Kollektif Öğrenme Dersi 2. Ödev Raporu

Yapılan çalışma ile birbirinden farklı 20 veriseti Bagging, Random Subspace ve Random Forest algoritmaları yardımıyla sınıflandırılmıştır. İlk adımda ZeroRule<0.8 kuralını sağlayan, sınıflandırmada kullanılacak farklı veri setleri seçilmiştir. Daha sonraki adımlar tek bir veri seti üzerinde açıklanacak olup 20 veri seti için benzer işlemler tekrarlanmıştır. Sınıflandırmada eğitim ve test için 5x2 fold cross validation yöntemi uygulanmıştır. Bunun için ikinci adımda veri seti seti iki ayrı parçaya bölünmüştür. Bu parçaların ilki modeli eğitmek amacıyla, ikincisi test seti olarak kullanılmıştır. Bir sonraki adımda eğitim ve test parçaları tersine çevrilerek benzer şekilde sınıflandırma işlemi uygulanmıştır. Böylece 2 fold cross validation yapılmıştır. Veri seti 5 defa rastsal olarak ikiye bölünerek benzer işlemler uygulanmıştır.

Kollektif öğrenme yöntemlerinden birisi olan Bagging için, M örnekten oluşan eğitim verisinden rastgele örnekler seçilerek, N (tekil öğrenici sayısı) tane M örnekli eğitim seti oluşturulmuştur. Tekrarlı verilerden oluşan bu eğitim setleri eğitilerek modeller elde edilmiştir. Test seti için her bir modelle tahmin yapılmıştır. N tane tekil öğrenicinin tahminleri demokrasi usulüyle birleştirilerek son tahminler elde edilmiştir.

Sonraki adımda, Random Subspace ile sınıflandırma yapmak için, her bir eğitim setinde N tane örnek alt uzayı oluşturulmuştur. Bu adımda veri setinin özellik sayısının yarısı kadar, eğitimde kullanılacak olan rastgele özellikler belirlenmiştir. N alt uzay için N defa özellik seçimi yapılmıştır. Bu alt uzayların her biri, rastgele seçilmiş özellikleri olan örneklerin tümünü içermektedir. N alt uzay tekil karar ağacı ile eğitilerek, oluşturulan modellerle yapılan tahminler demokrasi usulü birleştirilmiştir. Bagging yöntemindekinin aksine örneklerin bazıları değil tümü seçilmiştir.

Üçüncü yöntem olan Random Forest ile sınıflandırma için ilk adımda Bagging metoduyla M örnekten oluşan N tane eğitim seti oluşturulmuştur. Eğitim setleri N tane karar ağacıyla eğitilmiştir. Bu adımda her düğüm için $\log_2 N$ tane özellik rastgele seçilerek karar ağaçları oluşturulmuştur. N modelin tahmini demokrasi usulüyle birleştirilmiştir.

Tahminler test setinin gerçek etiketleriyle karşılaştılıp başarı oranları hesaplanmıştır.

Yukarıda açıklanan işlem adımları tüm eğitim ve test setlerine uygulanarak, her bir verisetinde toplamda 10 farklı sonuç elde edilmiştir. Yapılan çalışma 3 yöntem ve 20 farklı verisetiyle gerçekleştirilmiş olup toplamda 3x20x10 sonuç elde edilmiştir. Metotlar ve farklı verisetleri üzerindeki başarıları Tablo 1'de gösterilmiştir.

Tablo 1 incelendiğinde, her üç yöntemin de veri setlerini sınıflandırmada yakın başarıya sahip oldukları söylenebilir. Farklı veri setlerinde farklı yöntemler daha başarılıdır. Ancak başarı oranlarında az fark olduğu gözlenmektedir.

Dataset	Bagging										Random Subspace										Random Forest									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
diabetes	0,75	0,77	0,71	0,74	0,77	0,71	0,78	0,74	0,72	0,75	0,69	0,76	0,69	0,74	0,67	0,73	0,70	0,75	0,68	0,71	0,71	0,77	0,72	0,76	0,69	0,78	0,70	0,76	0,70	0,77
ringnorm	0,94	0,94	0,94	0,95	0,94	0,95	0,94	0,94	0,95	0,94	0,96	0,95	0,96	0,95	0,96	0,95	0,96	0,95	0,96	0,96	0,95	0,95	0,96	0,95	0,96	0,94	0,95	0,95	0,95	0,94
zoo	1	0,86	0,95	0,98	0,79	1	0,95	1	0,95	1	1	1	1	1	0,95	0,98	0,95	1	1	0,98	1	0,98	1	0,98	1	1	1	0,98	1	1
vote	0,95	0,97	0,97	0,95	0,95	0,95	0,97	0,93	0,95	0,93	0,93	0,95	0,91	0,95	0,93	0,96	0,93	0,96	0,95	0,93	0,95	0,97	0,96	0,96	0,97	0,94	0,97	0,96	0,96	0,94
vehicle	0,76	0,72	0,70	0,74	0,71	0,72	0,73	0,72	0,76	0,74	0,74	0,72	0,75	0,73	0,74	0,74	0,76	0,72	0,74	0,72	0,76	0,71	0,76	0,75	0,75	0,73	0,75	0,73	0,75	0,74
sonar	0,70	0,74	0,80	0,71	0,80	0,75	0,72	0,77	0,69	0,74	0,72	0,74	0,70	0,79	0,68	0,79	0,70	0,77	0,71	0,77	0,72	0,75	0,71	0,84	0,69	0,83	0,75	0,81	0,70	0,82
audiology	0,77	0,86	0,85	0,87	0,87	0,86	0,88	0,85	0,92	0,87	0,89	0,80	0,90	0,80	0,85	0,80	0,89	0,80	0,82	0,83	0,92	0,75	0,83	0,75	0,76	0,70	0,81	0,75	0,81	0,75
autos	0,62	0,73	0,79	0,70	0,73	0,74	0,73	0,70	0,75	0,73	0,80	0,68	0,76	0,72	0,73	0,68	0,82	0,68	0,76	0,68	0,69	0,67	0,70	0,66	0,69	0,65	0,70	0,63	0,67	0,65
labor	0,96	0,82	0,96	0,86	0,86	0,89	0,89	0,93	0,86	0,93	0,82	0,96	0,82	0,96	0,82	0,93	0,86	1	0,82	0,96	0,96	0,96	0,89	0,89	0,96	0,93	0,93	0,96	0,79	0,86
mushroom	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
soybean	0,93	0,89	0,91	0,90	0,92	0,93	0,91	0,93	0,93	0,88	0,92	0,90	0,93	0,91	0,93	0,89	0,93	0,88	0,93	0,92	0,92	0,88	0,93	0,87	0,93	0,86	0,93	0,88	0,92	0,89
glass	0,68	0,71	0,67	0,64	0,72	0,70	0,58	0,70	0,66	0,68	0,73	0,68	0,70	0,67	0,71	0,70	0,56	0,71	0,66	0,71	0,66	0,71	0,65	0,75	0,65	0,67	0,63	0,69	0,75	0,72
balance-scale	0,83	0,84	0,85	0,80	0,86	0,83	0,85	0,84	0,86	0,84	0,79	0,76	0,78	0,79	0,76	0,79	0,79	0,80	0,77	0,78	0,85	0,83	0,84	0,84	0,87	0,84	0,84	0,85	0,84	0,83
credit-a	0,86	0,86	0,86	0,85	0,85	0,86	0,87	0,86	0,86	0,87	0,84	0,86	0,86	0,85	0,85	0,86	0,84	0,86	0,84	0,84	0,86	0,84	0,86	0,85	0,85	0,85	0,83	0,85	0,85	0,86
heart-statlog	0,76	0,81	0,84	0,80	0,82	0,84	0,76	0,76	0,83	0,77	0,86	0,81	0,84	0,82	0,84	0,81	0,80	0,80	0,81	0,81	0,90	0,80	0,87	0,79	0,84	0,75	0,85	0,79	0,87	0,80
splice	0,96	0,94	0,94	0,95	0,94	0,95	0,95	0,95	0,95	0,94	0,96	0,95	0,96	0,95	0,95	0,95	0,96	0,95	0,95	0,95	0,90	0,90	0,93	0,91	0,91	0,89	0,91	0,92	0,92	0,92
vowel	0,79	0,80	0,86	0,78	0,81	0,85	0,84	0,78	0,80	0,82	0,81	0,82	0,81	0,80	0,79	0,81	0,78	0,83	0,82	0,81	0,81	0,84	0,83	0,82	0,81	0,82	0,80	0,83	0,83	0,84
abalone	0,24	0,26	0,24	0,25	0,25	0,26	0,23	0,23	0,25	0,25	0,26	0,24	0,25	0,26	0,26	0,25	0,25	0,24	0,25	0,25	0,25	0,25	0,23	0,24	0,24	0,25	0,25	0,25	0,26	0,24
iris	0,93	0,93	0,95	0,95	0,93	0,93	0,91	0,96	0,92	0,92	0,93	0,92	0,95	0,91	0,92	0,91	0,92	0,95	0,92	0,91	0,92	0,93	0,93	0,92	0,92	0,92	0,92	0,92	0,93	0,92
kr-vs-kp	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,93	0,97	0,91	0,98	0,96	0,97	0,93	0,93	0,95	0,95	0,97	0,98	0,97	0,97	0,97	0,98	0,97	0,98	0,98	0,97

Tablo 1. Bagging, Random Subspace ve Random Forest yöntemlerinin farklı veri setleri üzerindeki 5x2 CV başarı oranları.

Metotların başarılarını karşılaştırmak için yalnızca ortalama hata oranlarına bakmak yeterli değildir. Bu nedenle, yöntemlerin birbirinden daha başarılı olmaları arasında anlamsal bir bağ olup olmadığı t-test ile kontrol edilmiştir. T-test için güvenilirlik aralığını belirleyen alfa değeri 0.05 seçilmiştir. Alfa değerinin küçük olması testin güvenilirliğini arttırır. T-test temelde iki hipotez arasında seçim yapar:

- H0: Verilen iki sonuç arasında anlamlı bir farklılık yoktur ya da rastlantısal bir farklılık vardır.
- H1: Verilen iki sonuç arasında anlamlı bir farklılık vardır.

Uygulana t-teste başarı sonuçları H1 hipotezine dayanıyorsa, ortalaması büyük olan metod değerlendirmeye alınmıştır. Tablo 2'de metotların her bir veri seti üzerindeki ortalama başarıları verilerek, ikili t-test sonuçları renklendirilmiştir.

Dataset	Bagging	Random Subspace	Bagging	Random Forest	Random Subspace	Random Forest
diabetes	0,74	0,71	0,74	0,73	0,71	0,73
ringnorm	0,94	0,96	0,94	0,95	0,96	0,95
zoo	0,95	0,99	0,95	0,99	0,99	0,99
vote	0,95	0,94	0,95	0,96	0,94	0,96
vehicle	0,73	0,74	0,73	0,74	0,74	0,74
sonar	0,74	0,74	0,74	0,76	0,74	0,76
audiology	0,86	0,84	0,86	0,78	0,84	0,78
autos	0,72	0,73	0,72	0,67	0,73	0,67
labor	0,90	0,90	0,90	0,91	0,90	0,91
mushroom	1	1	1	1	1	1
soybean	0,91	0,91	0,91	0,90	0,91	0,90
glass	0,67	0,68	0,67	0,68	0,68	0,68
balance-scale	0,84	0,78	0,84	0,84	0,78	0,84
credit-a	0,86	0,85	0,86	0,85	0,85	0,85
heart-statlog	0,80	0,82	0,80	0,83	0,82	0,83
splice	0,95	0,95	0,95	0,91	0,95	0,91
vowel	0,81	0,81	0,81	0,82	0,81	0,82
abalone	0,25	0,25	0,25	0,25	0,25	0,25
iris	0,93	0,92	0,93	0,92	0,92	0,92
kr-vs-kp	0,99	0,95	0,99	0,97	0,95	0,97

Tablo 2. Bagging, Random Subspace ve Random Forest yöntemlerinin farklı veri setleri üzerindeki ortalama başarısı. İkili t-test sonuçları. ("Anlamlı farklılık vardır" (H1) hipotezi sarı ile renklendirilmiştir.)

T-test ile "Anlamlı farklılık vardır" (H1) hipoteziyle sonuçlanan ikililerin veri setleri üzerindeki başarısı Tablo 3'te özetlenmiştir.

	Random Subspace	*Random Forest*
Bagging →	3/16/1	5/14/1
Random Subspace →		5/9/6

Tablo 3. Bagging, Random Subspace ve Random Forest yöntemlerinin ikili karşılaştırması. (Satırlar bölmenin sağ tarafında, sütunlar sol tarafında gösterilmiştir.)

Tablo 3 incelendiğinde Bagging yönteminin Random Supspace ve Random Forest yöntemlerinden daha başarılı olduğu görülmektedir. Random Supspace ve Random Forest yöntemlerinin yaklaşık olarak eşit başarıda olduğu gözlenmiştir. Ancak, yöntemler arasındaki başarı durumu farklı veri setleriyle değişebilir. Bu nedenle sabit ölçütlerle en başarılı yöntem seçimi yapmak mümkün değildir.