FET445 Veri Madenciliği

Psikolojik Sağlık Geçmişi Sınıflandırma Sistemi

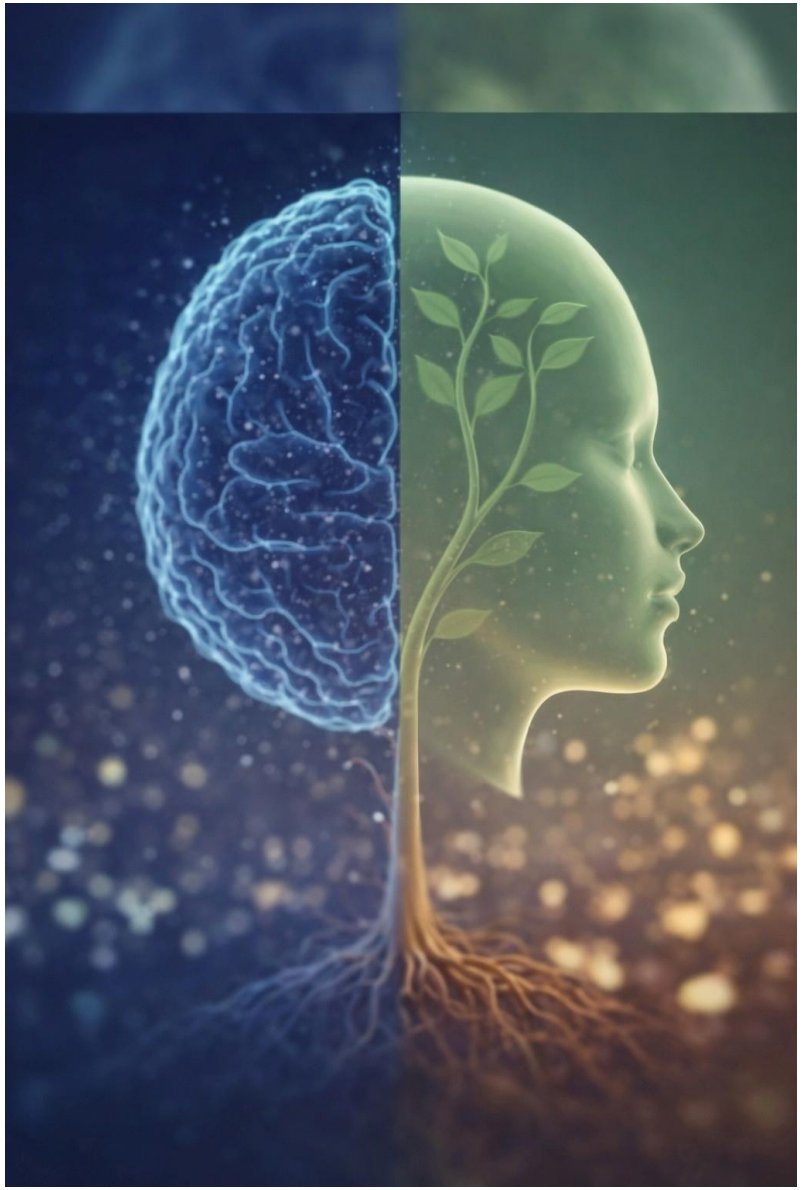
Digital Nomads

Wisam Almohamed
22040301146

Cüneyd Melledham
22040301244

Mustafa TÜRKMEN
22040301166

08.01.2026



PROBLEM TANIMI

Psikolojik Geçmişin Günlük Yaşamda Görünür Olup Olmadığı Sorusu ?

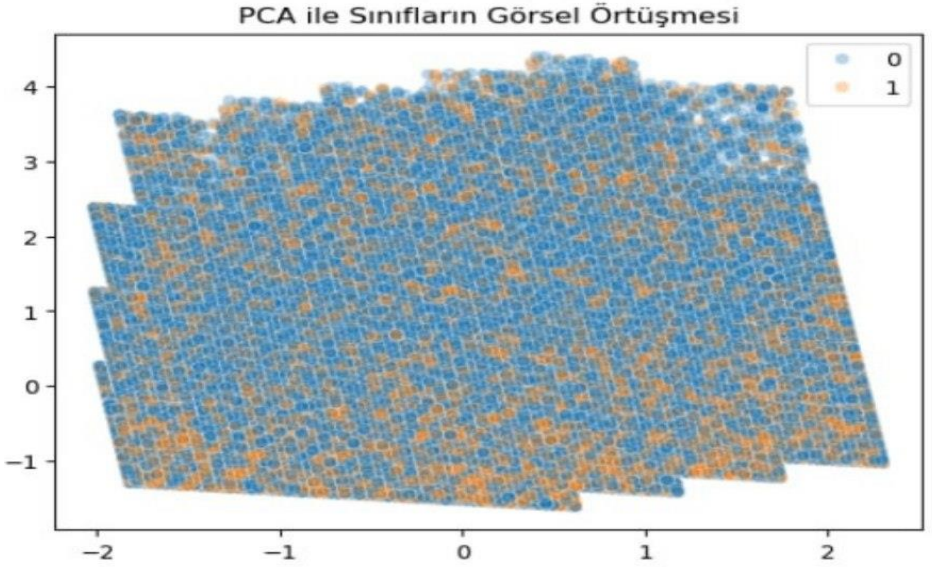
Veri Seti Açıklaması

9]:

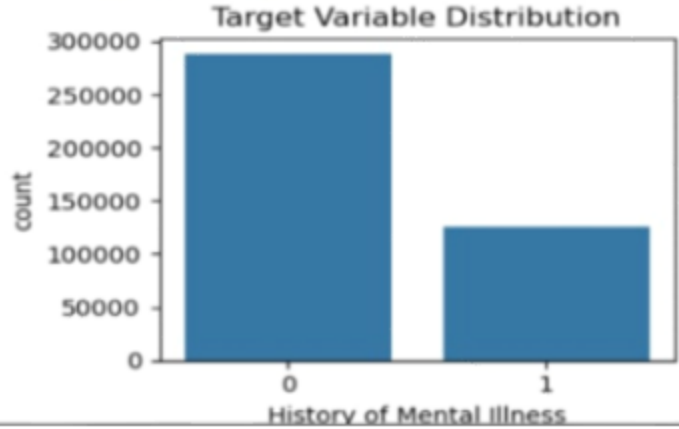
	Feature	Chi2_p_value	Cramers_V
4	Employment Status	0.000000e+00	0.140759
1	Education Level	5.678974e-302	0.058176
7	Sleep Patterns	3.304691e-129	0.037815
6	Dietary Habits	2.124259e-79	0.029591
5	Alcohol Consumption	2.712470e-17	0.013579
0	Marital Status	5.409032e-14	0.012519
3	Physical Activity Level	2.556256e-07	0.008566
10	Chronic Medical Conditions	1.043505e-04	0.006032
2	Smoking Status	3.273014e-03	0.005259
9	Family History of Depression	1.689708e-03	0.004881
8	History of Substance Abuse	2.666101e-01	0.001727

[83]:

	Feature	MannWhitney_p_value	Cohens_d
2	Income	0.000000e+00	-0.306509
0	Age	7.646928e-57	0.053569
1	Number of Children	1.742730e-02	0.005905




```
Target distribution (ratio):
History of Mental Illness
0    0.695904
1    0.304096
Name: proportion, dtype: float64
```



Veri Seti Açıklaması

Bu projede, Kaggle platformunda yer alan ve **Anthony Therrien** tarafından oluşturulmuş olan **Depression Dataset** kullanılmıştır. Bu veri seti sentetik (yapay) bir veri setidir ve gerçek klinik hasta verilerine dayanmamaktadır.

Veri seti, ruh sağlığı ile yaşam tarzı ve sosyoekonomik faktörler arasındaki ilişkileri incelemeyi kolaylaştırmak amacıyla tasarlanmıştır.

Kendisi bir **veri bilimci ve yazılım mühendisidir**.
Bu veri seti, **gerçek klinik veya tıbbi verilere dayanmadan**, ruh sağlığı, yaşam tarzı ve sosyoekonomik faktörler arasındaki ilişkinin incelenmesini kolaylaştırmak amacıyla tasarlanmıştır.

https://www.kaggle.com/datasets/anthonytherrien/depression-dataset/data

Veri Hazırlama ve Modelleme Stratejisi

•Çok aşamalı bir veri hazırlama ve modelleme stratejisi benimsenerek, sistemin ruh sağlığı açısından hassas vakaları ayırt etme kabiliyeti artırılmıştır.

- Hedef değişken History of Mental Illness, özel öneme sahip bir ikili sınıflandırma problemi olarak ele alınmış ve False Negative hataların azaltılmasına odaklanılmıştır.
- Temel modellerde, veri setindeki sınıf dengesizliği ve sınıflar arası yüksek benzerlik nedeniyle Recall değerlerinin düşük kaldığı gözlemlenmiştir.
- Bu durumu ele almak amacıyla, ilk aşamada RandomForest + SMOTETomek kullanılarak potansiyel vakaların yakalanması hedeflenmiş, ikinci aşamada ise XGBoost modeliyle Recall ve Precision arasında daha dengeli bir yapı oluşturulmuştur.
- Threshold değerleri manuel olarak belirlenmemiş, Precision–Recall eğrisi ve F1-skoru analizlerine dayanarak istatistiksel olarak optimize edilmiştir.
- Bu strateji, modelin hassasiyetini artırırken, klinik olmayan verilerin doğasından kaynaklanan gerçekçi sınırlamaları da dikkate alan metodolojik bir yaklaşımı yansıtmaktadır.

Kullanılan Baz modeller

[25]:

	Model	Accuracy	Recall (Positive)	F1-score (Positive)	ROC AUC	Balanced
0	KNN	0.6400	0.21	0.26	0.5400	No
1	Decision Tree	0.5800	0.34	0.33	0.5100	No
2	Random Forest	0.6500	0.17	0.23	0.5500	No
3	Linear SVM	0.7000	0.00	0.00	NaN	No
4	Linear SVM (balanced)	0.6100	0.47	0.42	NaN	Yes
5	Logistic Regression (baseline)	0.6959	0.00	0.00	0.5917	No
6	Logistic Regression (balanced)	0.6100	0.47	0.42	0.5900	Yes
7	SGD Classifier	0.6000	0.48	0.42	0.5900	No
8	SGD Classifier (balanced)	0.6000	0.48	0.42	0.5900	Yes
9	Passive Aggressive (balanced)	0.5900	0.26	0.28	NaN	Yes
10	Complement Naive Bayes	0.6000	0.51	0.43	0.5900	No
11	Gaussian Naive Bayes	0.6300	0.36	0.37	0.5900	No

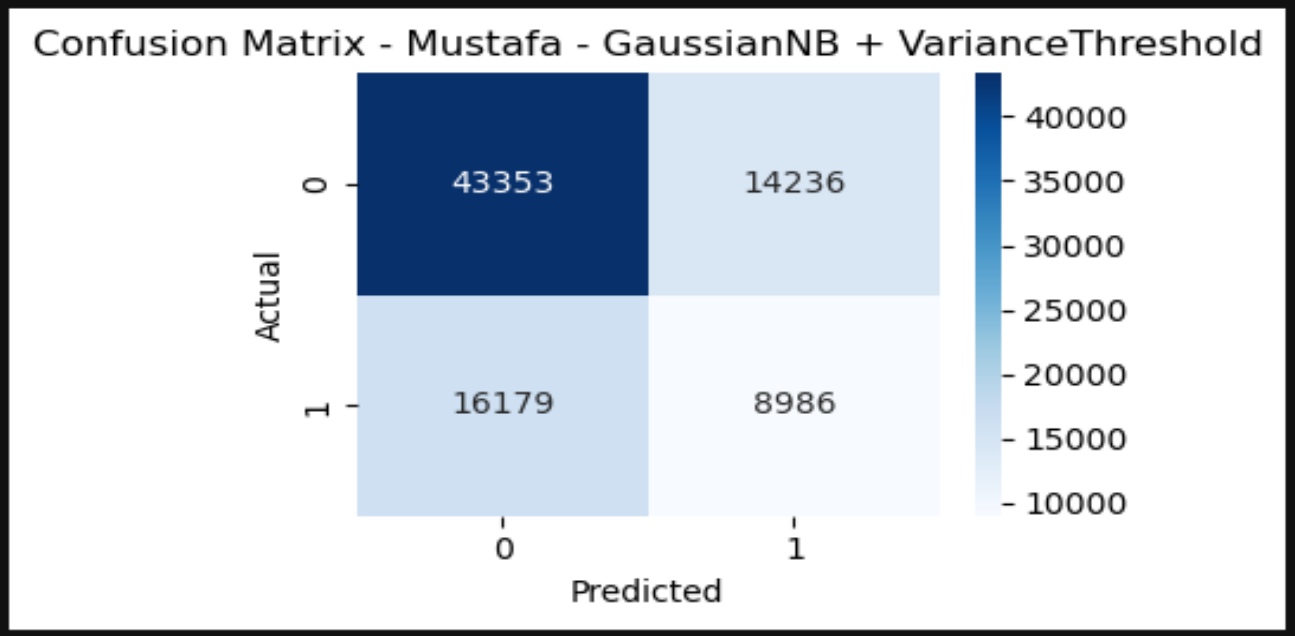
•Baz modeller karşılaştırıldığında, Complement Naive Bayes en yüksek Recall (0.51) ve F1-score değerlerini sağlayarak pozitif sınıfı yakalamada en dengeli performansı göstermiştir.

- Diğer modeller bazı durumlarda daha yüksek Accuracy üretmiş olsa da, pozitif sınıfı tespit etmede başarısız kalmışlardır.
- Bu nedenle, değerlendirmede Accuracy yerine Recall ve F1-score önceliklendirilmiştir.

Boyut İndirgeme ve Özellik Seçimi

Classification Report:

	precision	recall	f1-score	support
0	0.73	0.75	0.74	57589
1	0.39	0.36	0.37	25165
accuracy			0.63	82754
macro avg	0.56	0.55	0.56	82754
weighted avg	0.62	0.63	0.63	82754



•seti üzerinde boyut indirgeme ve özellik seçimi yöntemleri denenmiştir.

- Ancak sınıf dengesizliği nedeniyle bu yaklaşımlar Recall ve ayırt edicilik açısından anlamlı bir iyileşme sağlamamıştır.
- Bu nedenle bu yöntemler nihai çözüm olarak kullanılmamış, yalnızca örnek ve analiz amaçlı sunulmuştur.

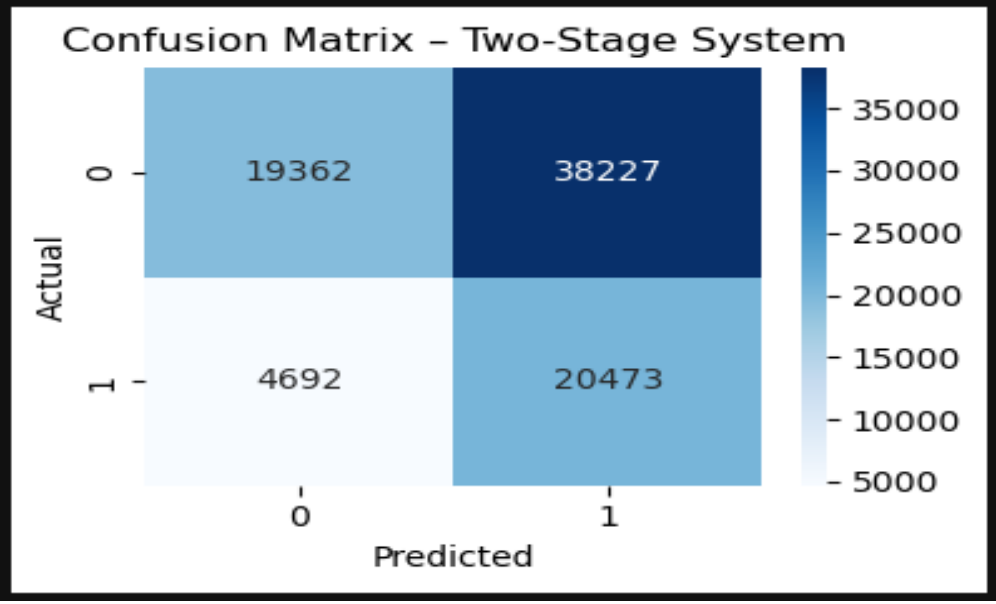
Sistemlerin Gerçek Sayılarla Karşılaştırılması

1 Gerçek Rakamlarla Karşılaştırma (Karmaşıklık Matrisi Analizi)

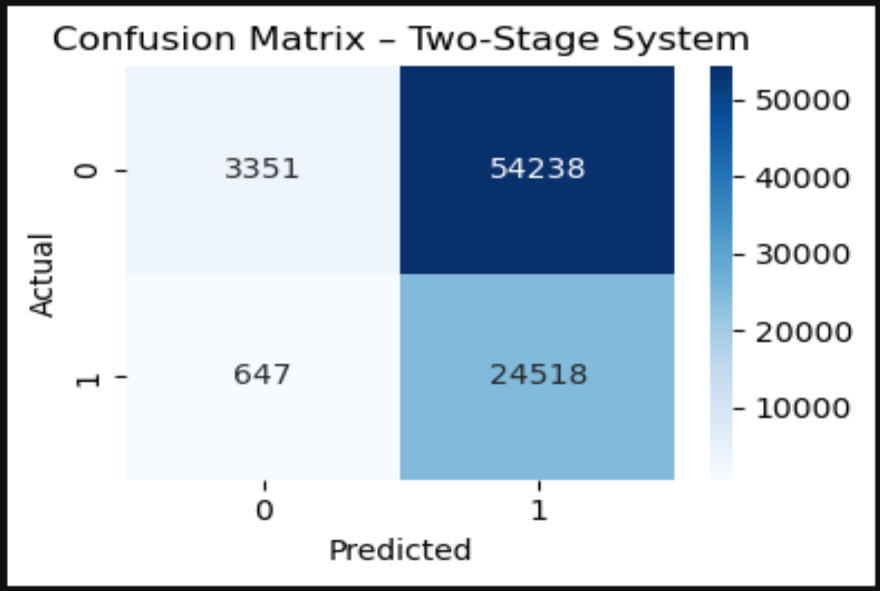
Metrik	Agresif Sistem (%97 Recall)	Dengeli Sistem (%81 Recall)	Değişimin Etkisi
Toplam Gerçek Hasta Sayısı	**25.165**	**25.165**	-
Sistemin Yakaladığı Hastalar (TP)	✅ **24.518** kişi	✅ **20.473** kişi	Tespit gücünde azalma
Sistemin Kaçırdığı Hastalar (FN)	❌ **647** kişi	❌ **4.692** kişi	Kayıp vakalarda artış
Toplam Gerçek Sağlıklı Sayısı	**57.589**	**57.589**	-
Doğru Tespit Edilen Sağlıklı Kişiler (TN)	✅ **3.351** kişi	✅ **19.362** kişi	Sağlıklıları korumada büyük artış
Hasta Denilen Ama Aslında Sağlıklı Olanlar (FP)	❌ **54.238** kişi	❌ **38.227** kişi	Yanlış alarmlarda ciddi azalma

İki Aşamalı Sistemlerde Davranış Farklılığı

```
=== Final Two-Stage System Results ===
              precision    recall  f1-score   support

           0       0.80      0.34      0.47     57589
           1       0.35      0.81      0.49     25165

    accuracy                           0.48     82754
   macro avg       0.58      0.57      0.48     82754
weighted avg       0.67      0.48      0.48     82754
```



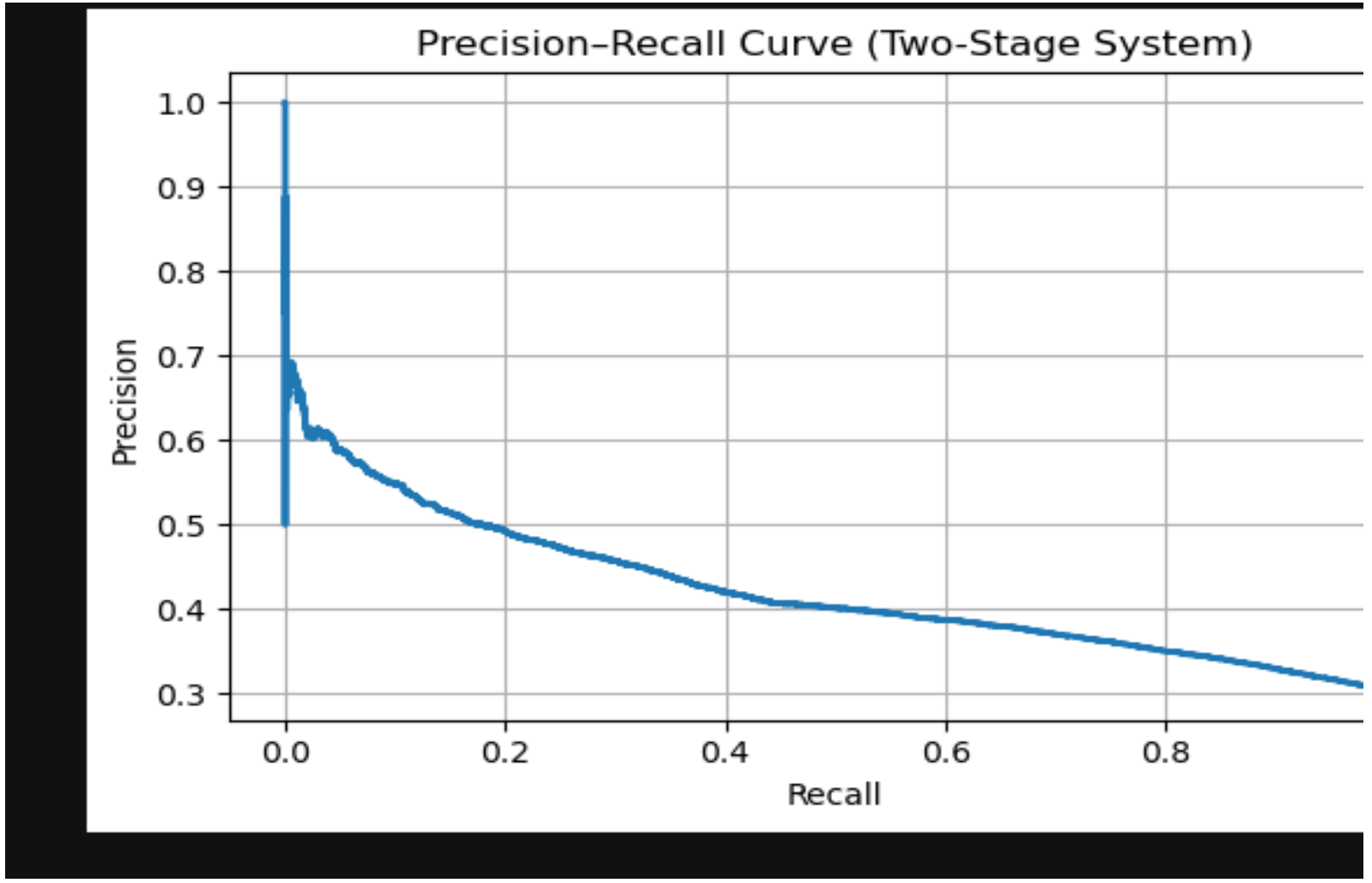

```
=== Final Two-Stage System Results ===
              precision    recall  f1-score   support

           0       0.84      0.06      0.11     57589
           1       0.31      0.97      0.47     25165

    accuracy                           0.34     82754
   macro avg       0.57      0.52      0.29     82754
weighted avg       0.68      0.34      0.22     82754
```



Threshold Seçiminin Precision–Recall Üzerindeki Etkisi

•Eğri üzerindeki her nokta, farklı bir Threshold değerini temsil etmektedir.

- Recall artırıldığında (daha fazla hassas vakanın yakalanması hedeflendiğinde),
- False Positive sayısının artması nedeniyle Precision düşmektedir.
- Tersi durumda ise Precision artarken Recall azalmaktadır.
- Bu grafik, mevcut verilerde Precision ve Recall değerlerinin aynı anda iyileştirilemeyeceğini görsel olarak ortaya koymaktadır.
- Bu nedenle Threshold seçimi, yalnızca teknik bir karar değil, sistemin amacına bağlı metodolojik bir tercihtir.



Problemin Çerçevelenmesi ve Ruh Sağlığında Recall Önceliği

- Bu projenin amacı, bireylerin davranışsal ve yaşam tarzına ilişkin genel verilerini kullanarak, geçmişte psikiyatrik bir hastalık veya ruhsal bozukluk yaşayıp yaşamadıklarını sınıflandırabilen bir sistem geliştirmektir.
- Bu çalışma klinik tanı koymayı değil, ruh sağlığı alanında ön tarama ve karar destek süreçlerini desteklemeyi hedeflemektedir.
- Problem, yalnızca istatistiksel bir tahmin görevi olarak değil, psikolojik açıdan hassas bir ikili sınıflandırma problemi olarak ele alınmıştır.
- Bu nedenle, ruhsal geçmişe sahip bireylerin gözden kaçırılması, sağlıklı bireylerin yanlışlıkla pozitif sınıflandırılmasından daha kritik kabul edilmiştir.
- Bu doğrultuda model değerlendirmesinde yalnızca doğruluk (Accuracy) değil, Duyarlılık (Recall) ve F1-Skoru metriklerine öncelik verilmiştir.
- Çünkü ruh sağlığı bağlamında False Negative hatalar, yani riskli bireylerin risksiz olarak değerlendirilmesi, daha ciddi sonuçlar doğurabilmektedir.
- Ayrıca birçok bireyin geçmişte yaşadığı ruhsal sorunlar resmî olarak teşhis edilmemiş olabilir veya bireyler farkında olmadan ruhsal rahatsızlığı olan kişilerle etkileşimde bulunmuş olabilir.
- Bu nedenle, ön tarama sistemlerinde yüksek Recall değerine sahip olmak, yanlış pozitiflerin artması pahasına bile olsa, metodolojik olarak gerekçelendirilmiş bir tercihtir.

Feature Engineering

Yüksek Boyutlu Özellik Uzayı ve Hesaplama Kısıtları

[32]:
```
from sklearn.ensemble import RandomForestClassifier

rf_pipe = Pipeline(steps=[
    ("preprocess", preprocessor_fe),
    ("model", RandomForestClassifier(
        class_weight="balanced_subsample",
        random_state=42,
        n_jobs=-1
    ))
])

rf_param_grid = {
    "model__n_estimators": [300, 600],
    "model__max_depth": [None, 8, 12, 16],
    "model__min_samples_split": [2, 5, 10],
    "model__min_samples_leaf": [1, 2, 5]
}

rf_search = GridSearchCV(
    rf_pipe,
    rf_param_grid,
    scoring=scoring,
    refit="mcc",
    cv=cv,
    n_jobs=-1,
    verbose=2
)

rf_search.fit(X_train_fe, y_train_fe)
print("Best RF params:", rf_search.best_params_)
print("Best RF MCC (CV):", rf_search.best_score_)
```

```
Fitting 5 folds for each of 72 candidates, totalling 360 fits
```

•Bu aşamada, öğrenilebilir psikolojik sinyaller üretmek amacıyla oluşturulan yüksek boyutlu özellik uzayı üzerinde RandomForest + GridSearchCV uygulanmıştır.

- Ancak, geniş parametre uzayı ve veri büyüklüğü nedeniyle bu adım çok yüksek hesaplama maliyetine yol açmış, süreç uzun süre sonuç veya hata üretmeden devam etmiştir.
- Bu nedenle, yöntem metodolojik olarak doğru olmasına rağmen pratik sınırlamalar nedeniyle tamamlanamamıştır.

Digital Nomads – Proje Ekibi