Ruh Sağlığı Verileriyle Depresyon Tahmini

Berkay ÜZER - 10009554206
Görkem YILDIZ - 28657235828

GitHub Link

Depresyon

+~300M

Dünya genelinde yaklaşık 300 milyondan fazla insanı etkileyen, en yaygın ruh sağlığı bozukluklarından biridir (WHO, 2021).

+~800k

Her yıl, depresyona bağlı olarak yaklaşık 800,000 kişi intihar sonucu hayatını kaybetmektedir. (WHO, 2021).

Erken tanı

Tedavi edilmediğinde yaşam kalitesini ciddi şekilde olumsuz etkileyen bu hastalık, erken tanı ve müdahale ile hem bireysel hem de toplumsal yükü azaltılabilir (Friedrich, 2017).

+%70

Erken teşhis ve tedavi ile depresyon hastalarının %70'inden fazlası belirgin iyileşme göstermektedir (Cuijpers et al., 2014).

Çözüm

Bireylerin çeşitli yaşam tarzı ve demografik özelliklerini kullanarak depresyon riskini tahmin eden bir model geliştirmek.

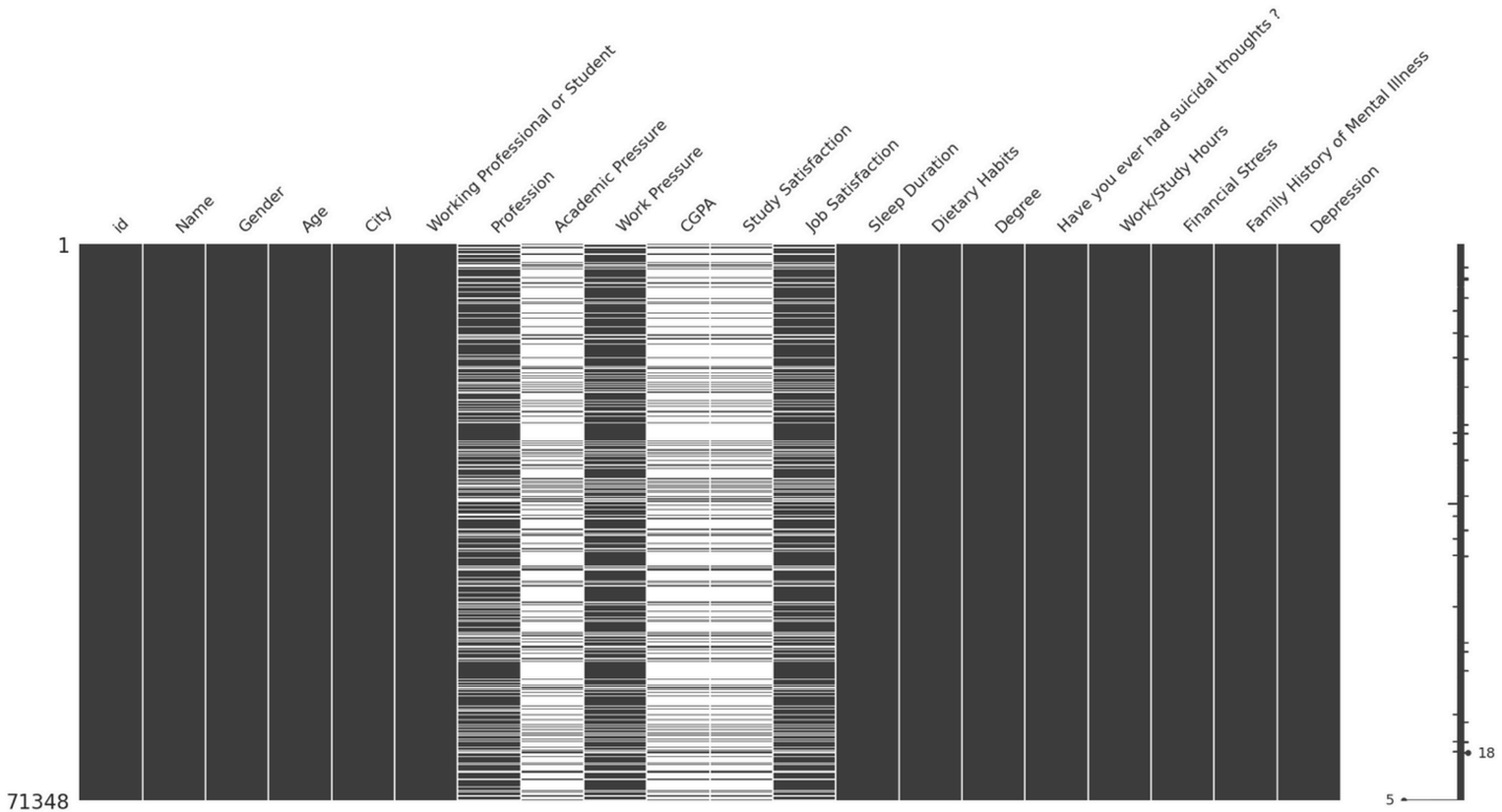
Mental Health Data
(Kaggle Playground Series S4E11)

Depresyon
Hedef Değişken

127,628
Toplam Veri Sayısı

Yaş, Cinsiyet, Şehir, Çalışma Durumu, Meslek, Akademik Baskı, İş Baskısı, Ağırlıklı Genel Not Ortalaması, Çalışma Tatmin Seviyesi, Uyku Düzeni, Diyet Düzeni, Eğitim, İntihara Eğilim, Çalışma Saati, Finansal Stres, Aile Geçmişi

Eksik Değerlerin Dağılımı



57177
Çalışma Memnuniyeti

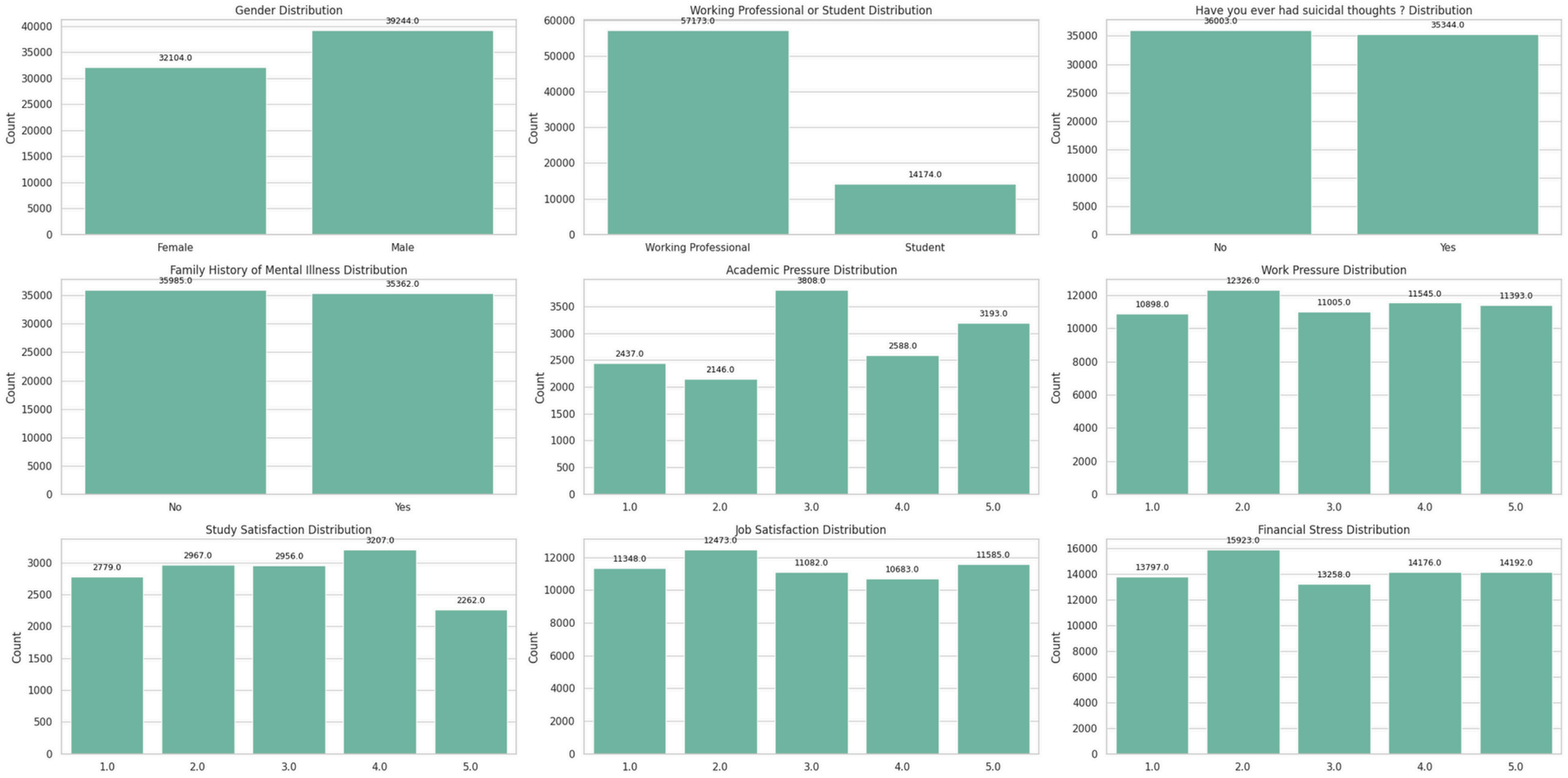
57176
Akademik Baskı

18653
Meslek

14181
İş Baskısı

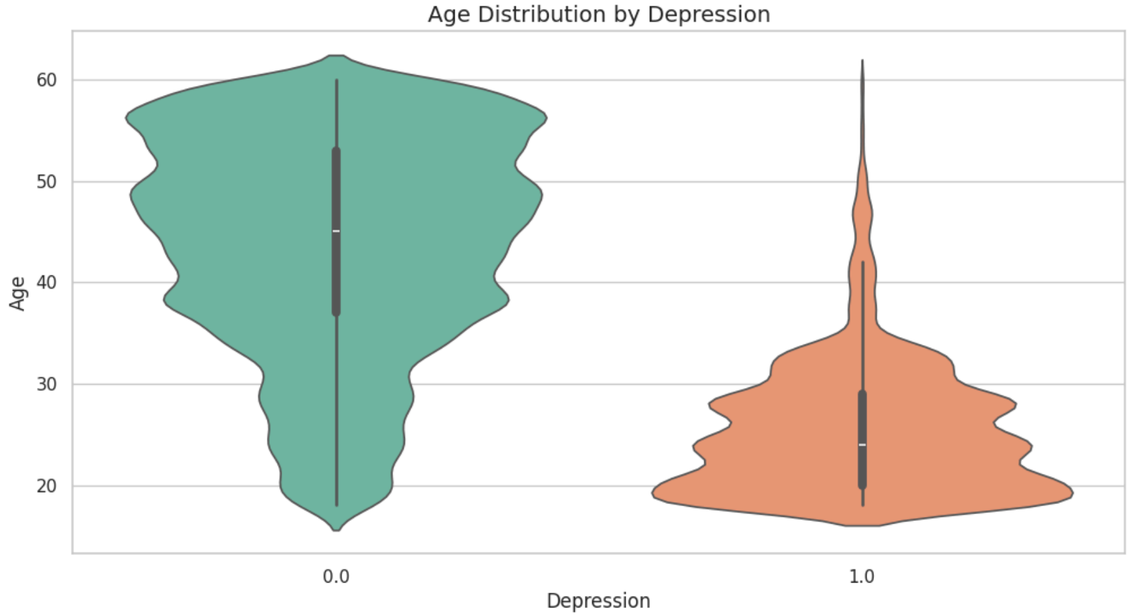
14177
İş Memnuniyeti

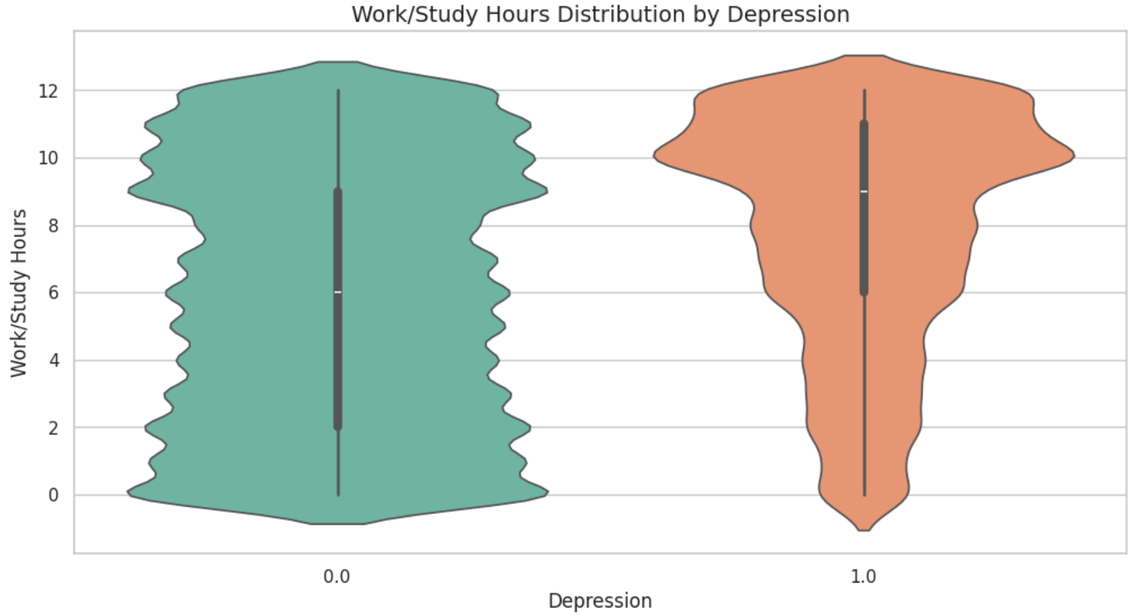
Kategorik Verilerin Dağılımı



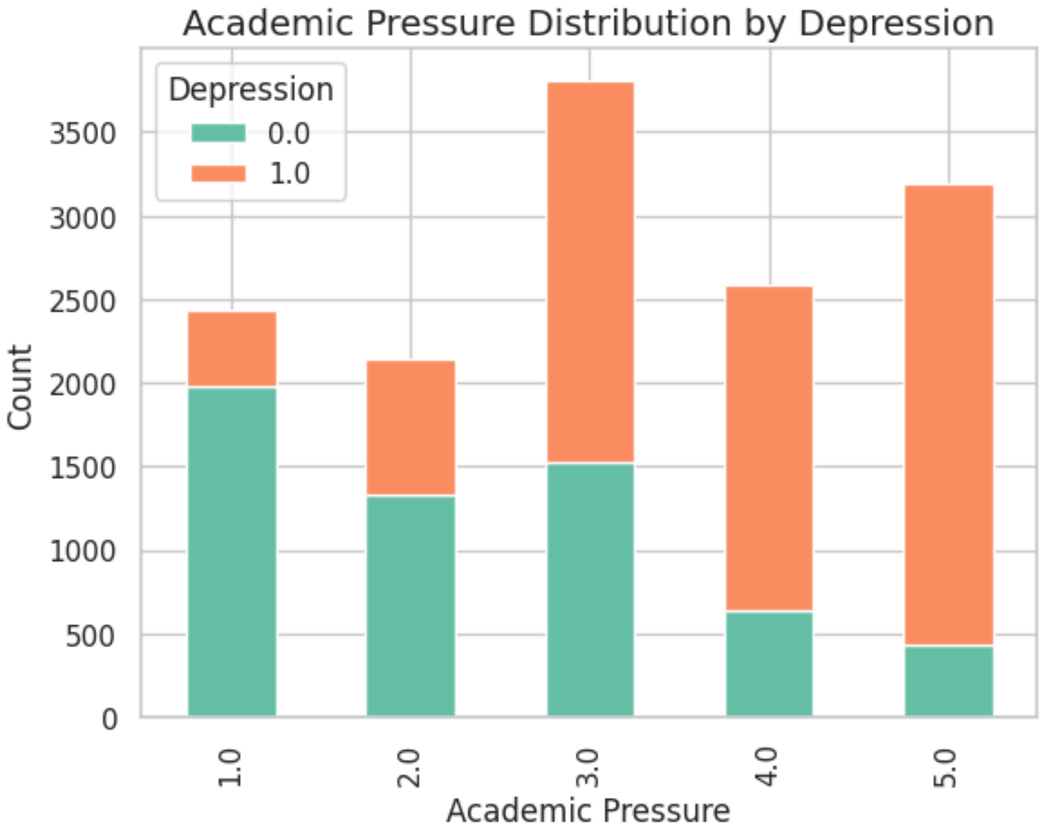
Cinsiyete Göre Depresyon ve İntihara Eğilim

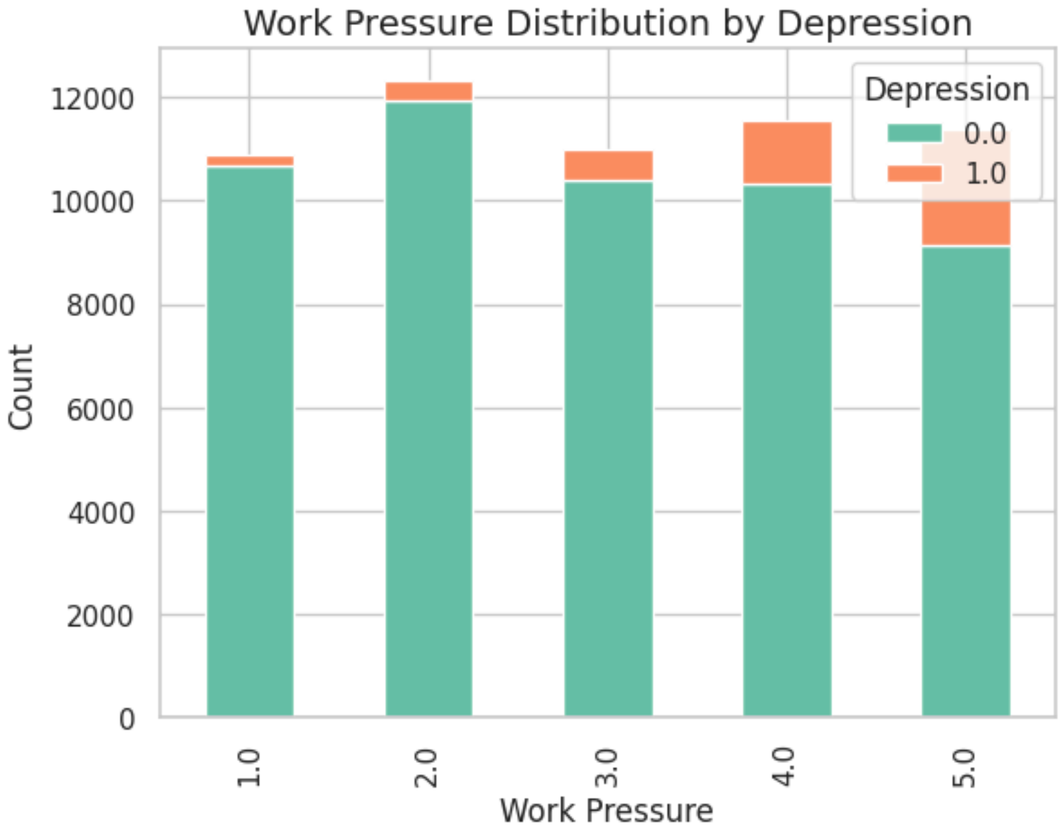
Yaşa ve Çalışma Saatine Göre Depresyon



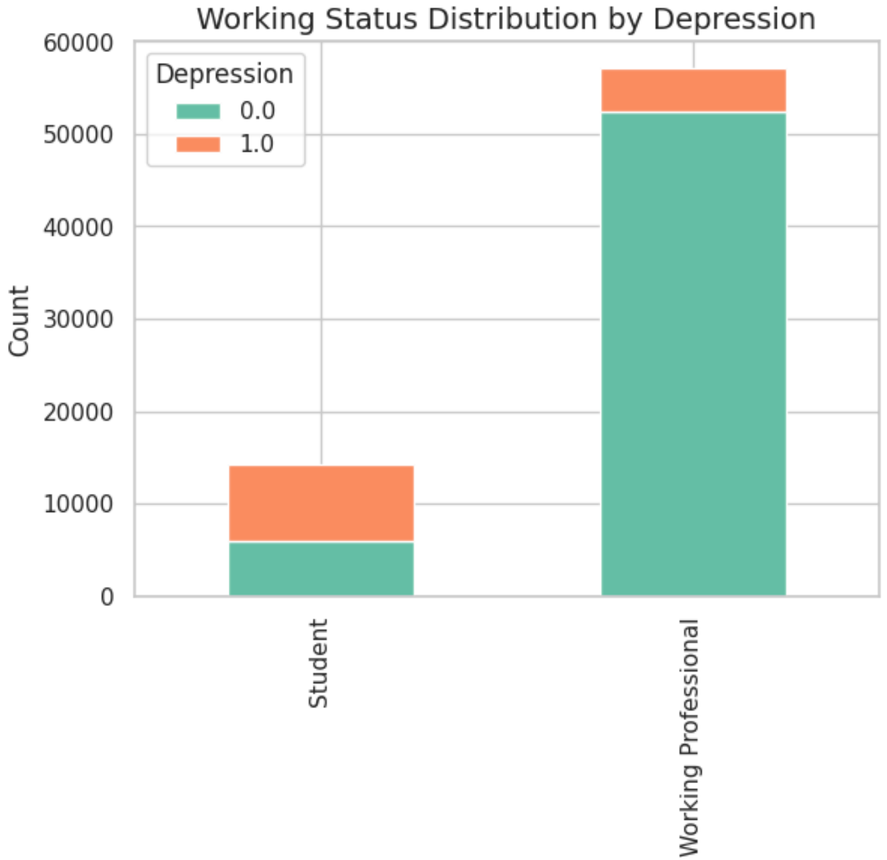


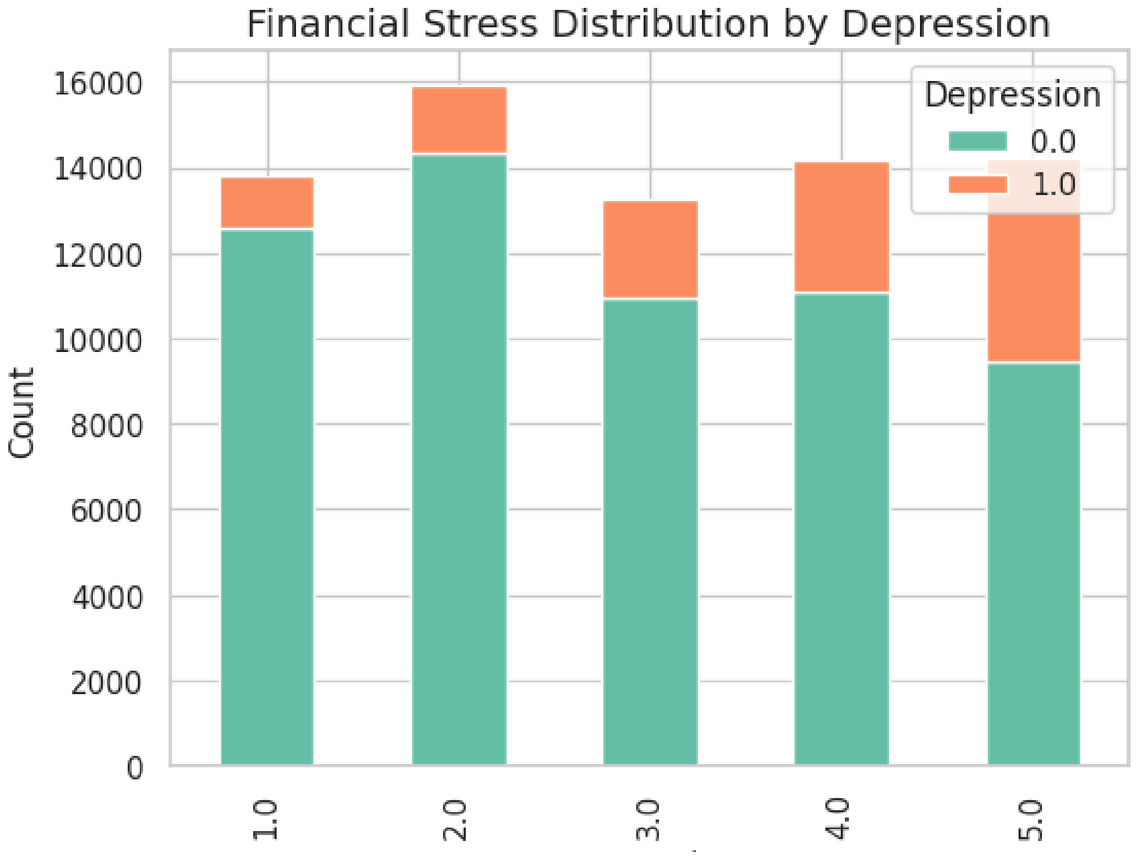
Akademik ve İş Baskısına Göre Depresyon



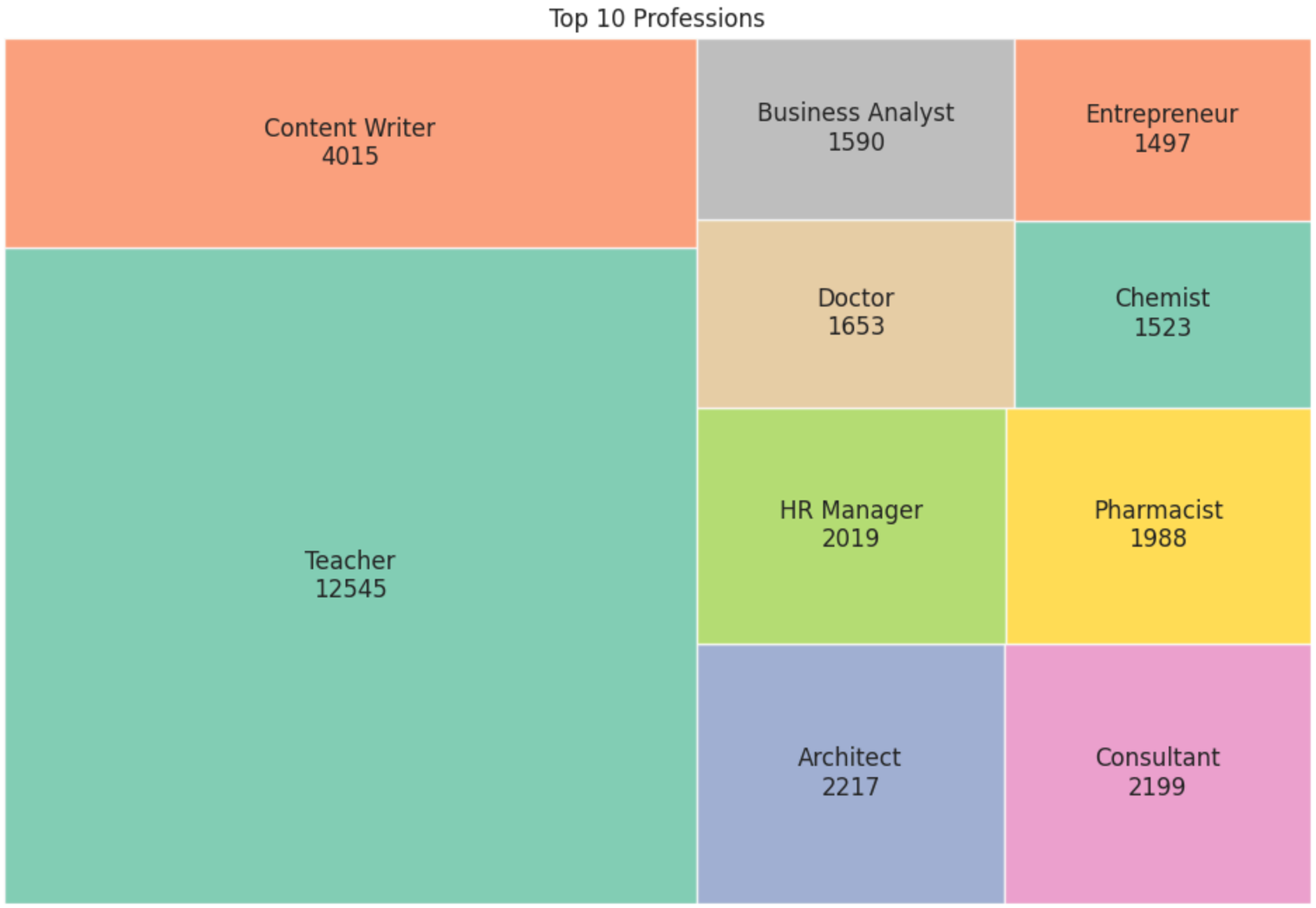


Çalışma Durumu ve Finansal Strese Göre Depresyon

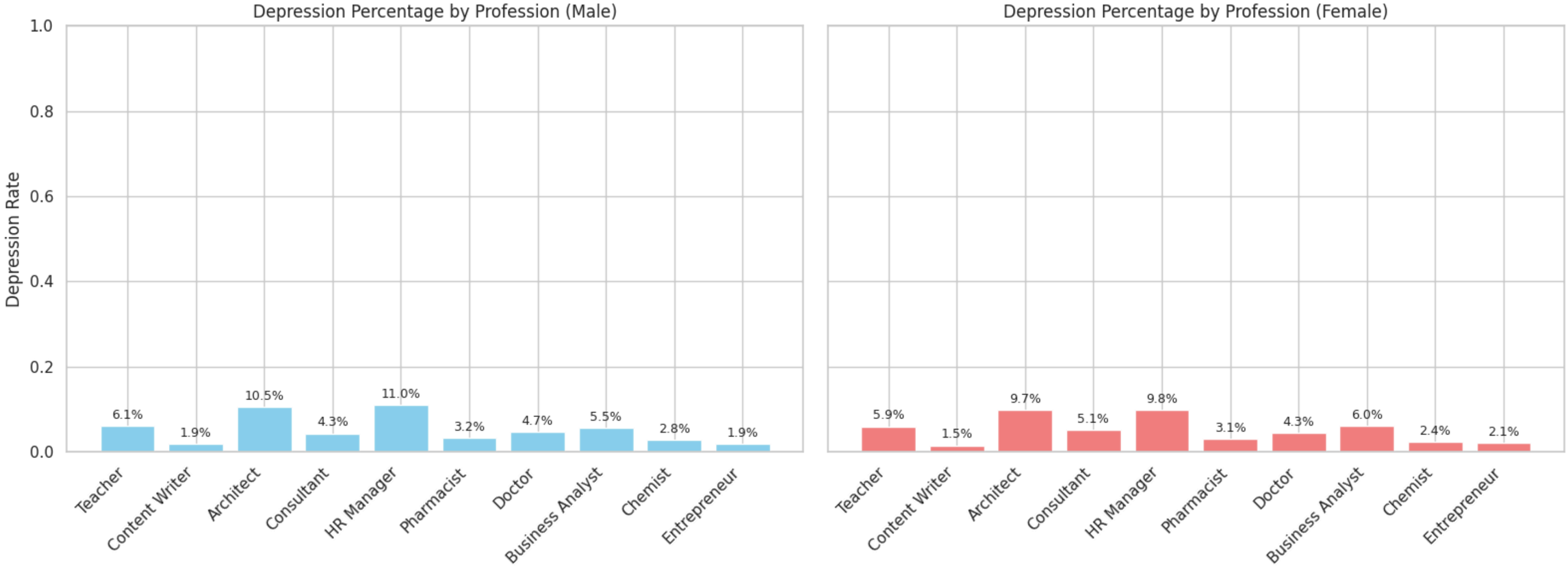




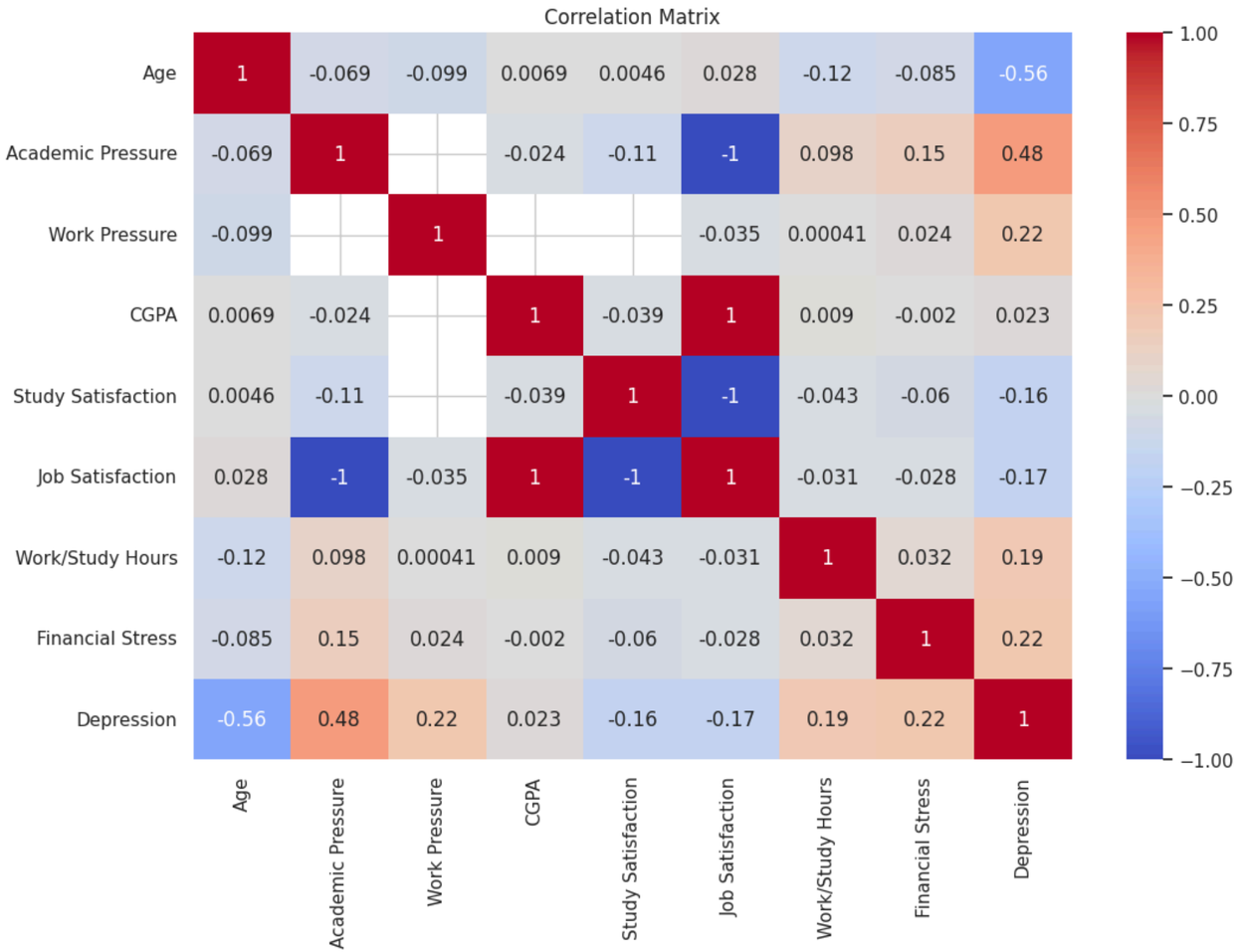
Meslek Dağılımı (İlk 10)



Mesleklere Göre Depresyon Oranı (Kadın - Erkek)



Korelasyon Matrisi



+0.48
Akademik Baskı ve Depresyon

+0.22
Finansal Stres ve Depresyon

−0.56
Yaş ve Depresyon

−0.17
İş Memnuniyeti ve Depresyon

Veri Temizleme

Meslek sütunundaki **“Öğrenci”** değerleri **Çalışma Durumu** sütunundaki **“Öğrenci”** değerleri ile birleştirildi.

Çalışanların akademik baskısı, **Öğrencilerin** iş bakısı **sıfıra** eşitlendi.

14 farklı **Diyet Düzeni** değeri **“Orta”, “Sağlıklı, “Sağlıksız”** ve **“NaN”** olmak üzere haritalandı.

Çalışma Memnuniyeti ve **İş Memnuniyeti** özellikleri tek sütunda birleştirildi.

Uzun isme sahip sütunlar daha kısa bir biçimde isimlendirildi.

23 farklı **Uyku Düzeni** değeri **“<5 saat”, “5-6 saat”, “6-7 saat”, “7-8 saat”** ve **“>8 saat”, “NaN”** ile haritalandı.

“Meslek”, “Şehir” ve **“Eğitim Seviyesi”** sütunlarında yer alan **5'ten az** olan değerler **“Diğer”** olarak sınıflandırıldı.

Özellik Mühendisliği

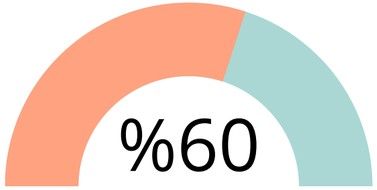
Mesleklere göre sentetik olarak **“Maaş”** sütunu oluşturuldu.

Mesleklere göre **“Çalışma Ortamı”** (ofis, uzaktan vs.) sütunu oluşturuldu.

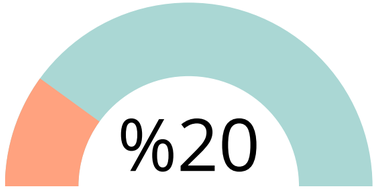
Maaş / Çalışma Saati değerlerine göre **“Saatlik Maaş”** sütunu oluşturuldu.

Maaş / Yaş değerlerine göre **“Yaşa Göre Gelir”** sütunu oluşturuldu.

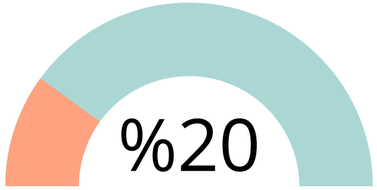
Verinin Bölünmesi



Eğitim Seti



Validasyon Seti



Test Seti

Veri Önişleme

Eksik değerler; *sayısal özellikler* için **“median”**, *kategorik değerler* için **“most frequent”** ile dolduruldu.

Kategorik özellikleri sayısal bir temsile dönüştürmek için **OrdinalEncoder** kullandık.

Veri Artırma

Model, **“Depresyon”** sınıfı az olduğundan vakaları tespit etmede zorlanabilirdi. Bu dengesizliği gidermek için eğitim verimizde **SMOTE (Synthetic Minority Over-sampling Technique)** tekniğini kullandık.

SMOTE, azınlık sınıfını artırmak için kullanılan bir yöntemdir. Bu yöntem, azınlık sınıfındaki veriler arasında benzerliklere dayalı olarak sentetik örnekler üretir. Var olan verileri birebir kopyalamak yerine, aralarındaki mesafeleri kullanarak yeni ve gerçekçi veriler oluşturur.

Depresyon	Orijinal	w/SMOTE
0.0	46714	46714
1.0	10363	46714

Model Geliştirme

XGBoost (Extreme Gradient Boosting)

Karar ağaçlarına dayalı popüler **boosting** algoritmalarından biridir. Hızlı, verimli ve yüksek doğruluk sunar. Eksik verilerle iyi başa çıkabilir ve düzenleme desteği sayesinde aşırı öğrenmeye karşı dirençlidir.

LightGBM (Light Gradient Boosting Machine)

Özellikle büyük veri setlerinde yüksek hızda ve düşük bellek kullanımıyla öne çıkar. Bilgiyi daha verimli ayıran **Leaf-wise** büyüme stratejisini kullanır. Kategorik değişkenleri otomatik işler.

CatBoost (Categorical Boosting)

Kategorik verilerle doğal olarak çalışmak üzere geliştirilmiş bir **boosting algoritmasıdır.** Ön işleme ihtiyaç duymaz ve **overfitting'**e karşı dayanıklıdır. Özellikle **dengesiz ve karmaşık veri setlerinde** güçlü sonuçlar verir.

Ensemble Learning

Her bir modelin en iyi performansını elde etmek için **Optuna** ile hiperparametre optimizasyonu yaptık.

Bu üç modelin tahminlerini birleştirmek için **"Soft Voting"** yöntemini kullanan bir **Voting Classifie**r oluşturduk.

Soft voting işleminde, her bir modelin sınıflar için ürettiği olasılıklar ortalaması alınır ve bu ortalama olasılığa göre tahmin yapılır.

Hiperparametre Optimizasyonu

XGBoost

- **n_estimators:** 766
- **max_depth:** 5
- **learning_rate:** 0.0929
- **subsample:** 0.822
- **colsample_bytree:** 0.695
- **reg_alpha:** 0.009
- **reg_lambda:** 0.382
- **gamma:** 3436
- **min_child_weight:** 7

CatBoost

- **iterations:** 826
- **depth:** 10
- **l2_leaf_reg:** 6.669
- **learning_rate:** 0.020
- **border_count:** 205
- **random_strength:** 8.612
- **bagging_temperature:** 0.825
- **od_type:** IncToDec

LightGBM

- **learning_rate:** 0.0192
- **num_leaves:** 41
- **max_depth:** 8
- **min_data_in_leaf:** 79
- **feature_fraction:** 0.5289
- **bagging_fraction:** 0.923
- **bagging_freq:** 6
- **lambda_l1:** 0.0433
- **lambda_l2:** 0.0843
- **n_estimators:** 429

Sonuçlar

Metrik	Validasyon Seti	Test Seti
Accuracy	91.18	91.41
Precision	73.13	74.20
Recall	81.32	80.83
F1-Score	77.01	77.38

Depresyon	Precision	Recall	F1-Score	Destek
0.0	0.96	0.94	0.95	23027
1.0	0.74	0.81	0.77	5113
accuracy			0.91	28140
macro avg	0.85	0.87	0.86	28140
weighted avg	0.92	0.91	0.92	28140

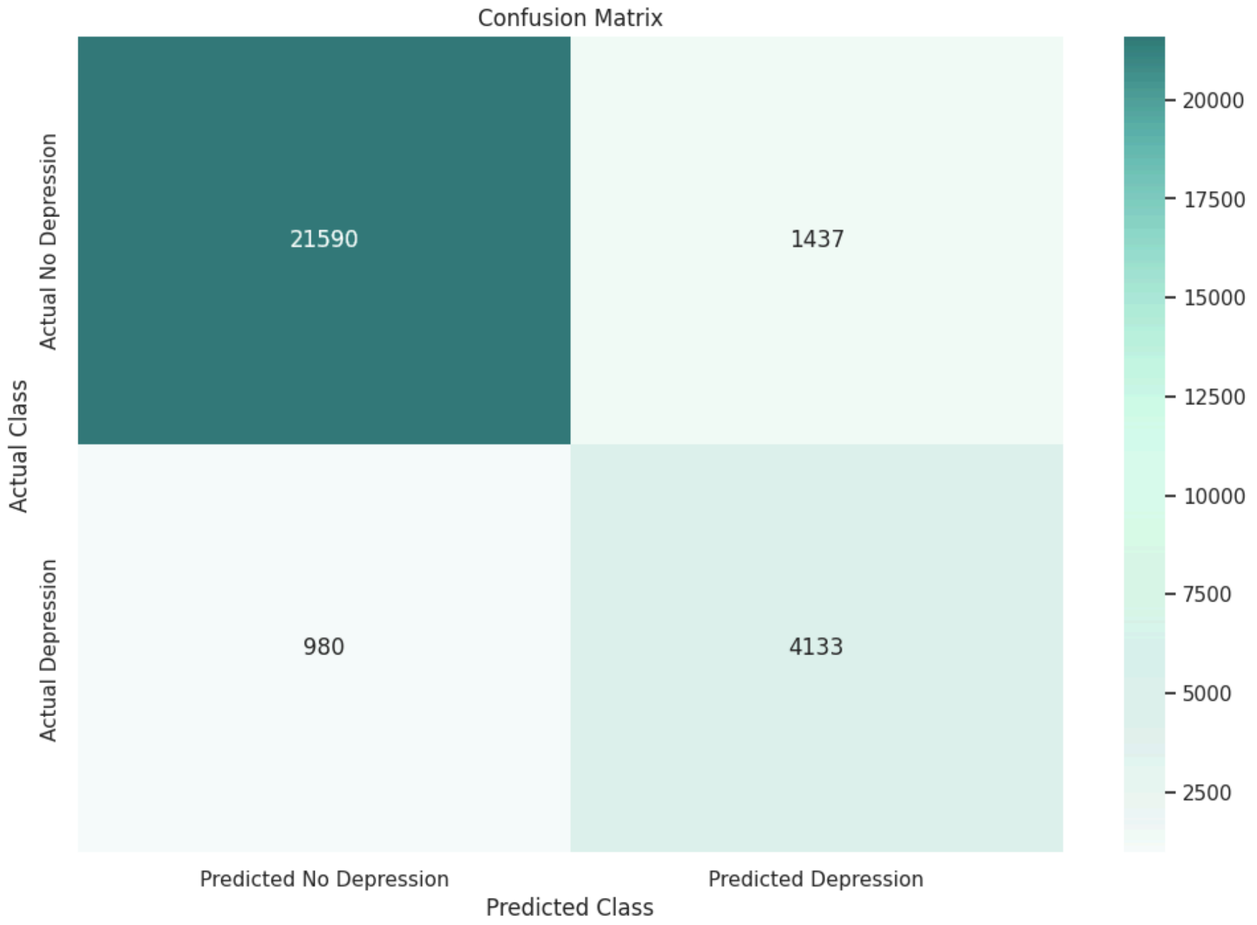
Sonuçlar

Modelimiz, SMOTE ve eşik optimizasyonu ile depresyon vakalarını tespit etmede **belirgin bir başarı gösterdi.**

Gerçek depresyon vakalarının **%80.83**'ünü doğru bir şekilde belirleyebildik.

Precision ve Recall dengesini gösteren F1 skorumuz **%77.38** ile modelimizin genel olarak güvenilir olduğunu kanıtlıyor.

Sonuçlar, modelimizin depresyon riski taşıyan bireyleri etkili bir şekilde işaretleyebileceğini gösteriyor.



Dinlediğiniz için teşekkürler!

GitHub Link