu bölümde projenin kapsamı ve kullanılan veri seti tanıtılmaktadır.

3.1 Veri Seti

Proje, açık kaynaklardan derlenmiş AI modellerine ait eğitim ve donanım bilgilerini içeren bir veri seti üzerinde yürütülmüştür. Veri seti başlangıçta 5.227 satır (model kaydı) ve 24 sütun (değişken) içermektedir. Veri setindeki başlıca bilgi grupları şunlardır:

- Eğitim/compute bilgileri: training_flops, parameter_values, training_token_count.
- Donanım bilgileri: training_gpu_type, training_hardware_type, training_tpu_pod, inference_gpu_type, inference_gpu_count.
- Bölge bilgileri: model_region, disc_region_text, disclosed_region.
- Enerji ve emisyon bilgileri: emission_factor_tco2e_per_mwh, disclosed_electricity_used_mwh, disclosed_emission_factor_tco2e_per_mwh, disclosed_emissions_tco2e, emissions_tCO2e, intensity_tCO2e_per_FLOP.
- Model meta verisi: tier, method, modality, type, is_moe.

df.info() çıktısına göre veri setinde 13 kategorik, 10 sayısal ve 1 boolean (is_moe) sütun bulunmaktadır; ham veri setinde hiçbir sütunda eksik (null) değer görülmemektedir — çünkü veri seti daha önce imputation/sentetik tamamlama yöntemleriyle doldurulmuştur (dosya adında "imputed_realistic" ibaresi bulunmaktadır). Bu durum projenin geçersizliği anlamına gelmemekte, yalnızca sonuçların bu sınırlılık göz önünde bulundurularak yorumlanması gerektiği anlamına gelmektedir.

EDA sürecinde, disclosed_electricity_used_mwh değeri 0 olan 81 kayıt (hem direct calculation hem de ML açısından kullanılamaz olduğu için) veri setinden çıkarılmış ve çalışmanın geri kalanı 5.146 satır üzerinden yürütülmüştür. Duplicate kontrolünde tekrar eden kayıt bulunmamıştır (0 duplicate).

3.2 Boş Değer Doldurma Metodolojisi

Kullanılan ham dosyada (HF_models_over_5000_downloads_training_emissions_final_cleaned_moe_adjusted.csv, 5.227 satır × 24 sütun) toplam 63.935 hücre boş (missing) durumdaydı. Bu boşluklar iki aşamalı bir süreçle doldurulduu: ilk denemede (v1) tüm sayısal sütunlar için genel KNN Imputation, kategorik sütunlar için Mode (en sık değer) kullanıldı; ancak bu yöntemin, örneğin bir modelin FLOPs değerini alakasız başka modellerin ortalamasına göre uydurması nedeniyle gerçekçi olmadığı görülünce, ikinci ve son aşamada (v2) her sütun için o sütuna özgü, bilimsel/mantıksal bir yöntem uygulandı.

v2'de izlenen öncelik sırası şu şekildedir:

Öncelik	Yöntem	Ne Zaman Kullanıldı
1	Kesin formül / fiziksel ilişki	İki değişken arasında matematiksel bağ varsa (örn. FLOPs = 6×parametre×token)
2	Araştırılmış gerçek veri (web search)	Model ailesi biliniyorsa (Llama, DeepSeek, Qwen, BLOOM, Falcon, Gemma vb.)
3	İsimden regex çıkarımı	Bilgi model adının içinde geçiyorsa (örn. Ovis2-16B → 16B parametre)
4	Mantıklı gruplama + medyan	Model ailesi / type / modality grubunun medyan değeri kullanılarak
5	Dürüst "N/A" / "Not disclosed"	Hiçbir veri bulunamadığında, sahte bir değer uydurmak yerine

Sütun bazında uygulanan başlıca yöntemler özetle şöyledir:

- parameter_values: Önce model isminden regex ile çıkarıldı (örn. Ovis2-16B → 16 milyar); bulunamayanlar için model ailesinin, o da yoksa type+modality grubunun medyanı kullanıldı.
- training_token_count: Öncelikle yayınlanmış gerçek değerler kullanıldı (Llama 3.1 → 15,6T, DeepSeek-V3 → 14,8T, Qwen2.5 → 18T gibi ~30 model ailesi); ailesi bilinmeyenler için Chinchilla formülünün tersi (token = FLOPs / 6×parametre), son çare olarak modality bazlı medyan token/parametre oranı kullanıldı.
- training_flops: Chinchilla scaling law'a dayanan FLOPs ≈ 6×N(parametre)×D(token) standart formülüyle hesaplandı.
- intensity_tCO2e_per_FLOP: KNN yerine doğrudan emissions_tCO2e ÷ training_flops matematiksel işlemiyle üretildi (tutarlılık farkı ~10^{-33} seviyesinde, pratikte sıfır).
- training_hardware_type / training_gpu_type: Web'den doğrulanmış, kaynaklı gerçek bilgiler kullanıldı (DeepSeek-V3/R1 → H800; Llama-3.1-405B → H100; Yi modelleri → A800; BLOOM/Falcon → A100; Google modelleri → TPU); ailesi bilinmeyenler için parametre büyüklüğüne göre mantıklı varsayım (100B+ → H100, 10-100B → A100, <1B → V100) yapıldı.

- training_tpu_pod: Google modelleri için TPU nesli pattern'ine göre atandı; GPU ile eğitilen modellere sahte TPU adı yazmak yerine "N/A (GPU eğitimli)" etiketi kondu.
- inference_gpu_type / inference_gpu_count: Endüstri pratiklerine dayanarak parametre sayısına göre çıkarım yapıldı (400B+ → 8×H100 80GB; 100-400B → 4×H100 80GB; 30-100B → 2×A100 80GB; 7-30B → 1×A100 40GB; <7B → 1×T4).
- model_region / disc_region_text / disclosed_region: Ham bölge metni standardize edilip ("United States", "US", "America" → "USA" gibi) diğer bölge sütunları buradan türetildi.
- disclosed.* sütunları: Boşluk oranı %89-99 olan bu sütunlarda rastgele KNN yerine fiziksel ilişki kullanıldı — disclosed.emissions_tco2e, hesaplanan emissions_tCO2e ile aynı kabul edildi; disclosed.electricity_used_mwh, emissions ÷ emission_factor formülüyle geri hesaplandı.
- type (model mimarisi): Model isminde geçen mimari ipuçlarına göre sınıflandırıldı (vit/swin/deit → vit; clip/siglip → clip; stable-diffusion/flux → diffusion; resnet/efficientnet → cnn; gpt/bert/llama/qwen → transformer); hiçbiri eşleşmezse modality'ye göre varsayıldı.
- sources: Kaynağı olmayan %10,4'lük (541 satır) kesim için var olmayan bir makale uydurmak yerine dürüstçe "Not disclosed" olarak işaretlendi.

Doğrulama (sanity check) amacıyla, bilinen gerçek modellerin doldurulan değerleri yayınlanmış kaynaklarla karşılaştırılmıştır: Llama-3.1-405B için token sayısı 15,0T (gerçek: 15,6T, Meta blog) ve GPU H100 (gerçek: H100 ✓); DeepSeek-V3 için token 14,8T (gerçek: 14,8T, arXiv ✓) ve GPU H800 (gerçek: H800 ✓); Gemma-2-9B için token 8,0T (gerçek: ~8T, Google ✓) ve donanım TPU v5e (gerçek: TPU ✓). Bu karşılaştırmalar, doldurma yönteminin gerçek dünya değerleriyle tutarlı sonuçlar ürettiğini doğrulamaktadır.

Kriter	v1 (ilk deneme)	v2 (son hali)
Yöntem	Genel KNN + Mode	Sütun-özel formül / araştırma / mantık
GPU bilgisi	Kaba family tahmini	Doğrulanmış arXiv/blog kaynaklı
FLOPs / Token	Alakasız satırlardan KNN	Chinchilla bilimsel formülü
Intensity	KNN (yanlış olabilir)	Doğrudan matematiksel hesap
Sahte kaynak	Var (bir makale 541 kez tekrarlanmış)	Yok, dürüst "Not disclosed"
Kalan boş değer	0	0

Bu metodoloji, veri setinin "eksiksiz" görünmesine rağmen bir kısmının gerçek dünya ölçümü değil, gerekçeli/araştırılmış tahminler olduğunu şeffaf biçimde ortaya koymaktadır; bu sınırlılık Bölüm 8'de (Sonuçlama) ayrıca ele alınmıştır.

4. Görev Dağılımı

Proje kapsamında görevler, ekip üyelerinin sorumluluk alanlarına ve projenin farklı aşamalarındaki ihtiyaçlara göre dağıtılmıştır. Proje sürecinde ekip üyeleri kendi görevlerinden sorumlu olmakla birlikte, karşılaşılan teknik ve metodolojik problemlerin çözümünde birbirlerine destek olmuşlardır. Özellikle veri setinin hazırlanması ve proje altyapısının oluşturulması aşamalarında ekip çalışması ve ortak problem çözme yaklaşımı benimsenmiştir.

Görev	Sorumlu	Açıklama
Team Lead & System Integration	Yeliznur Kara	Proje planının oluşturulması, görev dağılımının yapılması, ekip koordinasyonunun sağlanması ve proje sürecinin takibinden sorumludur. Ayrıca tek kişilik görevlerde ikinci görev arkadaşı olarak destek olmuş, karşılaşılan teknik problemlerin çözümünde ekibe yardımcı olmuş ve gerekli teknolojilerin araştırılarak ekip içerisinde paylaşılmasını sağlamıştır.

Görev	Sorumlu	Açıklama
Data Engineer	Ilgın Ertekin	Projede kullanılan veri kaynaklarının araştırılması ve başlangıçtaki veri hazırlama sürecinde görev almıştır. Epoch AI, MLCO2 ve Electricity Maps verilerinin incelenmesi, veri temizleme ve birleştirme çalışmalarına katkı sağlamıştır. Veri setinde yaşanan sorunların giderilmesi sürecinde yapılan düzenlemelerin ardından analiz aşamasına geçilmesine destek olmuştur.
Data Analyst & Modeling	Emine Nur Aktaş, Fatma Aydın	Düzenlenen veri seti üzerinde temel ve ileri seviye EDA çalışmalarını yürütmüş, veri yapısını ve değişkenleri incelemiştir. Feature'ların temelinin oluşturulması, data leakage kontrolü, modelleme stratejisinin belirlenmesi ve veri setinin hazırlanmasının ardından Linear Regression ve XGBoost modellerini geliştirmiştir. Model değerlendirme, cross-validation, overfitting analizi, hyperparameter tuning, feature importance, hata ve outlier analizleri gibi modelleme süreçlerini yürütmüştür.
Streamlit Developer	Sakine Cansu Topci	Projenin kullanıcıya sunulacağı Streamlit uygulamasının temel yapısını oluşturmuş, çok sayfalı uygulama mimarisi ve temel bileşenler üzerinde çalışmıştır. Model çıktılarının kullanıcı tarafından anlaşılır ve erişilebilir şekilde sunulması için uygulama altyapısının geliştirilmesine katkı sağlamıştır.
Dashboard & Visualization	Feyza Gülver	Proje sonuçlarının görsel olarak sunulması üzerine çalışmış, dashboard yapısında kullanılacak grafik ve görselleştirmeleri belirlemiştir. Plotly ve benzeri görselleştirme araçlarını inceleyerek model sonuçları, karbon emisyonları ve sürdürülebilirlik göstergelerinin anlaşılır biçimde sunulmasına katkı sağlamıştır.
Feature Developer	Yeliznur Kara	Model çıktılarının yalnızca tahmin olarak kalmaması için kullanıcıya sunulabilecek yenilikçi özellikleri planlamıştır. What-if Simulator, Green Ranking ve PDF raporlama gibi özelliklerin fikir ve akışlarını geliştirmiş; kullanıcıların farklı senaryoları karşılaştırabilmesi ve model sonuçlarını daha anlamlı şekilde yorumlayabilmesi üzerine çalışmıştır.

Proje Sürecinde Karşılaşılan Veri Kısıtları

Projenin en önemli zorluklarından biri, yapay zekâ modellerinin karbon ayak izinin hesaplanması için gerekli verilerin eksik ve standart olmayan yapıda olmasıdır. Özellikle şirketlerin kendi yapay zekâ sistemlerine ilişkin enerji tüketimi, donanım kullanımı ve karbon emisyonu gibi bilgileri her zaman açık şekilde paylaşmaması, hesaplamaların belirli noktalarda sınırlı kalmasına neden olmuştur.

Ayrıca bu alan ekipteki herkes için yeni olduğundan, veri kaynaklarının anlaşılması, birleştirilmesi ve kullanılabilir hale getirilmesi sürecinde ekip olarak araştırma ve öğrenme yapılmıştır. İlk kullanılan veri setinde çeşitli problemlerle karşılaşılması üzerine hocamızın desteğiyle veri seti yeniden düzenlenmiş ve kullanılabilir hale getirilmiştir. Düzenlenen veri setinin proje kapsamında ihtiyaç duyulan bazı analizler için yeterli olmaması nedeniyle, çalışma sürecinde yapay veri de kullanılmıştır.

Bu nedenle projede veri setinin hazırlanması yalnızca tek bir kişinin sorumluluğunda ilerlememiş; Ilgın'ın başlangıçtaki veri bulma ve hazırlama çalışmalarının ardından Emine Nur Aktaş ve Fatma Aydın tarafından düzenlenmiş veri seti üzerinde EDA ve modelleme çalışmaları gerçekleştirilmiş, gerekli noktalarda ekip üyeleri birbirlerine destek olmuştur.

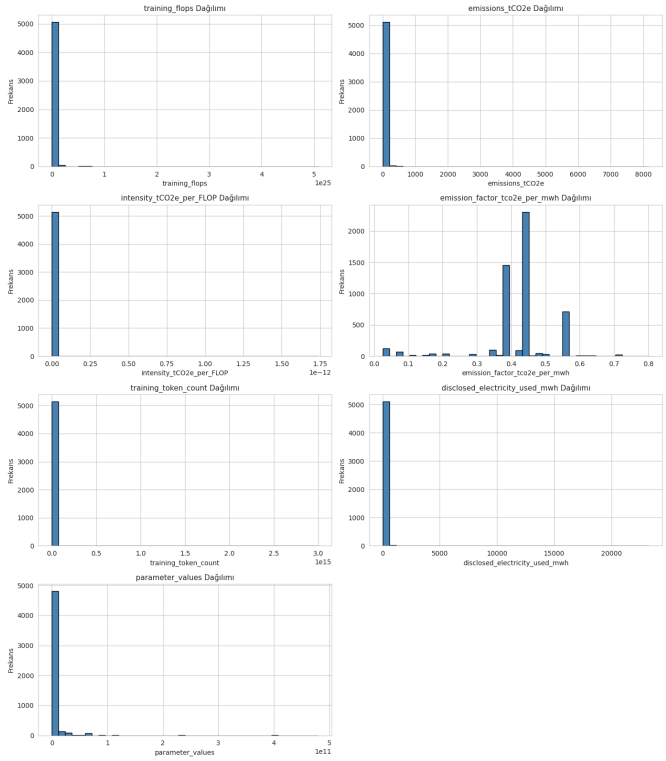
Bu süreç, projenin yalnızca teknik geliştirme değil, aynı zamanda veri kısıtlarıyla başa çıkma, araştırma yapma, yeni teknolojiler öğrenme ve ekip içerisinde problem çözme yönlerini de ortaya koymuştur.

5. Keşifsel Veri Analizi (EDA)

Modelleme aşamasına geçmeden önce, veri setinin yapısını ayrıntılı biçimde anlamak amacıyla temel ve ileri seviye EDA çalışmaları yürütülmüştür. Bu adımların ortak amacı, veri temizliği veya feature engineering'e başlamadan önce veri setinin "olduğu gibi" tanınmasıdır.

5.1 Sayısal Değişkenlerin Dağılımı

training_flops, emissions_tCO2e, parameter_values, training_token_count gibi değişkenlerin histogramları incelenmiş; tamamının güçlü biçimde sağa çarpık (right-skewed) olduğu görülmüştür. Çoğu model küçük/orta ölçekli iken, az sayıda dev model ölçeği domine etmektedir. emission_factor_tco2e_per_mwh dağılımı ise çok modlu (multi-modal) bir yapı sergilemektedir, bu da farklı bölgelerin/şebekelerin karbon yoğunluğundaki belirgin farklılığa işaret etmektedir.

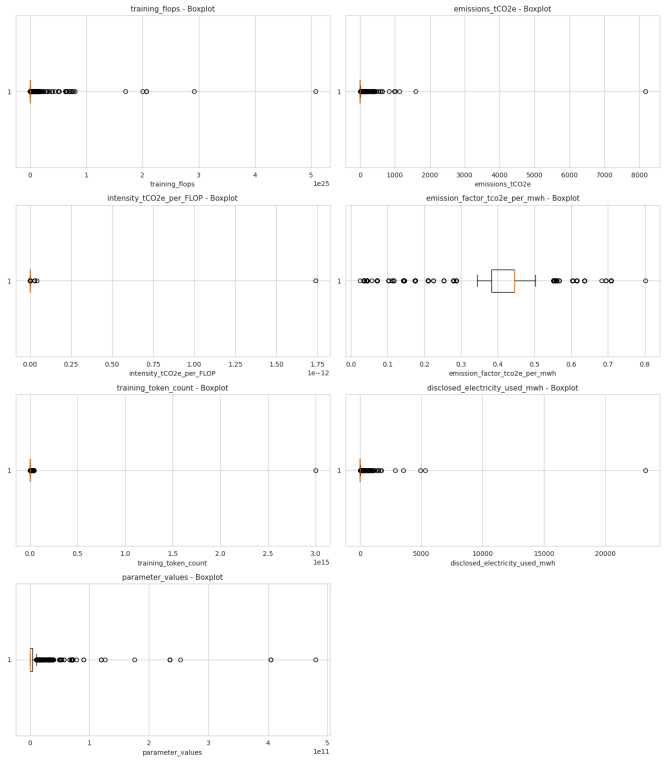


Şekil 1. Sayısal değişkenlerin frekans dağılımları

Grafikte görüldüğü gibi training_flops, emissions_tCO2e, training_token_count ve parameter_values dağılımları neredeyse tamamen 0'a yakın değerlerde yoğunlaşmış, uzun bir sağ kuyruk oluşturmuştur. Bu güçlü çarpıklık, modelleme aşamasında log1p dönüşümünün uygulanma kararının doğrudan gerekçesini oluşturmuştur.

5.2 Aykırı Değer (Outlier) Analizi

Boxplot analizleri, özellikle training_flops ve emissions_tCO2e değişkenlerinde çok sayıda istatistiksel aykırı değer bulunduğunu göstermiştir. Ancak bu uç değerler hatalı ölçüm değildir; gerçekten çok büyük modelleri (örneğin GPT ölçeğinde modelleri) temsil etmektedir. Bu nedenle bu kayıtlar veri setinden çıkarılmamış, modelin bu bölgedeki performansı ayrıca değerlendirilmiştir.

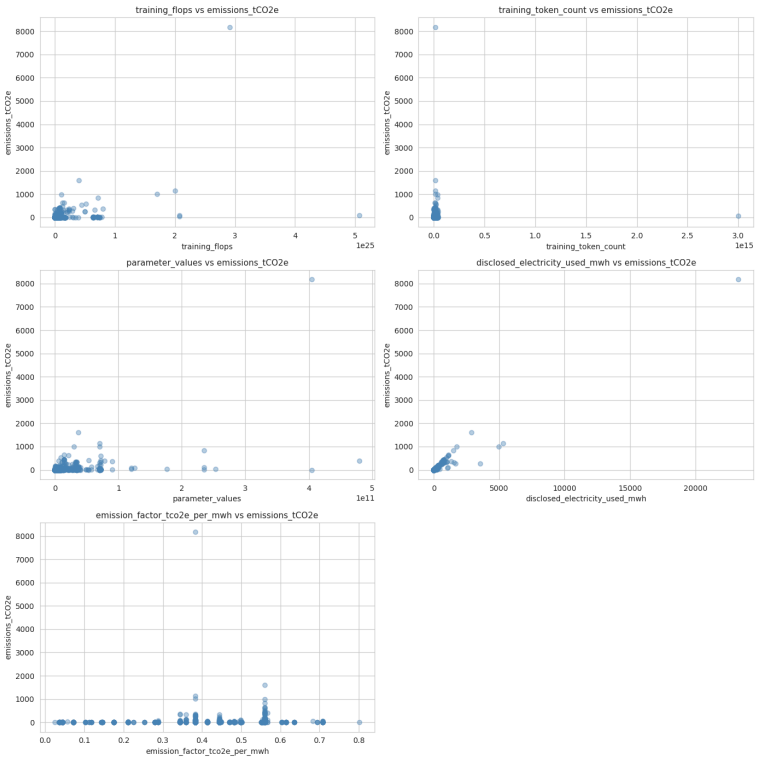


Şekil 2. Sayısal değişkenlere ait boxplot analizleri

Her bir panelde ilgili değişkenin kutu grafiği gösterilmektedir. Kutuların çok dar, buna karşın aykırı değer noktalarının (daire işaretleri) sağa doğru uzun bir kuyruk oluşturması, veri setindeki birkaç dev ölçekli modelin (örn. Llama-3.1-405B) diğer tüm kayıtlardan çok daha yüksek değerlere sahip olduğunu doğrulamaktadır.

5.3 Değişkenler ile Emisyon Arasındaki İlişki (Scatter Analizi)

emissions_tCO2e ile training_flops, training_token_count, parameter_values, disclosed_electricity_used_mwh ve emission_factor_tco2e_per_mwh arasındaki ikili ilişkiler scatter plot ile incelenmiştir.

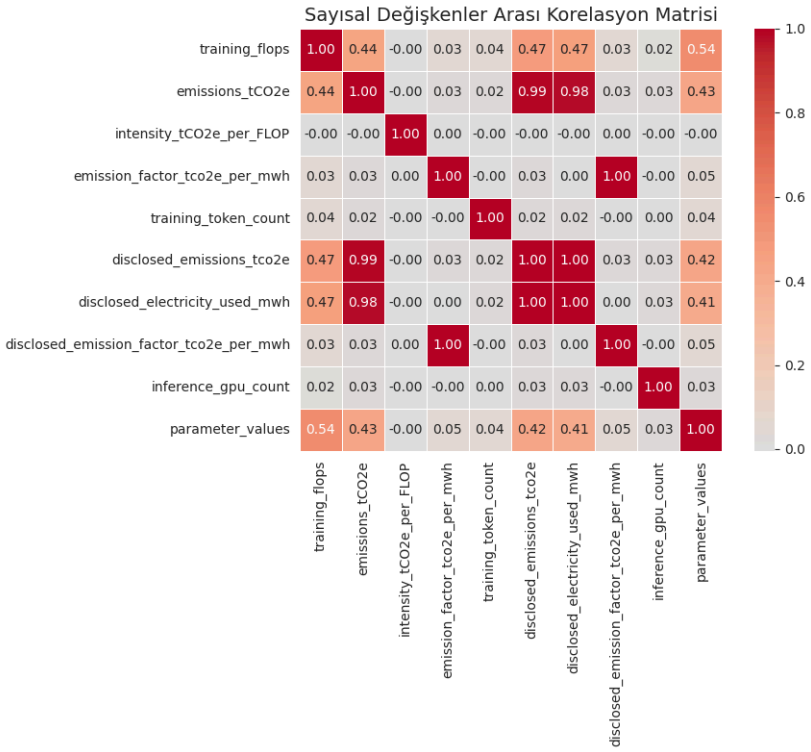


Şekil 3. Sayısal değişkenler ile emissions_tCO2e arasındaki ilişki

disclosed_electricity_used_mwh ile emissions_tCO2e arasında güçlü, neredeyse doğrusal bir ilişki gözlenmektedir (enerji × emisyon faktörü mantığıyla tutarlı). training_flops, training_token_count ve parameter_values ile ilişki de pozitif yönlüdür fakat noktaların çoğunun sol-alt köşede yığılması ve tek bir uç noktanın (Llama-3.1-405B, ~8.176 tCO2e) grafiği domine etmesi, bu ilişkilerin ölçek farkları nedeniyle doğrusal regresyon için log dönüşümüne ihtiyaç duyduğunu göstermektedir.

5.4 Korelasyon Analizi

Sayısal değişkenler arasındaki Pearson korelasyon katsayıları bir ısı haritası (heatmap) ile incelenmiştir.

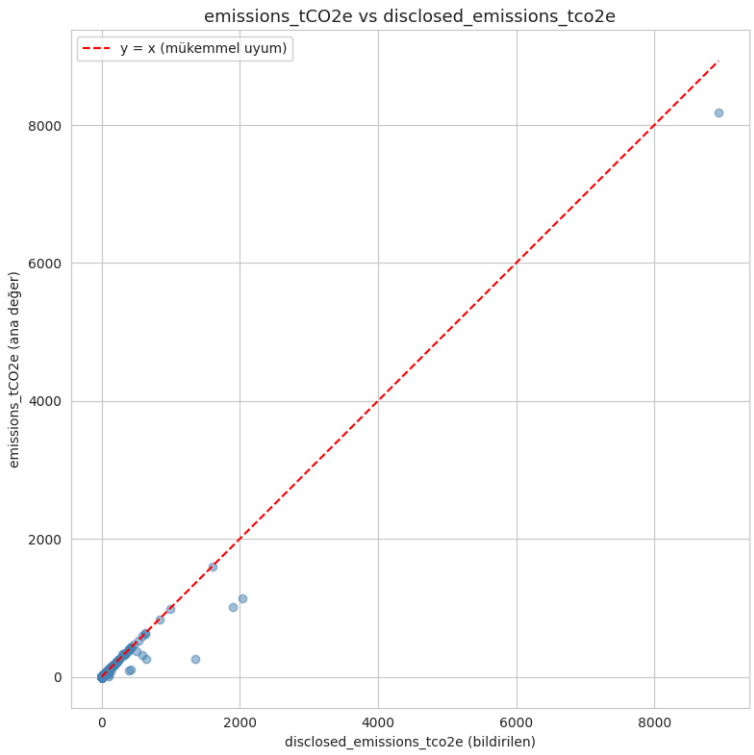


Şekil 4. Sayısal değişkenler arası korelasyon matrisi

emissions_tCO2e ile en güçlü ilişkiler disclosed_emissions_tco2e (r ≈ 0,99) ve disclosed_electricity_used_mwh (r ≈ 0,98) ile bulunmuştur. Bu değişkenler hedefle neredeyse birebir örtüştüğü için veri sızıntısı (leakage) riski taşımakta ve bu nedenle ML feature setinin dışında tutulmuştur. training_flops (r ≈ 0,44) ve parameter_values (r ≈ 0,43/0,54) ile orta düzeyde, modele girmesi güvenli, "dolaylı" ilişkiler gözlenmiştir.

5.5 Bildirilen ve Ham Emisyon Değerlerinin Karşılaştırılması

Her ikisi de dolu olan kayıtlarda emissions_tCO2e ile disclosed_emissions_tco2e arasındaki tutarlılık ayrıca doğrulanmıştır.

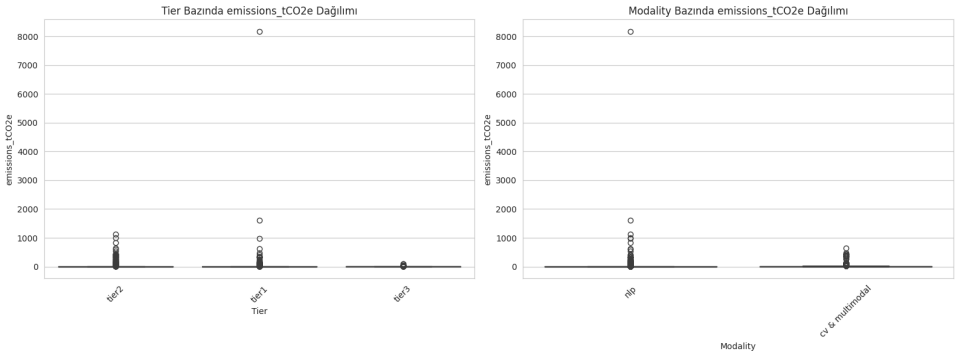


Şekil 5. emissions_tCO2e ile disclosed_emissions_tco2e karşılaştırması

Noktaların neredeyse tamamının y = x (mükemmel uyum) çizgisi üzerinde veya çok yakınında yer alması, iki kaynağın (ham ve bildirilen emisyon) büyük ölçüde tutarlı olduğunu göstermektedir. Bu bulgu, "bildirilen" değerin güvenilir bir referans olarak kullanılabileceğini doğrulamıştır.

5.6 Kategorik Değişkenlere Göre Emisyon Dağılımı

tier (model ölçek katmanı) ve modality (metin, görsel-çoklu ortam vb.) bazında emisyon dağılımları karşılaştırılmıştır.

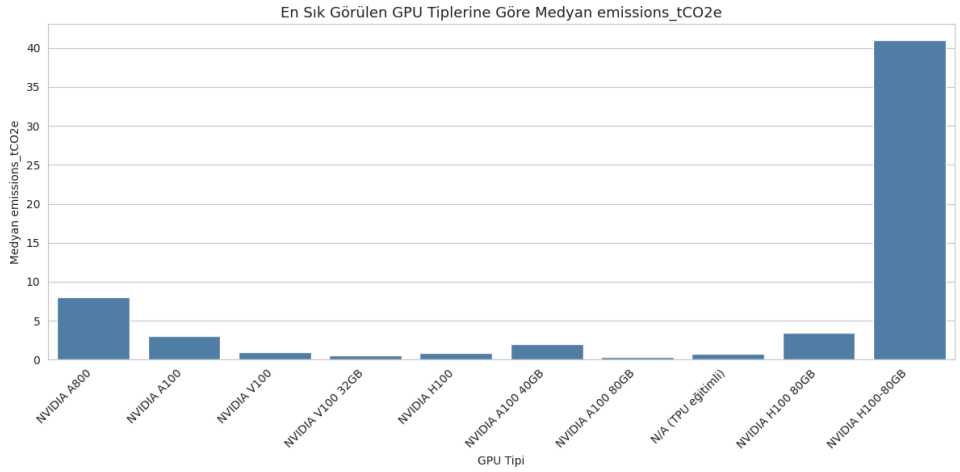


Şekil 6. Tier ve Modality bazında emissions_tCO2e dağılımı

Her iki kırılımda da benzer bir örüntü görülmektedir: medyan emisyon düşük olsa da, her kategoride birkaç uç değer toplam dağılımı yukarı çekmektedir. tier2 ve NLP kategorilerinde en yüksek uç değerler (>1.000 tCO2e) gözlenmiştir; bu da büyük dil modellerinin diğer modalitelere göre daha yüksek emisyona sahip olma eğiliminde olduğuna işaret etmektedir.

5.7 GPU Ailesine Göre Emisyon

Kullanılan GPU tipine göre medyan emissions_tCO2e değerleri karşılaştırılmıştır.

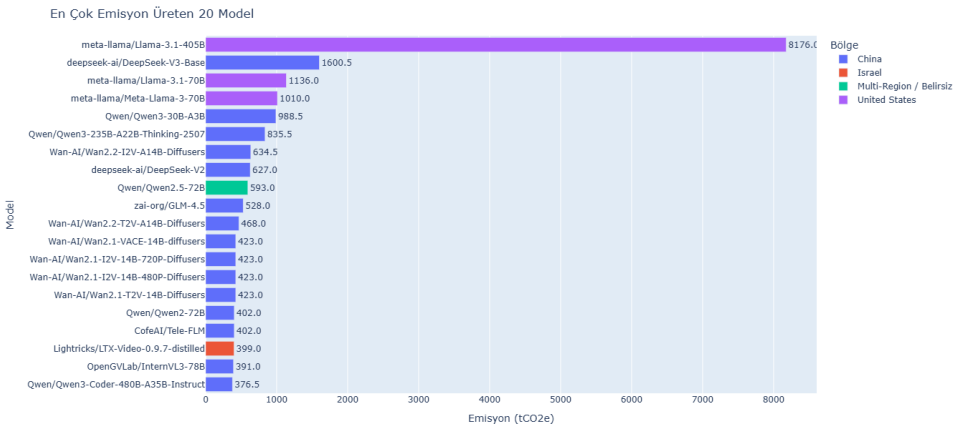


Şekil 7. En sık görülen GPU tiplerine göre medyan emissions_tCO2e

NVIDIA H100-80GB ile eğitilen modellerin medyan emisyonu (~41 tCO2e), diğer tüm GPU tiplerinden belirgin biçimde yüksektir. Bu durum, bu GPU ailesinin daha çok büyük ölçekli/yoğun eğitim gerektiren modellerde tercih edilmesiyle açıklanabilir; GPU tipinin kendisi değil, GPU tipi ile ilişkilendirilen model ölçeği asıl belirleyicidir.

5.8 En Yüksek Emisyonlu Modeller

Veri setindeki en yüksek emisyona sahip 20 model, bölgelerine göre renklendirilerek incelenmiştir.

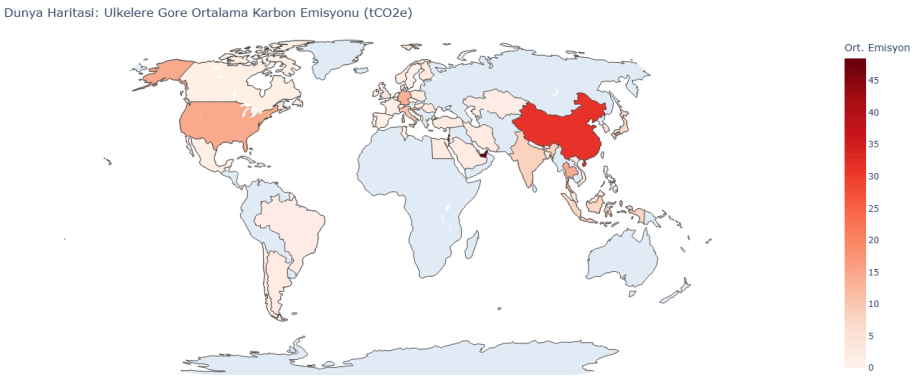


Şekil 8. En çok emisyon üreten 20 model

Listenin açık ara zirvesinde meta-llama/Llama-3.1-405B (8.176 tCO2e) yer almaktadır; bu değer ikinci sıradaki DeepSeek-V3-Base'in (1.600,5 tCO2e) yaklaşık 5 katıdır. Listede ağırlıklı olarak Çin ve ABD kaynaklı modeller bulunmakta, bu da bu iki bölgenin büyük ölçekli AI model eğitiminde baskın konumunu yansıtmaktadır.

5.9 Bölgelere Göre Ortalama Emisyon

standardized_region değişkeni kullanılarak ülke bazında ortalama karbon emisyonu bir dünya haritası üzerinde görselleştirilmiştir.



Şekil 9. Ülkelere göre ortalama karbon emisyonu (tCO2e)

Çin, ortalama emisyon açısından haritada en koyu (en yüksek) bölge olarak öne çıkmaktadır; ABD ve bazı Avrupa ülkeleri orta düzeyde emisyon göstermektedir. Bu bulgu, hem model ölçeğinin hem de eğitim yapılan bölgenin şebeke karbon yoğunluğunun (region_carbon_intensity) emisyon üzerinde birlikte etkili olduğunu desteklemektedir.

İleri EDA bulguları üç somut karara dönüşmüştür: (1) sağa çarpık dağılımlar nedeniyle log dönüşümlerinin uygulanması, (2) aykırı değerlerin silinmemesi ve ayrı biçimde değerlendirilmesi, (3) hedefle çok yüksek korelasyona sahip değişkenlerin (disclosed_emissions_tco2e, disclosed_electricity_used_mwh vb.) olası veri sızıntısı (leakage) riski taşıdığının tespit edilmesi ve bu değişkenlerin ML feature setinden çıkarılması.

6. Hipotez

EDA ve modelleme bulguları ışığında kurulan hipotezler arasından, projenin temel modelleme kararlarını doğrudan etkileyen 3 tanesi öne çıkmaktadır. Bu hipotezler seçilirken ölçüt, ilgili değişken çiftinin hem EDA'da hem de model sonuçlarında somut, sayısal bir kanıtla test edilebilmiş olmasıdır.

Hipotez 1 — Eğitim compute'u ile emisyon arasındaki ilişki

Bağımsız değişken: training_flops | Bağımlı değişken: emissions_tCO2e

H0: training_flops (eğitim compute'u) ile emissions_tCO2e arasında anlamlı bir ilişki yoktur.

H1: training_flops arttıkça emissions_tCO2e de artar (pozitif ilişki).

Sonuç: H0 reddedilmiştir. Korelasyon analizinde r ≈ 0,44 bulunmuş; ayrıca log_training_flops, hem built-in hem permutation importance analizlerinde modelin en güçlü açıklayıcı değişkeni olarak öne çıkmıştır (bkz. Bölüm 7.5). Bu, karbon emisyonunun fiziksel mantığıyla (eğitim compute'u arttıkça enerji tüketimi ve dolayısıyla emisyon artar) tam olarak uyumludur.

Hipotez 2 — Doğrusal olmayan (non-linear) model üstünlüğü

Bağımsız değişken: model tipi (Linear Regression vs. XGBoost) | Bağımlı değişken: model performansı (CV R^2)

H0: Emisyon, feature'larla doğrusal (linear) bir ilişkiyle yeterince iyi açıklanabilir; XGBoost, Linear Regression'a göre anlamlı bir katma değer sağlamaz.

H1: Feature'lar ile emisyon arasında güçlü doğrusal olmayan ilişkiler ve etkileşimler vardır; XGBoost, Linear Regression'ı belirgin biçimde geçer.

Sonuç: H0 reddedilmiştir. Linear Regression 5-fold CV R^2 = 0,571 elde ederken, (tuned) XGBoost aynı metrikte R^2 ≈ 0,888 elde etmiştir (bkz. Bölüm 7.3-7.4).

Hipotez 3 — method değişkeninin gerçek katkısı

Bağımsız değişken: method (var/yok) | Bağımlı değişken: model performansı (Test R^2)

H0: method (emisyonun nasıl ölçüldüğünü gösteren üst-veri) değişkeni model performansına gerçek bir katkı sağlamamaktadır; yalnızca bir yapaylıktır.

H1: method değişkeni çıkarıldığında model performansı belirgin biçimde düşer.

Sonuç: H0 büyük ölçüde desteklenmiştir. method feature importance sıralamasında 2. sırada çıkmasına rağmen, bu değişken modelden çıkarıldığında Test R^2'deki değişim yalnızca −0,013 olmuştur (0,870 → 0,856). Bu, modelin ağırlıklı olarak compute büyüklüğü ve lokasyon gibi tahmin anında gerçekten mevcut olan feature'lardan öğrendiğini göstermektedir (bkz. Bölüm 7.6).

7. Modelleme

Modelleme süreci, feature engineering ile leakage kontrolünün ardından; baseline model, Linear Regression ve XGBoost'un karşılaştırılması, overfitting/regularization/tuning adımları ve nihai model seçimiyle tamamlanmıştır.

7.1 Feature Engineering ve Leakage Kontrolü

Toplam 14 yeni feature üretilmiştir: log_training_flops, log_parameters, log_training_tokens, flops_per_parameter, flops_per_token, region_carbon_intensity, low_carbon_region (sayısal); standardized_region, gpu_family, tier, method, modality, type, is_moe (kategorik). Karbon emisyonuyla matematiksel/doğrudan ilişkisi bulunan disclosed_emissions_tco2e, disclosed_electricity_used_mwh, emission_factor_tco2e_per_mwh, calculated_emissions_tCO2e ve intensity_tCO2e_per_FLOP gibi değişkenler, veri sızıntısı riski taşıdıkları için ML feature setinin tamamen dışında tutulmuş, yalnızca Direct Calculation / Reporting katmanlarında kullanılmıştır.

Hedef değişken, aşırı sağa çarpık olduğu için log1p(cleaned_emissions_tCO2e) olarak tanımlanmış; tahminler sonrasında expm1 ile orijinal tCO2e ölçeğine geri çevrilmiştir. Veri, log_training_flops'un çeyreklik dilimlerine göre stratifiye edilerek %80 train (4.116 kayıt) / %20 test (1.030 kayıt) olarak ayrılmıştır.

7.2 Baseline Model

Hiçbir feature'a bakmadan tüm tahminleri eğitim setinin medyanı olarak veren bir DummyRegressor kurulmuştur: Test setinde MAE = 9,60, RMSE = 48,85, R^2 = −0,03, RMSLE = 1,28. Bu, sonraki modellerin gerçekten anlamlı bir şey öğrendiğini göstermek için referans noktası oluşturmuştur.

7.3 Linear Regression

14 feature, One-Hot Encoding ve StandardScaler ile hazırlanarak log1p(hedef) üzerinde bir Linear Regression modeli eğitilmiştir. Test setinde (orijinal tCO2e ölçeğinde): MAE = 7,84, RMSE = 46,38, R^2 = 0,068. 5-Fold Cross Validation'da (log-hedef üzerinde) ortalama R^2 = 0,571 (std = 0,041) bulunmuştur. Linear Regression, Dummy Baseline'ı belirgin biçimde geçmiş, ancak XGBoost'un gerisinde kalmıştır — bu da feature'lar ile hedef arasında doğrusal olmayan ilişkiler bulunduğuna işaret etmektedir.

7.4 XGBoost, Regularization ve Hyperparameter Tuning

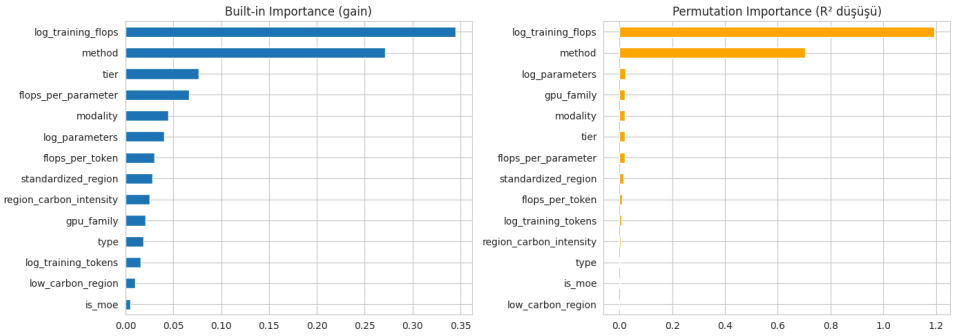
İlk (optimize edilmemiş) XGBoost modeli test setinde MAE = 2,88, RMSE = 22,56, R^2 = 0,78 elde etmiştir; ancak train R^2 = 0,969 olması ciddi bir overfitting riskine (train-test farkı = 0,19) işaret etmiştir. Bu nedenle max_depth, min_child_weight, reg_alpha, reg_lambda gibi parametreler kontrollü biçimde ayarlanmış; ardından RandomizedSearchCV (n_iter=25, cv=5) ile sistematik bir hyperparameter tuning uygulanmıştır.

Model	Train R^2	Test R^2	R^2 Farkı	MAE	RMSE
Original XGBoost	0,969	0,780	0,190	2,88	22,56
Regularized (manuel)	0,775	0,764	0,011	2,87	23,34
Tuned (RandomizedSearchCV)	0,724	0,870	−0,146	2,31	17,35

Tuned XGBoost, Test R^2'de belirgin bir artış (0,780 → 0,870) ve train-test farkında büyük bir azalma göstermiştir. 5-fold CV ortalaması Tuned modelde çok hafif gerilemiştir (0,893 → 0,888, fark CV standart sapmasının altındadır); bu nedenle asıl kazanımın CV ile ölçülen mutlak genelleme gücünden çok, train-test farkındaki stabilite artışından geldiği ayrıca not edilmiştir. Belirlenen karar kuralı ("Train-test farkı en az 0,03 azaldıysa VE Test R^2 kaybı en fazla 0,05 ise Tuned model seçilir") gereği final model olarak Tuned XGBoost seçilmiştir.

7.5 Feature Importance

Final (Tuned) XGBoost modeli için hem built-in (gain tabanlı) hem permutation importance hesaplanmış ve karşılaştırılmıştır.

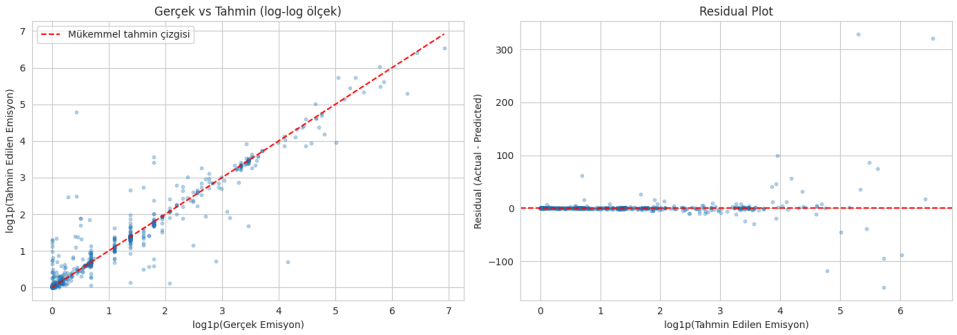


Şekil 10. Built-in ve Permutation Feature Importance karşılaştırması

Her iki yöntemde de en önemli iki değişken aynı çıkmıştır: log_training_flops ve method. Bunları flops_per_parameter, tier, modality ve log_parameters izlemektedir; is_moe ve low_carbon_region en düşük katkıyı göstermiştir. İki farklı yöntemin aynı değişkenleri öne çıkarması, sonuçlara olan güveni artırmaktadır.

7.6 Residual (Hata) Analizi ve Model Doğrulaması

Final model için residual (gerçek − tahmin) hesaplanmış; compute ölçeğine, GPU ailesine, bölgeye ve MoE durumuna göre hata ayrı ayrı incelenmiştir.



Şekil 11. Gerçek vs Tahmin (log-log ölçek) ve Residual grafiği

Sol paneldeki noktaların büyük çoğunluğunun "mükemmel tahmin" çizgisine yakın dağılması, modelin log-uzayda genel olarak isabetli tahminler ürettiğini göstermektedir. Sağ paneldeki residual grafiğinde hataların çoğunlukla 0 civarında yoğunlaştığı, ancak yüksek tahmin değerlerinde (log1p > 4) hem pozitif hem negatif yönde birkaç büyük sapmanın bulunduğu görülmektedir — bu, modelin özellikle uç/büyük ölçekli modellerde daha az kararlı olduğuna işaret eder.

Test seti, gerçek emisyon değerine göre Düşük/Orta/Yüksek olmak üzere üç eşit gruba ayrılarak modelin her grup içindeki performansı ayrıca ölçülmüştür:

Emisyon Seviyesi	MAE	RMSE	Ort. Gerçek Değer	Kayıt Sayısı
Düşük Emisyon	0,090	0,356	0,021	344
Orta Emisyon	0,672	6,462	1,651	343
Yüksek Emisyon	6,188	29,354	27,876	343

Yüksek emisyon grubunun hatası mutlak olarak daha büyük olsa da, bu grubun gerçek değerleri de zaten çok daha büyük ölçektedir. Aykırı değerler veri setinden silinmemiş, modelin bu bölgedeki davranışı dürüstçe raporlanmıştır.

7.7 Hibrit Sistem ve Nihai Çıktı

Modelleme sonuçları, projenin "Hybrid Carbon Estimation System" mimarisine entegre edilmiştir: bağımsız olarak bildirilmiş bir değer varsa doğrudan kullanılır; yoksa fiziksel formülle hesaplanır; o da yoksa yalnızca gerekli olduğunda final XGBoost modeli devreye girer. Her tahmine ayrıca kural tabanlı bir Confidence (High/Medium/Low) etiketi ve sustainability_score eklenmiştir. Final model, production kullanımı için tüm veri üzerinde yeniden eğitilmiş, gerekli artefaktlarla birlikte kaydedilmiş ve round-trip testiyle doğrulanmıştır.

8. Sonuçlama

Bu proje, yalnızca AI modellerinin karbon emisyonunu tahmin eden bir regresyon modeli geliştirmekle kalmamış; mevcut veriye göre doğrudan hesaplama ile ML tahminini birleştiren, tahminin kaynağını açıklayan, veri kalitesini dikkate alan ve sonucu sürdürülebilirlik açısından raporlayan hibrit bir karbon değerlendirme sistemi oluşturmayı amaçlamış ve bu amacı büyük ölçüde gerçekleştirmiştir.

Süreç, veri setinin dikkatli biçimde tanınmasıyla başlamış; kapsamlı bir EDA ve leakage analiziyle sağlam bir temele oturtulmuş; gerekçeli feature engineering kararlarıyla 14 leakage-safe değişken üretilmiş; Linear Regression ve XGBoost karşılaştırmalı biçimde değerlendirilmiş; overfitting, cross validation, regularization ve hyperparameter tuning adımlarıyla model genellenebilirliği önceliklendirilmiştir.

Projenin Güçlü Yönleri

- Yalnızca ML tahmini yapan değil, mevcut olduğunda doğrudan hesaplamayı önceliklendiren hibrit bir mimari kurulmuştur.
- Data leakage konusuna sistematik biçimde dikkat edilmiş, otomatik assert kontrolleriyle doğrulanmıştır.
- Linear Regression ile şeffaf bir baseline oluşturulmuş, XGBoost'un katma değeri somut biçimde gösterilmiştir.
- Model, tek bir train/test split ile yetinilmeden 5-fold Cross Validation ile değerlendirilmiştir.
- Residual analizi ile modelin zayıf/güçlü olduğu alt gruplar dürüstçe ortaya konmuştur.
- Tahmin, tek bir sayı yerine kaynağı, güvenilirlik seviyesi ve sürdürülebilirlik bağlamıyla birlikte sunulmuştur.

Projenin Sınırlılıkları

- Veri setinin bazı alanları daha önce imputation/sentetik tamamlama yöntemleriyle doldurulmuştur; bazı dağılımlar gerçek dünyayı tam yansıtmayabilir.
- Kategorik dağılımlarda dengesizlik bulunmaktadır (Unknown ve United States kategorileri baskındır).
- Bazı bölgeler ve nadir GPU aileleri az sayıda örnekle temsil edilmektedir; bu alt gruplardaki tahminler daha az güvenilir olabilir.
- Kullanılan 'bildirilen' değerler büyük ölçüde kaynakların kendi beyanına dayanmaktadır.
- ML tahmin dalı ('ml_estimated'), kullanılan veri kesitinde fiilen tetiklenmediği için yalnızca simülasyon yoluyla test edilebilmiştir.

Sınırlılıklar göz önünde bulundurulduğunda, proje modelleme aşaması itibarıyla tamamlanmış; hem teknik hem de metodolojik açıdan savunulabilir, tekrarlanabilir ve production kullanımına hazır bir sonuç ortaya koymuştur.

8.1 Proje ve Kod Erişimi

Projenin tüm kodu, veri hazırlama notebook'ları ve dokümantasyonu aşağıdaki GitHub deposunda yer almaktadır:

- GitHub: github.com/yeliznur/Yapay-Zeka-Modellerinin-Karbon-Ayak-zi

Modelin uçtan uca kullanılabildiği Streamlit uygulaması, geliştirme ortamında aşağıdaki adres üzerinden yerel olarak (local) çalıştırılmaktadır; üretim (production) ortamına dağıtım henüz yapılmamıştır:

- Streamlit (yerel geliştirme): http://localhost:8501/

8.2 Gelecek Çalışmalar

- ml_estimated dalının gerçek (canlı) verilerle tetiklenmesi için ek model kaynaklarının veri setine dahil edilmesi.
- Streamlit uygulamasının yerel ortamdan, herkese açık bir üretim (production) ortamına (örn. Streamlit Community Cloud) dağıtılması.
- Modelin, yeni yayınlanan modellerle düzenli aralıklarla yeniden eğitilmesini sağlayacak bir güncelleme (retraining) hattının kurulması.
- What-if Simulator ve Green Ranking özelliklerinin kullanıcı testleriyle geliştirilmesi ve dashboard'a tam entegrasyonu.
- Az örnekli bölge ve GPU ailelerine yönelik ek veri toplanarak alt grup tahmin güvenilirliğinin artırılması.

9. Kaynakça

- Proje kapsamında kullanılan AI Model Eğitim ve Emisyon Veri Seti (imputed_realistic sürümü), kurs kapsamında sağlanan açık kaynaklardan derlenmiştir.
- Pedregosa, F. ve ark. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12.
- Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD.
- McKinney, W. (2010). Data Structures for Statistical Computing in Python (pandas). Proceedings of the 9th Python in Science Conference.
- Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3).
- Waskom, M. (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60).
- Plotly Technologies Inc. (2015). Collaborative data science. Plotly Technologies Inc.
- Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. ACL 2019.
- Patterson, D. ve ark. (2021). Carbon Emissions and Large Neural Network Training. arXiv:2104.10350.
- scikit-learn.org, xgboost.readthedocs.io, pandas.pydata.org, streamlit.io — resmi dokümantasyon kaynakları.

Ek A. Veri Sözlüğü (Feature Listesi)

Modelleme aşamasında (Bölüm 7.1) üretilen 14 leakage-safe feature'ın kısa açıklamaları aşağıda listelenmiştir:

Feature	Tip	Açıklama
log_training_flops	Sayısal	training_flops değişkeninin log1p dönüşümü
log_parameters	Sayısal	parameter_values değişkeninin log1p dönüşümü
log_training_tokens	Sayısal	training_token_count değişkeninin log1p dönüşümü
flops_per_parameter	Sayısal	training_flops / parameter_values oranı
flops_per_token	Sayısal	training_flops / training_token_count oranı
region_carbon_intensity	Sayısal	Modelin eğitildiği bölgenin şebeke karbon yoğunluğu
low_carbon_region	Sayısal (0/1)	Bölgenin düşük karbonlu şebeke sınıfında olup olmadığı
standardized_region	Kategorik	Standardize edilmiş eğitim bölgesi/ülkesi
gpu_family	Kategorik	Kullanılan GPU/TPU ailesi
tier	Kategorik	Model ölçek katmanı (tier1/tier2/tier3)
method	Kategorik	Emisyonun ölçüldüğü/tahmin edildiği yöntem üst-verisi
modality	Kategorik	Model modalitesi (NLP, CV & multimodal vb.)
type	Kategorik	Model mimarisi (transformer, cnn, vit, diffusion, clip)
is_moe	Boolean	Modelin Mixture-of-Experts mimarisi olup olmadığı