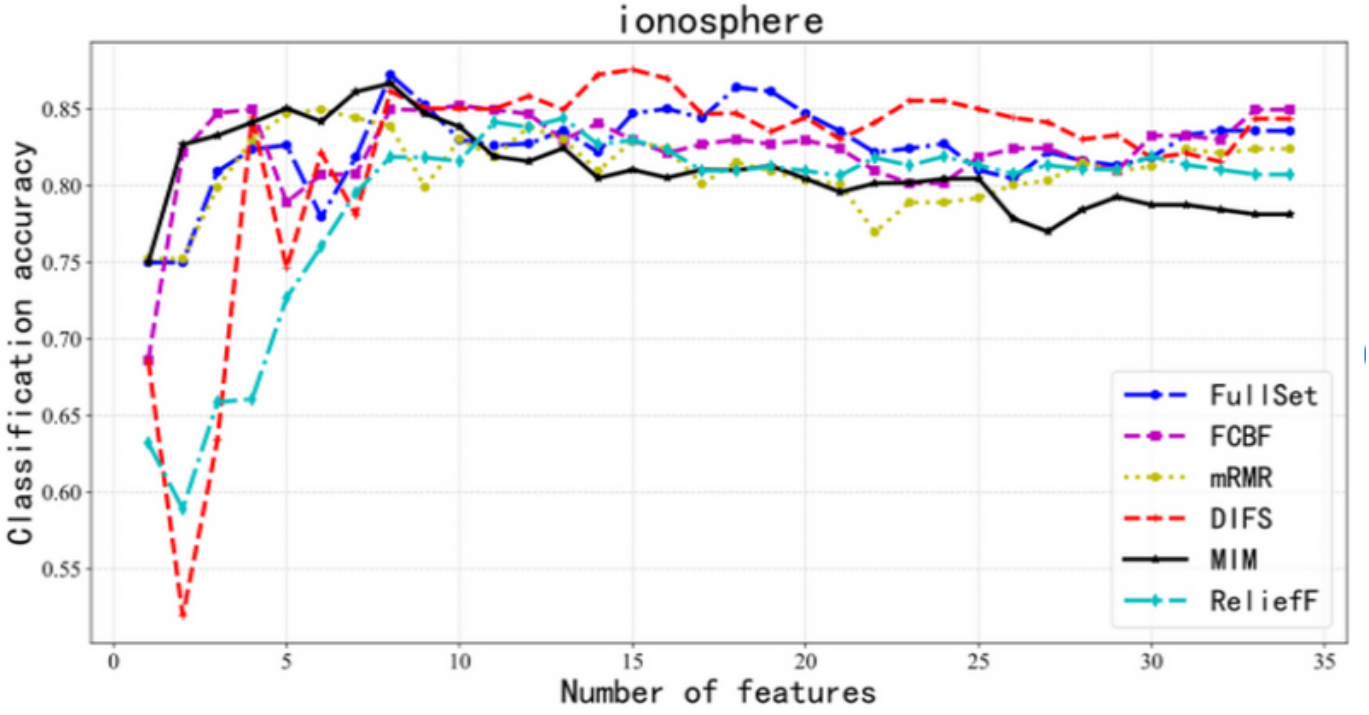
Bursa Teknik Üniversitesi Bilgisayar Mühendisliği Bölümü Veri Madenciliğine Giriş Dersi Proje Ödevi

19360859041
Çağla Uzundurukan

İstenen: Sınıflandırma yöntemlerinden birini benimseyerek UCI Machine Learning Repository'den (https://archive.ics.uci.edu/ml/index.php") istediğiniz bir veri seti üzerinde istediğiniz bir platformda sınıflandırma veya kümeleme yaparak elde ettiğiniz sonuçları yaygın değerlendirme(Accuracy, sensitivity, specificity, Fmeasure vb) ölçütleriyle (çeşitli görselleştirme araçlarıyla zenginleştirerek) sunmanız beklenmektedir. Repository'den seçtiğiniz veri seri üzerinde daha önceden yapılmış en az bir (akademik) çalışmayı bulup onun sonuçlarını da karşılaştırma amacıyla sunumunuza ekleyiniz.

Seçimlerim: K-Nearest-Neighbor (KNN) sınıflandırma yöntemi ve Ionosphere Data Set.

Sunum videomun youtube linki: https://www.youtube.com/watch?v=_v1MhSPNRYk



Average classification accuracy of KNN classifier on ionosphere data set

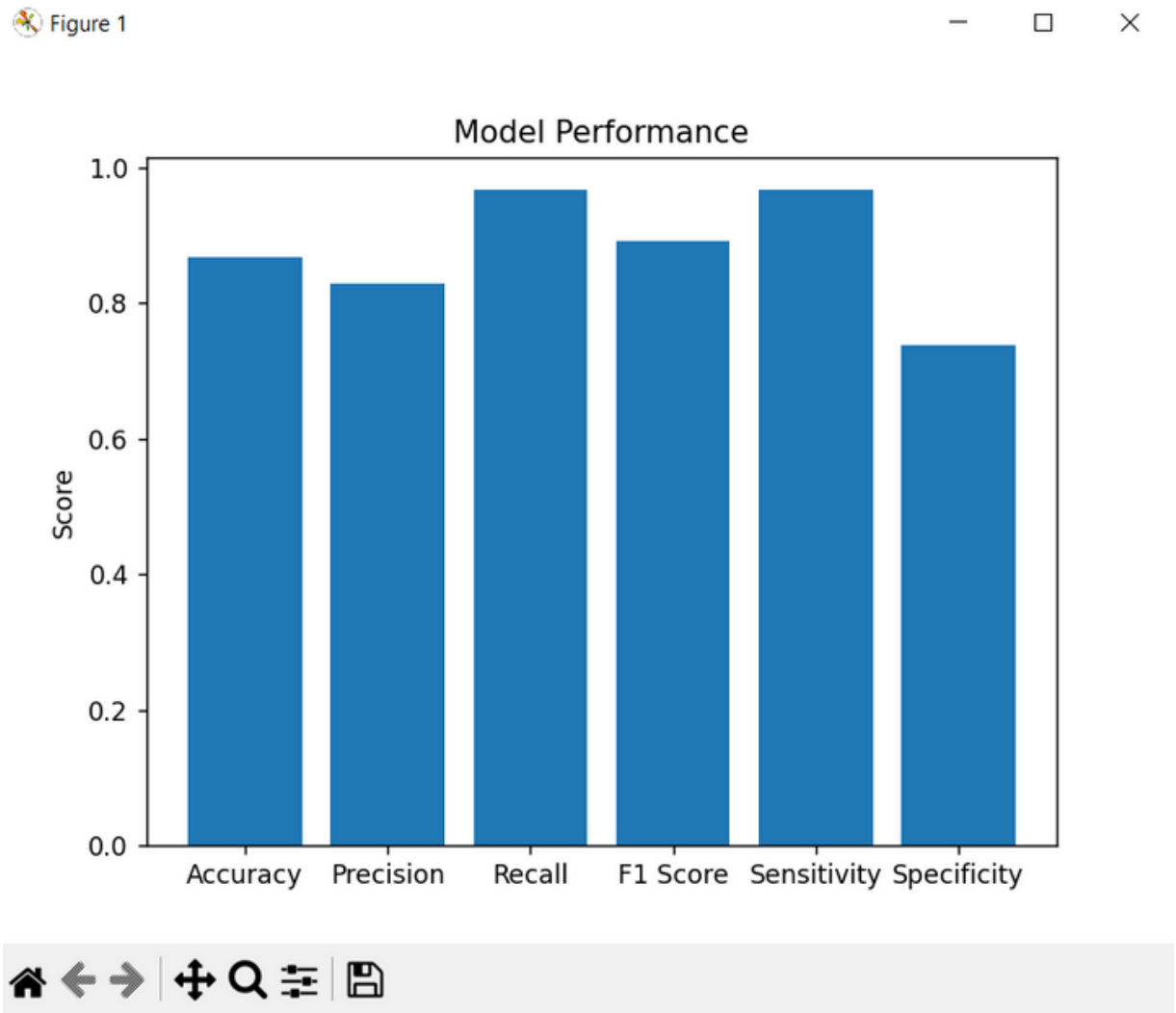
Proje Kodları:

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
import matplotlib.pyplot as plt


# Veri setini yukle
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/ionosphere/ionosphere.data'
df = pd.read_csv(url, header=None)

# Sinif etiketlerini kodla
df[34] = pd.Categorical(df[34])
df[34] = df[34].cat.codes

# Ozellikleri ve sinif etiketlerini ayir
X = df.iloc[:, :-1].values
y = df.iloc[:, -1].values

# Ozellikleri Olceklendir
sc = StandardScaler()
X = sc.fit_transform(X)

# Veri setini egitim ve test kumelerine ayir
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=40)

# Modeli egit
knn = KNeighborsClassifier(n_neighbors=5, weights='distance')
knn.fit(X_train, y_train)

# Test veri setiyle modeli dogrula ve performansi hesapla
y_pred = knn.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
sensitivity = recall_score(y_test, y_pred, pos_label=1)
specificity = recall_score(y_test, y_pred, pos_label=0)

results = {"Accuracy": accuracy, "Precision": precision, "Recall": recall, "F1 Score": f1,
"Sensitivity": sensitivity, "Specificity": specificity }

fig, ax = plt.subplots()
ax.bar(results.keys(), results.values())
ax.set_ylabel('Score')
ax.set_title('Model Performance')

# Performansi ekrana yazdir
print("Accuracy:", accuracy)
print("Precision:", precision)
print("Recall:", recall)
print("F1 Score:", f1)
print("Sensitivity:", sensitivity)
print("Specificity:", specificity)
plt.show()
```

Kodların Çıktıları:

Seç Komut İstemi - python main.py

```
C:\Users\ÇAĞLA\Desktop\veri_madenciligi>python main.py
Accuracy: 0.8679245283018868
Precision: 0.8285714285714286
Recall: 0.9666666666666667
F1 Score: 0.8923076923076922
Sensitivity: 0.9666666666666667
Specificity: 0.7391304347826086
```



Ionosphere Data Set Üzerinde KNN Algoritması ile Oluşturulan Akademik Çalışma:

https://www.sciencedirect.com/science/article/pii/S266730532200014X

Shashank Shekhar, Nazrul Hoque, Dhruba K. Bhattacharyya. PKNN-MIFS: A Parallel KNN Classifier over an Optimal Subset of Features.

Table 1. Symbols and their meanings.

Symbol	Meaning	Symbol	Meaning
D	Dataset	F	Original feature set
C_i	Class label for object O_i	f_i	i^th feature
C_d	Domination count	D_{train}	Train set
D_{test}	Test set	K	No. of neighbors
d	Total number of features	F'	Optimal feature set
FL	No. of features in F'	F_d	Dominated count
P	Precision	R	Recall
F_1	F_1 Score	Acc	Accuracy
MCC	Matthew's correlation coefficient		

Table 3. Comparison of PKNN-MIFS with KNN.

		KNN						PKNN-MIFS					
	Dataset	K=5						K=5					
SI		d	P	R	F1	MCC	Acc	FL	P	R	F1	MCC	Acc
1	Abalone	8	0.52	0.52	0.52	0.29	0.528	6	0.54	0.55	0.54	**0.32**	**0.553**
2	Acute	8	1	1	1	1	1	3	1	1	1	**1**	**1**
3	Avila	10	0.81	0.71	0.75	0.73	0.751	3	1	1	1	**1**	**1**
4	Breast Cancer 1	9	0.97	0.95	0.95	0.93	0.964	8	0.97	0.97	0.97	**0.94**	**0.971**
5	Breast Cancer 2	30	0.97	0.95	0.95	0.91	0.956	7	0.97	0.96	0.96	**0.93**	**0.965**
6	Breast Cancer 3	33	0.69	0.6	0.61	0.28	0.75	3	0.73	0.65	0.66	**0.37**	**0.775**
7	Cloud	10	1	1	1	1	1	1	1	1	1	1	1
8	Ecoli	7	0.63	0.63	0.63	0.65	0.867	6	0.64	0.64	0.64	**0.69**	0.867
9	Ionosphere	33	0.87	0.8	0.81	0.67	0.831	13	0.9	0.87	0.88	**0.78**	**0.887**
10	Iris	4	1	1	1	1	1	2	1	1	1	1	1
11	Monk2	6	0.83	0.82	0.82	0.8	0.9	3	0.86	0.85	0.85	**0.84**	**0.917**
12	Monk 3	6	0.95	0.96	0.95	0.91	0.955	3	0.96	0.96	0.96	**0.93**	**0.964**
13	Sonar	60	0.84	0.81	0.82	0.65	0.809	8	0.82	0.81	0.81	0.63	0.809
14	Wine	13	0.95	0.98	0.96	0.95	0.972	11	0.95	0.98	0.96	0.95	0.972

Yukarıda bahsedilen akademik çalışmada, Ionosphere veri setinde KNN algoritması kullanılarak bazı sonuçlar elde edilmiştir örneğin:
Accuracy: 0.831
Precision: 0.87
Recall: 0.80
F1 Score: 0.81

Benim projemde ise yine Ionosphere veri seti ve KNN algoritması kullanılarak şu sonuçlar elde edilmiştir:
Accuracy: 0.867
Precision: 0.82
Recall: 0.96
F1 Score: 0.89
Sensitivity: 0.96
Specificity: 0.73

Daha anlaşılır olması açısından bu terimlerin ne ifade ettiğini basit bir şekilde açıklamak isterim.

Accuracy, Precision, Recall ve F1 Score gibi metrikler makine öğrenimi modellerinin performansını ölçmek için kullanılır.

- Accuracy (Doğruluk): Doğru tahmin edilen örneklerin toplam örneklere oranıdır. Yanlış pozitifler ve yanlış negatifler eşit olarak hesaba katılır.
- Precision (Kesinlik): Modelin pozitif olarak tahmin ettiği örneklerin gerçekten pozitif olan örnekler arasındaki oranıdır. Bu, yanlış pozitiflerin azaltılması gerektiği durumlarda önemlidir.
- Recall (Duyarlılık): Gerçek pozitif örneklerin tüm pozitif örnekler içindeki oranıdır. Bu, yanlış negatiflerin azaltılması gerektiği durumlarda önemlidir.
- F1 Score: Precision ve Recall metriklerinin harmonik ortalamasını ifade eder. İkisi arasındaki dengeyi gösterir ve sadece tek bir metrik kullanmaktan daha iyi bir performans ölçüsüdür.

Accuracy, Precision, Recall ve F1 Score 1'e ne kadar yakınsa sonuç o kadar iyi kabul edilir. Ancak, her bir metrik için kabul edilebilir bir minimum değer yoktur ve veri seti ve problem bağlamına göre değişebilir. Örneğin, sınıf dengesi çok farklı olan bir veri setinde, sadece çoğunluk sınıfına tahmin yaparak yüksek bir Accuracy elde edebilirsiniz, ancak bu sonuç diğer metriklerde düşük olacaktır. Bu nedenle, doğru metriklerin seçilmesi ve sonuçların problem bağlamına göre yorumlanması önemlidir.

Kaynakça

Shashank Shekhar, Nazrul Hoque, Dhruba K. Bhattacharyya. PKNN-MIFS: A Parallel KNN Classifier over an Optimal Subset of Features.

https://www.sciencedirect.com/science/article/pii/S266730532200014X

https://archive.ics.uci.edu/ml/datasets/ionosphere

https://www.researchgate.net/figure/Average-classification-accuracy-of-KNN-classifier-on-ionosphere-data-set_fig1_345333755