Veri Madenciliği

PROJE ÖDEVİ

RECEP ONUR OKAN
20360859027

Bu proje ödevi için UCI Machine Learning sitesinde ecoli datasetini kullandım.Bu dataseti için karar ağacı modeli oluşturdum.İlgili dataset aşağıdaki linkte yer alıyor.

https://archive.ics.uci.edu/ml/machine-learning-databases/ecoli/

```
import pandas as pd
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/ecoli/ecoli.data'
df = pd.read_csv(url, header=None, delimiter='\s+',names = ["Sequence Name","mcg","gvh","lip","chg","aac","alm1","alm2","cp_pp"])
print(df)
```

```
    Sequence Name   mcg   gvh   lip  chg   aac  alm1  alm2 cp_pp
0       AAT_ECOLI  0.49  0.29  0.48  0.5  0.56  0.24  0.35    cp
1      ACEA_ECOLI  0.07  0.40  0.48  0.5  0.54  0.35  0.44    cp
2      ACEK_ECOLI  0.56  0.40  0.48  0.5  0.49  0.37  0.46    cp
3      ACKA_ECOLI  0.59  0.49  0.48  0.5  0.52  0.45  0.36    cp
4       ADI_ECOLI  0.23  0.32  0.48  0.5  0.55  0.25  0.35    cp
..            ...   ...   ...   ...  ...   ...   ...   ...   ...
331    TREA_ECOLI  0.74  0.56  0.48  0.5  0.47  0.68  0.30    pp
332    UGPB_ECOLI  0.71  0.57  0.48  0.5  0.48  0.35  0.32    pp
333    USHA_ECOLI  0.61  0.60  0.48  0.5  0.44  0.39  0.38    pp
334    XYLF_ECOLI  0.59  0.61  0.48  0.5  0.42  0.42  0.37    pp
335    YTFQ_ECOLI  0.74  0.74  0.48  0.5  0.31  0.53  0.52    pp

[336 rows x 9 columns]
```

Öncelikle pandası import ediyoruz.Ardından datasetin url'sini tanımladım ve dataseti okuyoruz.Atrribute isimlerini names kısmına kendimiz giriyoruz.

```
In [94]: df.head()
```

Out[94]:

	Sequence Name	mcg	gvh	lip	chg	aac	alm1	alm2	sabit
0	AAT_ECOLI	0.49	0.29	0.48	0.5	0.56	0.24	0.35	cp
1	ACEA_ECOLI	0.07	0.40	0.48	0.5	0.54	0.35	0.44	cp
2	ACEK_ECOLI	0.56	0.40	0.48	0.5	0.49	0.37	0.46	cp
3	ACKA_ECOLI	0.59	0.49	0.48	0.5	0.52	0.45	0.36	cp
4	ADI_ECOLI	0.23	0.32	0.48	0.5	0.55	0.25	0.35	cp

Burda sadece baş kısmını gösteriyor.

```
In [106]: df.tail()
```

Out[106]:

	Sequence Name	mcg	gvh	lip	chg	aac	alm1	alm2	cp_pp
331	TREA_ECOLI	0.74	0.56	0.48	0.5	0.47	0.68	0.30	pp
332	UGPB_ECOLI	0.71	0.57	0.48	0.5	0.48	0.35	0.32	pp
333	USHA_ECOLI	0.61	0.60	0.48	0.5	0.44	0.39	0.38	pp
334	XYLF_ECOLI	0.59	0.61	0.48	0.5	0.42	0.42	0.37	pp
335	YTFQ_ECOLI	0.74	0.74	0.48	0.5	0.31	0.53	0.52	pp

Burda ise son kısmı gösteriyor.

Şimdi ise veri analizi kısmına gelelim.Sutunlarımızı bağımlı ve bağımsız değişken olarak atamalıyız.

```
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
X = df.iloc[:, 1:-1]
y = df.iloc[:, -1]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Bağımsız değişken son sutun yani cp_pp sutunu olarak atadım diğer özellikler bağımlı değişken olacak.

```
model = DecisionTreeClassifier()
model.fit(X_train, y_train)
test_pred = model.predict(X_test)
train_pred = model.predict(X_train)
test_accuracy = accuracy_score(y_test, test_pred)
train_accuracy = accuracy_score(y_train,train_pred)
print("Test verileri doğruluk değeri:", test_accuracy)
print("Eğitim verileri doğruluk değeri:",train_accuracy)
```

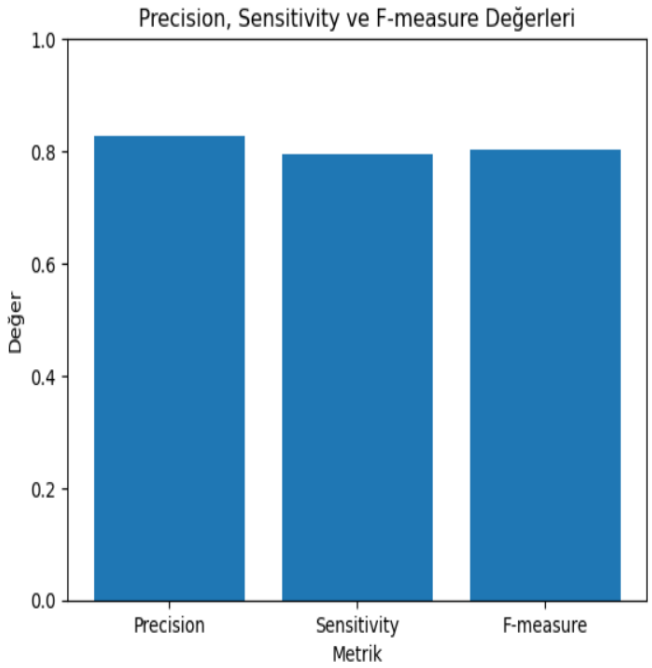
```
Test verileri doğruluk değeri: 0.8088235294117647
Eğitim verileri doğruluk değeri: 1.0
```

Ardından modelimizi karar ağaçlarıyla kuruyoruz.Önceden tanımladığımız değişkenlerle modelimizi eğitelim ve tahmini değerleri tanımlayarak doğruluk değerlerimizi hesapladım.

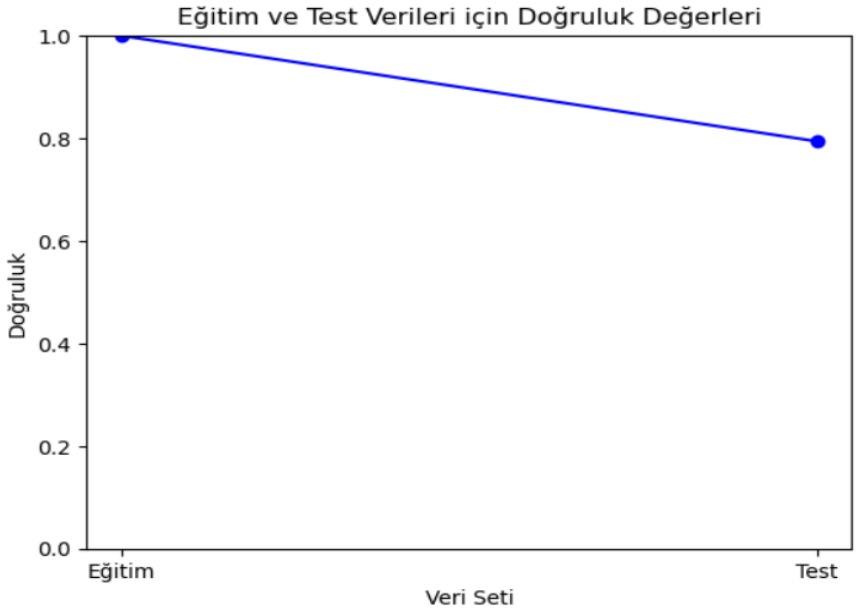
```
: from sklearn.metrics import confusion_matrix, recall_score, f1_score,precision_score
cm = confusion_matrix(y_test, y_pred)
sensitivity = recall_score(y_test, y_pred, average='weighted')
f_measure = f1_score(y_test, y_pred, average='weighted')
precision = precision_score(y_test, y_pred, average='weighted')
print("Precision değeri:", precision)
print("Sensitivity değeri:", sensitivity)
print("F-measure değeri :", f_measure)
```

```
Precision değeri: 0.827431757016532
Sensitivity değeri: 0.7941176470588235
F-measure değeri : 0.8017030284293712
```

Sonrasında precision,sensitivity ve f-measure gibi değerleri hesapladım ve bunları görselleştirdim.



```
plt.show()
```



Bu da doğruluk değerleri tablosu

Akademik çalışma :Decision tree with minimal costs.

https://www.csd.uwo.ca/~xling/cs860/ICML04-Ling.pdf

Ve sonuçları:

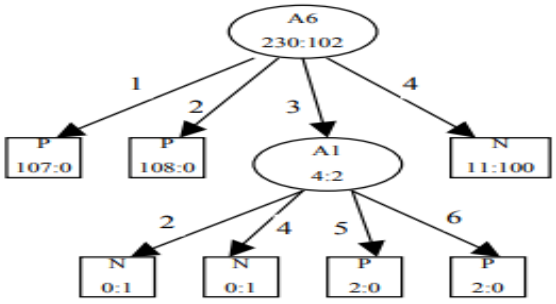


Figure 3. A decision tree built from the Ecoli dataset (costs are set as in Table 2).

Table 2. Test and misclassification costs set for Ecoli dataset.

A1	A2	A3	A4	A5	A6	FP/FN
50	50	50	50	50	20	800/800

Table 3. An example test case with several unknown values. The true values are in parenthesis and can be obtained by performing the tests (with costs list in Table 2).

A1	A2	A3	A4	A5	A6	Class
? (6)	2	? (1)	2	2	? (3)	P

with missing values.

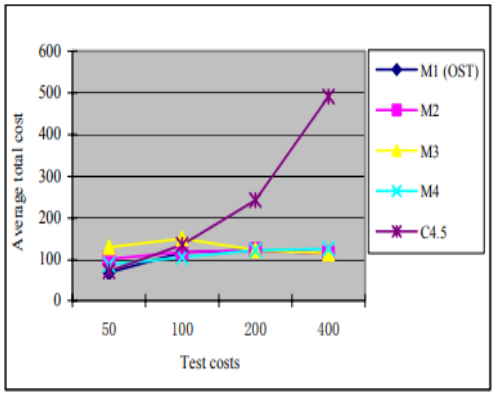


Figure 5. Comparison under different test costs.

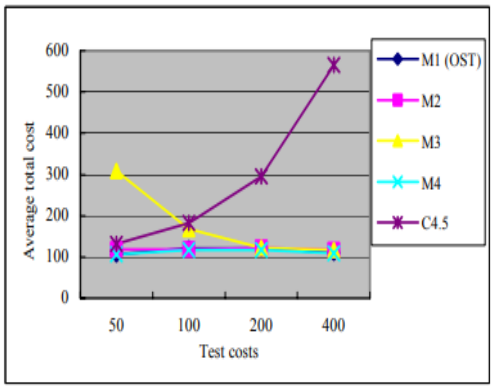


Figure 6. Comparing unbalanced misclassification costs.