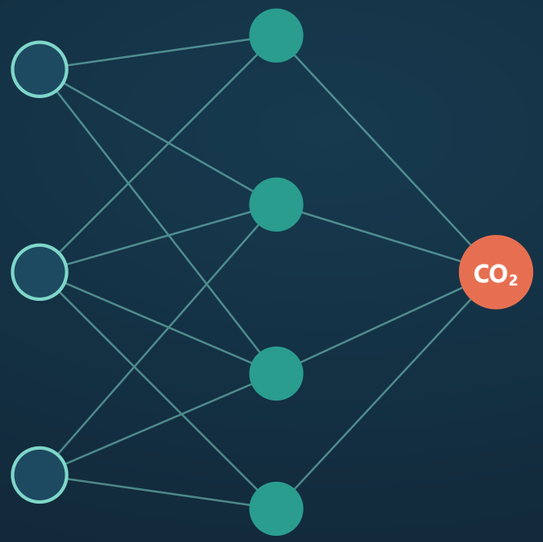
DERİN ÖĞRENME · DÖNEM PROJESİ

Araç Özelliklerinden CO_2 Emisyonu Tahmini

Klasik makine öğrenmesi ile yapay sinir ağlarının karşılaştırılması,
veri sızıntısı, sınıf dengesizliği ve model yorumlama



Semih Özdaş
Sunan

Dr. Sefa Aras
Ders sorumlusu

Kasım 2026
Kaggle · Kanada araç verisi

01 · PROBLEM

Neden araç emisyonunu tahmin etmek istiyoruz?

Ulaşım payı

Karayolu taşımacılığı enerji kaynaklı CO_2'nin en büyük kalemlerinden biri. Araç seçimi doğrudan karbon ayak izini belirliyor.

Ölçüm pahalı

Emisyon laboratuvar testleriyle ölçülüyor. Tasarım aşamasında, yalnız teknik özelliklerden tahmin yapabilmek değerli.

Araştırma sorusu

Motor, silindir, şanzıman, yakıt türü, sınıf ve marka bilgisinden CO_2 (g/km) ne kadar doğru tahmin edilir? Derin öğrenme klasik modelleri geçer mi?

Regresyon: CO_2 (g/km)

Sınıflandırma: Düşük / Orta / Yüksek

Yorumlama: SHAP

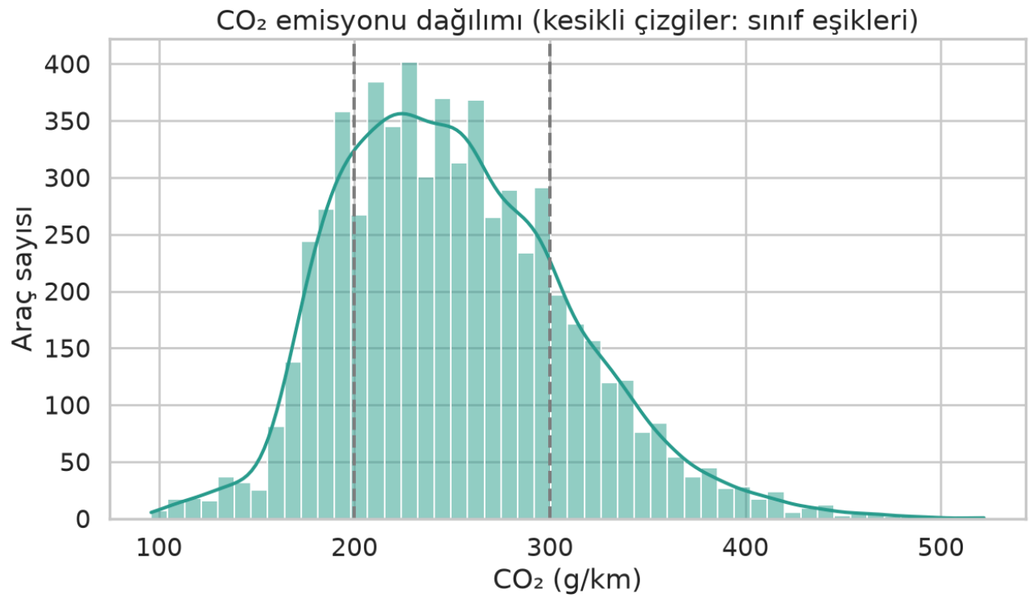
Uygulama: hibrit tasarrufu

02 · VERİ SETİ

7.385 araç, 12 sütun: temizlikten sonra 6.281

ADIM	SATIR
Ham veri (Kanada hükümeti, Kaggle)	7.385
Birebir tekrarlar silindi	6.282 (−1.103)
Tek örnekli doğalgaz aracı silindi	**6.281**

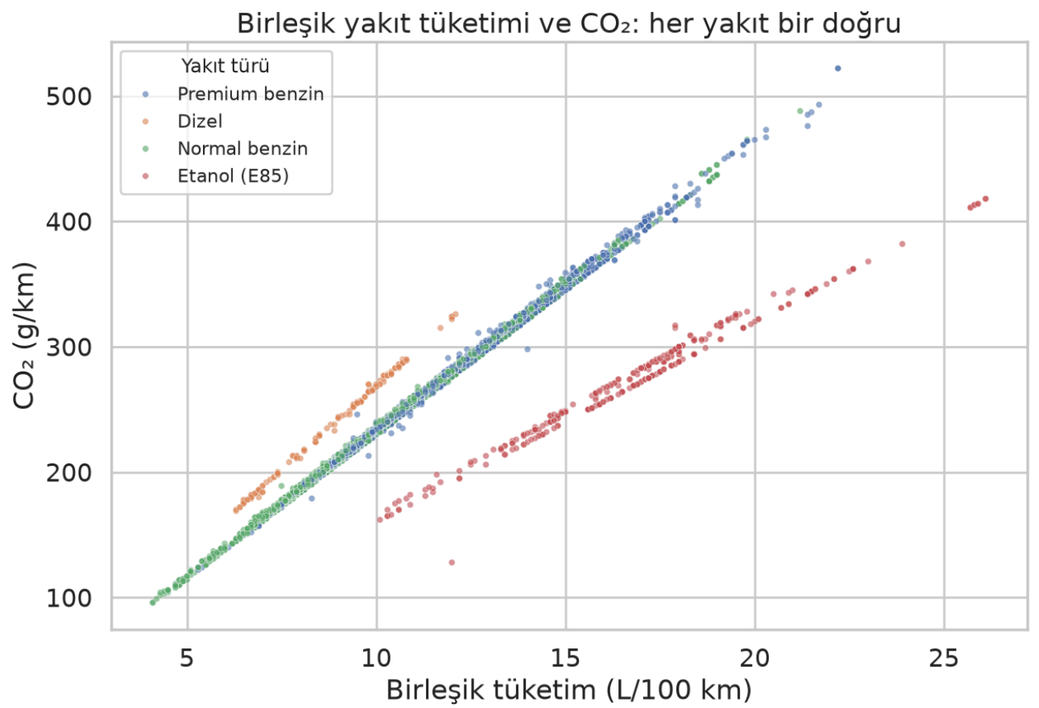
- **%15 tekrar:** aynı araç hem eğitimde hem testte olsaydı model ezberler, skor şişerdi.
- **Özellik mühendisliği:** "AS6" → şanzıman tipi (AS) + vites sayısı (6) + CVT bayrağı.
- 42 marka, 16 araç sınıfı, 4 yakıt türü. Hedef ortalaması **251 g/km** (σ = 59).



Hedef değişkenin dağılımı

03 · KRİTİK BULGU

Yakıt tüketimi = veri sızıntısı



Her yakıt türü kendi doğrusu üzerinde: ilişki fiziksel, istatistiksel değil

CO_2, yanan yakıt miktarıyla **kimyasal olarak orantılı**. Oran yakıt türü içinde neredeyse sabit:

YAKIT	CO_2 / TÜKETİM	Σ
Etanol (E85)	16,31	0,44
Benzin (normal / premium)	23,3	0,24
Dizel	26,89	0,20

Bu sütunu modele vermek, cevabı soruya yazmak demek. Bu yüzden iki senaryo kurduk:
A = tüketim var (sızıntılı) · **B** = tüketim yok (gerçekçi)

04 · YÖNTEM

Deney düzeni: her model aynı koşullarda



Ön işleme sızıntısı yok

Ölçekleme ve kodlama Pipeline içinde: istatistikler her katta yalnız eğitim kısmından hesaplanıyor.

Metrikler

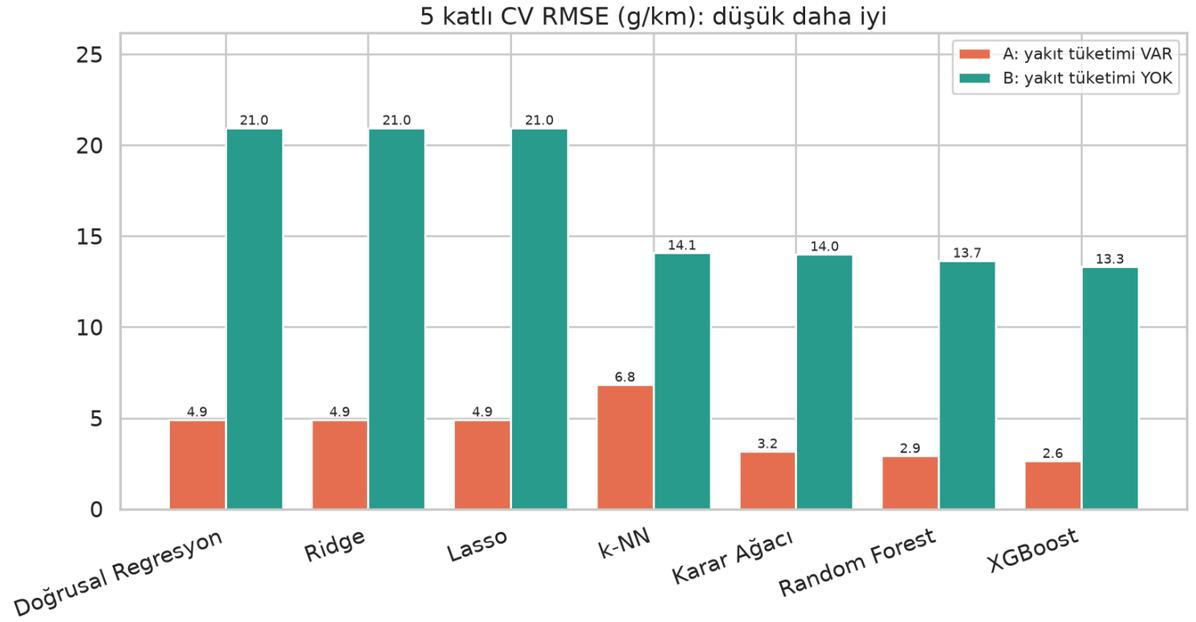
MAE, RMSE, R^2 (CV ortalaması ± std) + hiç dokunulmamış test kümesi. Sınıflandırmada makro F1.

Tekrarlanabilir

Tüm rastgelelik tohum 42. Sinir ağları da klasik modellerle **aynı bölme ve aynı CV katları** üzerinde.

05 · KLASİK MODELLER

Sızıntı başarıyı 5 kat şişiriyor



2,6

g/km RMSE · XGBoost, A
R^2 = 0,998

13,3

g/km RMSE · XGBoost, B
R^2 = 0,949

- Doğrusal modeller en çok etkilenen: R^2 0,993 → 0,873.
- Tüketim yokken ilişki doğrusal değil; ağaçlar ve k-NN etkileşimleri yakalıyor.
- CV ≈ test (13,32 / 13,18): aşırı öğrenme yok.

Soru: Bir sinir ağı XGBoost'u yakalayabilir mi?

Çok katmanlı algılayıcı (MLP)

$$\mathrm{h}_{\ell} = \mathrm{ReLU}(\mathrm{BN}(\mathrm{W}_{\ell}\,\mathrm{h}_{\ell-1} + \mathrm{b}_{\ell}))$$

Evrensel yaklaşım teoremi: yeterli genişlikte bir gizli katman her sürekli fonksiyonu yaklaşık temsil edebilir. Ama **tablo verisinde** ağaç toplulukları genelde hâlâ çok güçlü.

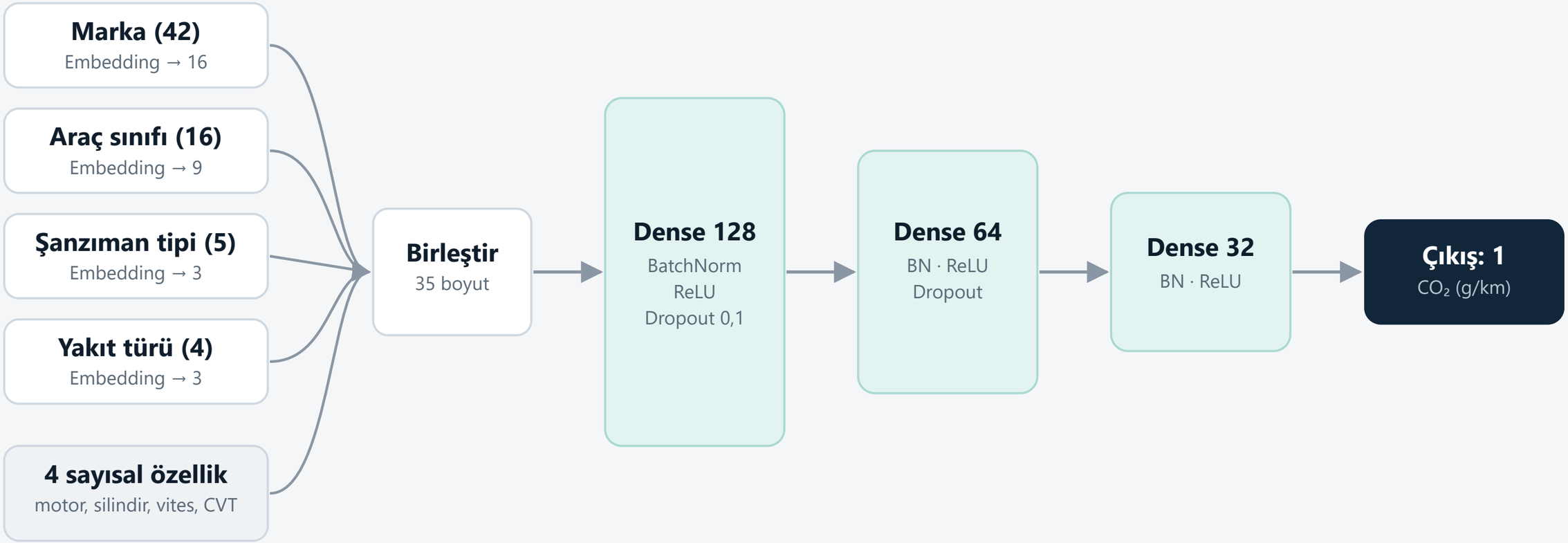
Tablo verisinin zorlukları

- Az örnek: ~5.000 eğitim satırı
- Yüksek kardinaliteli kategorik değişken: **42 marka**
- Düzensiz, eşik benzeri ilişkiler (ağaçların doğal gücü)

Çözüm denemesi: kategorikler için **öğrenilebilir gömme (entity embedding)**.

07 · MİMARİ

MLP + gömme ağı (PyTorch)



Gömme boyutu = min(16, (kategori sayısı + 1) / 2). Görülmemiş kategori için ayrılmış "bilinmeyen" indeksi var. Toplam ≈ 16.300 öğrenilebilir parametre.

One-hot mu, öğrenilebilir gömme mi?

One-hot

BMW	`[0 0 1 0 0 … 0]`
MERCEDES	`[0 0 0 0 1 … 0]`

- 42 boyutlu, seyrek
- Her marka diğerlerine **eşit uzaklıkta**
- Benzerlik bilgisi yok

Entity embedding

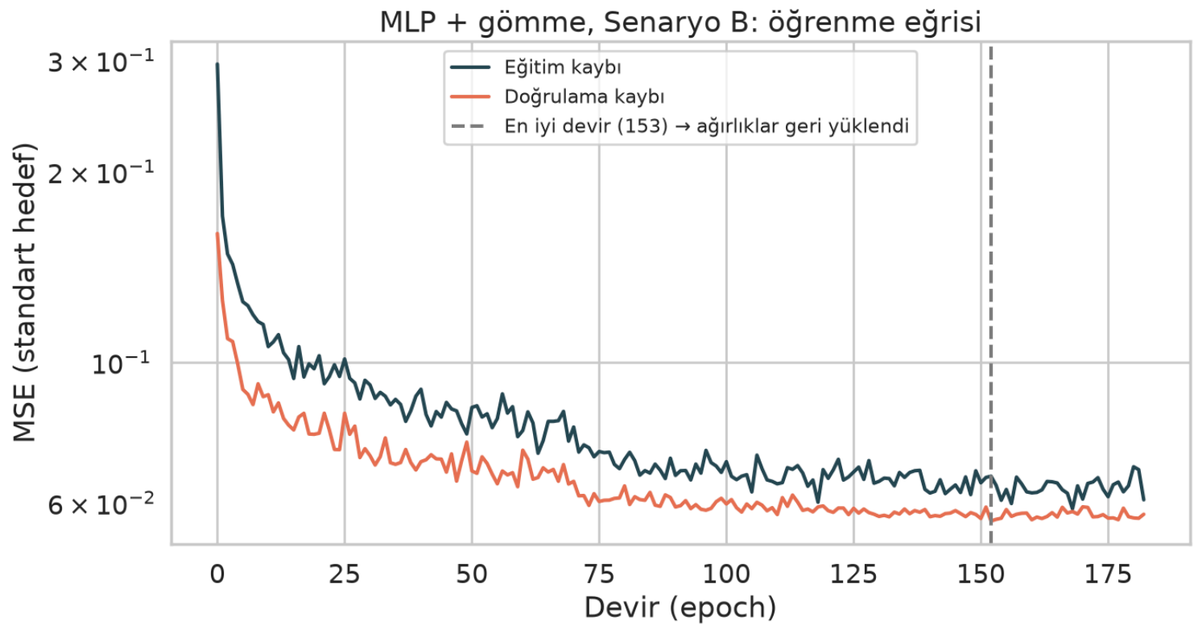
BMW	`[0,41 −1,2 0,07 …]`
MERCEDES	`[0,38 −1,1 0,15 …]`

- 16 boyutlu, yoğun, **geri yayılımla öğrenilir**
- Benzer davranan markalar yakınlaşır
- Word2Vec fikrinin kategorik değişkene uyarlanmışı (Guo & Berkhahn, 2016)

Vektörlerdeki sayılar temsilîdir; gerçek öğrenilmiş gömmeler 12. slayttaki grafikte.

09 · EĞİTİM

Eğitim süreci ve erken durdurma

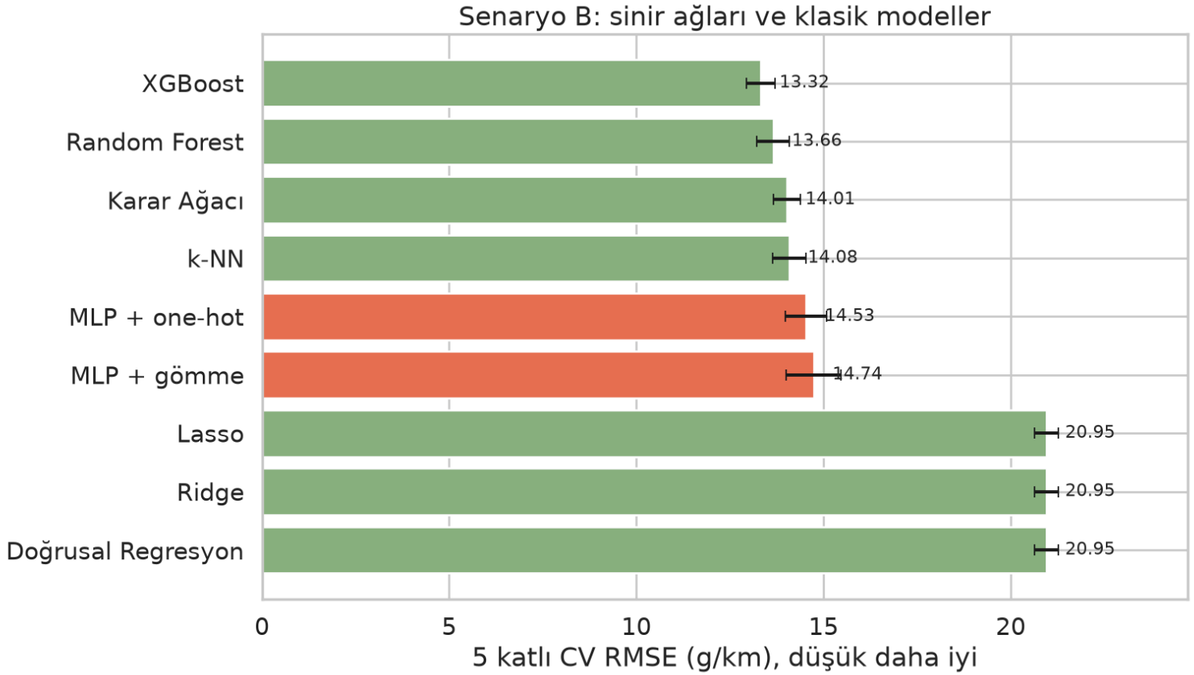


Kayıp	MSE (standart hedef)
Optimizasyon	AdamW, lr $3 \cdot 10^{-3}$
Weight decay	10^{-4}
Mini-batch	128
LR zamanlayıcı	ReduceLROnPlateau ×0,5
Erken durdurma	sabır 30 devir

Doğrulama kümesi her katın **eğitim kısmından** ayrılıyor (%15). En iyi devirdeki ağırlıklar geri yükleniyor; aşırı öğrenmeye başlayan kısım atılıyor.

Eğitim kaybının doğrulamadan yüksek olması beklenen durum: Dropout yalnız eğitimde aktif, eğitim kaybı da devir boyunca ortalanıyor.

Sinir ağı rekabetçi, ama XGBoost hâlâ önde

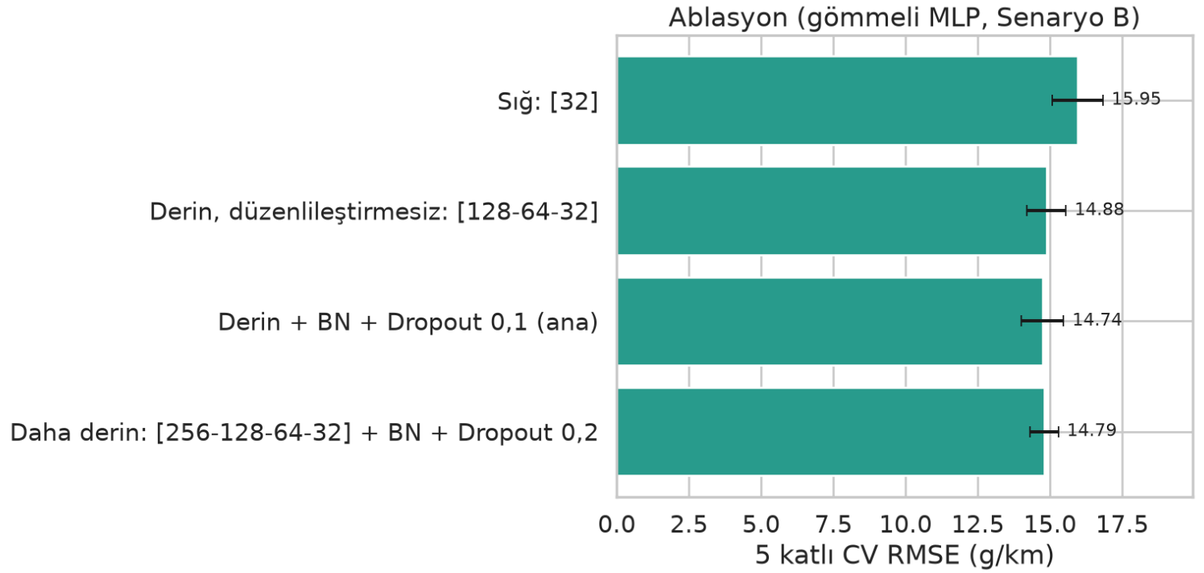


MODEL (B)	CV RMSE	TEST R^2
XGBoost	**13,32**	**0,953**
Random Forest	13,66	0,950
MLP + one-hot	14,53	0,948
MLP + gömme	14,74	0,946
Doğrusal Regresyon	20,95	0,878

- Sinir ağı doğrusal modeli **%30** geçiyor, k-NN ve karar ağacı düzeyinde.
- Testte fark küçük: MLP 13,8 / XGBoost 13,2 g/km.
- Gömme bu veri boyutunda kazanç getirmedi (fark std içinde).
- Literatürle uyumlu: orta boy tablo verisinde ağaç toplulukları güçlü (Grinsztajn vd., 2022).
- Senaryo A'da MLP de sızıntıyı yakalıyor: RMSE 3,1, R^2 0,997.

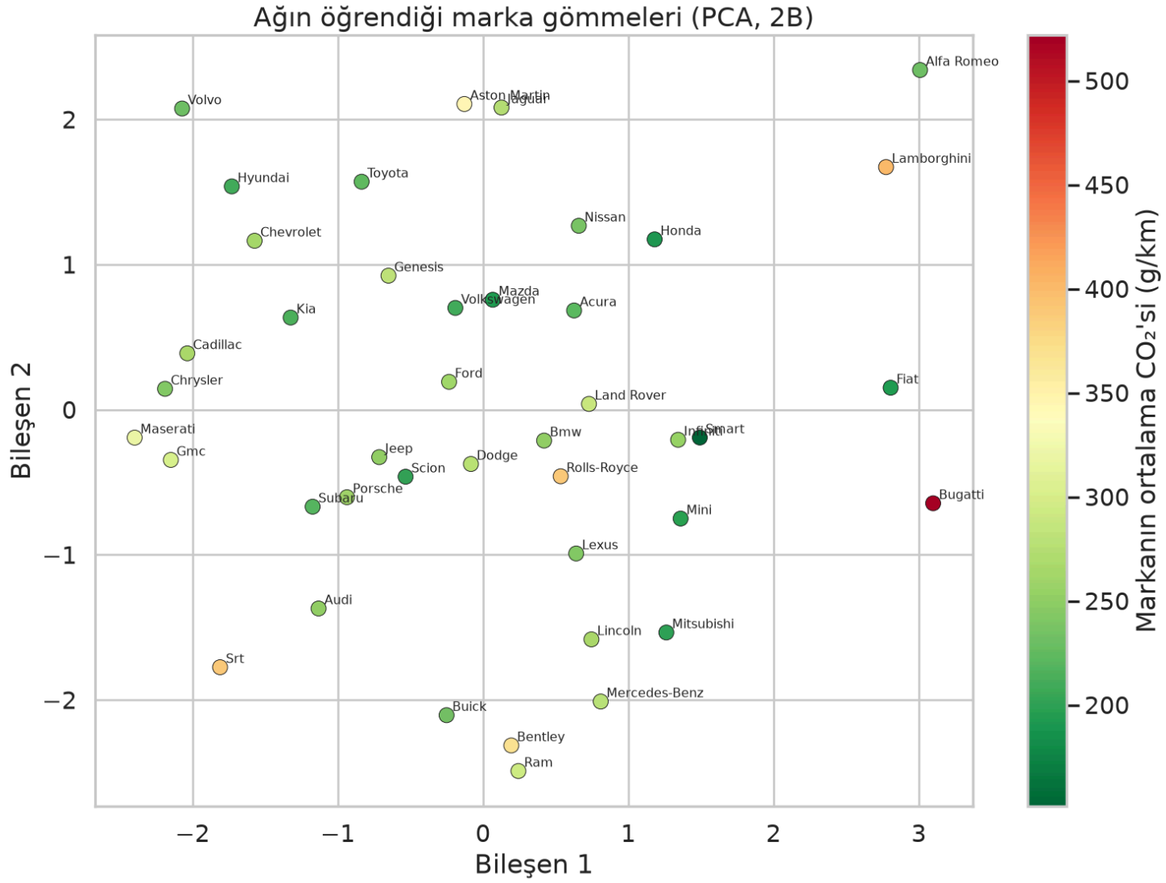
11 · ABLASYON

Derinlik ve düzenlileştirme ne kazandırıyor?



- **Derinlik işe yarıyor:** tek gizli katman (32) → 15,95; üç katman → 14,88 g/km (%7 iyileşme).
- **BatchNorm + Dropout** küçük ama tutarlı katkı: 14,88 → 14,74.
- **Daha derin ağ kazandırmıyor** (14,79): ~5.000 satırda kapasite değil veri sınırlayıcı.
- Daha derin ağın std'si en düşük (±0,49): güçlü düzenlileştirme katlar arası oynaklığı azaltıyor.
- Hiçbir mimari XGBoost'un 13,32'sine inemedi; fark mimari seçiminden büyük.

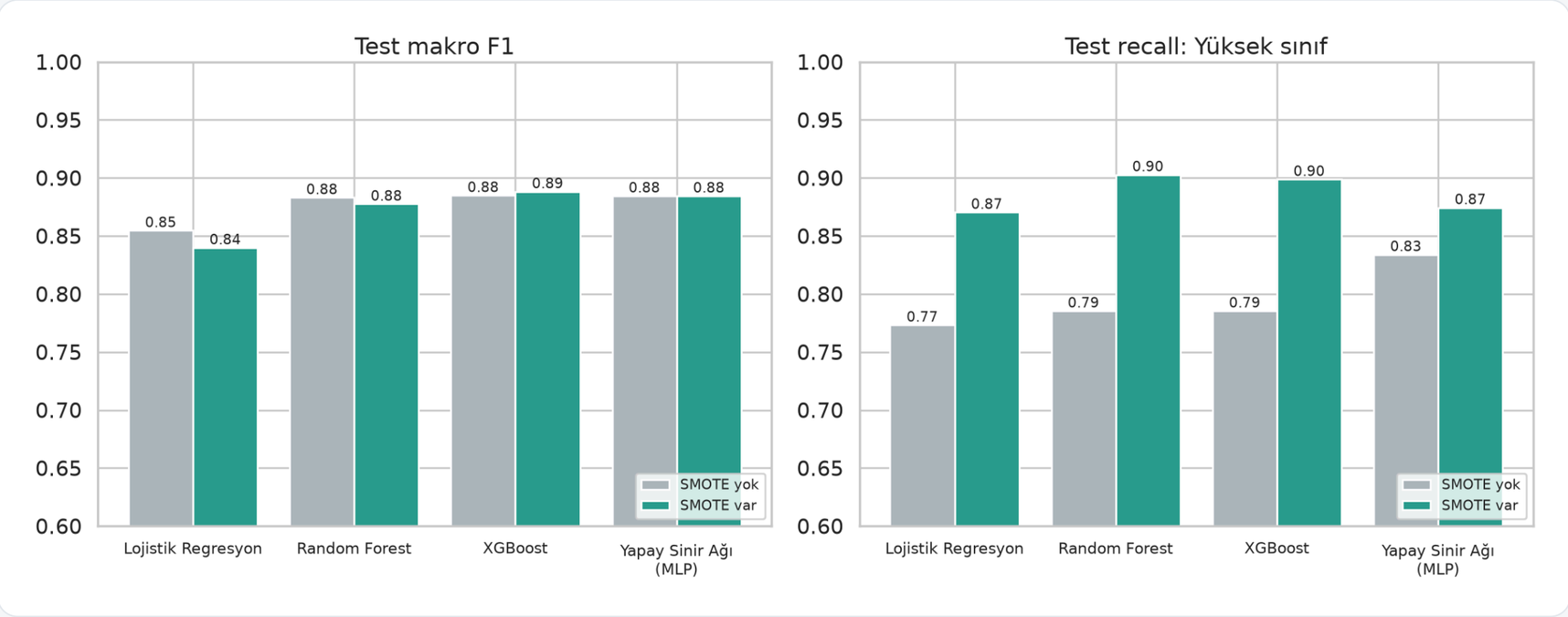
Marka gömmeleri: ağ markaları nasıl temsil etti?



16 boyutlu marka vektörleri PCA ile 2 boyuta indirildi. Renk, markanın gerçek ortalama emisyonu.

- Ağa markaların emisyonu **hiç söylenmedi**; yalnız araç bazında hata geri yayıldı.
- Gömmeler ham ortalamaya göre sıralanmıyor: motor hacmi ve sınıf zaten verildiği için ağ markanın **artık etkisini** kodluyor (SHAP'ta marka payı yalnız %10).
- Gömmeler başka modellere (ör. XGBoost) hazır özellik olarak da verilebilir.

Dengesiz sınıflar ve SMOTE



%20 / %60 / %20

Hep "Orta" diyen model %60 doğruluk alır ama makro F1 = 0,25. **Doğruluk yanıltıcı.**

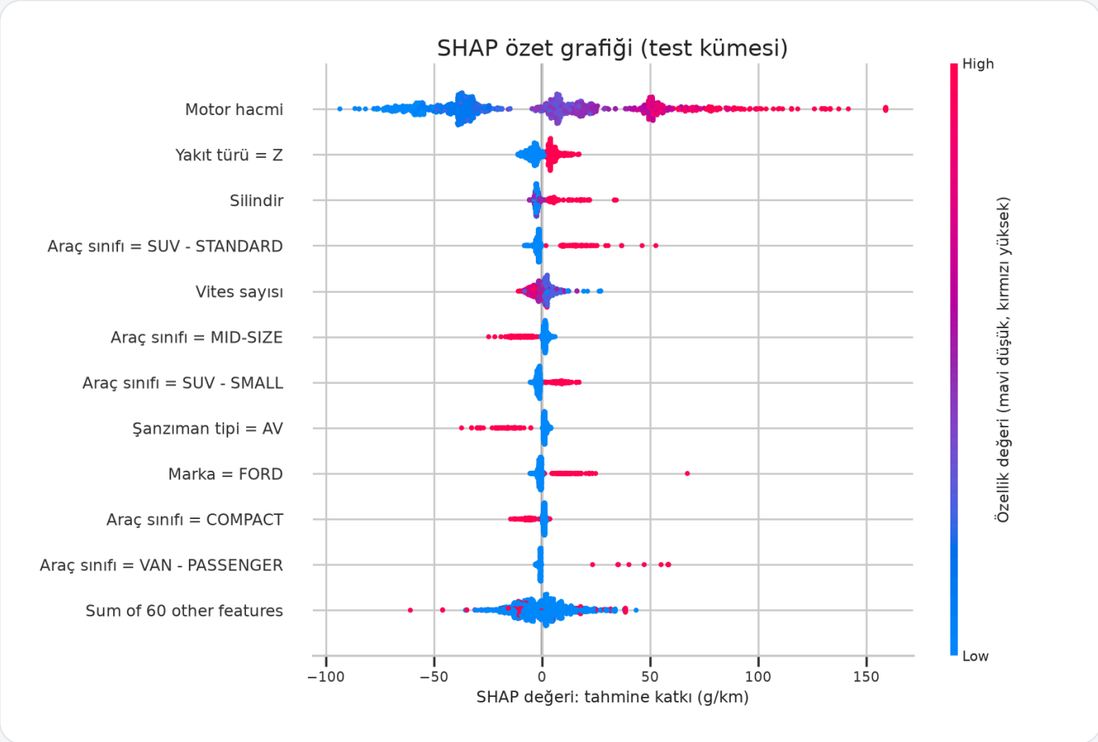
SMOTENC, yalnız eğitim katında

imblearn Pipeline'ı: sentetik örnekler teste sızmıyor; kategorikler (marka, sınıf) geçerli değer alıyor.

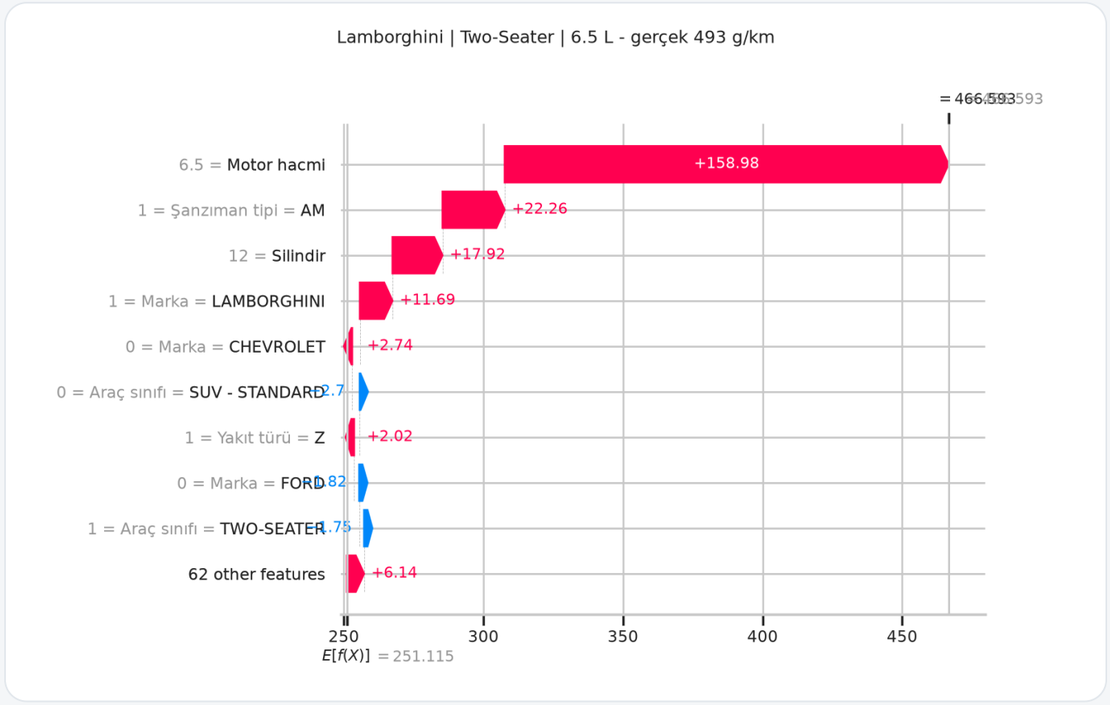
Yüksek sınıf recall: 0,79 → 0,90

Makro F1 korunurken en kritik sınıfı kaçırma oranı yarıya iniyor. En iyi: XGBoost + SMOTE (makro F1 0,888).

SHAP: model neye bakıyor?



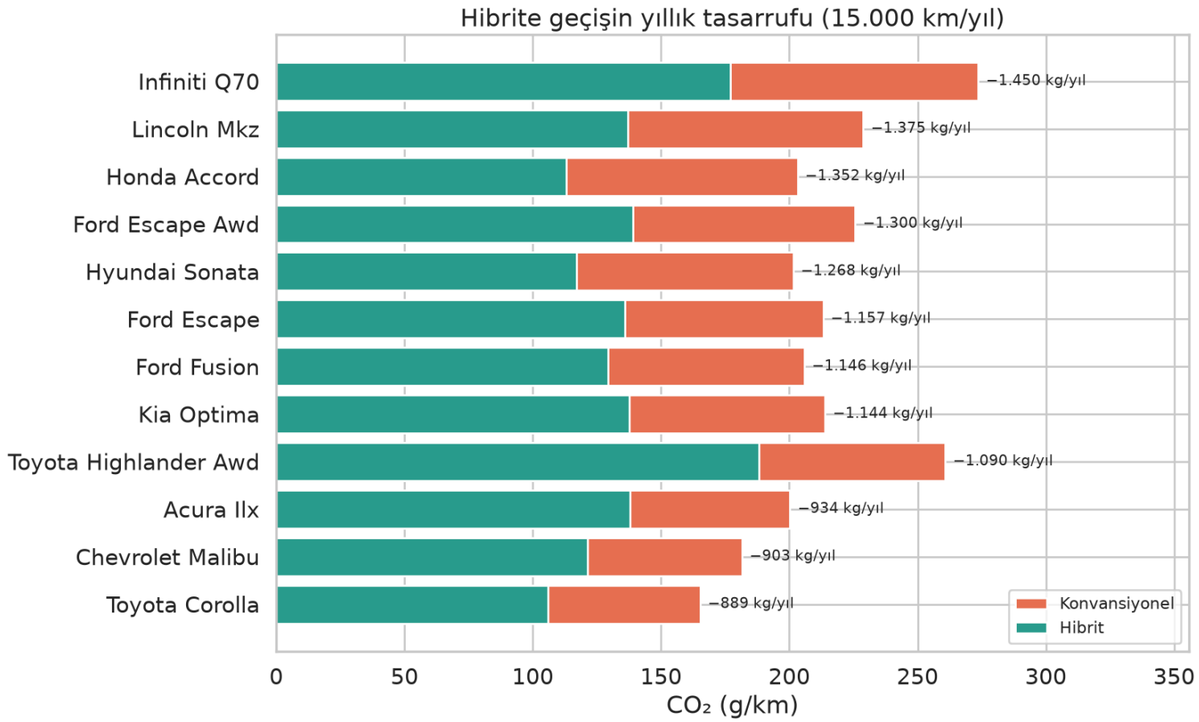
Her nokta bir test aracı; sağ = emisyonu artırıyor



Tek araç: 251 g/km ortalamadan 467 g/km tahmine

XGBoost (B) için ortalama |SHAP| payı: **motor hacmi %54**, araç sınıfı %14, marka %10, yakıt türü %7. RF önem skorunda motor hacmi %79'du; SHAP etkiyi sınıf ve markaya daha dengeli dağıtıyor.

Benzinli yerine hibrit: yılda ~1 ton CO_2



951
kg CO_2 / yıl
ortalama tasarruf

%29
ortalama emisyon
azalması (63 g/km)

Aynı marka + model + sınıf + yakıt türündeki hibrit ve konvansiyonel sürümler eşleştirildi (**22 çift**). Varsayım: 15.000 km/yıl. Konvansiyonel tarafta farklı motor seçeneklerinin ortalaması alındı.

$$\text{tasarruf} = \Delta(\text{g/km}) \times 15.000 \,/\, 1000$$

Ne öğrendik?

1. **Önce veriyi sorgula.** R^2 = 0,998 bir başarı değil, sızıntı işaretiydi. Gerçekçi hata ~13 g/km (ortalamanın %5'i).

2. **Derin öğrenme her zaman kazanmaz.** MLP rekabetçi (R^2 0,948) ama ~5.000 satırlık tablo verisinde XGBoost önde. Model seçimi veriye göre yapılmalı.

3. **Düzenlileştirme ve erken durdurma şart.** BatchNorm + Dropout + en iyi devre dönüş ağın aşırı öğrenmesini engelledi; CV ≈ test.

4. **Dengesiz veride makro F1 + SMOTE.** Yüksek emisyon sınıfında recall 0,79 → 0,90.

5. **SHAP ile açıklanabilir.** Motor hacmi tek başına katkının yarısından fazlası.

Gelecek çalışma: iç içe (nested) CV, marka-model grubuyla GroupKFold, TabNet / FT-Transformer gibi tabloya özel derin mimariler.

TEŞEKKÜRLER

Sorular?

Tarayıcıda canlı çalışan model: **yapayogrenme.semihozdas.com**
Bu sunum: **derinogrenme.semihozdas.com**

---

scikit-learn · XGBoost
Klasik modeller

PyTorch
Sinir ağları

imbalanced-learn · SHAP
SMOTE, yorumlama