Yapay Zeka 2021-2022
Bahar Dönemi
2. Ödev

Ömer Buğrahan Çalışkan - 17011076
Ömer Aras Kaplan - 17011039

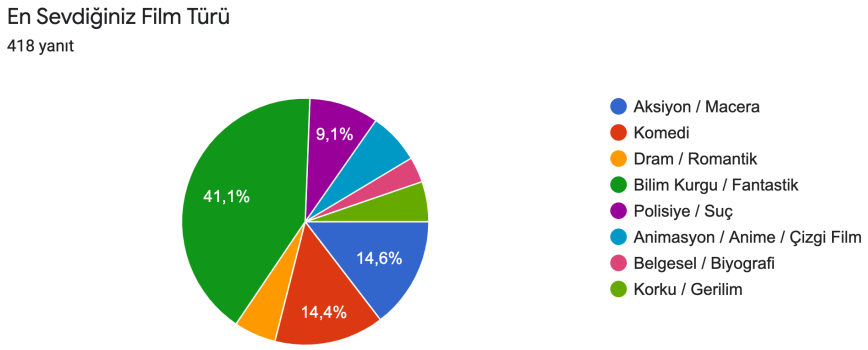
Verisetinin Hazırlanması

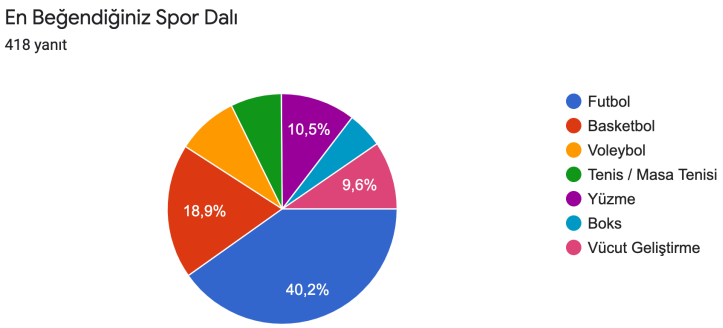
Verisetimizi hazırlayabilmek için Google forms üzerinden 11 soruluk bir anket oluşturduk. Formumuzu %78.7 si Erkek, %21.3'ü Kadın olmak üzere toplamda 418 kişi doldurmuştur. Formumuzu arkadaş çevremize, çalıştığım şirketten çalışma arkadaşlarıma ve bir canlı yayın platformu üzerinde yayın yapmakta olan bir yayıncının kitlesine yayın esnasında ulaştırarak birçok farklı kesimden birçok sayıda veri elde ettik.

Anketimizin soruları aşağıda verildiği üzeredir.

- Cinsiyet
- Yaşınız
- Kilo Aralığınız
- En Çok Vakit Harcadığınız Sosyal Medya
- Ne Sıklıkla Dışarıda Yemek Yiyorsunuz / Sipariş Veriyorsunuz
- En Sevdiğiniz Renk
- En Sevdiğiniz Film Türü
- En Çok Dinlediğiniz Müzik Türü
- En Sevdiğiniz Yiyecek Türü
- En Beğendiğiniz Spor Dalı
- Aşağıdakilerden En Çok Gezmek İstediğiniz Ülke / Ülkeler / Kıta

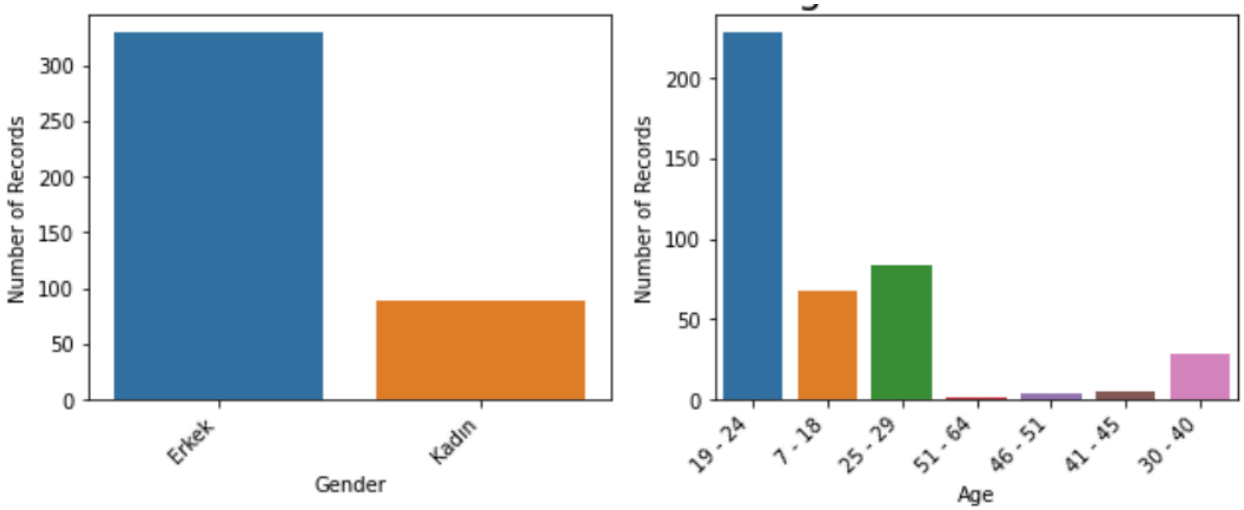
Sorularımızın cevaplanabilecek örnek birkaç şıkkı ve dağılımı aşağıda görüldüğü üzeredir.



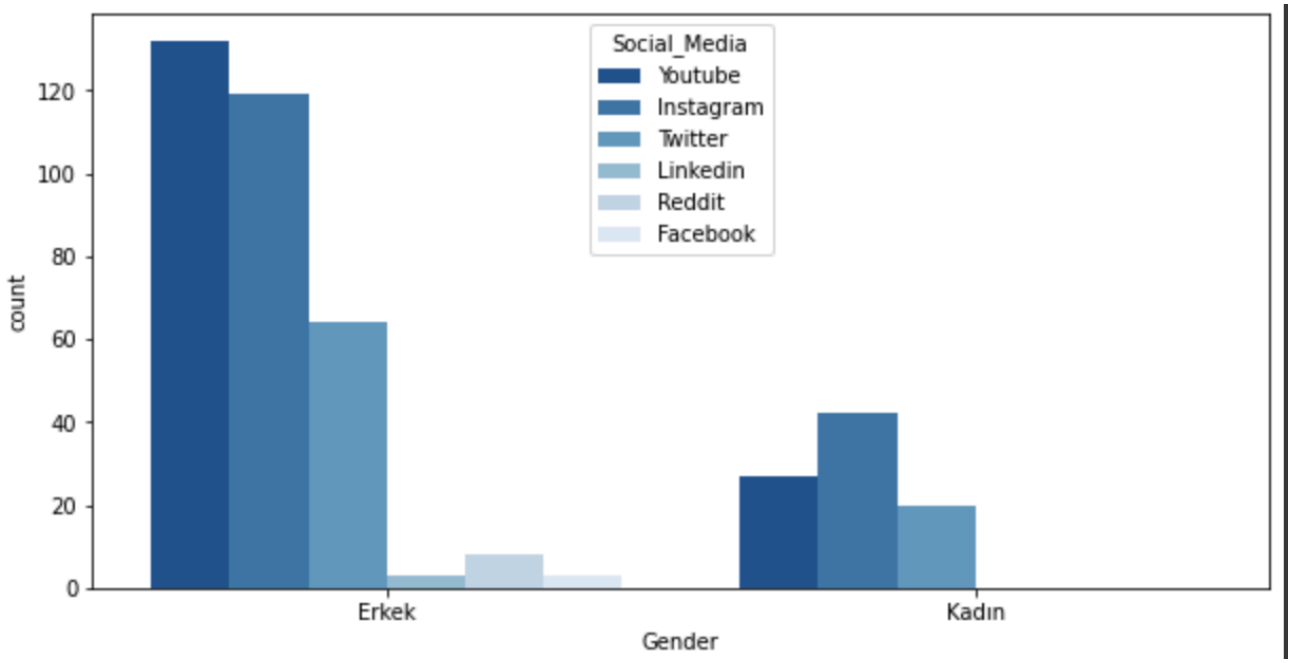


Verilere Dair Bulgular

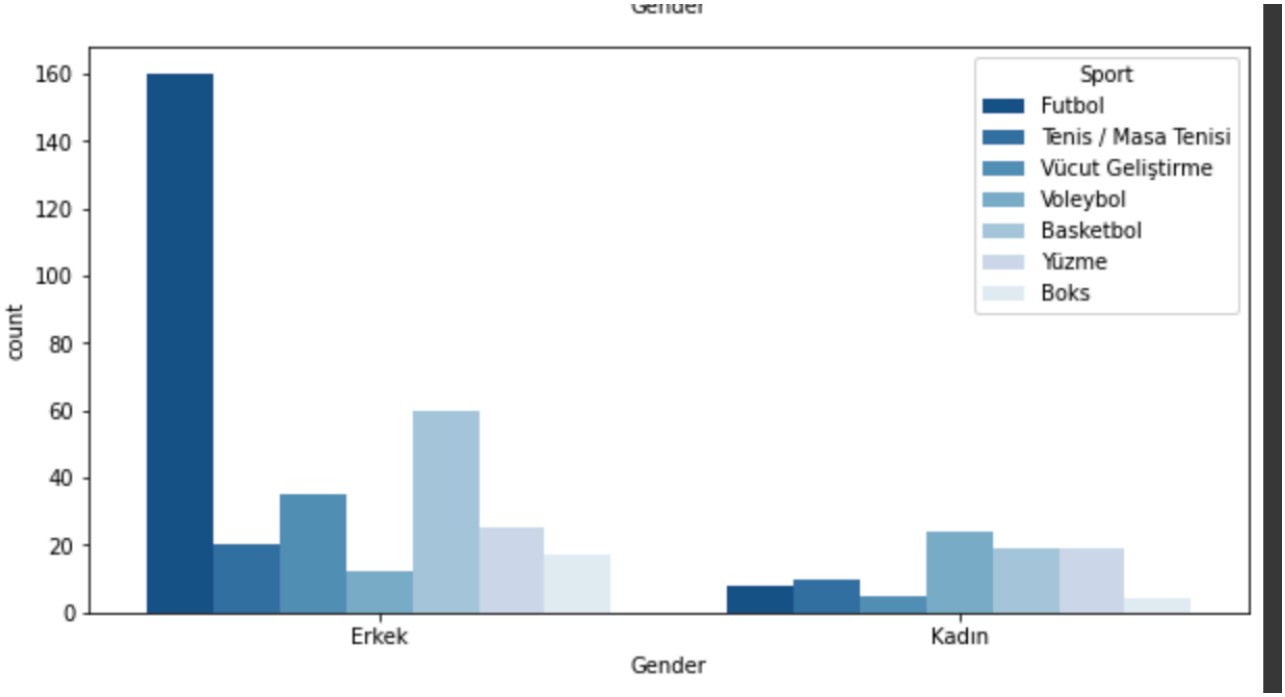
Verilerimizin birkaç özelliğinde baskınlık bulunmasını ulaştırdığımız çevrelerden kaynaklı olduğu sonucuna vardık. Aşağıdaki grafiklerden görüldüğü üzere verilerimiz %78.7 ilk Erkek ve %54.5 oranında 19-24 yaş aralığındaki insanlardan oluşmaktadır.



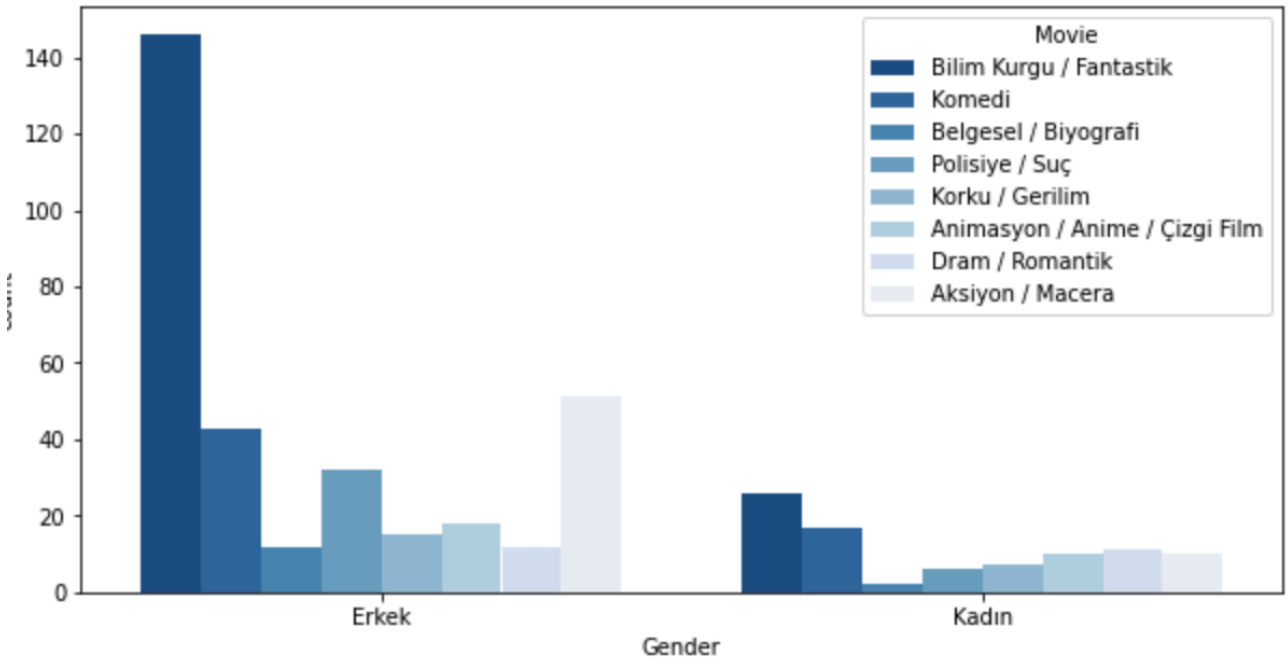
Aşağıda verilen dağılım oranlarına dayanarak Erkeklerde Youtube uygulaması çoğunlukla kullanılmaktayken Kadınlarda ilk sırayı Instagram almıştır. Erkeklerde Reddit, Linkedin ve Facebook kullanımları da göze çarpmaktayken kadınlarda bu uygulamalarda herhangi bir kullanıma rastlanmamıştır. Veri sayıları da dikkate alındığında Kadınlarda görülen Twitter kullanma oranının Erkeklerin oranından daha yüksek olduğu görülmektedir.



Bir diğer dağılım ise spor dallarında oluşmaktadır. Erkeklerde futbol açık ara bir şekilde üstünlük kurmaktayken kadınlarda en yüksek verinin Voleybolda olduğu görülmektedir. Hem erkek hem kadınlarda 2. Sıranın basketbolda olduğu görülmektedir.



Bir diğer inceleme ise film türleri dağılımındadır. Birinciliği her iki cinsiyette de Bilim Kurgu / Fantastik türü almaktadır fakat erkeklerde bu farkın daha bariz olduğu görülmektedir. İkinci sırada ise kadınlar komedi türü seçerken erkekler Aksiyon / Macera türüne yönelmiştir. Kadınlarda bu türün sayısının oldukça az olduğu görülmektedir.



Verisetinin Makine Öğrenmesi Modellerine Hazırlanması

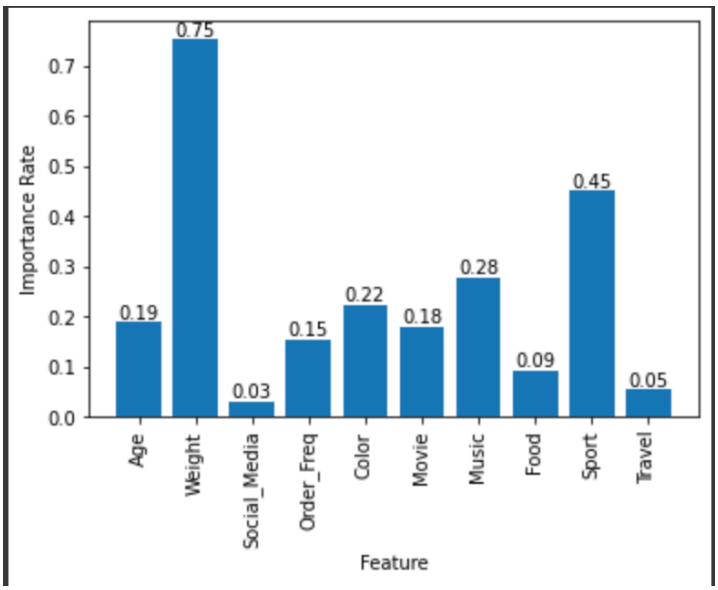
Verisetini çeşitli classification ve clustering işlemlerine sokabilmek için kategorik verileri (nominal) sayı verilerine(ordinal) dönüştürmek gereklidir. Bu aşamada sırasıyla her eşsiz nitelik için bir değer ataması yapılmış ve sonucunda aşağıda görülen verişti oluşmuştur.

	Gender	Age	Weight	Social_Media	Order_Freq	Color	Movie	Music	Food	Sport	Travel
0	Erkek	19 - 24	86 - 100	Youtube	Haftanın Her Günü	Kırmızı	Bilim Kurgu / Fantastik	Rock	Hamburger	Futbol	Hollanda
1	Kadın	19 - 24	58 - 65	Instagram	Haftada 1 Günden Az	Mor	Komedi	Rock	Hamburger	Tenis / Masa Tenisi	İtalya
2	Kadın	19 - 24	51 - 57	Instagram	Haftada 1-3 Gün	Siyah	Bilim Kurgu / Fantastik	Tekno	Pizza	Vücut Geliştirme	Uzakdoğu Ülkeleri
3	Erkek	7 - 18	76 - 85	Instagram	Haftada 1 Günden Az	Siyah	Belgesel / Biyografi	Rock	Dünya Mutfağı	Voleybol	İtalya
4	Erkek	19 - 24	101+	Youtube	Haftanın Her Günü	Kırmızı	Polisiye / Suç	Rock	Pizza	Basketbol	Hollanda

	Gender	Age	Weight	Social_Media	Order_Freq	Color	Movie	Music	Food	Sport	Travel
0	1.0	2.0	7.0	4.0	4.0	3.0	4.0	2.0	1.0	1.0	5.0
1	0.0	2.0	4.0	1.0	1.0	8.0	2.0	2.0	1.0	4.0	3.0
2	0.0	2.0	3.0	1.0	2.0	1.0	4.0	10.0	4.0	7.0	9.0
3	1.0	1.0	6.0	1.0	1.0	1.0	7.0	2.0	9.0	3.0	3.0
4	1.0	2.0	8.0	4.0	4.0	3.0	5.0	2.0	4.0	2.0	5.0

Ağırlıklı Özellik Seçimi

Verisetinde makine öğrenmesi işlemleri gerçekleştirilmeden önce tahmin edilecek bir seçenek bulunması gereklidir. Biz bu noktada kullanıcının formda doldurmuş olduğu değerlerden yola çıkarak cinsiyetini tahmin etmeye çalıştık. Tahmin edeceğiniz üzere bazı özelliklerin bu tahminde daha ön planda olması beklenebilir. Örneğin gezmek istediği ülkenin Almanya olması cinsiyet açısından kritik bir öneme sahip değildir fakat sporun futbol olması cinsiyet açısından erkek cinsiyetine yaklaşması öngörülebilir.



Temel Bileşen Analizi (PCA)

Temel bileşen analizi adımında yeni kaç tane özellik kullanılması gerektiğine karar verebilmek için her bir özellik sayısı için verisetinin ne kadarının temsil edildiğini gördük ve %76'lık oranda karar kılıp özellik sayısını 5 özellikle temsil etmeye karar verdik.

```
from sklearn.decomposition import PCA
# step of deciding on the number of components
pca_temp = PCA()
principalComponents = pca_temp.fit_transform(df_features)
print(np.cumsum((pca_temp.explained_variance_ratio_)))
```

```
[0.20133288 0.3705898  0.52235528 0.65871483 0.76001191 0.84520167
 0.91185489 0.96177935 0.98267407 1.        ]
```

K - Fold

Veriseti k adet kümeye bölünür, her iterasyonda kümelerden biri test kümesi olarak seçilir. K-1 adet küme üzerinden model eğitilir, seçilen küme üzerinde model test edilir ve sonucunda bir doğruluk değeri hesaplanır. İterasyonun her çevriminde seçilen küme değişir. Kümelerin doğruluk değerleri bir diziye alınır. İterasyon bittiğinde bu değerlerin ortalaması doğruluk değeri olarak kabul edilir.

K - Fold yaklaşımı modelin yüksek/düşük performansının rastgele olup olmadığını ölçmemizi sağlar.

```
k_fold=KFold(n_splits=10, shuffle=False, random_state=None)
clf = GaussianNB()
cross_val = cross_val_score(clf, df_features, df_target, cv=k_fold, n_jobs=1)
print(cross_val)
```

```
[0.80952381 0.88095238 0.83333333 0.85714286 0.85714286 0.83333333
 0.83333333 0.83333333 0.87804878 0.80487805]
```

```
print(f'Mean K-Fold:{np.mean(cross_val)}')
```

```
Mean K-Fold:0.842102206736353
```

Sınıflandırma işlemlerde kullanılan verisetleri orijinal verisetinden türetilen 4 farklı yöntemle hesaplanmıştır:

- Manuel olarak %90 train - %10 test olarak bölünen
- Sadece 4 ağırlıklı özelliğin alındığı
- PCA uygulanarak 5 komponentli, verinin %76 sini temsil eden
- K-Fold ile 10 katlı çapraz geçerleme uygulanıp ortalaması hesaplanarak

Toplamda 5 farklı classification, 3 farklı clustering modeli uygulanmıştır.

Classification Yöntemleri

Naive Bayes: Naive Bayes sınıflandırma yöntemleri, olasılık teorisinden Bayes Teoremi'nden türetilmiştir. Basitçe söylemek gerekirse, Naive Bayes, özellik kümelerine bağlı olarak olasılık fonksiyonlarını kullanarak nesneleri sınıflandırır. Naive Bayes yaklaşımı oldukça ölçeklenebilirdir. Yani bağımsız özelliklere sahip belirli miktarda test değerine ihtiyaç duyar.

K Nearest Neighbour: K en yakın komşu algoritması, "k" en yakın komşuyu bulmak için farklı uzaklık ölçümleri kullanır. Projede kullanılan mesafe ölçüsü euler mesafesidir. Model, en yakın komşuları bulduktan sonra, komşular arasında en sık kullanılan etiketin etiketini atar.

Random Forest Classification: Rastgele orman sınıflandırıcıları, çok sayıda karar ağacından oluşur. Ormandaki her ağaç bir sınıf tahmini üretir. Tahmin aşamasından sonra en sık yapılan tahmin etiket olarak atanır. Rastgele orman ağaçları, fazla takmaya karşı daha dayanıklıdır, ancak bu, doğruluklarına bir maliyet olarak gelir.

Support Vector Machines: SVM'ler, sınıflandırma ve regresyon analizi için kullanılan denetimli öğrenme modelleridir. Destek vektör makinesi, veri noktalarını belirgin bir şekilde sınıflandıran N boyutlu bir uzayda bir hiperdüzlem bulur.

Decision Tree Classifier: Karar Ağacı sınıflandırıcıları, insanlara benzer kararlar vermek için bir dizi sorgu kullanır. Her karar verildiğinde, özellik alanı bölgelere ayrılır. Mükemmel bir düğüm, bir etiketi diğerlerinden ayırmalıdır. Ağaçtan aşağı inerken olası sınıfların miktarı azalır. Ağacın son yaprağı sınıflara göre saf değilse, nesneye en yaygın sınıf atanır.

Clustering Yöntemleri

K-Means Clustering: KNN K-Means algoritmasına benzer şekilde, mesafeleri hesaplayarak çalışır. K-araçları, veri alanını k kümeye ayırır. Bunu yapmak için algoritma, birbirine daha yakın araçlara sahip kümeleme sınıfları önerir..

Hierarchical Clustering: Hiyerarşik kümeleme, her bir gözlemi ayrı bir küme olarak ele alarak başlar. Bundan sonra en benzer küme çiftini birleştirir. Bu iterasyon, tüm kümeler birleştirilene kadar yapılır. Birleşmelerden sonra hiyerarşik sıra, birleşme sırası ile gösterilir. Birincisi zincirde daha düşük ve sonuncusu daha yüksek. Benzerlik, öklid mesafesi ile çıkarılan mesafe ile ölçülür. Bu durumda kullandığımız hiyerarşik kümeleme, tam bağlantılı öklid yakınlığıdır. Hangi iki kümenin tüm gözlemleri arasındaki maksimum mesafeleri kullanır..

Agglomerative Clustering: Aglomeratif kümeleme, bağlantı farkıyla Hiyerarşik kümelemenin bir alt kümesidir. Bu bağlantı farkı birleştirilen kümelerin varyansını en aza indirir. Agglomerative aşağıdan yukarıya kümeleme yaparken Hierarchical kümelemede yukarıdan aşağı bir yöntem söz konusudur.

Classification Sonuçları

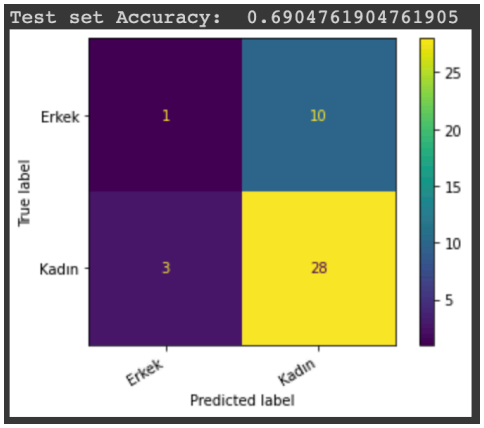
* Sol yukarı görsel manuel olarak %90-%10 olarak bölünerek elde edilen verisetinin sonuçlarını

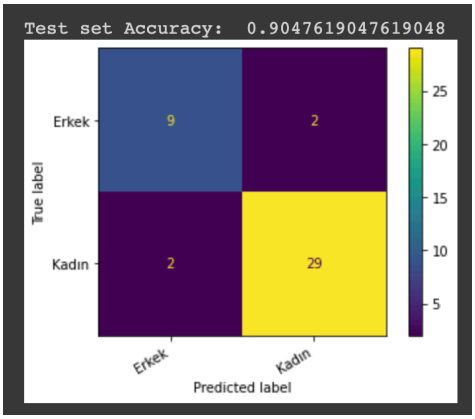
* Sağ üst görsel feature importance aşamasında belirlenen ağırlığı en yüksek 4 özellik baz alınarak elde edilen verisetinin sonuçlarını - ***Color, Weight, Music, Sport -***,

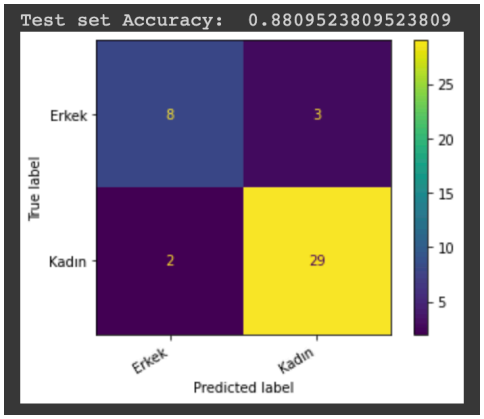
* Sol aşağı görsel ise PCA sonucu elde edilen verisetinin sonuçlarını,

* Sağ aşağı ekran görüntüsü ise 10 katlı çapraz geçerleme sonucu elde edilen değerlerin ortalamasını ifade etmektedir.

Naive Bayes Classification

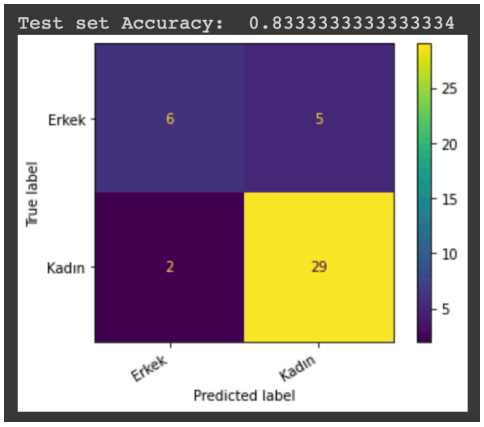


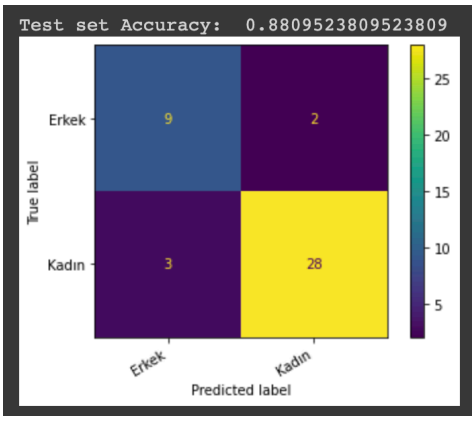


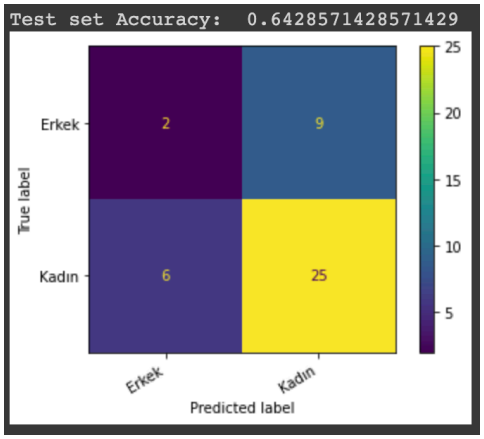



```
Mean K-Fold:0.842102206736353
```

K Nearest Neighbour Classification

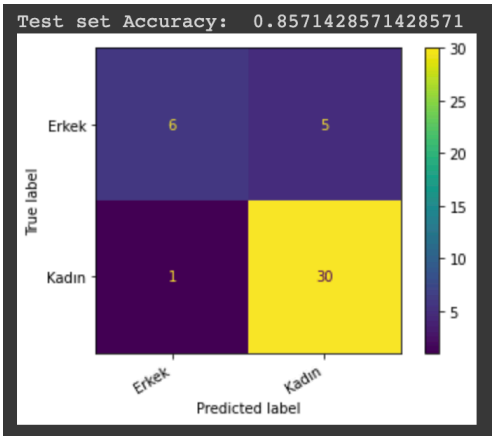


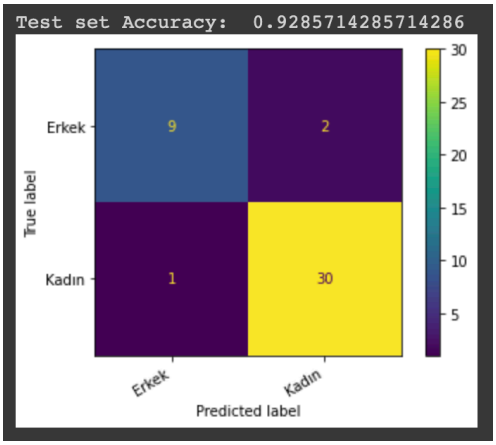


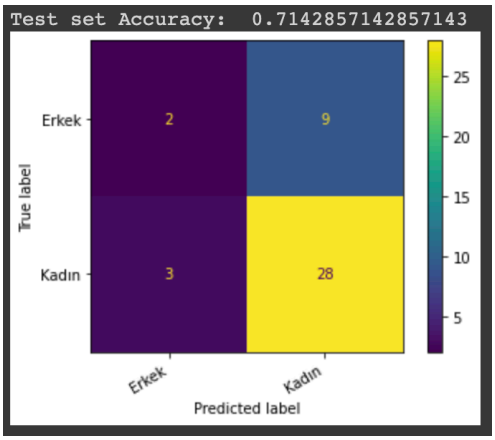



```
Mean K-Fold:0.798896631823461
```

Random Forest Classification

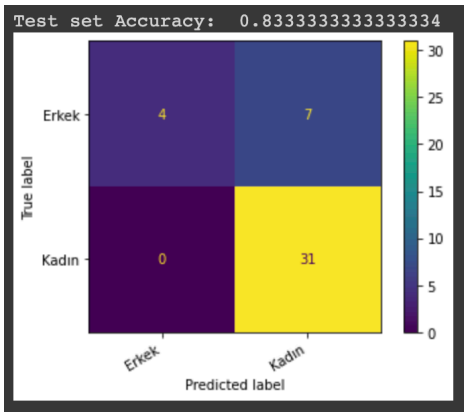


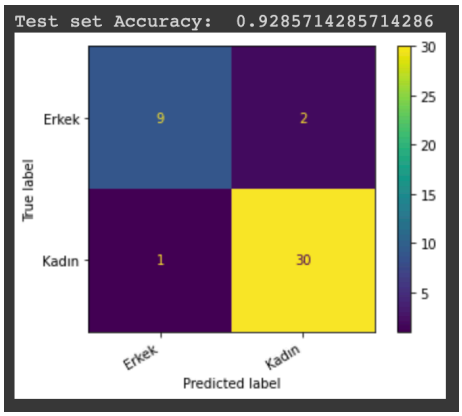


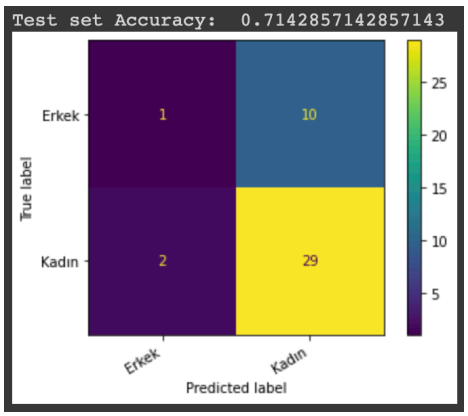



```
Mean K-Fold:0.8659698025551684
```

Support Vector Machine Classification

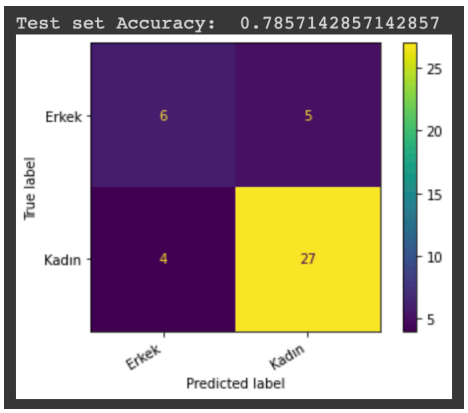


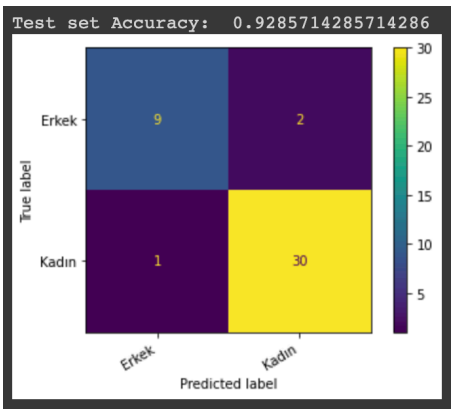


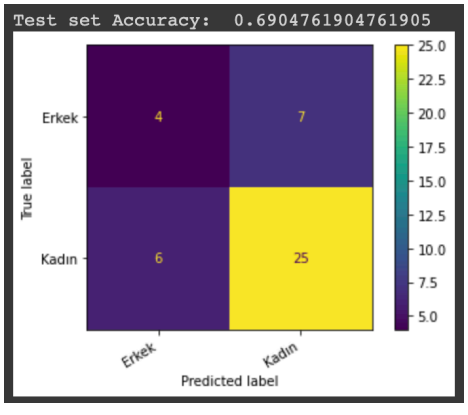



```
Mean K-Fold:0.83739837398374
```

Decision Tree Classification







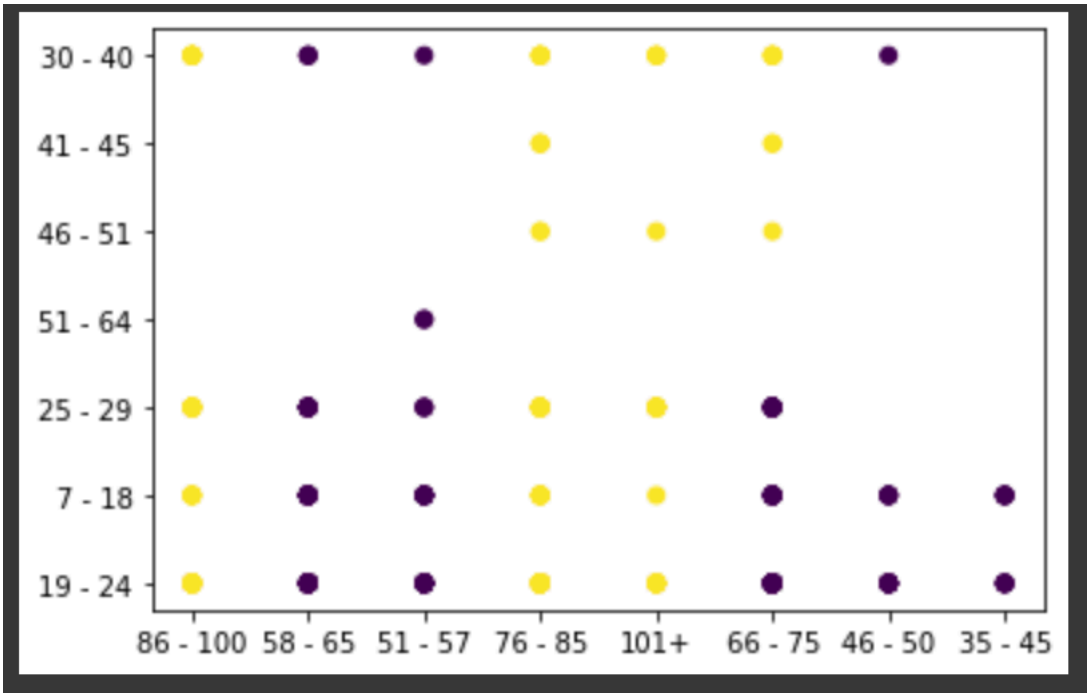

```
Mean K-Fold:0.8060975609756097
```

Aşağıdaki tabloda görüldüğü üzere doğruluk oranı en yüksek sınıflandırma algoritması 0.88 ile **Naive Bayes**, ağırlıklı özellikler göz önüne alındığında ise **Random Forest, SVM ve Decision Tree** değerleri 0.93 ile ilk sırayı paylaşmaktadır, PCA verişti göz önüne alındığında ise **SVM ve Random Forest** en yüksek doğruluk oranlarını vermektedir. K-Fold ortalamasında ise **Random Forest** algoritması en başarılı sonucu vermektedir.

Method	Normal Acc	Important Feature Acc	PCA	K-Fold
Naive Bayes	0.88	0.90	0.69	0.84
K Nearest Neighbour	0.83	0.88	0.64	0.80
Random Forest	0.86	0.93	0.71	0.87
Support Vector Machine	0.83	0.93	0.71	0.84
Decision Tree	0.79	0.93	0.69	0.81

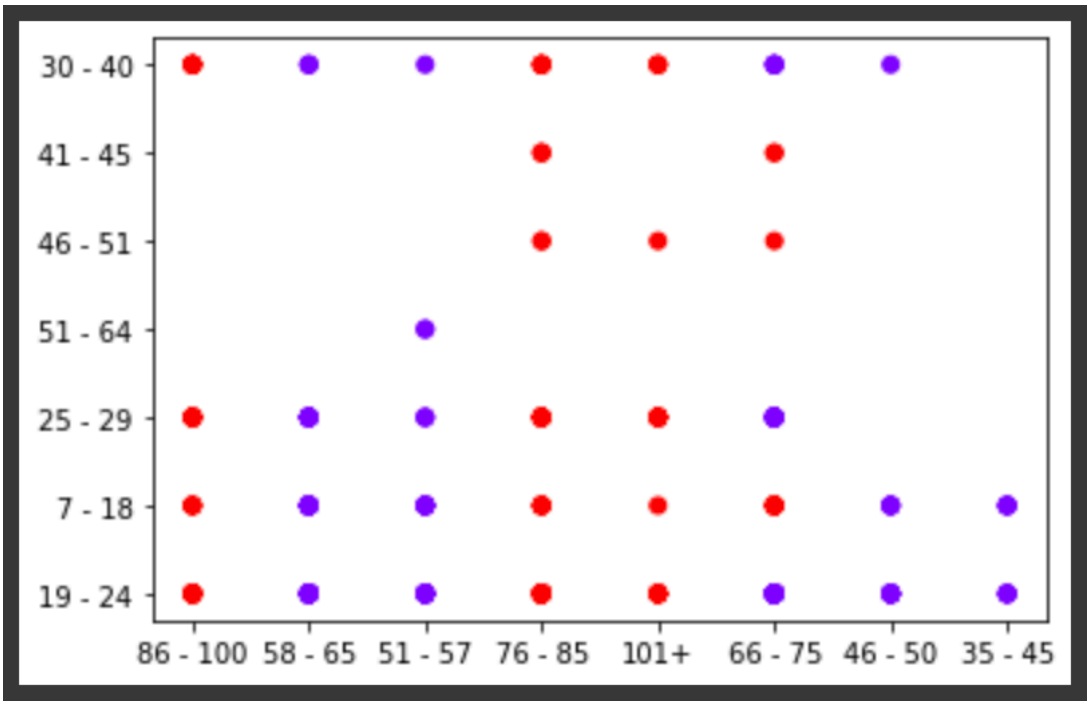
Clustering Sonuçları

K - Means Clustering



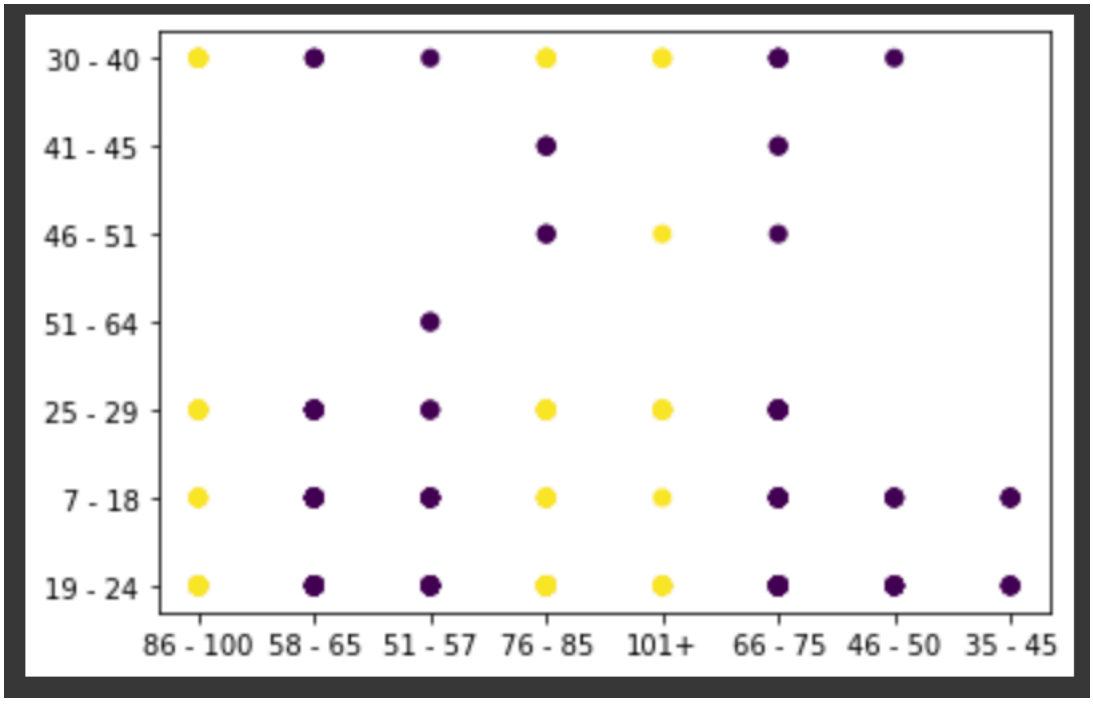

```
0.5190412263490642
```

Hierarchical Clustering




```
0.5263502116966714
```

Agglomerative Clustering




```
0.5128567002857044
```

Aşağıdaki tabloda görüldüğü üzere her 3 clustering işleminin de doğruluk oranı birbirine çok yakındır fakat yine de en yüksek değerin **Hierarchical Clustering** olduğu görülmektedir.

Method	Accuracy
K - Means	0.519
Hierarchical	0.526
Agglomerative	0.513