feature-selection-22mcb1002

April 9, 2023

[61]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import OrdinalEncoder
from sklearn.preprocessing import LabelEncoder
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
from sklearn.feature_selection import mutual_info_classif
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

[106]:
```
df = pd.read_csv('/content/Student.csv')
df.dropna(inplace = True)
df = df.iloc[:, 1:5]
```

[107]:
```
dataset = df.values
dataset
```

[107]:
```
array([['group B', "bachelor's degree", 'standard', 'incompleted'],
       ['group C', 'some college', 'standard', 'completed'],
       ['group B', "master's degree", 'standard', 'incompleted'],
       …,
       ['group C', 'high school', 'free', 'completed'],
       ['group D', 'some college', 'standard', 'completed'],
       ['group D', 'some college', 'free', 'incompleted']], dtype=object)
```

[110]:
```
# Splitting data intp input and output variables
X = dataset[:,:-1]
Y = dataset[:,-1]
```

[111]:
```
# Formatting fields as strings
X = X.astype(str)
```

[112]:
```
# Splitting data in training and testing set
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.33,
                                                    random_state = 1)
```

```
print("Train", X_train.shape,Y_train.shape)
print("Test", X_test.shape, Y_test.shape)
```

```
Train (670, 3) (670,)
Test (330, 3) (330,)
```

[113]:
```
# Preparing input variable
def prepare_inputs(X_train, X_test):
  oe = OrdinalEncoder()
  # Fitting encoding on training set
  oe.fit(X_train)
  # Applying on train set
  X_train_enc = oe.transform(X_train)
  # Applying on test set
  X_test_enc=oe.transform(X_test)
  return X_train_enc, X_test_enc
```

[114]:
```
# Preparing target variable
def prepare_target(Y_train, Y_test):
  le = LabelEncoder()
  le.fit(Y_train)
  Y_train_enc = le.transform(Y_train)
  Y_test_enc = le.transform(Y_test)
  return Y_train_enc, Y_test_enc
```

[115]:
```
X_train_enc, X_test_enc = prepare_inputs(X_train, X_test)
Y_train_enc, Y_test_enc = prepare_target(Y_train, Y_test)
print('Train', X_train_enc.shape, Y_train_enc.shape)
print('Test', X_test_enc.shape, Y_test_enc.shape)
```

```
Train (670, 3) (670,)
Test (330, 3) (330,)
```

1 Chi- Square feature selection

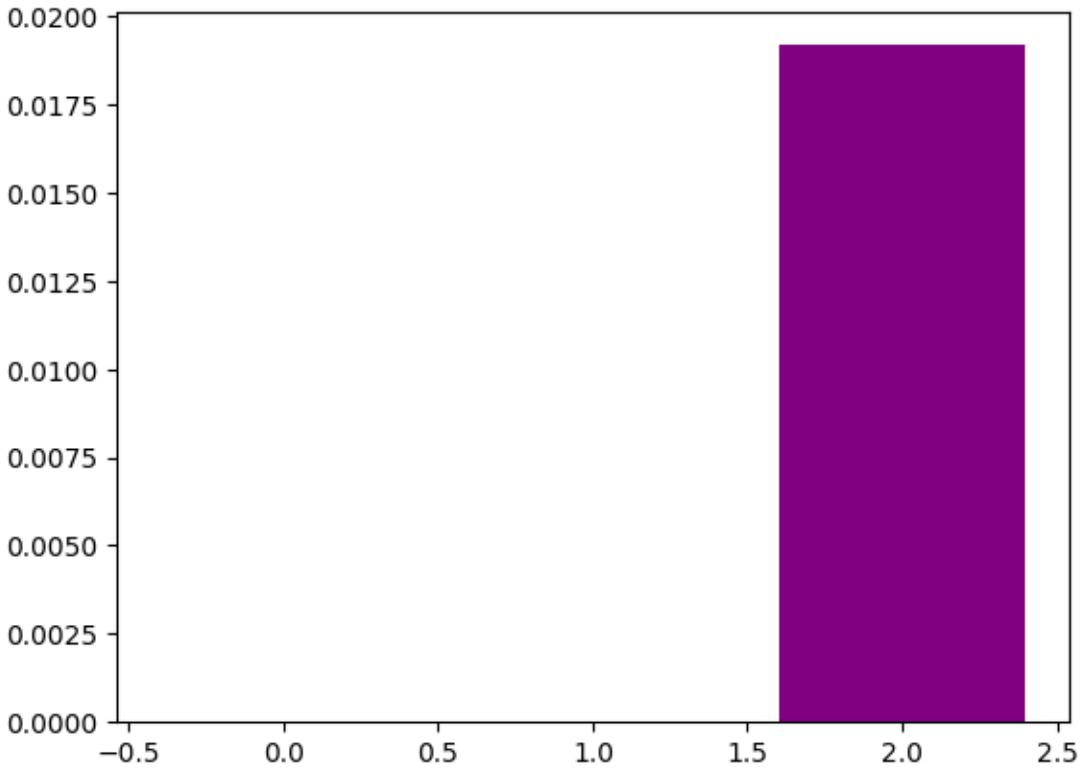
[116]:
```
def select_features(X_train, Y_train, X_test):
  fs = SelectKBest(score_func=chi2, k='all')
  fs.fit(X_train, Y_train)
  X_train_fs = fs.transform(X_train)
  X_test_fs = fs.transform(X_test)
  return X_train_fs, X_test_fs, fs
```

[117]:
```
X_train_fs, X_test_fs, fs = select_features(X_train_enc, Y_train_enc,␣
 ↪X_test_enc)
```

[118]:
```
fs.scores_
```

[118]:
```
array([0.00916221, 6.05500635, 0.52178586])
```

[128]:
```
plt.bar([i for i in range (len(fs.scores_))],fs.scores_, color = 'purple')
plt.show()
```



2 Mutual Information feature selection

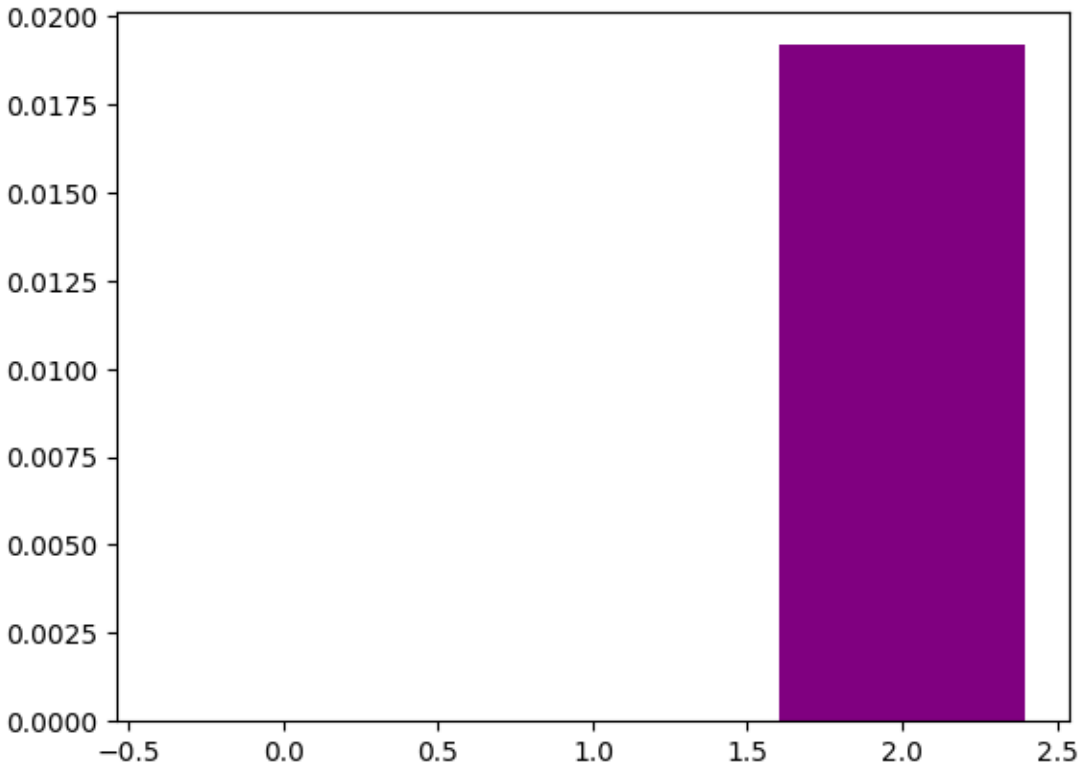
[120]:
```
def select_features_2(X_train, Y_train, X_test):
  fs=SelectKBest(score_func=mutual_info_classif, k='all')
  fs.fit(X_train, Y_train)
  X_train_fs_2=fs.transform(X_train)
  X_test_fs_2=fs.transform(X_test)
  return X_train_fs_2, X_test_fs_2, fs

#Calling feature selection function
X_train_fs_2, X_test_fs_2, fs = select_features_2(X_train_enc, Y_train_enc,␣
 ↪X_test_enc)

fs.scores_
```

[120]:
```
array([0.        , 0.        , 0.01917452])
```

[125]:
```
plt.bar([i for i in range (len(fs.scores_))],fs.scores_, color = 'purple')
plt.show()
```



3 Model built using all features

[122]:
```
model = LogisticRegression(solver='lbfgs')
model.fit(X_train_enc, Y_train_enc)

LogisticRegression(C=1.0, class_weight = None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio = None, max_iter=100,
multi_class = 'auto', n_jobs=None, penalty = 'l2',
random_state=None, solver='lbfgs', tol=0.0001, verbose=0, warm_start=False)

#predict the model
yhat = model.predict(X_test_enc)
#Evaluate the prediction
accuracy = accuracy_score(Y_test_enc, yhat)
print("Accuracy: %.2f" %(accuracy*100))
```

```
Accuracy: 63.94
```

4 Model built using Chi-squared features

[123]:
```
model1 = LogisticRegression(solver='lbfgs')
#fit the model
model1.fit(X_train_fs, Y_train_enc)
#evaluate the model
yhat = model1.predict(X_test_fs)
#evaluate the performance
accuracy = accuracy_score(Y_test_enc, yhat)
print("Accuracy: %.2f" %(accuracy*100))
```

```
Accuracy: 63.94
```

5 Model built using Mutual Information

[124]:
```
model2 = LogisticRegression(solver = 'lbfgs')
#fit the model
model2.fit(X_train_fs_2, Y_train_enc)
#evaluate the model
yhat = model2.predict(X_test_fs_2)
#evaluate the performance
accuracy = accuracy_score(Y_test_enc, yhat)
print("Accuracy: %.2f" %(accuracy*100))
```

```
Accuracy: 63.94
```

Accuracy of chi-square and mutual information is same