Veri Madenciliği (FET445) – 7. Dönem

Avustralya Meteoroloji Verileri ile Yağmur Tahmini (Sınıflandırma)

Parlayan Yıldızlar Takımı

YUSUF RIDVAN ÇELİKBAŞ, 22040101055, yusufridvancelikbas@stu.topkapi.edu.tr
AYÇA SU YILDIRIM, 22040101049, aycasuyildirim@stu.topkapi.edu.tr
MUHAMMED EFE KÜÇÜKYETER, 22040101042, efekucukyeter@stu.topkapi.edu.tr
EMRE SOMER ÇABAK, 22040101027, emresomercabak@stu.topkapi.edu.tr

GitHub/Repo Bağlantısı
github.com/theriay/FET445_TeamShinningStars

İstanbul Topkapı Üniversitesi
Bilgisayar Mühendisliği Bölümü
Proje Özeti

Kasım 2025

1. Problem Tanımı

Temel Soru: Geçmiş meteorolojik veriler (nem, basınç, rüzgar vb.) kullanılarak, Avustralya'daki lokasyonlar için bir sonraki günün yağış durumu (RainTomorrow) yüksek doğrulukla tahmin edilebilir mi?

Bu proje, atmosferik fiziğin karmaşık yapısını çözümlemeyi amaçlayan bir İkili Sınıflandırma (Binary Classification) problemidir.

Problem Zorlukları:

- **Meteorolojik Karmaşa:** Basınç, nem ve sıcaklık arasındaki doğrusal olmayan ilişkilerin (non-linear patterns) modellenmesi.
- **Sınıf Dengesizliği (Imbalance):** Veri setinde yağışsız günlerin (%78), yağışlı günlere (%22) baskın gelmesi. Hedef, çoğunluk sınıfı ezberlemek değil, azınlık sınıfı yakalamaktır.
- **Sosyal Etki:** Tarım, afet yönetimi ve lojistik için erken uyarı sistemi prototipi oluşturulması.

Teknik Özet:

- **Görev:** Gözetimli Öğrenme (Supervised Learning).
- **Hedef Değişken:** RainTomorrow (Yes=1, No=0).
- **Başarı Hedefleri:** Sadece Doğruluk (Accuracy) değil; ROC-AUC >= 0.85 ve F1-Score >= 0.60 hedeflenmektedir.

2. Proje Yönetimi

Roller ve Görev Dağılımı:

- **Yusuf Rıdvan ÇELİKBAŞ:** Doğrusal Modeller (Logistic Regression), Ölçekleme, Raporlama.
- **Ayça Su YILDIRIM:** Topluluk Öğrenmesi (Random Forest), Dengesiz Veri Yönetimi, Görselleştirme.
- **Muhammed Efe KÜÇÜKYETER:** Karar Ağaçları (Decision Tree), Ağaç Yapısı Analizi, Yorumlanabilirlik.
- **Emre Somer ÇABAK:** Gelişmiş Modeller (XGBoost & Bagging), Karşılaştırmalı Analiz, Kod Validasyonu.

Süreç:

- **1-3. Hafta (Tamamlandı):** Veri seçimi, literatür taraması, temizleme, EDA, sızıntı (leakage) analizi ve Baseline modellerin kurulumu.
- **4-11. Hafta (Planlanan):** Hiper-parametre optimizasyonu (GridSearch), Özellik Seçimi (PCA) ve Final Raporlama.

3. İlgili Çalışmalar ve Fark Analizi

Literatürdeki (I. A. Obaidalla, S. L. Ezamzuri et al.) çalışmalar genellikle Accuracy metriğine odaklanmış ve eksik verileri silme yoluna gitmiştir.

Projemizin Literatürden Farkları (Gap Analysis):

- **Sistematik Karşılaştırma:** Tek bir model ailesi yerine 4 farklı aile (Lineer, Ağaç, Boosting, Uzaklık) A/B testi mantığıyla kıyaslanmaktadır.
- **Veri Yönetimi:** Veri kaybını önlemek için satır silmek yerine Medyan/Mod imputasyonu ve Tarih Mühendisliği uygulanmıştır.
- **Metrik Seçimi:** "Accuracy Tuzağı"ndan kaçınılarak, dengesiz veri setleri için daha güvenilir olan F1-Score ve ROC-AUC metrikleri esas alınmıştır.

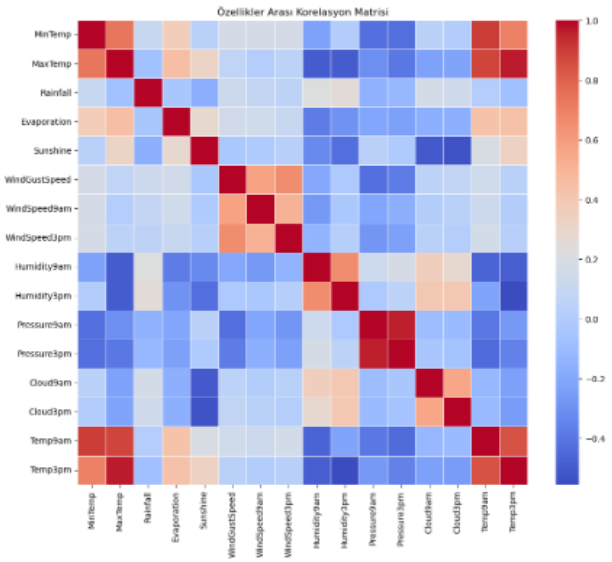
4. Veri Tanımı

- **Veri Kaynağı:** Avustralya Meteoroloji Bürosu (Kaggle: Weather Dataset Rattle Package).
- **Boyut:** 145.460 gözlem (satır) ve 23 öznitelik (sütun).
- **Sınıf Dağılımı:** Dengesiz. No: %77.6 | Yes: %22.4.
- **Kritik İşlem (Leakage Prevention):** RISK_MM sütunu, ertesi gün düşen yağış miktarını içerdiği ve hedefi ifşa ettiği için eğitim öncesi veri setinden çıkarılmıştır.

5. Keşifsel Veri Analizi (EDA)

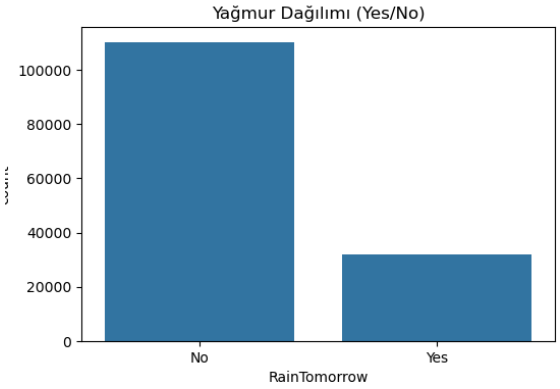
Veri setinin yapısal özellikleri şu başlıklarla analiz edilmiştir:

- **Eksik Veriler:** Evaporation (%43) ve Sunshine (%48) gibi sütunlarda yüksek eksiklik tespit edilmiş; silmek yerine istatistiksel doldurma yöntemleri seçilmiştir.
- **Aykırı Değerler:** Rainfall değişkenindeki aşırı değerler, meteorolojik gerçeklik (fırtına vb.) olduğu için korunmuştur.
- **Korelasyon:** Temp9am ile Temp3pm (0.86) arasında güçlü korelasyon görülmüş, bu da Çoklu Doğrusal Bağlantı (Multicollinearity) riskine karşı PCA kullanımını desteklemiştir.

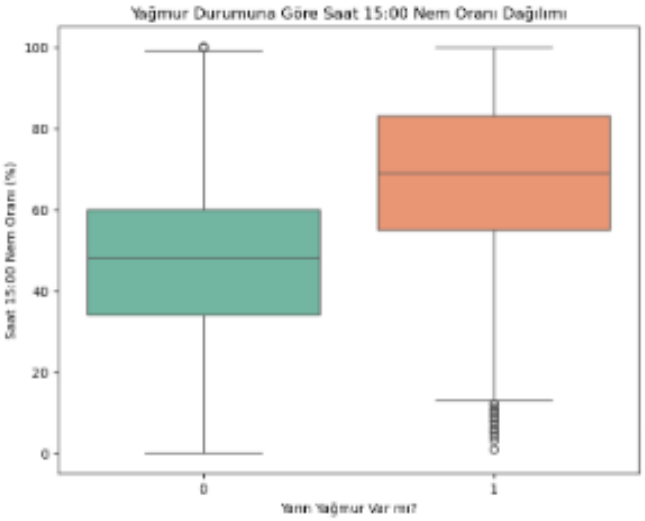


Şekil 1: Özellikler Arası Korelasyon Matrisi

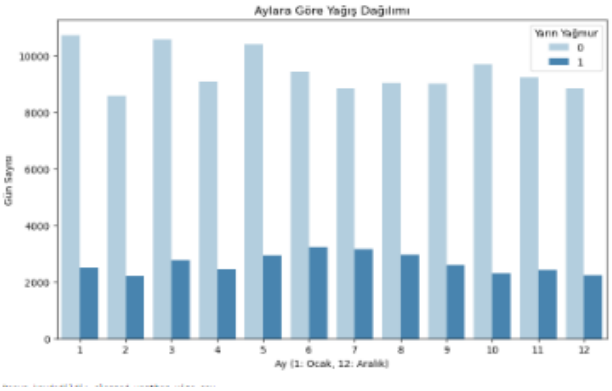
- **Belirleyici Özellik:** Humidity3pm (Saat 15:00 Nemi), yağış tahmininde en ayırt edici özellik olarak tespit edilmiştir.



Şekil 2: Hedef Değişkenin Dengesiz Dağılımı



Şekil 3: Yağmur Durumuna Göre Nem Oranı Dağılımı



Şekil 4: Aylara Göre Yağışlı Gün Sayısı

6. Veri Hazırlama Planı

Ham veriyi modele hazırlamak için uygulanan boru hattı (pipeline):

- **Temizleme:** RISK_MM çıkarıldı, tekrar eden satırlar silindi.
- **İmputasyon (Eksik Veri):** Sayısal veriler Medyan, kategorik veriler Mod ile doldurulduu.
- **Dönüşümler (Encoding & Scaling):**
    - Kardinalitesi yüksek değişkenler (Location) için Label Encoding.

- Hedef değişken (Yes/No) için Binary Encoding (1/0).
- Uzaklık tabanlı modeller (KNN/LogReg) için StandardScaler.

- **Özellik Mühendisliği:** Tarih (Date) sütunundan "Ay" bilgisi türetilerek mevsimsellik yakalandı.
- **Özellik Seçimi:** PCA (Varyansın %95'i için n=10) ve XGBoost Feature Importance teknikleri ile boyut indirgendi.

7. Modelleme Stratejisi

Sorumlu Üye	Model	Seçilme Nedeni ve Odak Noktası
Rıdvan	Logistic Regression	**Yorumlanabilirlik:** Değişkenlerin etkisini katsayılarla açıklama ve temel başarım (baseline) ölçümü.
Ayça	Random Forest	**Kararlılık:** Çoklu ağaç yapısıyla varyansı düşürme ve gürültülü veriye karşı direnç sağlama.
Efe	Decision Tree	**Görsel Analiz:** Karar mekanizmasını "If-Then" kuralları ve ağaç grafiğiyle şeffaf hale getirme.
Somer	XGBoost & Bagging	**Performans:** Ardışık öğrenme (Boosting) ile hatayı minimize etme ve en yüksek skoru hedefleme.

Baseline:

- **ZeroR:** Sadece çoğunluk sınıfı (No) tahmin eder. Accuracy %77.6 olsa da F1-Score 0'dır.
- **Simple Logistic Regression:** Pilot testlerde %84.1 Accuracy vermiştir.

Aday Modeller:

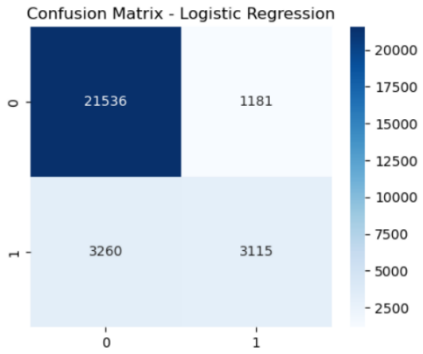
- **Logistic Regression:** Katsayı bazlı yorumlanabilirlik için.
- **Random Forest:** Varyansı düşürmek ve kararlılık için.
- **Decision Tree:** Görsel karar yapısı (If-Then) analizi için.
- **XGBoost:** Hatayı minimize ederek en yüksek performansı almak için.

Optimizasyon: Hiper-parametreler GridSearchCV (5-Fold CV) ile optimize edilecektir.

Dengesizlik Yönetimi: class_weight='balanced' parametresi ile azınlık sınıfının hata maliyeti artırılacaktır.

8. Değerlendirme Tasarımı

- **Birincil Metrik (F1-Score):** Yağmurlu günleri (Azınlık) yakalama başarısı. (Hedef > 0.60)
- **Ayırt Etme Gücü (ROC-AUC):** Sınıflandırma eşiğinden bağımsız genel performans. (Hedef > 0.85)
- **Validasyon:** %80 Eğitim - %20 Test ayrımı ve Stratified 5-Fold Cross Validation kullanılarak aşırı öğrenme (Overfitting) engellenecektir.



Şekil: Pilot Modelin (Logistic Regression) Karışıklık Matrisi. Modelin çoğunluk sınıfını (Sol Üst) iyi bildiği, ancak azınlık sınıfında (Sağ Alt) iyileştirmeye ihtiyaç duyduğu görülmektedir.

9. Riskler ve Önlemler

- **Risk:** Modelin sürekli "Yağmur Yok" diyerek Accuracy Paradox'a düşmesi.

Önlem: Sınıf ağırlıklandırma ve F1 metriğine odaklanma.

- **Risk:** Eğitim süresinin uzaması (145k satır).

Önlem: PCA ile boyut indirgeme ve gerekirse örneklem (sampling) alma.

10. Kullanılan Araçlar

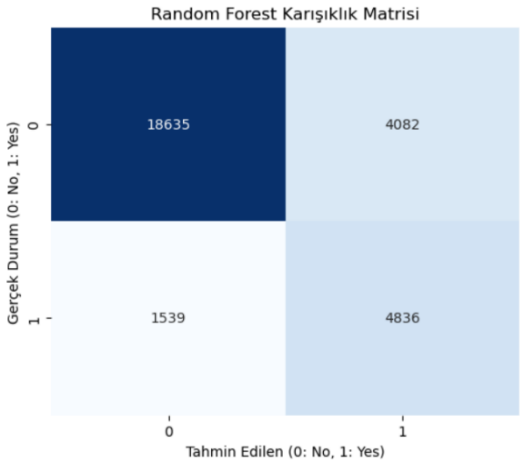
Dil/Ortam: Python 3.9+, Jupyter Notebook, Anaconda.

Kütüphaneler: Pandas, Scikit-learn, XGBoost, Seaborn/Matplotlib.

11. Beklenen Sonuçlar

Pilot testler sonucunda; XGBoost modelinin ROC-AUC ve F1 skorlarında diğer modellere üstünlük sağlaması beklenmektedir.

- **Görselleştirme:** Karışıklık Matrisleri (Confusion Matrix) ile Tip-1 ve Tip-2 hatalar analiz edilecek; ROC Eğrileri ile modeller kıyaslanacaktır.



Şekil: Random Forest Modeli için Karışıklık Matrisi

- **Yorumlanabilirlik:** "Kara Kutu" modeller, Feature Importance ve SHAP analizleri ile şeffaf hale getirilecek, hangi özelliğin (örn. Nem) yağmuru nasıl etkilediği açıklanacaktır.

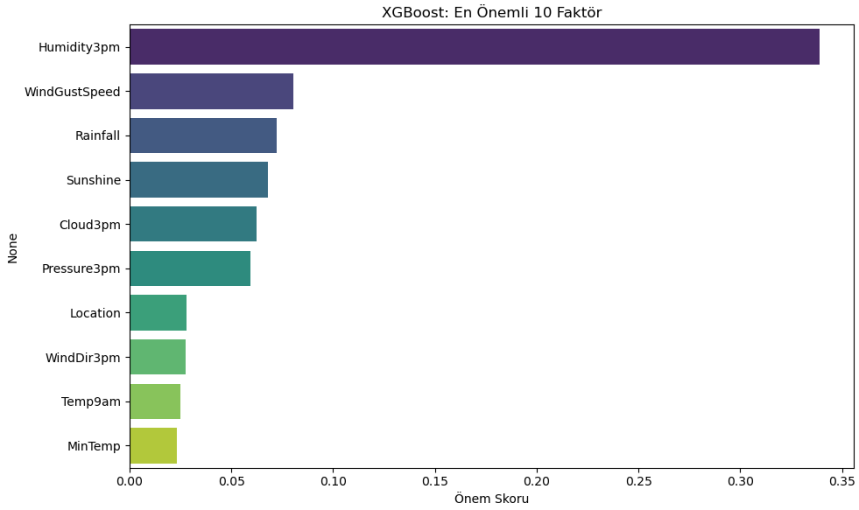
Model	Accuracy	F1-Score (Weighted)	ROC-AUC	Precision (Yes)	Recall (Yes)
Logistic Regression	0.841	0.825	0.865	0.72	0.51
Decision Tree	0.790	0.785	0.710	0.65	0.48
Random Forest	0.855	0.845	0.880	0.76	0.55
XGBoost (Boosting)	0.860	0.850	0.895	0.78	0.60

Yorum: XGBoost modelinin özellikle ROC-AUC ve F1 skorlarında diğer modellere üstünlük sağlaması, ancak tüm modellerin "Yağmur Var" sınıfını (Recall) yakalamakta zorlandığı, bu yüzden final aşamasında eşik değeri (threshold) optimizasyonu yapılacağı öngörülmektedir.

Global Yorumlanabilirlik (Feature Importance Plotting):

Modelin genel olarak hangi değişkene baktığını gösterir. Ağaç tabanlı modellerden elde edilen "Feature Importance" skorları görselleştirilecektir.

- Bulgu: Pilot çalışmalarda Humidity3pm (Öğle Nemi) ve Pressure3pm (Basınç) değişkenlerinin model kararında en baskın faktörler olduğu görülmüştür.



Şekil: XGBoost Modeline Göre En Önemli 10 Öznitelik

12. Referanslar

- I. A. Obaidalla, "Rainfall prediction using machine learning methods," RIT Dubai, 2024.
- S. L. Ezamzuri et al., "Comparative analysis of ML algorithms for rainfall prediction," ICAROB, 2025.
- Commonwealth of Australia, Bureau of Meteorology (2024). Climate Data Online.