**SİBER GÜVENLİKTE YAPAY ZEKA ÖDEV RAPORU**

Uygulama için UCI Machine Learning Repository içerisinden “Adult Data Set” kullanılmıştır.

(<https://archive.ics.uci.edu/ml/datasets/Adult>)

Dataset 6 adet sürekli ve 8 adet kategorik veri içermekte olup içerikleri aşağıda sunulmuştur.

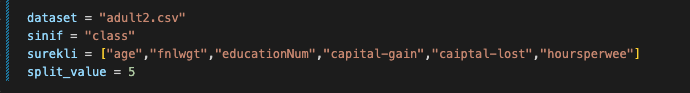
Kategorimiz ise yetişkinlerin yıllık gelirlerinin >50K, <=50K olacak şekilde sınıflandırılmasıdır.

age: continuous.  
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.  
fnlwgt: continuous.  
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.  
education-num: continuous.  
marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.  
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.  
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.  
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.  
sex: Female, Male.  
capital-gain: continuous.  
capital-loss: continuous.  
hours-per-week: continuous.  
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

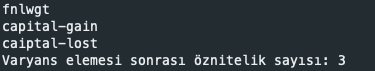
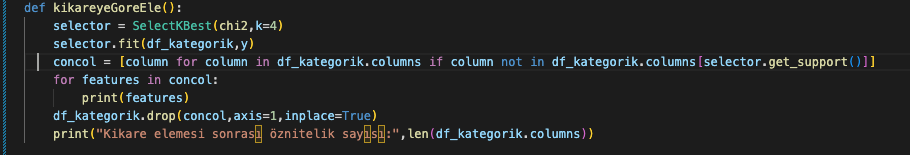
Öncelikle dataset içerisinde bulunan boş veri içeren satırlar manuel olarak silinmiştir. Yine de kodumda boş verileri kontrol etmesi için ayrıca kontrol yapılmıştır.

Müteakiben dataset içerisinde bulunan kategorik veriler string değerinde kaydedildiği için manuel olarak sayısal olacak şekilde veri düzenlemesi yapılmıştır. Örnek olarak cinsiyet bayan 0, bay 1 olacak şekilde düzenlenmiştir.

Ödevin ekinde sunulan python kodu ile ;

* Datasetimiz, sınıfımız ve sürekli verilerimizi içeren sütunlar ve cross validation split sayısı ilgili değişkenlere verilmiştir.
* 
* Sınfımız ve veriler y değişkenine tanımlanmıştır.
* Kalan veriler pandas kütüphanesi sayesinde df\_surekli ve df\_kategorik olacak şekilde 2 ayrı dataframe’e bölünmüştür.
* Text

  Description automatically generated
* Surekli verilerimiz normalize edilerek , 0.01 varyansa göre eleme yapılmış 3 adet öznitelik belirtilen varyans değerinin altında kaldığı için çıkarılmıştır.
* Text

  Description automatically generated
* 
* Kategorik verilerimiz ise en iyi 4 öznitelik kalacak şekilde chi2 testi ile eleme yapılmış ekran görüntüsünde belirtilen öznitelikler çıkarılmıştır.
* 
* Text

  Description automatically generated with low confidence
* Müteakiben elden kalan öznitelikler birleştirilerek X değişkeninde birleştirilmiştir.



* Elde edilen verilerimiz 5 split olacak şekilde cross validation yapılmıştır.
* Text

  Description automatically generated

**SONUÇ :**

Cross Validation sonucunda;

1. **TUR:**

[[ 850 652]

[ 710 3821]]

Accuracy: 0.7742416708105421

Balanced Accuracy: 0.7046069082906011

precision recall f1-score support

0 0.54 0.57 0.56 1502

1 0.85 0.84 0.85 4531

accuracy 0.77 6033

macro avg 0.70 0.70 0.70 6033

weighted avg 0.78 0.77 0.78 6033

1. **TUR:**

[[ 864 638]

[ 692 3839]]

Accuracy: 0.7795458312613956

Balanced Accuracy: 0.7112536775067217

precision recall f1-score support

0 0.56 0.58 0.57 1502

1 0.86 0.85 0.85 4531

accuracy 0.78 6033

macro avg 0.71 0.71 0.71 6033

weighted avg 0.78 0.78 0.78 6033

1. **TUR:**

[[ 859 642]

[ 664 3867]]

Accuracy: 0.7834880636604774

Balanced Accuracy: 0.7128695634529529

precision recall f1-score support

0 0.56 0.57 0.57 1501

1 0.86 0.85 0.86 4531

accuracy 0.78 6032

macro avg 0.71 0.71 0.71 6032

weighted avg 0.78 0.78 0.78 6032

1. **TUR:**

[[ 870 631]

[ 659 3872]]

Accuracy: 0.7861405835543767

Balanced Accuracy: 0.7170855418832821

precision recall f1-score support

0 0.57 0.58 0.57 1501

1 0.86 0.85 0.86 4531

accuracy 0.79 6032

macro avg 0.71 0.72 0.72 6032

weighted avg 0.79 0.79 0.79 6032

1. **TUR**

[[ 866 636]

[ 697 3833]]

Accuracy: 0.7790119363395226

Balanced Accuracy: 0.7113507229507088

precision recall f1-score support

0 0.55 0.58 0.57 1502

1 0.86 0.85 0.85 4530

accuracy 0.78 6032

macro avg 0.71 0.71 0.71 6032

weighted avg 0.78 0.78 0.78 6032

Average Accuracy: 0.7804856171252628

Average Balanced Accuracy: 0.7114332828168534

Veri setimizde bulunan 30162 kayıttan geliri >50 K olan 7508 kişi var iken <=50K olan 22654 kişi bulunmaktadır. Bu durumda veri setimizin dengesiz olduğunu gördüğümüz için başarımız balanced accuracy değeri olan 0.71 olarak bulunmuştur.