

# **YILDIZ TEKNİK ÜNİVERSİTESİ**

## **YAPAY ZEKAYA GİRİŞ 2. ÖDEV RAPORU**



Ahmet Enis ŞİMŞİR 19011077

Kutay ALPTEKİN 20011615

# VERİ KÜMESİ

## Veri Kümesi Nasıl Elde Edildi?

Veri kümesi kendi oluşturduğumuz bir anket yardımıyla elde edilmiştir. Toplam 231 kişi tarafından doldurulmuştur. Sonrasında bu yanıtlar csv formatlı bir dosya biçimine getirilmiştir. Dosyada her satır bir kişinin cevabını temsil eder ve cevaplar virgüllerle ayrılmıştır.

Ankette 10 adet soru vardır. Ekstradan 11. soru ise direkt olarak kişinin hangi sınıftan olduğunu sormaktadır.

## Anket Neye Göre Sınıflandırma Yapıyor?

Anket kişinin introvert (içe dönük) veya extrovert (dışa dönük) olması üzerinedir. İlk 10 soruda istenilen bilgiler direkt olarak kişinin introvert veya extrovert olduğu bilgisine ulaşılmasını sağlamazlar ama introvert veya extrovert bireylerin stereotipik olarak daha yatkın bulunduğu durumlar içermektedirler. İki introvert bir soruda farklı şıkkı işaretleyebilir veya bir extrovert veya bir introvert aynı şıkta kesin karar kılabilir. Lakin topluca bakıldığında bir kümeleme yapmak mümkündür.

## Anket Soruları

Aşağıda anket soruları verilmiştir.

Geniş bir sınıfa veya odaya girdiğimde \*

- ☐ ortalarda oturmayı tercih ederim
- ☐ duvara yakın oturmayı tercih ederim

Grup projeleri konusunda \*

- ☐ Projeyi geniş bir grupla yapmayı tercih ederim
- ☐ Projeyi ufak bir grupla yapmayı tercih ederim
- ☐ Projeyi tek başıma yapmayı tercih ederim

Yolda yürürken karşı kaldırımda bir sınıf arkadaşınızı gördünüz ve sizi fark etmedi: \*

- ☐ Yanına gidip selam veririm
- ☐ Beni görmediyse selam vermekle uğraşmam

Diyelim ki doktor randevusuna gittiniz ve yarım saattir bekleme odasında tutuluyorsunuz \*

- ☐ Gidip açıkça şikayet ederim. Oturup sessiz kalmam
- ☐ Bekleyen diğer insanlara doktorun hep geç kaldığını anlatırım
- ☐ Sessizce doktorun çağırmasını beklerim

Ders dinlerken cana yakın olduğunu bildiğiniz bir hocanızın hata yaptığını fark ettiniz: \*

- ☐ El kaldırıp uyarırım
- ☐ Sessiz kalırım

Başınıza kötü bir olay geldi ve çevrenizdeki insanların yardım etmek için yapabilecekleri bir şey yok \*

- ☐ Yine de konuyu onlara açıp en azından beni dinlemelerini isterim
- ☐ Konuyu yardımcı olmayacak kimseye anlatmam

İnsanlarla tanışmayı sever misiniz? \*

- ☐ Modumdaysam evet
- ☐ Hayır, mümkün olduğunca az insan tanımak isterim

Evden çıkacaksınız ama kapının dışından gelen seslerle komşunuzun da evden çıkmakta olduğunu fark ettiniz: \*

- ☐ Kapıyı açarım ve komşumla selamlaşırım
- ☐ Önce komşumun gitmesini beklerim

İnsan içinde olmak \*

- ☐ Beni iyi hissettiriyor
- ☐ Yalnız olmak daha iyi geliyor

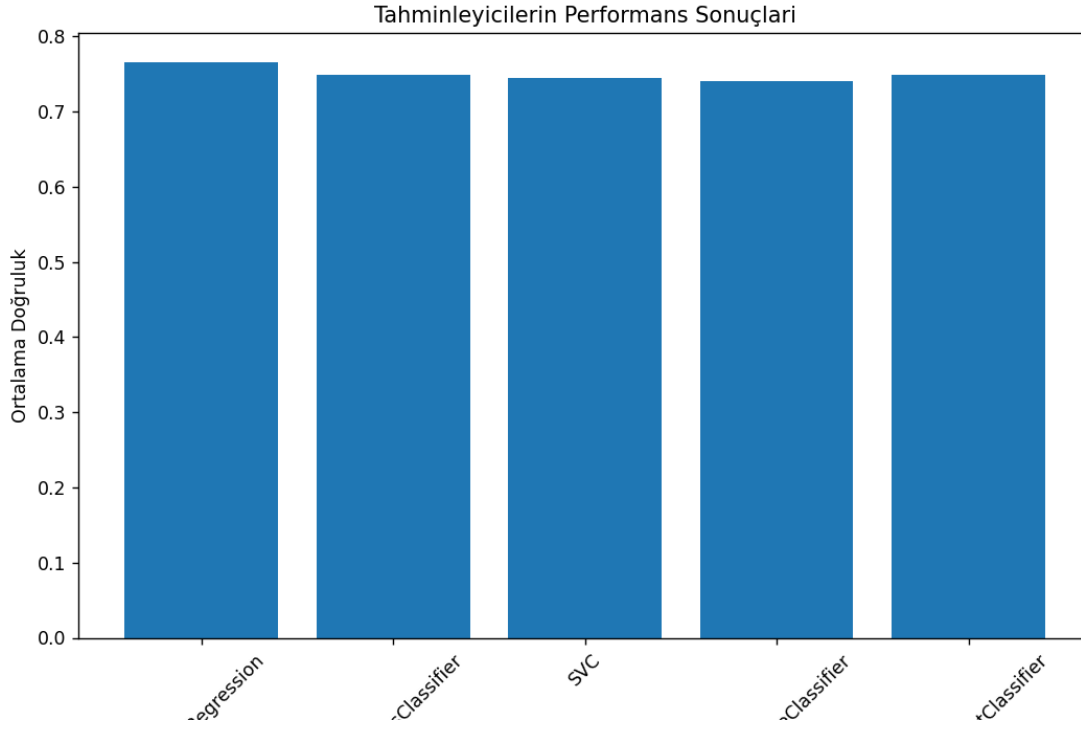
Yeni tanıştığınız birisi sizi nasıl tanımlar? \*

- ☐ Arkadaş canlısı ve konuşkan
- ☐ Sessiz ve sakın

Son olarak: Kendinizi nasıl tanımlarsınız? \*

- ☐ Introvert (İçe Dönük)
- ☐ Extrovert (Dışa Dönük)

## BULGULAR



Grafikte sırasıyla LogisticRegression, KNeighborsClassifier, SVC, DecisionTreeClassifier ve RandomForestClassifier tahminleyicilerinin ortalama doğruluğu verilmiştir. Yakın görünseler de aralarında farklar bulunmaktadır.

Daha detaylı sonuçlar:

```
Tahminleyici: LogisticRegression
Ortalama Doğruluk: 0.7664855072463768
Tüm Sonuçlar: [0.79166667 0.75 0.83333333 0.75 0.79166667 0.79166667
0.86956522 0.73913043 0.52173913 0.82608696]
Tahminleyici: KNeighborsClassifier
Ortalama Doğruluk: 0.7492753623188405
Tüm Sonuçlar: [0.70833333 0.75 0.83333333 0.70833333 0.79166667 0.875
0.73913043 0.69565217 0.56521739 0.82608696]
Tahminleyici: SVC
Ortalama Doğruluk: 0.7452898550724638
Tüm Sonuçlar: [0.75 0.75 0.875 0.70833333 0.75 0.75
0.82608696 0.60869565 0.60869565 0.82608696]
Tahminleyici: DecisionTreeClassifier
Ortalama Doğruluk: 0.7409420289855072
Tüm Sonuçlar: [0.70833333 0.70833333 0.875 0.70833333 0.75 0.83333333
0.82608696 0.56521739 0.60869565 0.82608696]
Tahminleyici: RandomForestClassifier
Ortalama Doğruluk: 0.7494565217391305
Tüm Sonuçlar: [0.70833333 0.75 0.875 0.70833333 0.75 0.83333333
0.82608696 0.60869565 0.60869565 0.82608696]
```

## Karşılaştırmalar:

```
Karsilastirma: LogisticRegression vs. KNeighborsClassifier
T istatistigi: 0.8926103019845693
P degeri: 0.39530525673365813
Istatistiksel olarak anlamlı fark yok.
```

```
-----
Karsilastirma: LogisticRegression vs. SVC
T istatistigi: 1.1485009359359806
P degeri: 0.28037026120271064
Istatistiksel olarak anlamlı fark yok.
```

```
-----
Karsilastirma: LogisticRegression vs. DecisionTreeClassifier
T istatistigi: 1.0969856514652356
P degeri: 0.3011298093659277
Istatistiksel olarak anlamlı fark yok.
```

```
-----
Karsilastirma: LogisticRegression vs. RandomForestClassifier
T istatistigi: 0.8347580498069052
P degeri: 0.42545666242406
Istatistiksel olarak anlamlı fark yok.
```

```
-----
Karsilastirma: KNeighborsClassifier vs. SVC
T istatistigi: 0.1947623056632602
P degeri: 0.8499046029704805
Istatistiksel olarak anlamlı fark yok.
```

```
-----
Karsilastirma: KNeighborsClassifier vs. DecisionTreeClassifier
T istatistigi: 0.43708545856153885
P degeri: 0.6723490264283374
Istatistiksel olarak anlamlı fark yok.
```

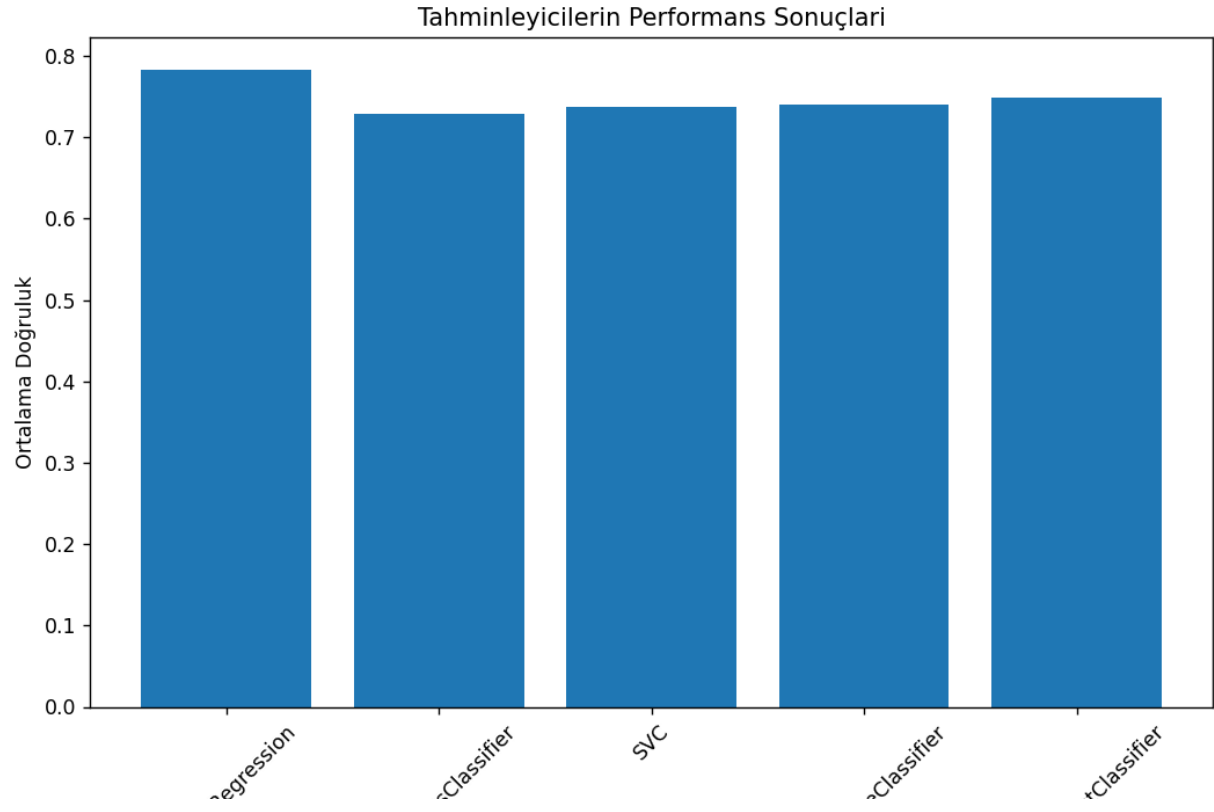
```
-----
Karsilastirma: KNeighborsClassifier vs. RandomForestClassifier
T istatistigi: -0.011529170315115132
P degeri: 0.9910527791314974
Istatistiksel olarak anlamlı fark yok.
```

```
-----
Karsilastirma: SVC vs. DecisionTreeClassifier
T istatistigi: 0.37469519624322317
P degeri: 0.7165638081576089
Istatistiksel olarak anlamlı fark yok.
```

```
-----
Karsilastirma: DecisionTreeClassifier vs. RandomForestClassifier
T istatistigi: -1.4995757798495437
P degeri: 0.16795856684045526
Istatistiksel olarak anlamlı fark yok.
```

Kodda istatistiksel fark oranı 0,05 olarak seçilmiştir. Yani P değerlerinin farkı oranı 0,05'ten küçük olduğu için "istatistiksel olarak anlamlı fark yok" çıktısı elde edilmiştir. Lakin detaylıca incelediğimizde en iyi sonucun LogisticRegression tahminleyicisine ait olduğunu görürüz.

Normalizasyon Kaldırınca:





Karsilařtırma: LogisticRegression vs. KNeighborsClassifier  
T istatistięi: 1.962823956036113  
P deęeri: 0.08127462854514783  
İstatistiksel olarak anlamlı fark yok.

-----  
Karsilařtırma: LogisticRegression vs. SVC  
T istatistięi: 2.263575177831789  
P deęeri: 0.049884246494669146  
İstatistiksel olarak anlamlı fark var.

-----  
Karsilařtırma: LogisticRegression vs. DecisionTreeClassifier  
T istatistięi: 2.1029744201087754  
P deęeri: 0.06480469361938995  
İstatistiksel olarak anlamlı fark yok.

-----  
Karsilařtırma: LogisticRegression vs. RandomForestClassifier  
T istatistięi: 1.8572050505191746  
P deęeri: 0.09624083085136895  
İstatistiksel olarak anlamlı fark yok.

-----  
Karsilařtırma: KNeighborsClassifier vs. SVC  
T istatistięi: 0.007941372602930517  
P deęeri: 0.993837012387506  
İstatistiksel olarak anlamlı fark yok.

-----  
Karsilařtırma: KNeighborsClassifier vs. DecisionTreeClassifier  
T istatistięi: -0.1659905999243111  
P deęeri: 0.8718333924579009  
İstatistiksel olarak anlamlı fark yok.

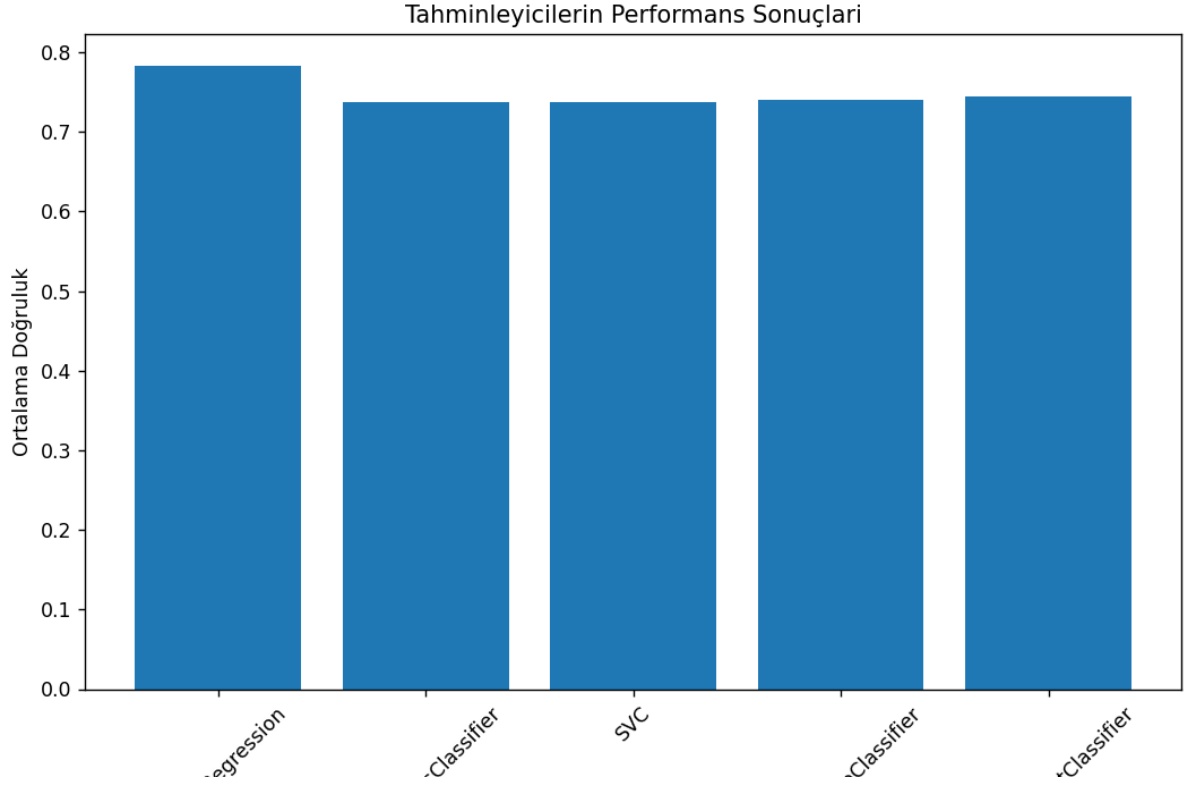
-----  
Karsilařtırma: KNeighborsClassifier vs. RandomForestClassifier  
T istatistięi: -0.3471079355909155  
P deęeri: 0.7364887051635435  
İstatistiksel olarak anlamlı fark yok.

-----  
Karsilařtırma: SVC vs. DecisionTreeClassifier  
T istatistięi: -0.5570860145311576  
P deęeri: 0.5910512317836034  
İstatistiksel olarak anlamlı fark yok.

-----  
Karsilařtırma: DecisionTreeClassifier vs. RandomForestClassifier  
T istatistięi: -1.0000000000000002  
P deęeri: 0.34343639613791355  
İstatistiksel olarak anlamlı fark yok.

-----  
Tahminleyici Ortalama Doğruluk  
0 LogisticRegression 0.783696  
1 KNeighborsClassifier 0.736957  
2 SVC 0.736775  
3 DecisionTreeClassifier 0.740942  
4 RandomForestClassifier 0.745109

## Özellik Dönüşümü (PCA) Kaldırınca:



Karsilařtırma: LogisticRegression vs. KNeighborsClassifier

T istatistięi: 3.0185528512476214

P deęeri: 0.014513302118328705

İstatistiksel olarak anlamlı fark var.

Karsilařtırma: LogisticRegression vs. SVC

T istatistięi: 2.263575177831789

P deęeri: 0.049884246494669146

İstatistiksel olarak anlamlı fark var.

Karsilařtırma: LogisticRegression vs. DecisionTreeClassifier

T istatistięi: 2.1029744201087754

P deęeri: 0.06480469361938995

İstatistiksel olarak anlamlı fark yok.

Karsilařtırma: LogisticRegression vs. RandomForestClassifier

T istatistięi: 1.914524723929527

P deęeri: 0.08782319829787728

İstatistiksel olarak anlamlı fark yok.

Karsilařtırma: KNeighborsClassifier vs. SVC

T istatistięi: -0.40169268369379724

P deęeri: 0.6972816858584727

İstatistiksel olarak anlamlı fark yok.

Karsilařtırma: KNeighborsClassifier vs. DecisionTreeClassifier

T istatistięi: -0.6941425087743892

P deęeri: 0.5051139019065413

İstatistiksel olarak anlamlı fark yok.

Karsilařtırma: KNeighborsClassifier vs. RandomForestClassifier

T istatistięi: -1.3321312205961306

P deęeri: 0.21556334365242094

İstatistiksel olarak anlamlı fark yok.

Karsilařtırma: SVC vs. DecisionTreeClassifier

T istatistięi: -0.5570860145311576

P deęeri: 0.5910512317836034

İstatistiksel olarak anlamlı fark yok.

Karsilařtırma: SVC vs. RandomForestClassifier

T istatistięi: -1.415176595213591

P deęeri: 0.1906732612416367

İstatistiksel olarak anlamlı fark yok.

Karsilařtırma: DecisionTreeClassifier vs. RandomForestClassifier

T istatistięi: -1.4995757798495437

P deęeri: 0.16795856684045526

İstatistiksel olarak anlamlı fark yok.

	Tahminleyici	Ortalama Doğruluk
0	LogisticRegression	0.783696
1	KNeighborsClassifier	0.728442
2	SVC	0.736775
3	DecisionTreeClassifier	0.740942
4	RandomForestClassifier	0.749457

Görüldüğü gibi normalizasyon ve özellik dönüşümü kaldırılınca tahminleyicilerde istatistiksel olarak anlamlı farklar oluşmuştur.

Normalleştirme işlemi daha tutarlı ve karşılaştırılabilir bir ölçekte veri seti oluşturmamıza sebep olduğu için onu kaldırmak verimi düşürmüştür. Aynı zamanda PCA işlemi de orijinal veri setinin boyutunu azaltıp en verimli ve değişkenliği en iyi açıklayan ana bileşenleri elimizde tutmamızı sağladığı için onu kaldırmak da kötü bir etkiye sahip olmuştur.