

İnsülin Avcıları Project Outline

Bölüm	Kontrol Listesi Öğesi	Durum/Açıklama
I. Problem Tanımı & Hedefler		
Problem Tanımı	Diyabet hastalarının hastaneye tekrar yatış yapma olasılığının (readmission) yüksekliğini tahmin etmek.	
Hedef Belirleme	Hastanede kalış verilerini kullanarak, bir hastanın 30 gün içinde yeniden yatış yapıp yapmayacağını ikili sınıflandırma ile tahmin etmek.	
Başarı Ölçütü	Naive Bayes modelinin kanıtlandığı gibi, ROC AUC skorunu en az 0.70 seviyesine çıkarmak ve F1 Skoru ile modelin hem pozitif hem de negatif sınıflarda dengeli performans göstermesini sağlamak.	
II. Literatür İncelemesi & Kaynaklar		
Referans Karşılaştırması	2–4 referans karşılaştırıldı.	
Eksiklik Tespiti	Literatürde genellikle karmaşık Neural Network çözümlerine odaklanılmasına karşın, projemiz basit ve hızlı çalışan Naive Bayes modelinin bu veri setinde ne kadar etkili olabileceği eksikliğine odaklanmıştır.	
III. Veri Seti Detayları		
Veri Seti Kaynağı	Diabetes 130-US Hospitals Veri Seti (Figshare)	
Boyut & Şema	100.000 gözlem; 50'den fazla özellik (klinik, demografik, laboratuvar sonuçları, ilaçlar).	

Lisans & Etik Risk	Lisans belirtildi. Etik Risk: Irk (Race) ve Cinsiyet (Gender) gibi özelliklerin önyargı yaratma riski not edildi.	
IV. Keşifçi Veri Analizi (EDA)		
EDA Planı	Eksiklikler, aykırı değerler, sızıntılar, sınıf dengesi (imbalance) incelendi.	
Sızıntı (Leakage) Kontrolü	Readmission hedefine doğrudan bağlı olan 'discharge disposition' gibi değişkenler sızıntı riski açısından değerlendirildi.	
Sınıf Dengesi	Yeniden yatış yapmayan hasta sayısı, yapanlardan daha fazla olduğu için veri dengesizliği (imbalance) tespit edildi.	
V. Veri Ön İşleme (Preprocessing)		
Temizleme & Kodlama	Kategorik değişkenler One-Hot Encoding ile kodlandı; gereksiz yüksek kardinaliteli değişkenler çıkarıldı.	
İmputasyon	Eksik değerler (örneğin 'race' ve 'gender' için) en sık tekrar eden değer (mode) veya NaN olarak ele alındı.	
Ölçeklendirme	Sayısal özellikler StandardScaler (Pipeline içinde) kullanılarak ölçeklendirildi.	
Özellik Fikirleri	<i>Has_Circulatory</i> , <i>Has_Diabetes</i> gibi genel hastalık kategorilerine ait yeni boolean (ikili) özellikler türetildi.	
VI. Model Geliştirme Planı		

Temel Değerler (Baselines)	Rastgele tahmin (AUC: 0.50) ve Sınıf Dağılımına Dayalı Basit Kural (Çoğunluk Sınıfı).	
Aday Modeller	Naive Bayes (Şampiyon), Perceptron, LDA, Lojistik Regresyon Pipeline.	
Ayarlama Stratejisi	Grid Search veya Random Search ile hiperparametre ayarlaması yapıldı.	
Dengesizlik Yönetimi	Modellerde class_weight='balanced' parametresi kullanıldı.	
VII. Değerlendirme & Riskler		
Çapraz Doğrulama (CV)	Stratified K-Fold CV (5-kat) kullanıldı.	
Ölçütler	ROC AUC, F1 Skoru, Precision, Recall, Accuracy.	
Hata Analizi Planı	Karışıklık Matrisi (Confusion Matrix) ve Yanlış Pozitif/Negatif analizleri ile Naive Bayes modelinin neden bu kadar iyi çalıştığı derinlemesine incelendi.	
Riskler	Modelin hastane dışındaki verilerde genelleme yapma yeteneğinin düşük olması (Overfitting).	
Risk Azaltma	Düzenleştirme (Regularization) ve Basit Model (Naive Bayes) tercihi.	
VIII. Proje Yönetimi		
Zaman Çizelgesi	EDA, Temizleme, Modelleme, Karşılaştırma, Raporlama aşamaları için zaman ayrıldı.	

Roller	Veri Temizleme, Modelleme ve Raporlama rolleri belirlendi.	
Depo & Ortam	GitHub Deposu belirtildi Python, Scikit-learn (Sürüm Uyuşmazlığı Not Edildi), Pandas ortamı.	
Tohumlar (Seeds)	Tekrarlanabilirlik için random state/seed (42) kullanıldı.	
Klasör Yapısı	Veri, Notebook, Modeller ve Çıktılar klasörleri belirlendi.	