MYZ 307 - ÖDEV 3 RAPORU

Dersten Sorumlu Akademisyen: Prof. Dr. Bilge Günsel

Ata Ataş - 040230771

17 Kasım 2025

Contents

1 Giriş

1.1 Eğitim ve Test Veri Seti Hakkında

Ödevde hem eğitim hem test veri seti olarak "$breast-cancer.arff$" veri setini kullandım. Weka'dan aldığım standart data klasörü içerisinde birden fazla aynı konsepte ait veri setine ulaşamadığım için bu şekilde ilerledim. Veri kümesi 286 örnek, 10 öznitelikten oluşmaktadır.

1.2 Kullanılan Parametreler ve Seçim Nedenleri Hakkında

Ödevin ilk aşamasında "age", "$tumor-size$", "$deg-malig$" özdeğerlerinin Gauss dağılımına yakın bir dağılıma sahip olması sebebiyle, birden fazla modelin birbiriyle olan kıyasında Gauss temelli sınıflandırıcılar açısından fazla elverişsiz bir ortam yaratılmayacağı ve daha adaletli bir karşılaştırma yapılabileceği için "$breast-cancer.arff$" veri setini seçtim. İlk sonuçları ve açıklamaları, eğitilen modelleri yine bu veri seti üzerinde test ederek çıkardım ve raporladım.

2 (a) Naive Bayes Sınıflandırıcı

Değerlendirme modu: Eğitim verisi (WEKA "Use training set" ayarı kullanıldı).

Genel performans

- Doğru sınıflandırılan örnek: 215 / 286
- Accuracy: 75.17%
- Ağırlıklı Precision: 0.753
- Ağırlıklı Recall: 0.752
- Ağırlıklı F1-skoru: 0.743

Confusion Matrix

	Tahmin: no-rec.	Tahmin: rec.
Gerçek no-rec.	174	27
Gerçek rec.	44	41

Burada "no-rec." = no-recurrence-events, "rec." = recurrence-events. "$breast-cancer.arff$" veri seti içinde belirtilen iki sınıftır.

3 (b) K-Ortalamalı Öbekleyici (SimpleKMeans)

Veri kümesi: Meme kanseri (286 örnek). Bu yöntem denetimsiz (unsupervised) olduğu için karmaşıklık matrisi veya doğruluk hesaplanmaz.

Model özellikleri

- Algoritma: SimpleKMeans
- Küme sayısı: $k = 2$ (veri setinde verilen sınıf sayısına uygun olması için 2 alındı)
- Uzaklık ölçütü: Öklid Uzaklığı
- Maksimum iterasyon: 500

Küme büyüklükleri

- Küme 0: 225 örnek ($\approx 79\%$)
- Küme 1: 61 örnek ($\approx 21\%$)

Etiketlenmiş veride sınıf 0 201 iken sınıf 1 85 elemanlıdır, ama doğrudan eleman sayısına göre kümeleme verimini değerlendirmek uygun olmayacağı için öklid uzaklığı cinsinden kayıp baz alınmalıdır.

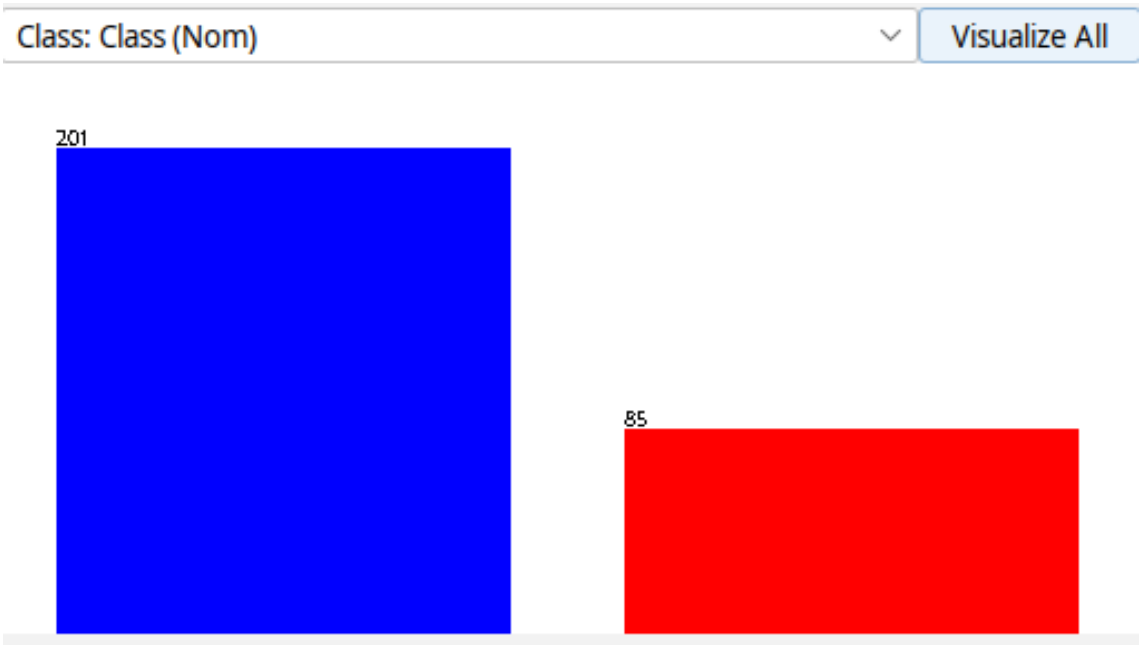

Figure 1: Sınıf Görselleştirmesi

Küme içi hata

Toplam küme içi kareler toplamı kayıp değeri(WCSS): 1177.0.

Öbekleyici öbeklerine karşın veri setinde tanımlanmış sınıfları gösteren grafik "$figure\ 2'de$" verilmiştir.

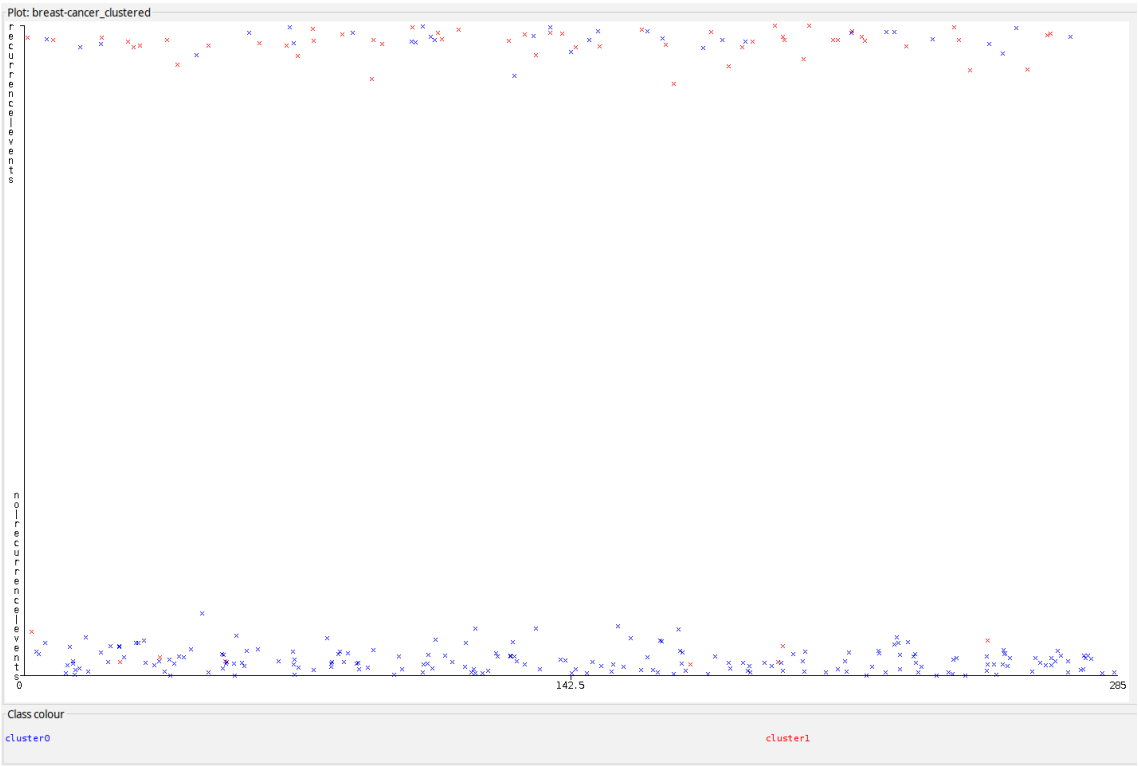

Figure 2: Öbek-Sınıf Grafiği

4 (c) k-NN Sınıflandırıcı

Veri kümesi: Meme kanseri (286 örnek). Uzaklık: Öklid Uzaklığı. Arama yöntemi: LinearNNSearch.

4.1 KNN ($k = 1$)

Değerlendirme modu: Eğitim verisi.

Genel performans

- Doğru sınıflandırılan örnek: 280 / 286
- Doğruluk: 97.90%
- Ağırlıklı Precision: 0.979
- Ağırlıklı Recall: 0.979
- Ağırlıklı F1: 0.979

Confusion Matrix

	Tahmin: no-rec.	Tahmin: rec.
Gerçek no-rec.	200	1
Gerçek rec.	5	80

4.2 KNN ($k = 5$)

Genel performans

- Doğru sınıflandırılan örnek: 221 / 286
- Doğruluk: 77.27%
- Ağırlıklı Precision: 0.791
- Ağırlıklı Recall: 0.773
- Ağırlıklı F1: 0.730

Confusion Matrix

	Tahmin: no-rec.	Tahmin: rec.
Gerçek no-rec.	197	4
Gerçek rec.	61	24

4.3 KNN ($k = 11$)

Değerlendirme modu: 10 katlı çapraz doğrulama (10-fold CV). 286 elemanlı bir veri seti için $k = 11$ çok fazla olabileceğinden cross-validation aktive ederek aldığım sonuçları ekledim.

Genel performans

- Doğru sınıflandırılan örnek: 207 / 286
- Doğruluk: 72.37%
- Ağırlıklı Precision: 0.706
- Ağırlıklı Recall: 0.724
- Ağırlıklı F1: 0.658

Confusion Matrix

	Tahmin: no-rec.	Tahmin: rec.
Gerçek no-rec.	194	7
Gerçek rec.	72	13

5 (d) AdaBoost Sınıflandırıcı

AdaBoost üç farklı temel öğrenici ile test edilmiştir: RandomForest, KNN (IB1 / $k = 1$) ve Naive Bayes.

5.1 AdaBoost + RandomForest

Genel performans

- Doğru sınıflandırılan örnek: 280 / 286

- Doğruluk: 97.90%
- Ağırlıklı Precision: 0.979
- Ağırlıklı Recall: 0.979
- Ağırlıklı F1: 0.979

Confusion Matrix

	Tahmin: no-rec.	Tahmin: rec.
Gerçek no-rec.	199	2
Gerçek rec.	4	81

5.2 AdaBoost + KNN (IB1 / $k = 1$)

Değerlendirme modu: 10-fold CV.

Genel performans

- Doğru sınıflandırılan örnek: 196 / 286
- Doğruluk: 68.53%
- Ağırlıklı Precision: 0.689
- Ağırlıklı Recall: 0.685
- Ağırlıklı F1: 0.667

Confusion Matrix

	Tahmin: no-rec.	Tahmin: rec.
Gerçek no-rec.	168	33
Gerçek rec.	57	28

5.3 AdaBoost + Naive Bayes

Genel performans

- Doğru sınıflandırılan örnek: 185 / 286
- Doğruluk: 64.68%
- Ağırlıklı Precision: 0.679
- Ağırlıklı Recall: 0.647
- Ağırlıklı F1: 0.645

Confusion Matrix

	Tahmin: no-rec.	Tahmin: rec.
Gerçek no-rec.	159	42
Gerçek rec.	59	26

6 (e) RandomForest Sınıflandırıcı

Veri kümesi: Meme kanseri. Tüm modeller eğitim verisi üzerinde değerlendirilmiştir. Üç farklı Random Forest modeli üç farklı bag-size yüzdesi verilerek oluşturulmuştur.

Bag size yüzdesi, random forest modelindeki her bir ağacın veri setinin yüzde kaçı üzerinden eğitileceğini belirler, daha düşük yüzdeler Bias-Variance alışverişinde her bir ağacın daha düşük variance sahibi olmasını sağlar ve ağaçlar toplu öğrenme için bir araya geldiklerinde tahmin yeteneği %100 bag size ile kıyaslanınca daha düşük olsa bile overfit olasılığı da aynı şekilde daha düşük olur. Aynı ilişkinin tahmin tutarlılığı daha yüksek ve overfit'e daha yatkın olan hali de %100 bag size için kurulabilir. Bu veri seti küçük bir veri seti olduğu için doğrudan bu çıkarımlara varmak doğru olmasa da, %100 bag size model en yüksek doğruluk oranına sahip olup bu overfit'e işaret olabilir.

6.1 RandomForest (Bagging Yüzdesi=%50)

Genel performans

- Doğru sınıflandırılan örnek: 271 / 286
- Doğruluk: 94.76%
- Ağırlıklı Precision: 0.949
- Ağırlıklı Recall: 0.948
- Ağırlıklı F1: 0.946

Confusion Matrix

	Tahmin: no-rec.	Tahmin: rec.
Gerçek no-rec.	199	2
Gerçek rec.	13	72

6.2 RandomForest (Bagging Yüzdesi=%80)

Genel performans

- Doğru sınıflandırılan örnek: 280 / 286
- Doğruluk: 97.90%
- Ağırlıklı Precision: 0.979
- Ağırlıklı Recall: 0.979
- Ağırlıklı F1: 0.979

Confusion Matrix

	Tahmin: no-rec.	Tahmin: rec.
Gerçek no-rec.	199	2
Gerçek rec.	4	81

6.3 RandomForest (Bagging Yüzdesi=%100)

Genel performans

- Doğru sınıflandırılan örnek: 280 / 286
- Doğruluk: 97.90%
- Ağırlıklı Precision: 0.979
- Ağırlıklı Recall: 0.979
- Ağırlıklı F1: 0.979

Confusion Matrix

	Tahmin: no-rec.	Tahmin: rec.
Gerçek no-rec.	198	3
Gerçek rec.	3	82

7 Tüm Modellerin Genel Karşılaştırması

Aşağıdaki accuracy, precision, recall, F1 karşılaştırılmıştır.

Table 1: Sınıflandırıcıların genel karşılaştırması

Model	**Accuracy**	**Prec.**	**Recall**	**F1**
Naive Bayes (train)	0.752	0.753	0.752	0.743
KNN $k = 1$ (train)	0.979	0.979	0.979	0.979
KNN $k = 5$	0.773	0.791	0.773	0.730
KNN $k = 11$ (CV)	0.724	0.706	0.724	0.658
AdaBoost + RF (train)	0.979	0.979	0.979	0.979
AdaBoost + KNN (CV)	0.685	0.689	0.685	0.667
AdaBoost + Naive (CV)	0.647	0.679	0.647	0.645
RF (P=50, train)	0.948	0.949	0.948	0.946
RF (P=80, train)	0.979	0.979	0.979	0.979
RF (P=100, train)	0.979	0.979	0.979	0.979

Yorum

- Eğitim verisi üzerinde yapılan değerlendirmelerde (KNN $k = 1$, AdaBoost+RF, RF P=80/100) doğruluk yaklaşık **98%** olup overfit'e işaret etmektedir.
- Gerçekçi sonuçlar çapraz doğrulama (CV) ile elde edilen modellerdir: KNN($k = 11$), AdaBoost+KNN ve AdaBoost+Naive Bayes. CV ile en iyi sonuç %72 doğruluk ve F1 ≈ 0.66 ile KNN($k = 11$)'dedir.
- RandomForest, eğitim verisinde en güçlü performansı göstermektedir.
- K-Means denetimsiz olduğu için denetimli modellerle doğruluk bazında doğrudan karşılaştırılamaz.