

# Hane Halkı Gelir Analizi

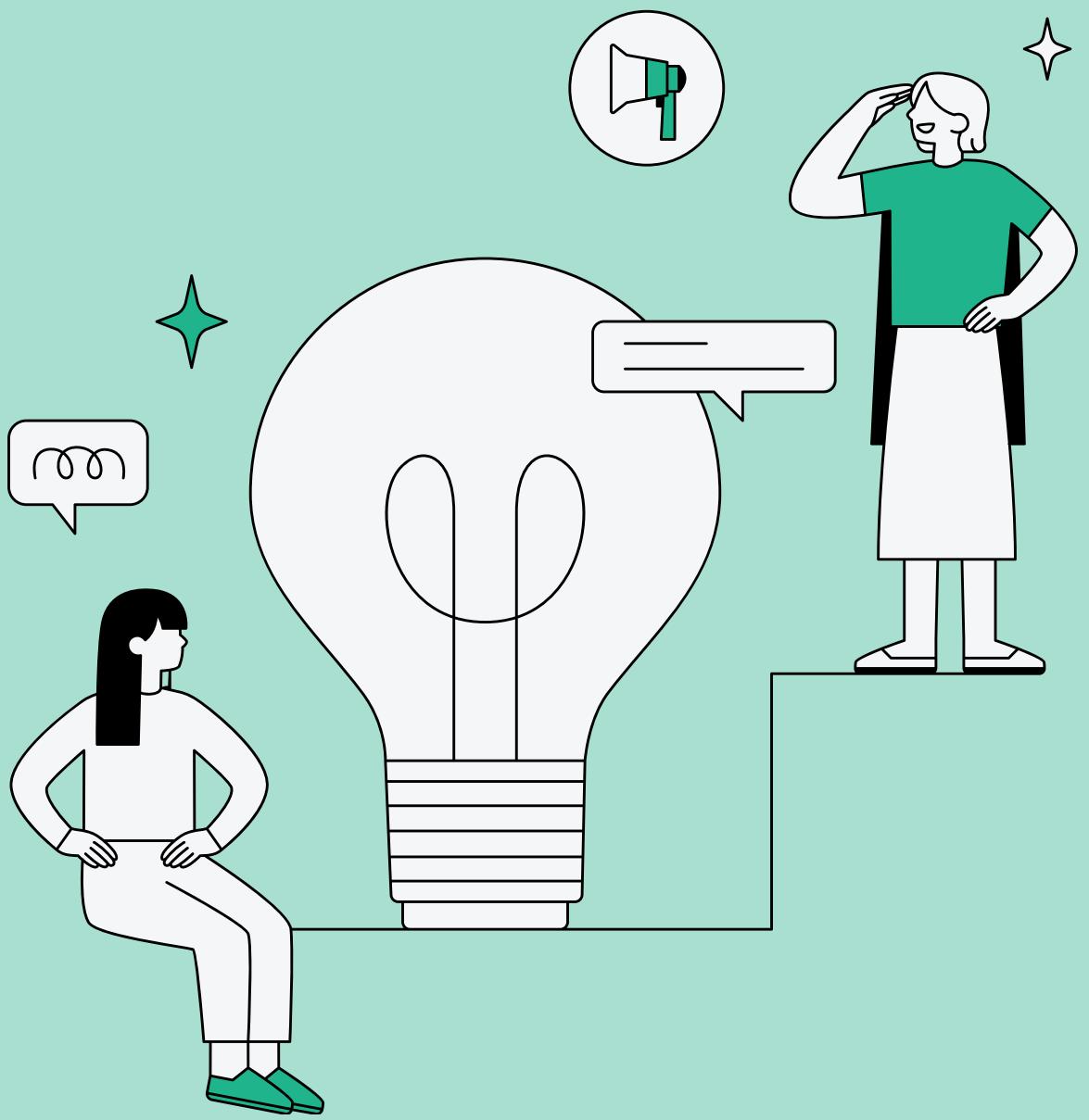
Abdullah BUZKAN  
Gizem Yağmur ERGELEN  
Hasret BAG  
Zeynep USLUBAŞ

21120205067  
21120205063  
21120205053  
21120606052



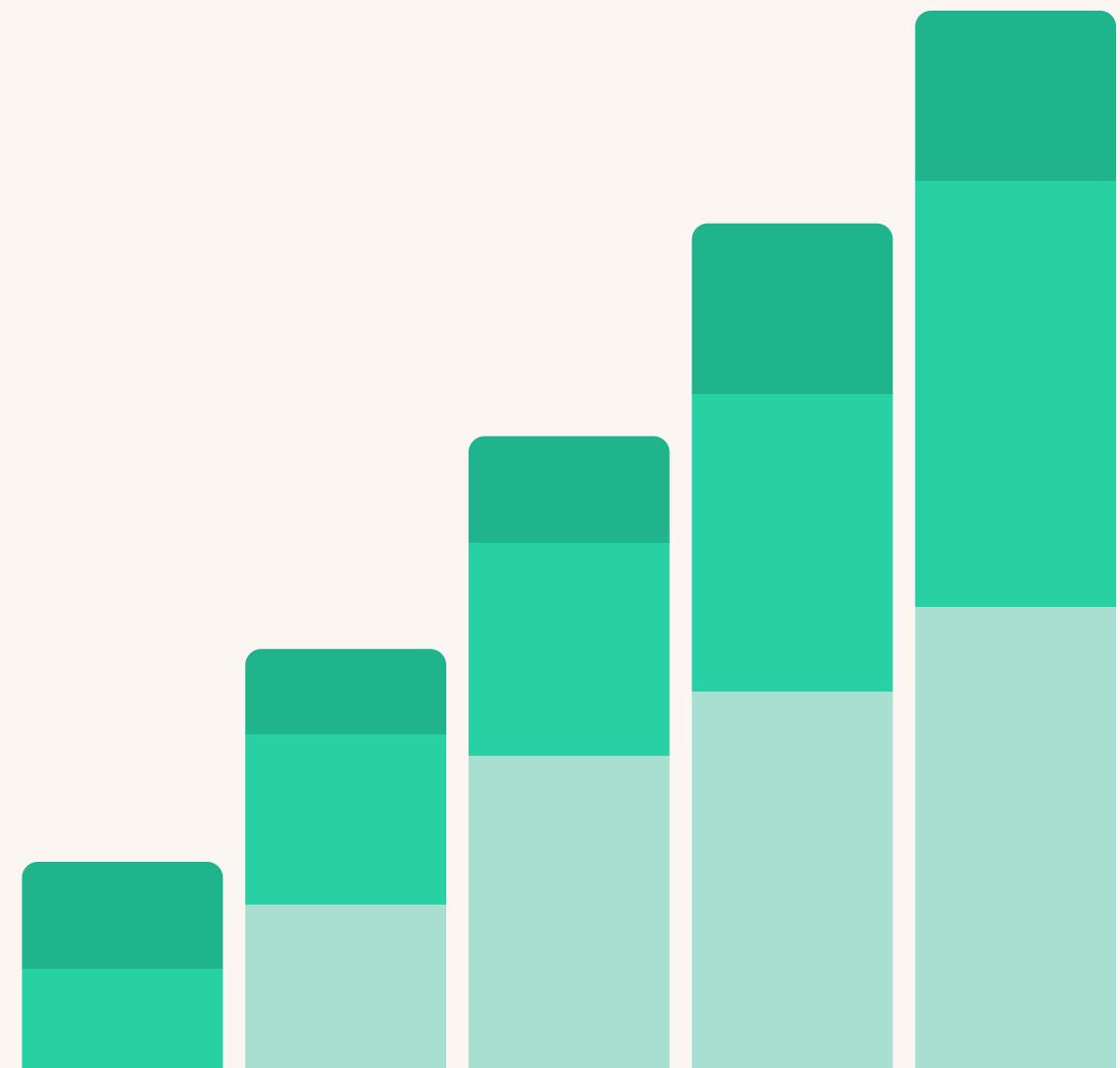
# 1. Problem Tanımı

Gelir seviyesi, bireylerin yaşam standartlarını, sosyal ve ekonomik durumlarını belirlemede önemli bir göstergedir. Bu çalışmada, bireylerin belirli demografik ve sosyo-ekonomik özelliklerine dayanarak yıllık gelirlerinin 50.000 dolardan yüksek veya düşük olduğunu tahmin etmek hedeflenmiştir.



## 2. Veri Toplama

Bu çalışmada kullanılan veri seti, "Kaggle" platformundan alınmıştır. Veri seti, çeşitli özelliklerden (örneğin, yaş, cinsiyet, eğitim seviyesi, meslek, çalışma saatleri vb.) oluşmakta ve önceden toplanmış, düzenlenmiş bir biçimde sağlanmıştır. Dolayısıyla, bu çalışma kapsamında doğrudan veri toplama işlemi gerçekleştirilmemiştir.



# 3. Veri Ön İşleme

## 3.1 Veri Setinin İncelenmesi

Veri seti 48842 satır ve 15 sütundan oluşmaktadır.

## 3.2 Tekrarlayan Satırların Kontrolü ve Temizlenmesi

Veri setinde 52 adet tekrarlayan (duplicate) satır tespit edilmiştir.

Bu satırlar modelin doğruluğunu etkileyebileceği için temizlenmiştir.

## 3.3 Eksik Değerlerin Kontrolü

Veri setinde NULL değerler bulunmasa da, bazı sütunlarda "?" simbolü eksik veri olarak kullanılmıştır.

"?" simbolü, aşağıdaki sütunlarda tespit edilmiştir:

- workclass: 2795 adet
- occupation: 2805 adet
- native-country: 856 adet

Bu değerler, NULL (NaN) olarak değiştirilmiştir.

## 3.4 Eksik Değerlerin Silinmesi

- Eksik değerlerin olduğu satırlar, modelin doğruluğunu artırmak amacıyla veri setinden silinmiştir.
- İşlem sonrasında:
- Satır sayısı: 45,175
- Sütun sayısı: 15 olarak güncellenmiştir..



### 3.5 Değişkenlerin Gruplanması

- Veri setindeki değişkenler iki ana gruba ayrılmıştır:
  - Kategorik Değişkenler: Metinsel veya kategorik verileri içerir.
  - Nümerik Değişkenler: Sayısal verileri içerir.
- Bu işlem, her bir grup için uygun analiz yöntemlerini kullanmayı kolaylaştırmıştır.

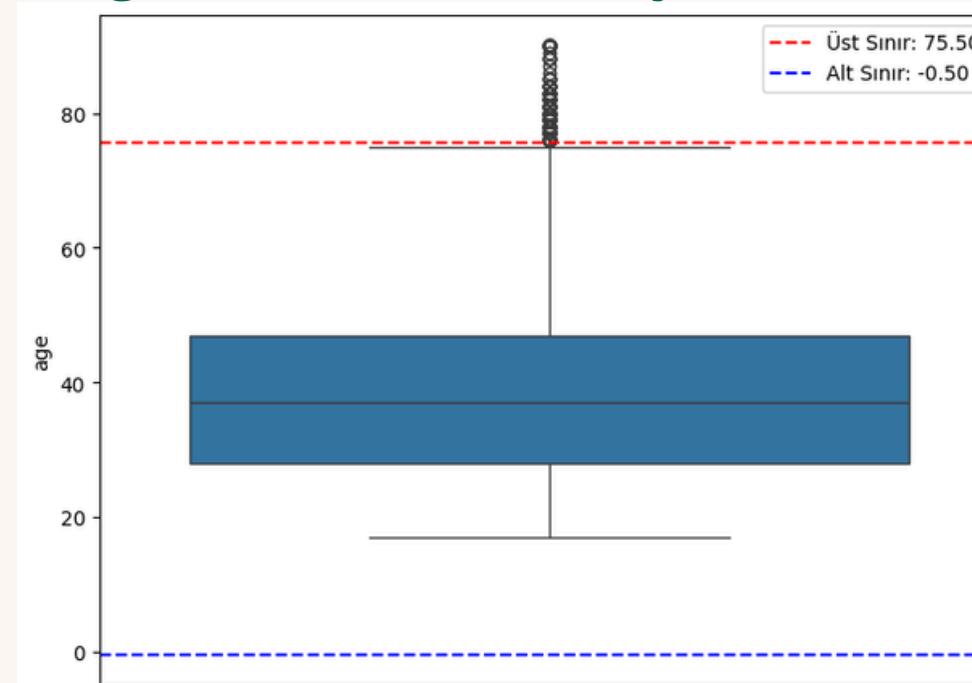
### 3.6 Aykırı Değerlerin Tespiti ve IQR Analizi

- Aykırı değerlerin tespiti için IQR (Interquartile Range) yöntemi kullanılmıştır.
- Aykırı değerler, bu sınırların dışında kalan veriler olarak tanımlanmıştır.
- Tespit edilen aykırı değerler, kutu grafikleri (Boxplot) kullanılarak görselleştirilmiştir.

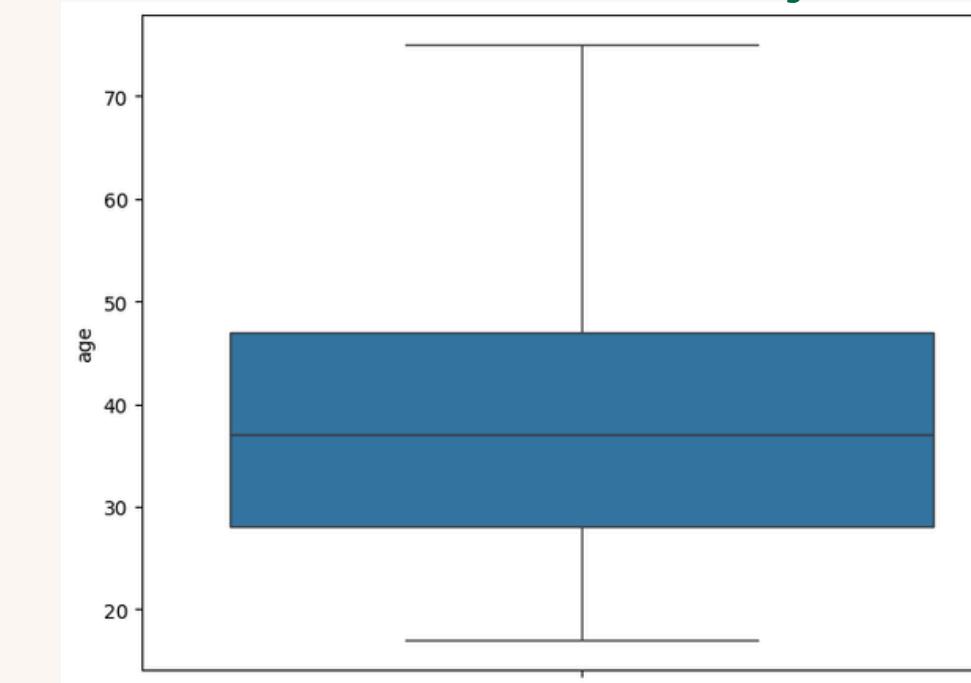
### 3.7 Normalizasyon

- Amaç: Değişkenleri belirli bir aralığa (genellikle 0 ile 1 arasına) ölçekleyerek, veri setindeki değişkenlerin büyüklik farklılıklarını minimize etmek.
- Yöntem:
  - MinMaxScaler kullanılarak sayısal veriler normalize edildi.

#### Boxplot ile age Sütununun Aykırı Değerlerinin Görselleştirilmesi



#### Boxplot ile age Sütununun Aykırı Değerlerinin Silindikten Sonra Görselleştirilmesi



### 3.8 Encoding

- Amaç: Kategorik değişkenleri sayısal değerlere dönüştürmek ve analiz sırasında modellerin bu verileri işlemesini sağlamak.
- Aşamalar:
  - Kategorik Sütunların Belirlenmesi:
    - Kategorik sütunlar object veri tipine göre seçildi ve değişken değerleri liste halinde gözlemlendi.
  - Mantıksal Sıralama Tanımlanması:
    - Her kategorik değişken için mantıksal sıralamalar belirlendi (örneğin, eğitim seviyeleri "Preschool"dan "Prof-school'a kadar).
  - Encoding İşlemi:
    - Listedeki sıraya göre kategorilere sayısal değerler atanarak encoding işlemi yapıldı.
    - native-country kategorisi diğer sütunlardan ayrı tutuldu.
  - Sonuç:
    - Encoding sonrası kategorik değişkenler sayısal değerlere dönüştü.
    - Yeni oluşturulan veri seti incelendi ve ilk satırlar görüntülendi.



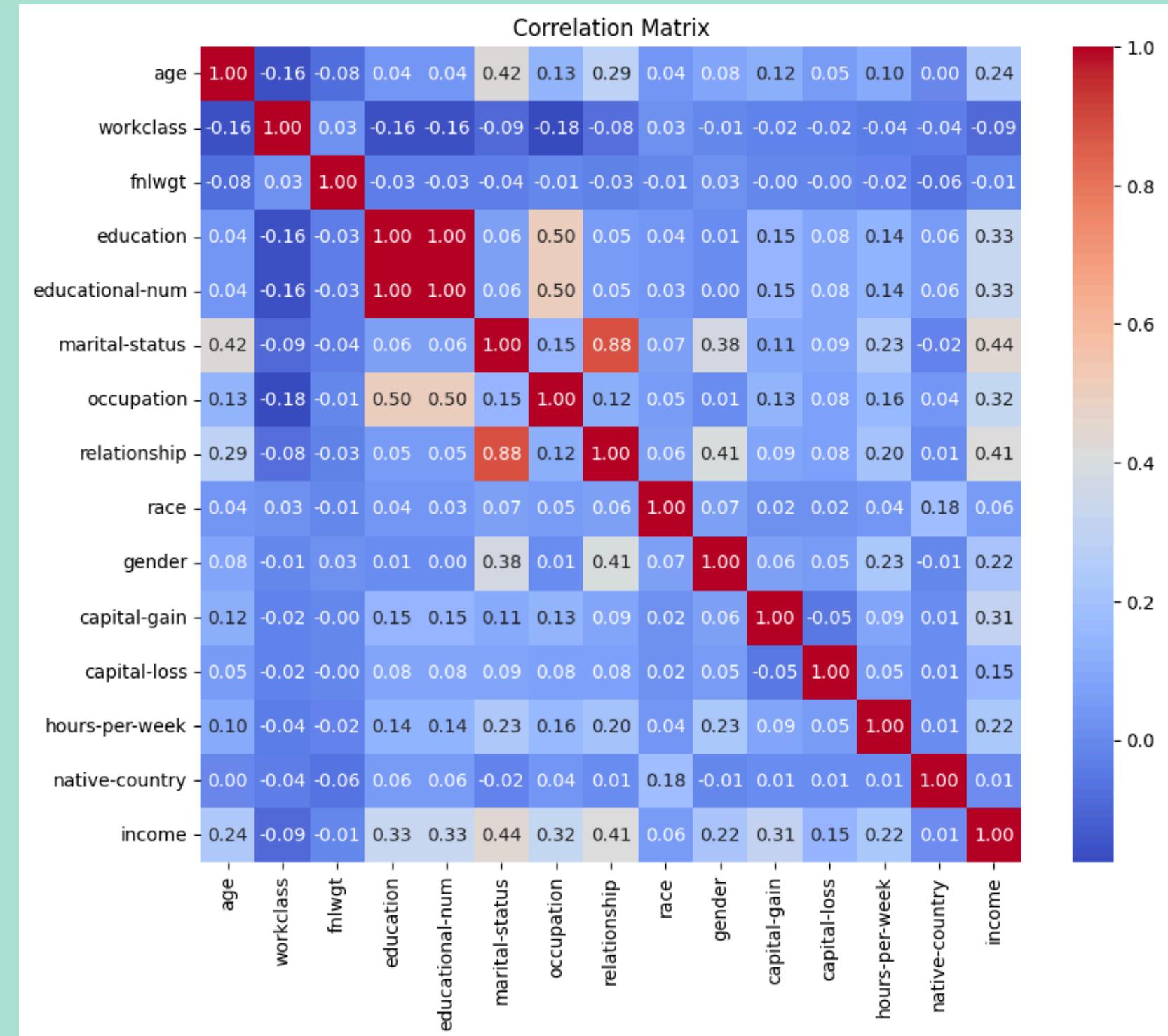
### 3.9 Korelasyon Matrisi

Korelasyon matrisi, veri setindeki sayısal değişkenler arasındaki doğrusal ilişkileri göstermektedir. Korelasyon katsayıları -1 ile 1 arasında değer alır:

- **Pozitif değerler (+):** Değişkenler arasında pozitif bir doğrusal ilişki olduğunu gösterir (biri arttığında diğer de artar).
- **Negatif değerler (-):** Değişkenler arasında negatif bir doğrusal ilişki olduğunu gösterir (biri arttığında diğer azalır).

### Veri Setinde Korelasyon Katsayısı Yüksek Olan Sütunlar

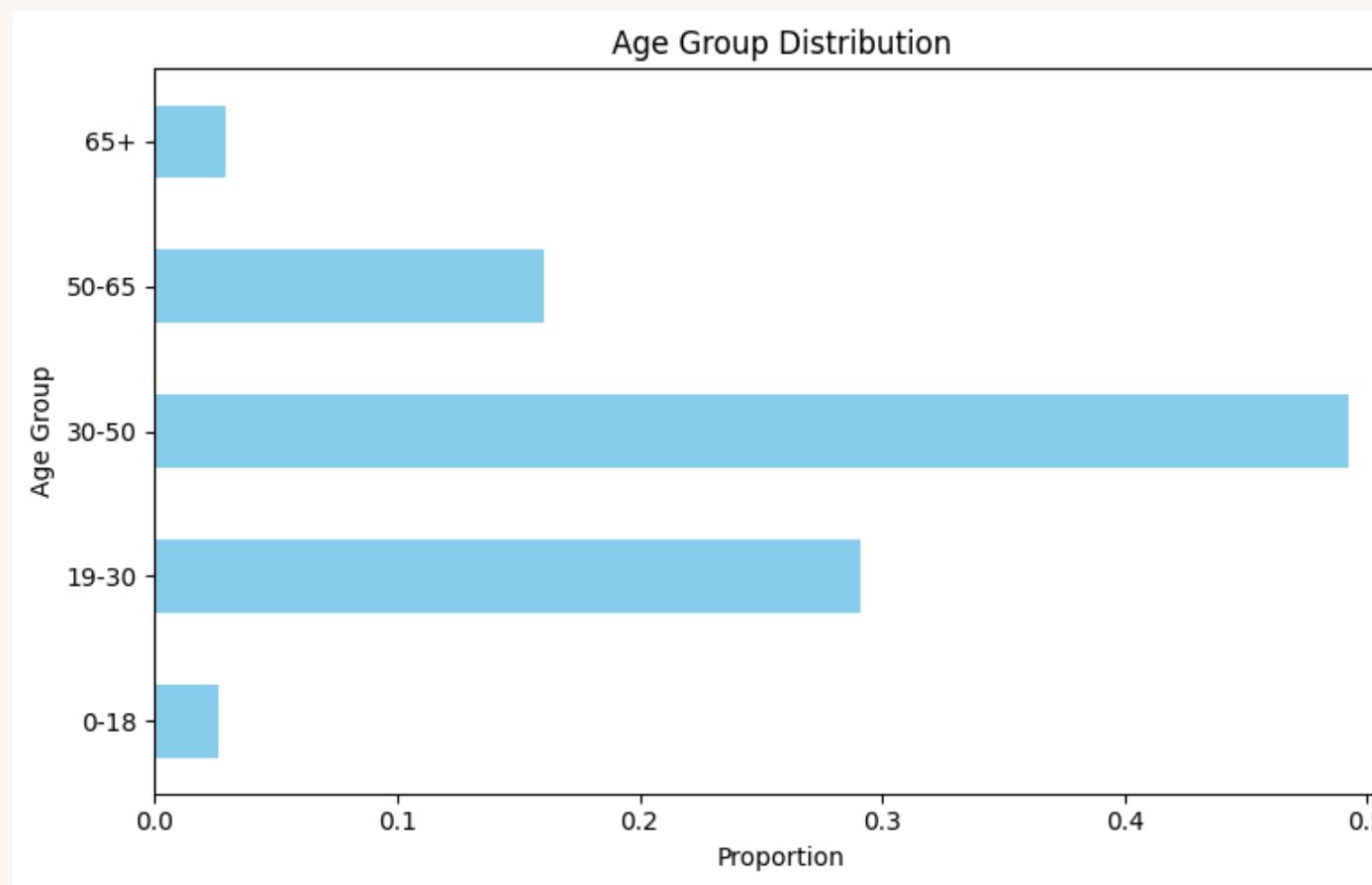
- Educational-num ve Education:
  - tam korelasyon (1.00):
    - Çünkü bunlar aynı veriyi farklı şekilde temsil eden sütunlardır.
- Relationship ve Marital-status:
  - yüksek korelasyon (0.88):
    - Medeni durumun aile ilişkileriyle çok güçlü bir bağı bulunmaktadır.



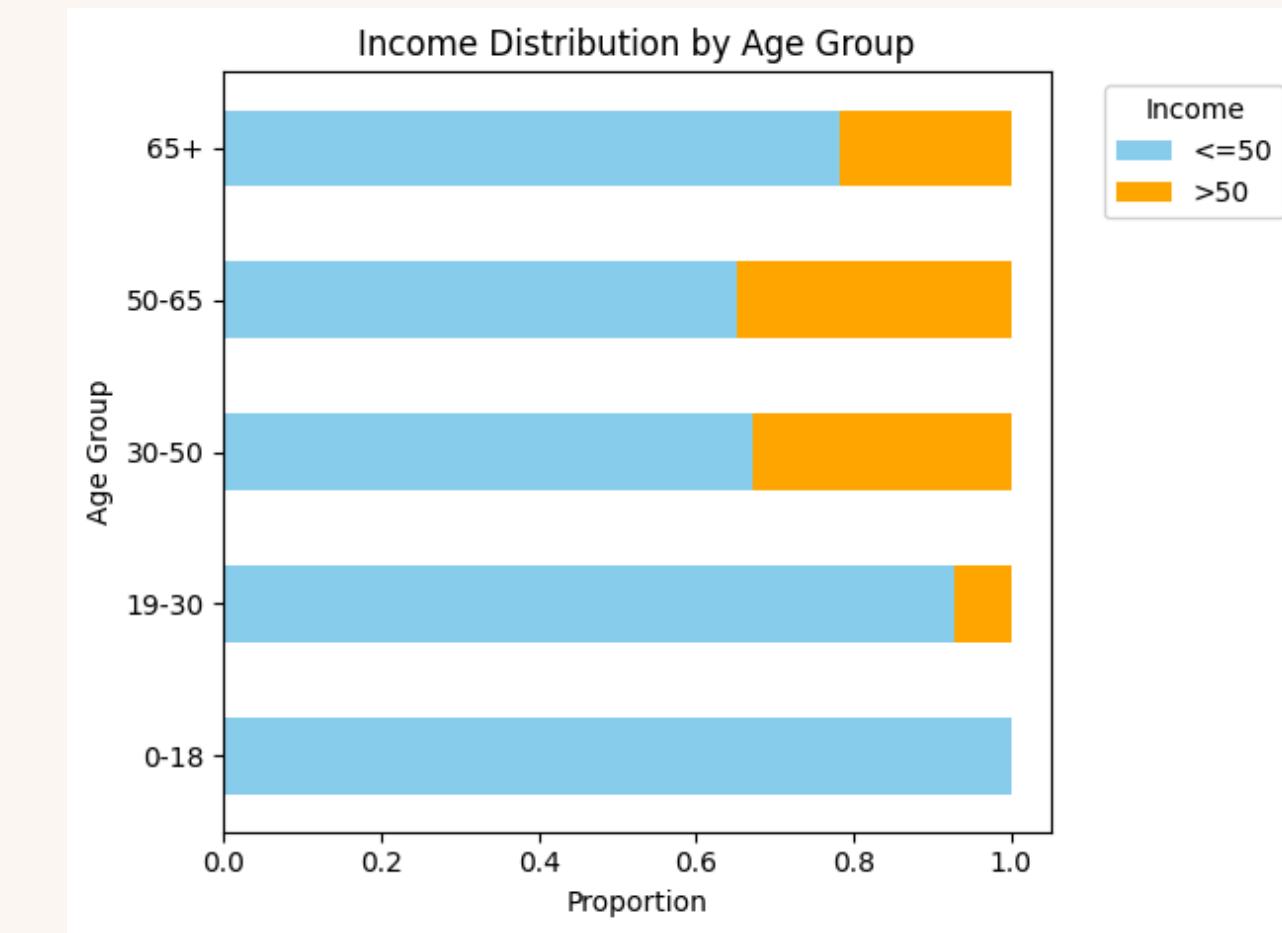
# 3. Keşifsel Veri Analizi (EDA)

Keşifsel Veri Analizi (EDA), veri setini anlamak, veri özelliklerini keşfetmek ve veri setindeki kalıpları, ilişkileri ve olası problemleri görselleştirerek analiz etmek için kullanılan bir veri analizi yöntemidir.

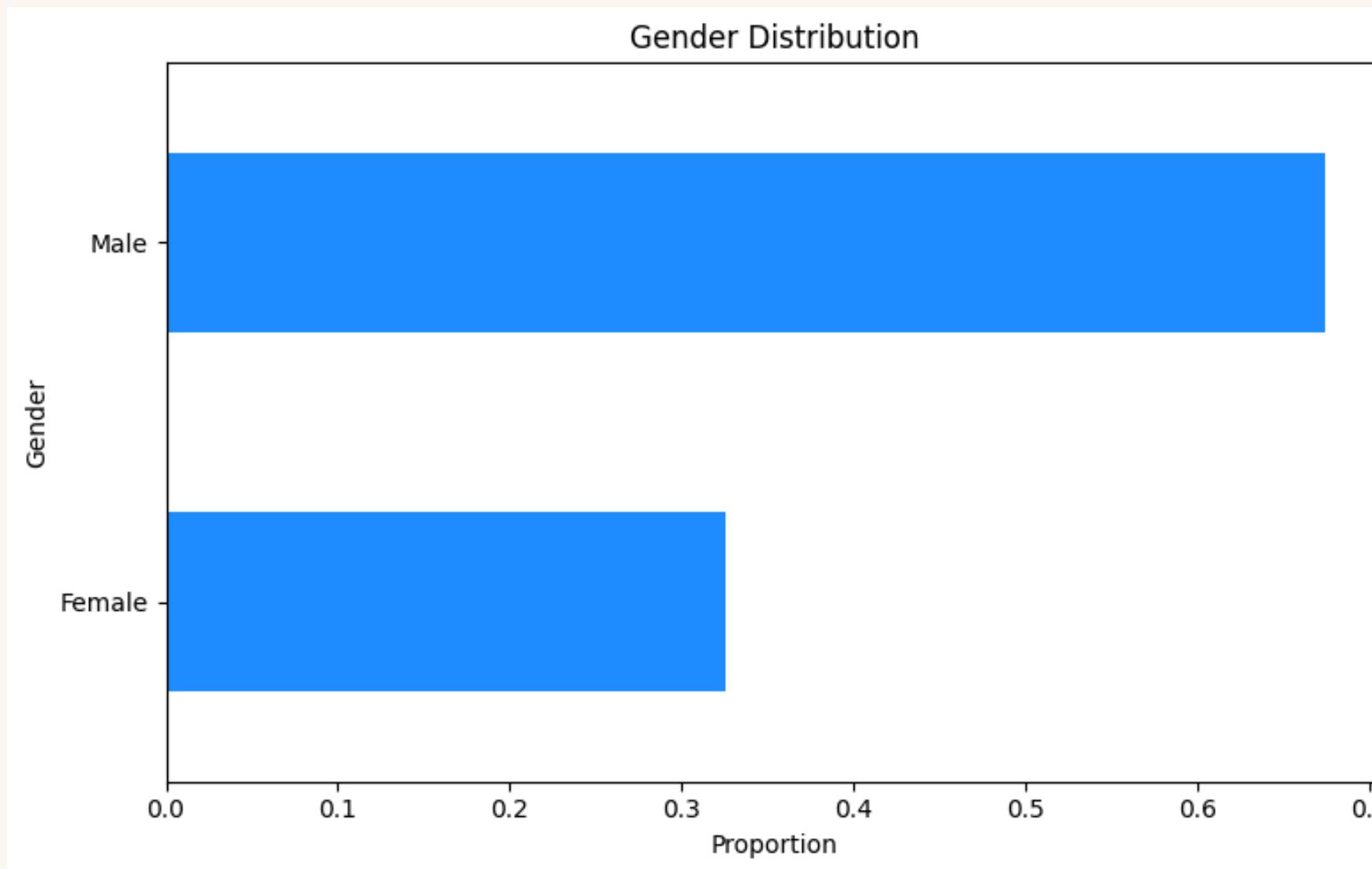
Aşağıda age Sütununa Ait Tek Değişkenli Grafik



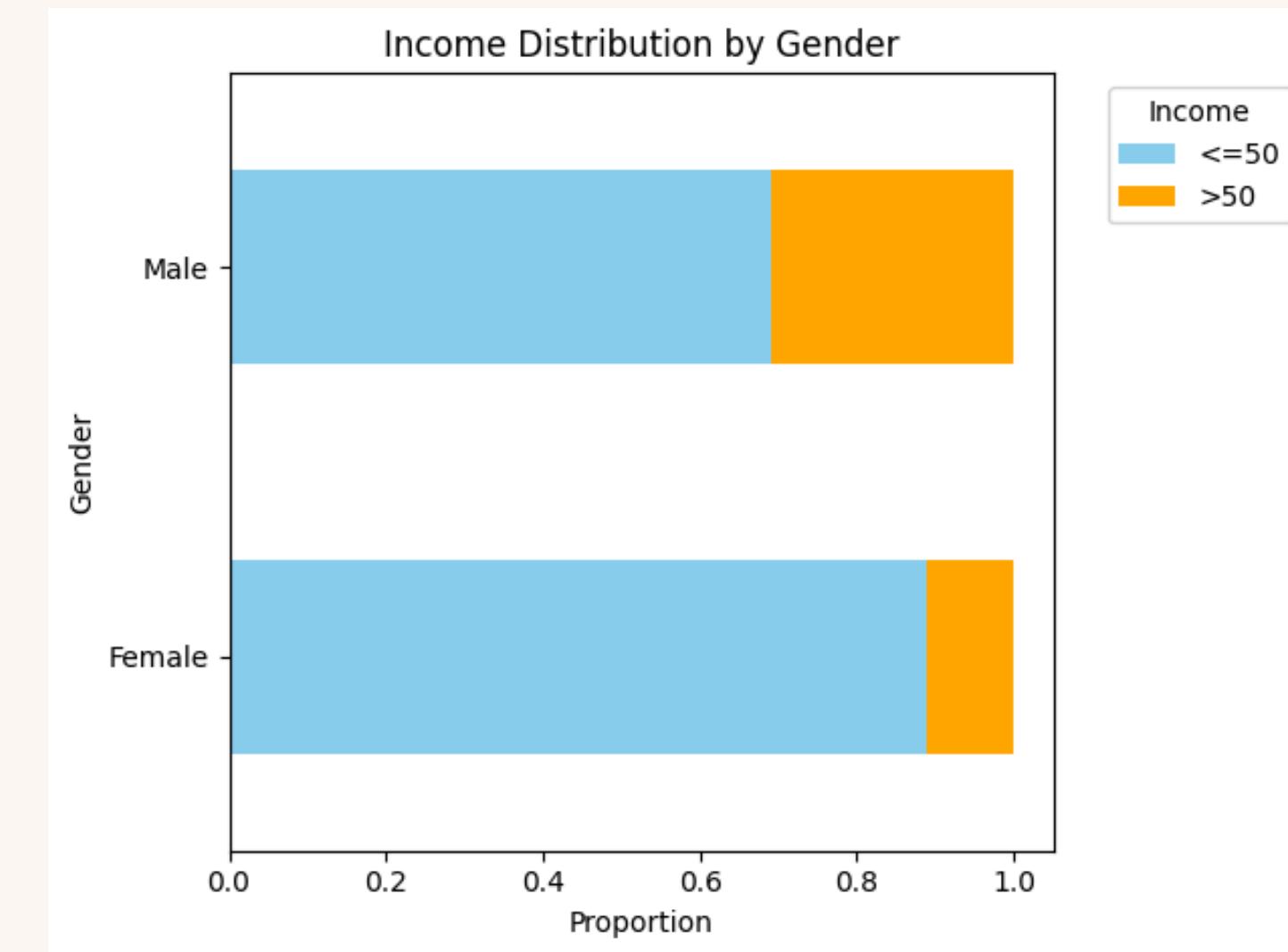
Aşağıda age Sütununa Ait Çift Değişkenli Grafik

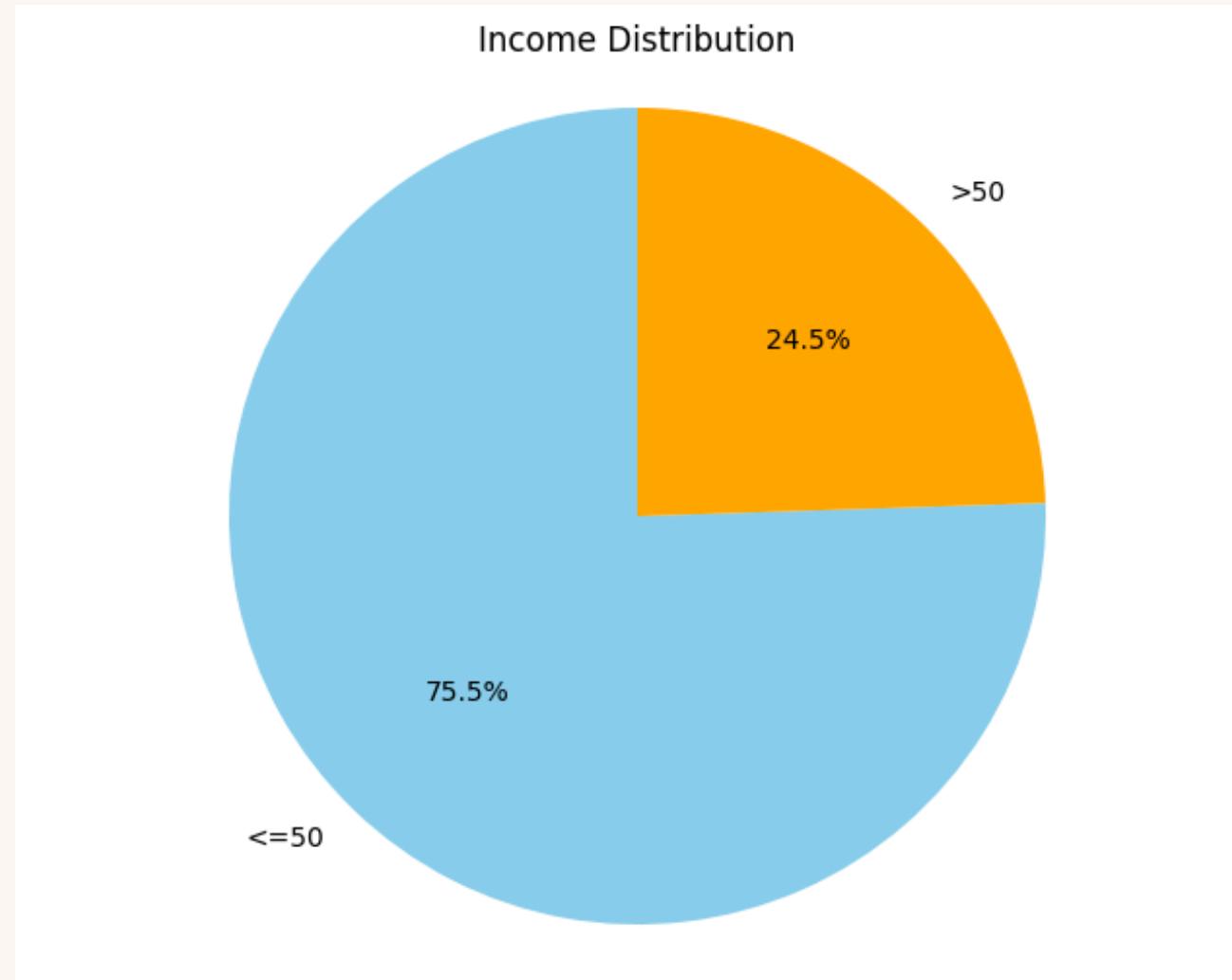


## Aşağıda gender Sütununa Ait Tek Değişkenli Grafik



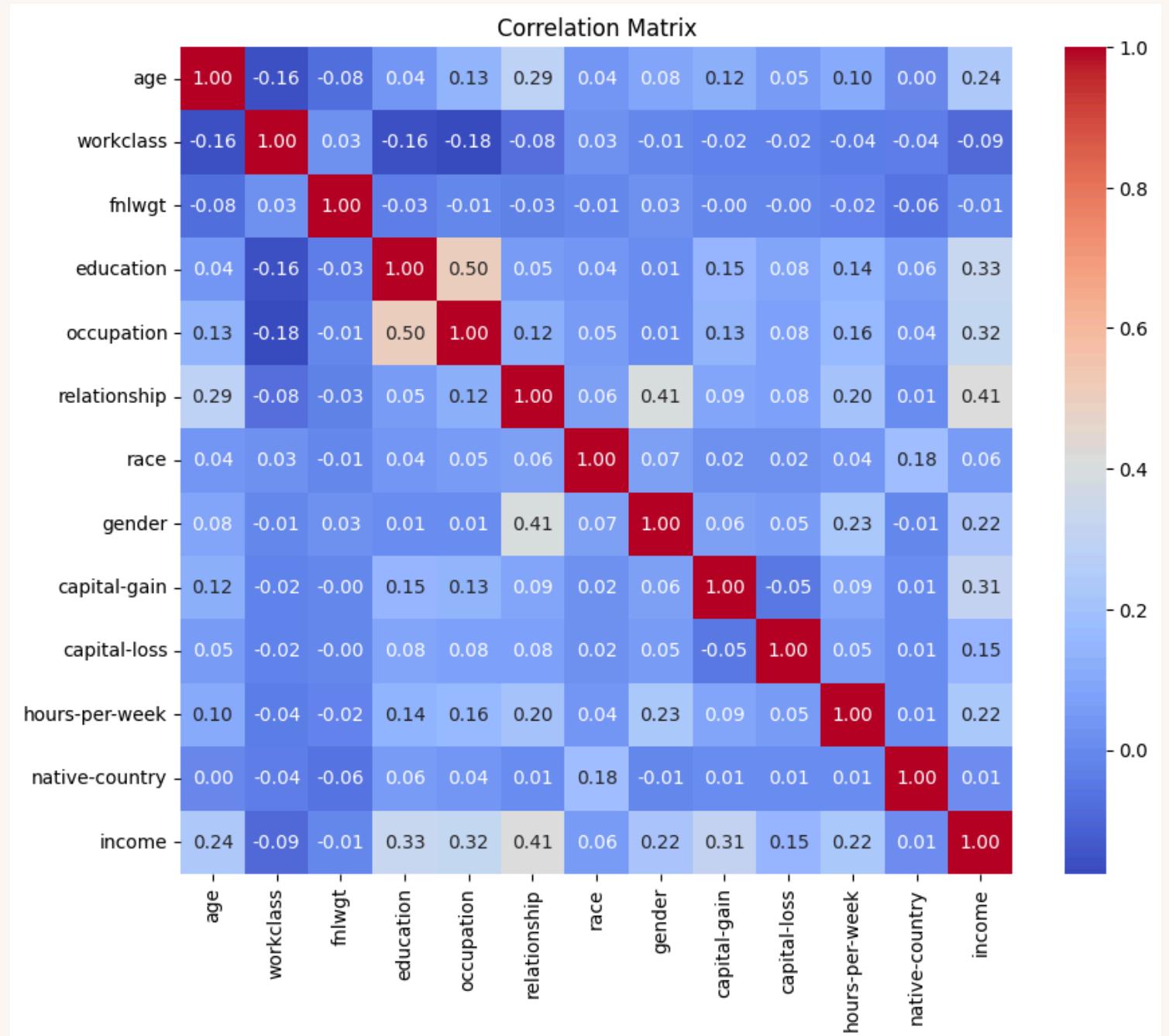
## Aşağıda gender Sütununa Ait Çift Değişkenli Grafik





### Çıkarımlar:

- **Veri seti dengesizdir, çünkü düşük gelir grubu ( $\leq 50$ ) açık ara baskın durumdadır.**
- **Bu durum, veri modellemesi yapılmırken dikkat edilmesi gereken bir unsurdur.**
- **Örneğin, sınıf dengesizliği, tahmin modellerinde düşük gelir grubuna fazla ağırlık verilmesine neden olabilir.**



## Korelasyon Matrisinden Çıkarılacak Yorumlar

### 1. Yüksek Korelasyonlu Değişkenler:

- Education ve educational-num gibi yüksek korelasyonlu değişkenlerden biri çıkarılabilir. Fazla bilgi tekrarından kaçınılmalıdır.

### 2. Orta Düzey Korelasyonlar:

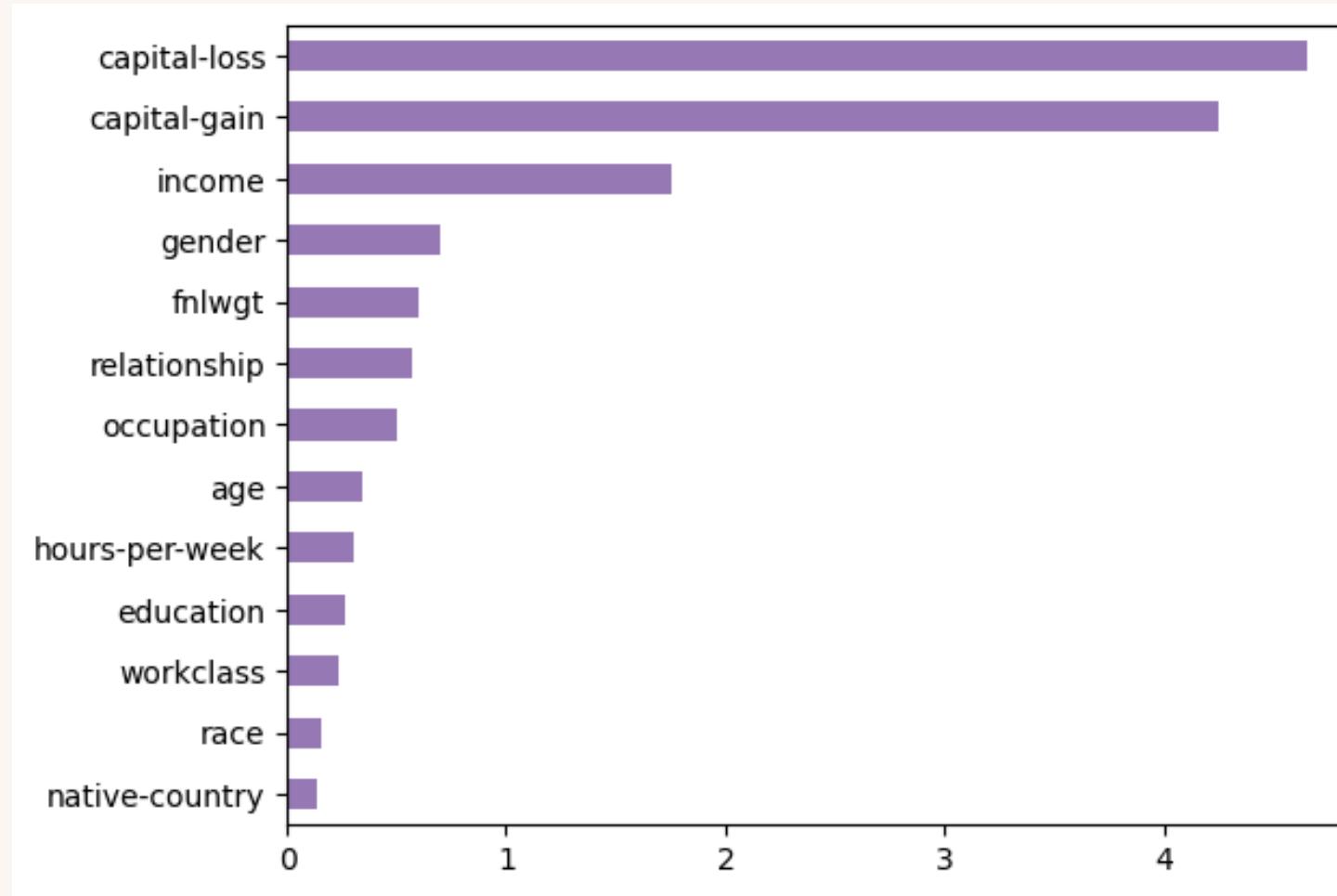
- Relationship, occupation, education gibi değişkenler modelleme sırasında daha fazla dikkate alınabilir.

### 3. Düşük Korelasyonlar:

- Korelasyonu düşük olan değişkenler (örneğin, race veya workclass), çıkarılmadan önce etkilerinin model performansına katkısı test edilebilir.

Bu korelasyon analizi, değişkenler arasındaki ilişkileri anlamak ve modelleme sırasında hangi değişkenlere öncelik verileceğini belirlemek için kritik bilgiler sağlar.

## Varyasyon Katsayı Değeri



### En Yüksek Değişim Gösteren Değişkenler:

- Capital-loss ve capital-gain: Bu değişkenler en yüksek değişim katsayısına sahiptir (yaklaşık %424 ve %465). Bu, bu değişkenlerin ortalamalarına kıyasla oldukça büyük varyasyona sahip olduğunu gösterir.
- Neden? Bu değişkenlerin çoğu birey için sıfır olabileceği, ancak bazı bireylerde yüksek değerlere sahip olduğu için büyük varyasyon göstermektedir.

### En Düşük Değişim Gösteren Değişkenler:

- Native-country, race, ve workclass: Bu değişkenler en düşük değişim katsayısına sahiptir (yaklaşık %14-%23). Bu, bu değişkenlerin daha sabit ve ortalama değerlerine daha yakın olduğunu gösterir.
- Neden? Bu değişkenler genellikle kategorik özelliklere sahiptir ve dağılımları daha homojendir.

age	34.307003
workclass	23.130042
fnlwgt	59.946104
education	26.938794
occupation	50.615930
relationship	56.739928
race	16.075662
gender	69.498313
capital-gain	424.804197
capital-loss	465.899127
hours-per-week	30.024235
native-country	14.000457
income	175.608611
dtype: float64	

# 4. Veri Modelleme

## Algoritma Seçimi

Mevcut problem ikili sınıflandırma problemidir. Bu probleme uygun algoritmalar araştırıldı

XGBoost:

- Karmaşık karar sınırlarını öğrenme yeteneğine sahip olduğundan, doğrusal olmayan ve karmaşık veri setlerinde iyi performans gösterir.
- Hangi özelliklerin tahminleme sürecine daha fazla katkı sağladığını belirleyebilir.
- Class weights veya sampling tekniklerini desteklemesi ve modelin her iki sınıfı da etkili bir şekilde öğrenmeye çalışması ile veri dağılımı dengesiz ise çok daha etkili olabilir.

## Model Eğitimi

Öncelikle veri seti eğitim ve test olmak üzere ikiye ayrıldı. Sonrasında K-Fold Cross Validation yöntemi kullanıldı. Veri seti K eşit parçaya bölünür; her seferinde bir parça test verisi, kalanlar ise eğitim verisi olarak kullanılır. Bu işlem K kez tekrarlanır ve sonuçlar ortalanarak modelin genel başarımı hesaplanır. Bu yöntem, modelin genellenebilirliğini değerlendirme açısından oldukça önemlidir.

```
for fold, (train_idx, val_idx) in enumerate(kf.split(X_train, y_train)):  
    print(f"Fold {fold+1}")  
  
    # Eğitim ve validasyon verilerini ayır  
    X_train_fold, X_val_fold = X_train.iloc[train_idx], X_train.iloc[val_idx]  
    y_train_fold, y_val_fold = y_train.iloc[train_idx], y_train.iloc[val_idx]  
  
    # Modeli eğit  
    model.fit(X_train_fold, y_train_fold)
```



# 5. Model Değerlendirme

## Değerlendirme Metrikleri

Bu çalışmada, veri setimiz 10 parçaya (10-fold) ayrılmıştır.

Her bir fold için aşağıdaki metrikler hesaplanmıştır:

- Doğruluk (Accuracy): Tüm sınıflar için doğru tahminlerin oranıdır.
- Kesinlik (Precision): Pozitif sınıfın doğruluğunu ifade eder.
- Duyarlılık (Recall): Pozitif sınıfın ne kadar iyi tahmin edildiğini gösterir.
- F1 Skoru (F1 Score): Kesinlik ve duyarlılığın harmonik ortalamasıdır.

## Classification Metrics Formulas

		Predicted		Row Totals
Actual	Positive	Negative		
Positive	TP	FN	TotActPos	
Negative	FP	TN	TotActNeg	
Col Totals	TotPredPos	TotPredNeg	Total	

$$\text{Precision} = \frac{\text{TP}}{\text{TotPredPos}}$$

$$\text{Accuracy} = \frac{\# \text{ Right}}{\text{Total}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TotActPos}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TotAcNeg}}$$

$$\text{Error} = \frac{\# \text{ Wrong}}{\text{Total}}$$

$$F = 2 * \frac{\text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

```
Fold 4
Fold 4 doğruluk skoru (Accuracy): 0.8644
Fold 4 kesinlik (Precision): 0.7625
Fold 4 duyarlılık (Recall): 0.6479
Fold 4 F1 Skoru: 0.7006
```

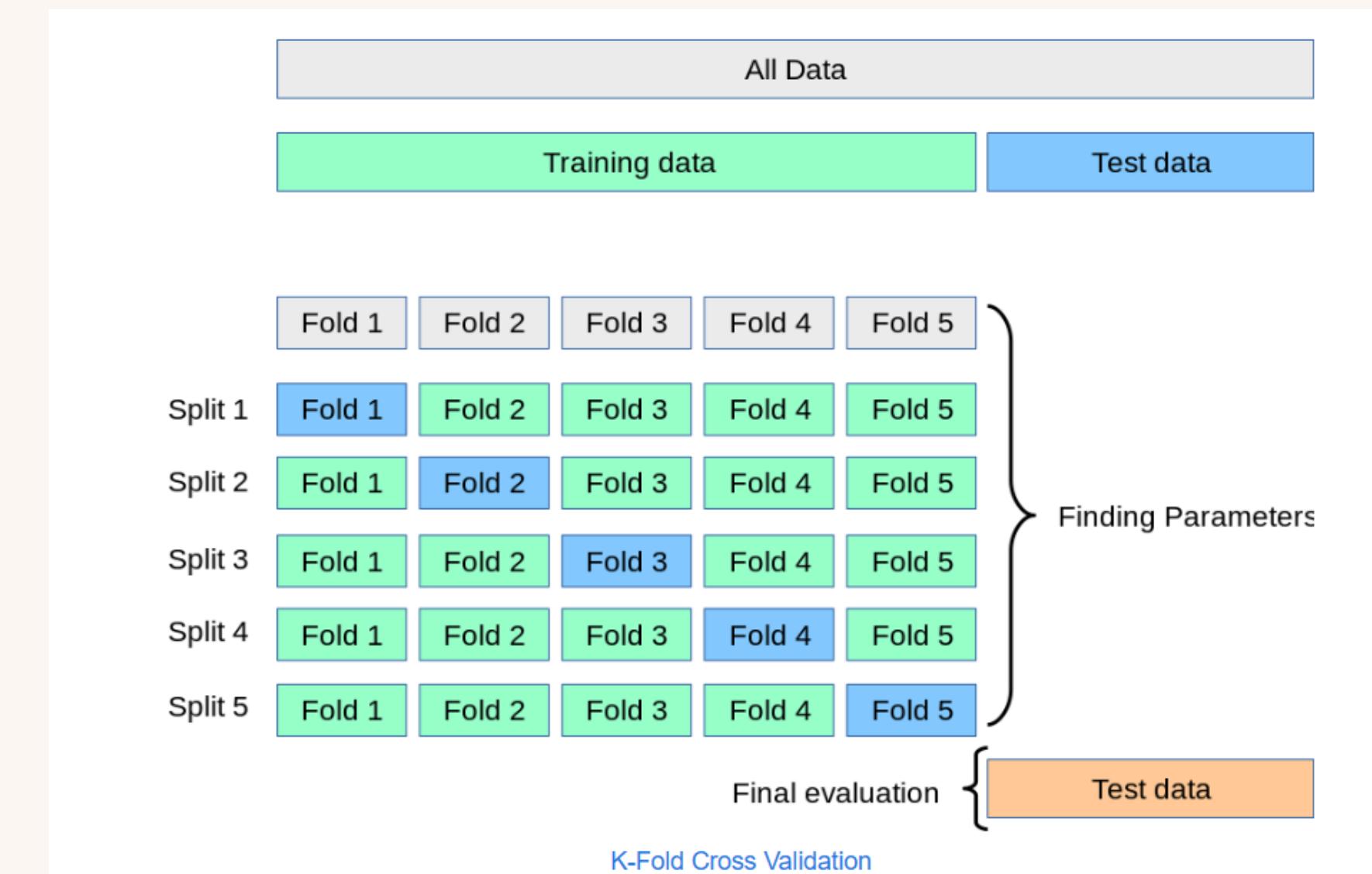
## Overfitting ve Underfitting Kontrolü

Overfitting: Modelin eğitim verisine aşırı uyum sağlama, yeni verilerde düşük performans göstermesi.

Underfitting: Modelin ne eğitim verisini ne de yeni verileri iyi öğrenememesi, düşük performans göstermesi.

K-Fold Cross Validation yöntemi, overfitting'i kontrol etmek için etkili bir yöntemdir çünkü:

- Eğitim ve validasyon setleri sürekli olarak değişir.
- Model her bir fold'da farklı veri setlerinde test edilir.
- Eğer model, tüm fold'larda tutarlı bir şekilde yüksek performans sergiliyorsa, bu genelleme yeteneğinin iyi olduğunu gösterir ve overfitting riskini azaltır.



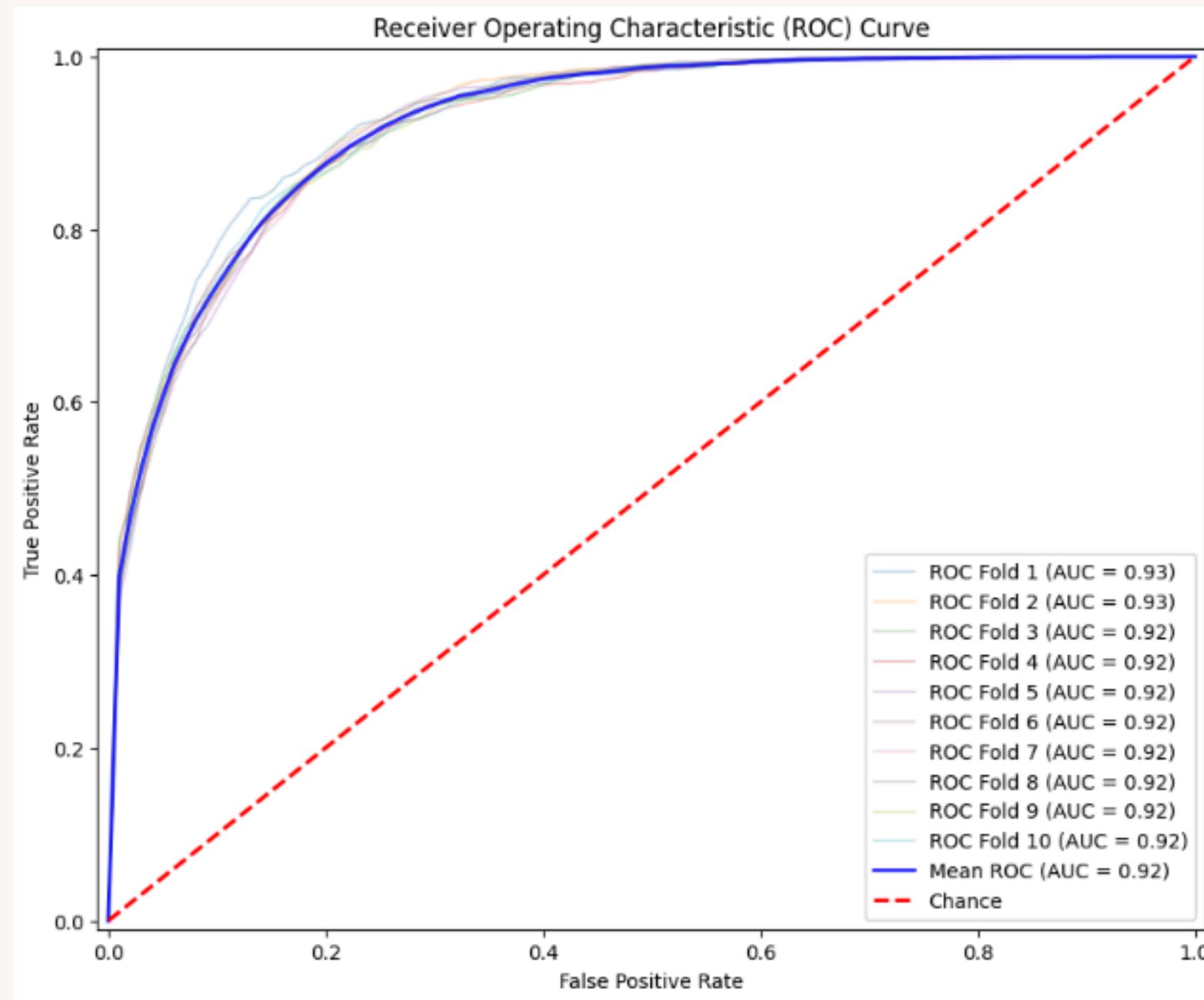
<https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/>

# 6. Sonuçların Yorumlanması ve Görselleştirilmesi

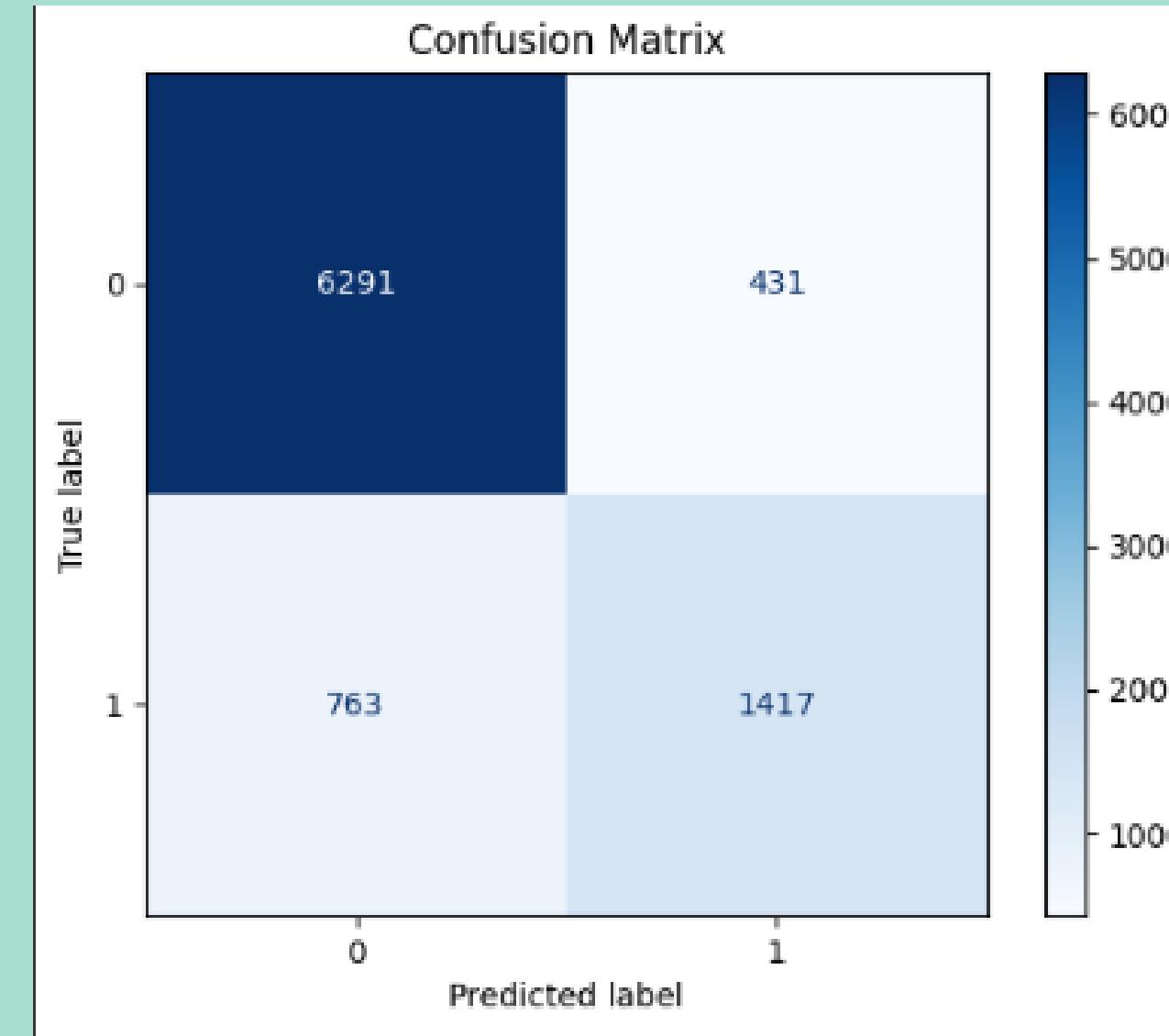
ROC eğrisi, sınıflandırma modelinin performansını ölçmek için kullanılır. Eğri, True Positive Rate (TPR) (Duyarlılık) ve False Positive Rate (FPR) değerlerini farklı eşik değerleri için grafiğe döker. Model, ideal olarak, sol üst köşeye yakın bir eğri çizmeli ve yüksek bir AUC değerine sahip olmalıdır.

Confusion matrix, bir modelin sınıflandırma performansını değerlendirmek için kullanılan bir tablodur. Gerçek ve tahmin edilen sınıfların sayısını True Positive (TP), True Negative (TN), False Positive (FP) ve False Negative (FN) olarak özetler.





- Her fold'un AUC değeri yaklaşık 0.92-0.93 arasında değişmektedir. Bu, modelin tutarlı bir şekilde iyi performans gösterdiğini, sınıflandırma başarısının yüksek olduğunu göstermektedir.
- Mavi kalın çizgi, tüm fold'ların ortalama ROC eğrisini göstermektedir. Ortalama AUC değeri 0.92 olarak hesaplanmıştır.
- Model, her fold üzerinde tutarlı bir şekilde iyi bir performans sergilemiştir ve farklı veri dağılımlarında genel olarak başarılı olma eğilimindedir. Bu durum, modelin hem eğitim hem de test verileri üzerinde iyi bir genelleme yapabildiğini ifade eder.



- True Negatives (TN): 6291 (Model, 0 sınıfını doğru tahmin etmiş).
- False Positives (FP): 431 (Modelin 1 olarak tahmin ettiği ancak gerçekten 0 olan örnekler).
- False Negatives (FN): 763 (Modelin 0 olarak tahmin ettiği ancak gerçekten 1 olan örnekler).
- True Positives (TP): 1417 (Model, 1 sınıfını doğru tahmin etmiş).

Test Verisi Performansı:  
 Doğruluk Skoru (Accuracy): 0.8659  
 Kesinlik (Precision): 0.7668  
 Duyarlılık (Recall): 0.6588  
 F1 Skoru: 0.7036

Detaylı Sınıflandırma Raporu:				
	precision	recall	f1-score	support
0	0.89	0.94	0.91	6722
1	0.77	0.65	0.70	2180
accuracy			0.87	8902
macro avg	0.83	0.79	0.81	8902
weighted avg	0.86	0.87	0.86	8902

- Accuracy:

Modelin doğruluk skoru 0.8659, yani test verilerindeki örneklerin %86.59'u doğru sınıflandırılmıştır.

Pozitif sınıf için kesinlik nispeten düşük, bu da modelin yanlış pozitiflere (False Positives) dikkat etmesi gerektiğini gösterir.

- Duyarlılık (Recall):

Sınıf 0 için: 0.94 (negatif sınıfların %94'ü doğru tahmin edilmiş).

Sınıf 1 için: 0.65 (pozitif sınıfların sadece %65'i doğru tahmin edilmiş).

Pozitif sınıfıta duyarlılığın daha düşük olması, modelin bazı pozitif örnekleri atladığını (False Negatives) gösterir.

- F1 Skoru:

F1 skoru (kesinlik ve duyarlılığın harmonik ortalaması) genel olarak dengeli bir değerlendirme sağlar.

Sınıf 0 için: 0.91

Sınıf 1 için: 0.70

Pozitif sınıfıta F1 skoru, modelin bu sınıfıta iyileştirmeye ihtiyaç duyduğunu vurgulamaktadır.

- Destek (Support):

Sınıf 0: 6722 örnek

Sınıf 1: 2180 örnek

Veri setinde sınıflar arasında bir dengesizlik olduğu görülmektedir.

Test Verisi Performansı:					
Doğruluk Skoru (Accuracy): 0.8659					
	precision	recall	f1-score	support	
0	0.89	0.94	0.91	6722	
1	0.77	0.65	0.70	2180	
accuracy			0.87	8902	
macro avg	0.83	0.79	0.81	8902	
weighted avg	0.86	0.87	0.86	8902	

# Tesekkürler...

