YILDIZ TECHNICAL UNIVERSITY

FACULTY OF ELECTRICAL AND ELECTRONICS

Yapay Zeka

BLM 4510

Ödev 2 Raporu

19011501– Doğa GÜNDOĞAR

19011502 – Gülsüm İrem BAŞ

doga.gundogar@std.yildiz.edu.tr

gulsum.bas@std.yildiz.edu.tr

DEPARTMENT OF COMPUTER ENGINEERING

Giriş:

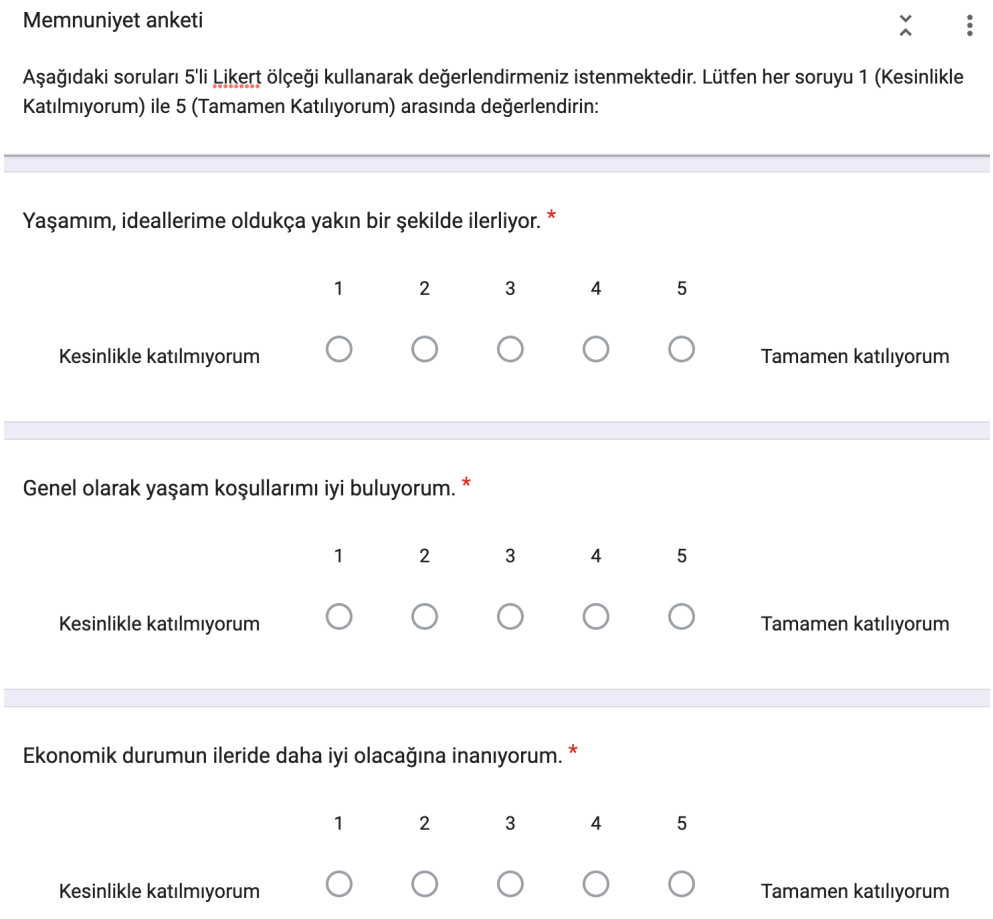
Bu çalışma, çeşitli faktörlerin insanların yaşam memnuniyeti üzerindeki etkisini anlamaya yönelik bir veri analizi örneğidir. Çalışmada kullanılan veri seti, katılımcılardan aldığımız bir dizi anket yanıtını içermektedir. Ankette, katılımcılara ekonomi, istihdam, eğitim, sosyal çevre, gelir ve genel beklentileri hakkında çeşitli sorular sorulmuştur. Ayrıca, yaşam memnuniyeti ölçeği ile genel yaşam memnuniyetleri de değerlendirilmiştir.

Bu veri analizi çalışmasının amacı, yaşam memnuniyetine en çok etki eden faktörleri belirlemektir. Aynı zamanda verileri makine öğrenmesi algoritmalarıyla eğiterek, tahmin etme başarıları ölçülmüştür. Ayrıca, bu faktörler arasındaki ilişkiler de anlamaya çalışılarak analiz boyunca, veri seti temizlenerek gerekli hesaplamaları yapılarak sonuçlar görselleştirilmiştir.

Veri Toplama:

Veri toplama süreci, projenin başarısı için hayati önem taşıyan bir aşamadır. Bu özel durumda, kullanılan veri seti, genel yaşam memnuniyetini ve çeşitli faktörlerin yaşam memnuniyeti üzerindeki etkisini değerlendiren bir dizi anket yanıtından toplandı. Anket ile, katılımcılara ekonomi, istihdam, eğitim, sosyal çevre, gelir ve genel beklentileri hakkında çeşitli sorular sorulmuştur. Bu şekilde geniş yelpazedeki bireylerden yanıtlar toplanarak çeşitlilik elde edilmiştir. Bu çeşitlilik, veri setinin genel popülasyonu temsil etme kabiliyetini artırır. Ankette sorulan sorular genel olarak 5'li Likert ölçeği kullanılarak yanıt verilmesi istenmiştir. Aşağıda soruların bir kaçına dair örnek bulunmaktadır, aynı zamanda anketten elde edilen verisetinin de bir örneği verilmiştir.

1- Anket örneği:



2- Veri seti örneği:

	Yaş	Cinsiyet	Bölüm	Not_Ortalaması	Aile_Geliri	İlişki_Durumu	Gelecek_Beklentileri	Çalışma_Durumu	Etkinliklere_Katılım	İdeal	...
0	18-24	Kadın	Bilgisayar Mühendisliği	3.00 - 3.50	20000 - 50000	İlişkisi var	8	Yarı zamanlı	4	3	...
1	18-24	Erkek	Kontrol ve Otomasyon Mühendisliği	3.00 - 3.50	10000 - 20000	İlişkisi var	6	Stajyer	4	1	...
2	18-24	Erkek	Bilgisayar Mühendisliği	3.00 - 3.50	5000 - 10000	İlişkisi var	7	Çalışmıyor	1	1	...
3	18-24	Kadın	Bilgisayar Mühendisliği	3.00 - 3.50	20000 - 50000	İlişkisi var	4	Çalışmıyor	5	3	...
4	18-24	Kadın	Endüstri Mühendisliği	3.50 - 4.00	5000 - 10000	İlişkisi yok	3	Çalışmıyor	4	2	...

5 rows × 30 columns

...	Gelecek_Yaşam	Gelecekte_Hedefleri	Gelecek_Mutluluk	Hedefler	İyi_İş	İdeal_Yaşam_Yakınlık	İyi_Koşullar	Memnuniyet	İstekler	Değişim
...	4	4	4	4	4	4	4	4	4	2
...	4	3	3	3	3	4	3	4	3	2
...	3	4	4	4	4	3	5	5	3	2
...	4	2	3	2	2	2	3	3	4	1
...	3	3	3	3	3	2	3	3	2	2

Ankette sorulan soruların bir örneği ve nasıl gruplandırdığımıza dair bir resim aşağıda verilmiştir.

Ekonomik_Durum_G	Ekonomik durumun ileride daha iyi olacağına inanıyorum.	Ekonomi
İş_G	Mezun olduğumda hızlı bir şekilde iş bulabileceğime inanıyorum.	İstihdam
İş_Bulma	Bölümümden mezun olanların iş bulmada zorlandığını görüyorum.	İstihdam
Eğitim	Aldığım eğitimin kaliteli olduğunu düşünüyorum.	Eğitim
İş_Başarı	Aldığım eğitim sayesinde iş hayatında başarılı olabileceğimi düşünüyorum.	Eğitim
Yaşanılan_Yer	Yaşadığım yerden memnuniyet duyuyorum.	Sosyal Çevre
Arkadaşlar	Arkadaşlarımın bana iyi davrandığını düşünüyorum.	Sosyal Çevre
Aile	Ailemin bana iyi davrandığını düşünüyorum.	Sosyal Çevre
Harcama_Kısıtlaması	Harcamalarımı kısıtlamak zorunda olduğumu hissediyorum.	Gelir
Gelecek_Yaşam	Gelecekte rahat ve huzurlu bir yaşamım olacağına inanıyorum.	Beklenti
Gelecekte_Hedefleri	Gelecekte hedeflerime ulaşabileceğimi düşünüyorum.	Beklenti
Gelecek_Mutluluk	Gelecekte mutlu olacağıma inanıyorum.	Beklenti
Hedefler	Hedeflerimin gerçekleşeceğine inanıyorum.	Beklenti
İyi_İş	İyi bir iş bulabileceğime inanıyorum.	Beklenti
İdeal_Yaşam_Yakınlık	Yaşamımın birçok yönüyle ideallerime yakın olduğunu düşünüyorum.	Yaşam_Memnuniyeti_Ölçeğ
İyi_Koşullar	Yaşam koşullarımın iyi olduğunu düşünüyorum.	Yaşam_Memnuniyeti_Ölçeğ
Memnuniyet	Yaşamımdan genel olarak memnuniyet duyuyorum.	Yaşam_Memnuniyeti_Ölçeğ
İstekler	Şu ana kadar istediğim her şeyi başarılı bir şekilde elde ettiğimi düşünüyorum.	Yaşam_Memnuniyeti_Ölçeğ
Değişim	Eğer yeniden dünyaya gelseydim, yaşamımdan hemen hemen hiçbir şeyi değiştirmezdim.	Yaşam_Memnuniyeti_Ölçeğ

Yukarıdaki memnuniyet ölçeği sorularına verilen yanıtlar veri setinde en soldaki sütun isimlerine göre değiştirilmiştir.

Veri İnceleme:

Oluşturulan veri setinde yukarıdaki gibi sütun değişiklikleri yapıldıktan sonra bazı negatif anlamlı özellikler için puanları 5 ‘ ten çıkarma işlemi yapılmıştır. Daha sonrasında sorular anlamlarına göre gruplandırılmıştır. Yapılan gruplandırma aşağıdaki gibidir.

```python
# sorular anlamlarına göre gruplandırıldı
gruplar = {
    'Ekonomi': ['Ekonomik_Durum_G'],
    'İstihdam': ['İş_G', 'İş_Bulma'],
    'Eğitim': ['Eğitim', 'İş_Başarı'],
    'Sosyal Çevre': ['Yaşanılan_Yer', 'Arkadaşlar', 'Aile'],
    'Gelir': ['Harcama_Kısıtlaması'],
    'Beklenti': ['Gelecek_Yaşam', 'Gelecekte_Hedefleri', 'Gelecek_Mutluluk', 'Hedefler', 'İyi_İş'],
    'Yaşam_Memnuniyeti_Ölçeği': ['İdeal_Yaşam_Yakınlık', 'İyi_Koşullar', 'Memnuniyet', 'İstekler', 'Değişim
}
```

Bu gruplar doğrultusunda verilerin istatistiksel değerlerini ve birbirleriyle olan ilişkisi aşağıda incelenmiştir.

	std_dev	mean
Ekonomi	1.044368	3.765363
İstihdam	1.415729	2.888268
Eğitim	1.107213	3.351955
Sosyal Çevre	0.990565	3.988827
Gelir	1.031243	1.122905
Beklenti	0.911721	3.768715
Yaşam_Memnuniyeti_Ölçeği	1.131465	3.410056

Yukarıdaki verileri yorumlayacak olursak,

Ekonomi grubunun ortalama puanı yaklaşık 3.77'dir. Bu, ankete katılanların genel olarak ekonomik durumlarını nötrden biraz daha olumlu bir şekilde değerlendirdiklerini gösterir. Standart sapmanın 1.04 olması, değerlendirmelerin bir miktar dağılım gösterdiğini, yani farklı bireylerin ekonomik durumları hakkında biraz farklı düşüncelere sahip olduğunu gösterir.

İstihdam grubunun ortalama değeri 2.89'dur. Bu değer, katılımcıların iş durumlarını ve iş bulma olasılıklarını nötrden biraz daha düşük bir şekilde değerlendirdiklerini gösterir. Standart sapma 1.42 olarak hesaplanmış, bu da katılımcılar arasında istihdam durumu hakkında geniş bir görüş çeşitliliği olduğunu gösterir.

Eğitim için ortalama değer de 3.35 bulunmuş ve ekonomi gibi nötrden daha olumlu değerlendirilmiştir.

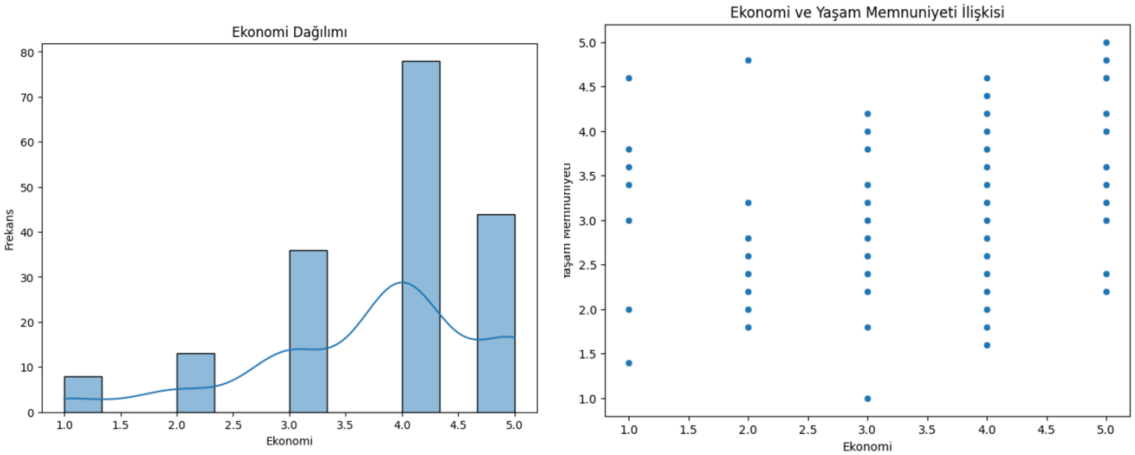
Sosyal çevre grubu, ortalama değeri yaklaşık 3.99 olan tüm gruplar arasında en yüksek puanı alır. Bu, katılımcıların genel olarak sosyal çevrelerinden memnun olduklarını gösterir. Standart sapma 0.99'dur, bu da sosyal çevre hakkındaki görüşlerin daha homojen olduğunu gösterir.

Gelir grubunun ortalama değeri 1.12'dir. Bu, katılımcıların genel olarak gelir durumlarının harcamalarını önemli ölçüde kısıtladığını gösterir. Standart sapma 1.03'tür, bu da gelirle ilgili görüşlerin geniş bir aralıkta olduğunu gösterir.

Yaşam memnuniyeti ölçeği grubu için ortalama değer 3.41'dir. Aynı şekilde beklenti grubu ortalaması da 3.41 hesaplanmıştır. Bu, katılımcıların genel yaşam memnuniyetinin ve beklentilerinin nötrden biraz daha olumlu olduğunu gösterir. Standart sapma 1.13'tür, bu da yaşam memnuniyeti hakkındaki görüşlerin biraz dağıldığını gösterir. Beklenti için görüşler nispeten daha homojen dağılmıştır.

Aşağıda ise verilen her grubun yaşam memnuniyeti ile ilişkisine dair görseller verilmiştir.

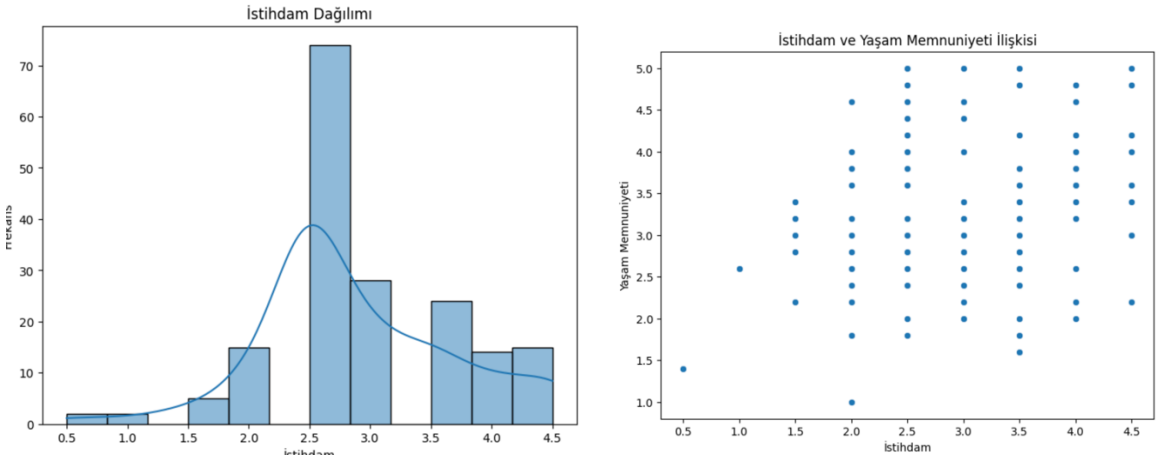
1- Ekonomi:



İlk tabloyu yorumlayacak olursak, verilerin genel olarak ortalama değer olan 3.76 etrafında dağıldığı görülebilir. Farklı değerlerin olasılığı düşüktür.

İkinci tabloda ise Yaşam Memnuniyeti Ölçeği"ne göre scatter plot (nokta dağılım grafiği) çizilmiştir. Scatter plot, iki değişken arasındaki ilişkiyi göstermek için kullanılır. X eksenindeki her bir grup için hesaplanan ortalama değerler ve Y ekseninde "Yaşam Memnuniyeti" ölçeğinin ortalaması bulunmaktadır. Bu grafikler, her bir grup ile yaşam memnuniyeti arasındaki potansiyel ilişkiyi anlamak için kullanılır. Eğer noktalar bir çizgi etrafında toplanıyorsa, bu iki değişken arasında bir ilişki olabileceğini gösterir.

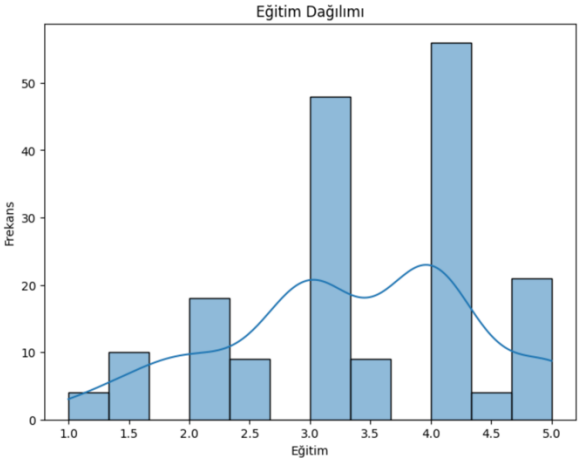
2-

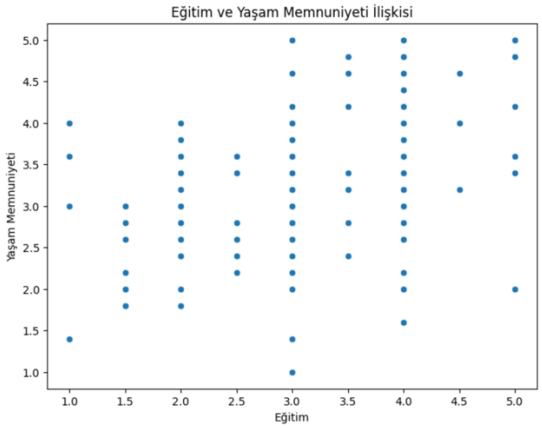


Yukarıdaki tablodan da anlaşılacağı üzere Ekonomi grubuna göre İstihdam için veriler çok daha ortalama değer etrafında toplanmıştır. Verilerin yayılımı baya azdır. Ortalama değer olan 2.88 kısmındaki değerlerin frekansı çok yüksek çıkmıştır.

İstihdam için Yaşam Memnuniyeti noktasında daha yüksek bir korelasyon olduğunu 2. Tablodan görebiliriz. Yüksek istihdam değerleri olan kullanıcılar çoğunlukla Yaşam memnuniyetlerine yüksek vermişlerdir.

3- Eğitim

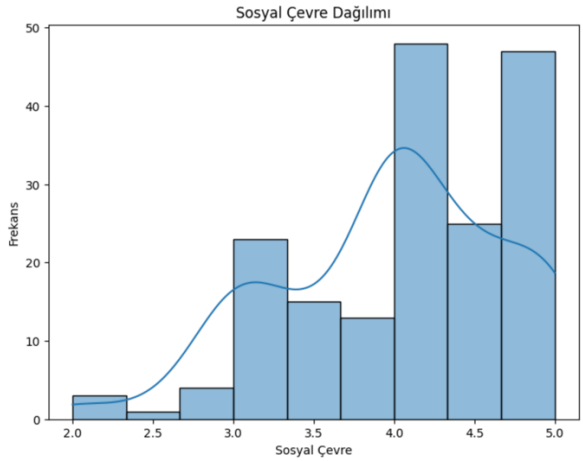


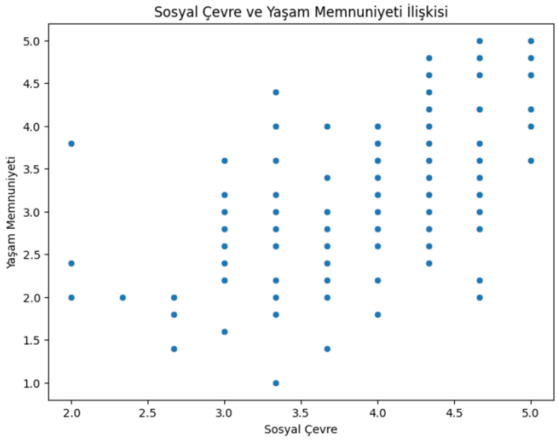


Eğitim grubu için ilk tablo incelendiğinde verilerin iki noktada toplandığı, tablonun iki noktada tepe yapmasından görülebilir. Genel olarak 3 ve 4 değerleri etrafında toplanmış fakat verilerin yayılımı diğer gruplara göre daha fazladır.

İkinci tablo incelendiğinde ise verilerde çok fazla korelasyon olmadığı gözlemlenebilir. Eğitim verilerini düşük işaretleyen kullanıcılardan da yüksek yaşam memnuniyeti olduğu gözlenebilir. Fakat tamamiyle korelasyon yoktur denilemez.

4- Sosyal Çevre

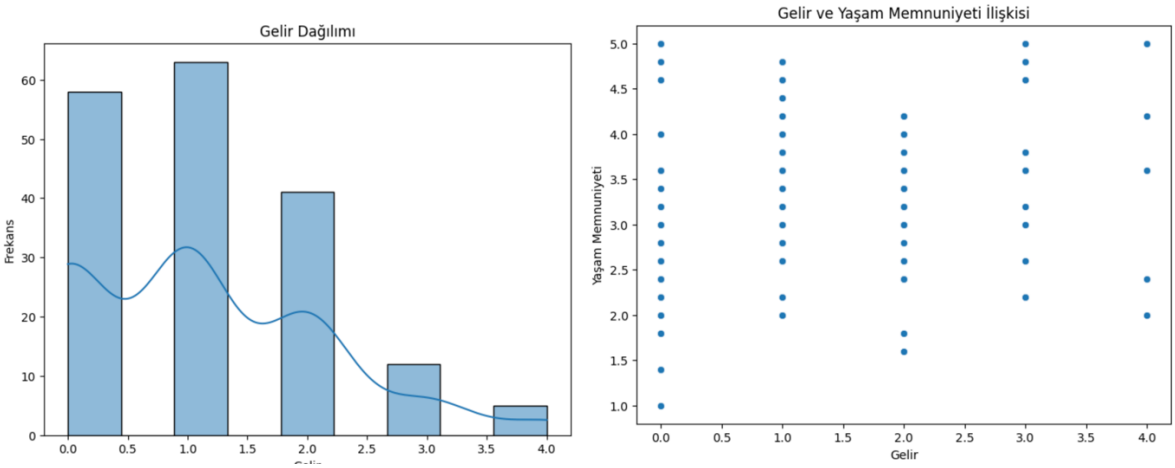




Sosyal Çevre için oluşturulan tablolar incelendiğinde verilerin genellikle 4 – 4.5 ve 5 civarına toplandığı gözlemlenebilir. Düşük değerler çok az işaretlenmiştir. Dağılım azdır.

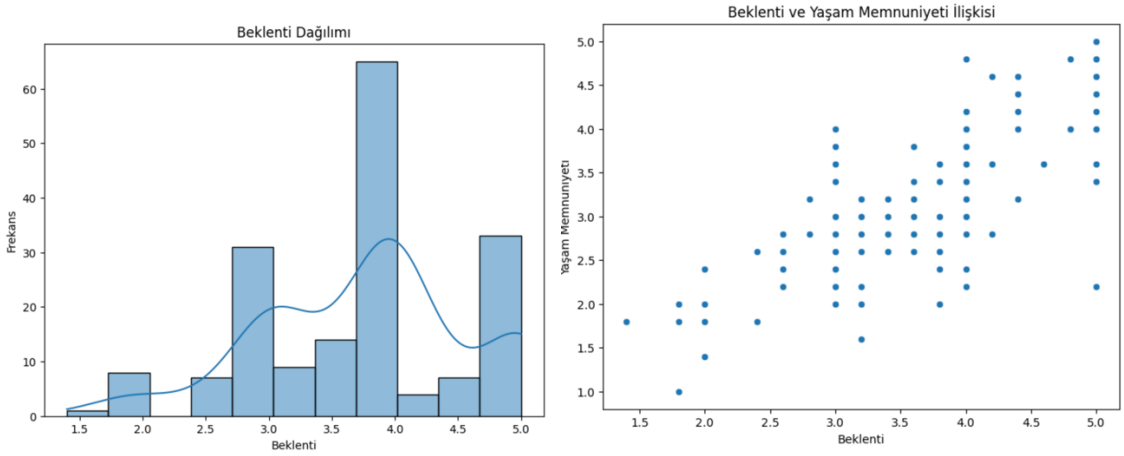
İkinci tablodan da görüleceği üzere diğer gruplara göre daha yükse bir korelasyon vardır. Sosyal çevre puanları arttıkça yaşam memnuniyetleri puanları da yükselme eğilimi göstermektedir.

5- Gelir



Gelir grubu için oluşturulan tablolarda ise verilerin genel olarak düşük puanlar üstünde toplandığı görülmektedir. Dağılımın az olduğu ve anketin eriştiği topluluğun genel olarak gelirinden memnun olmadığı anlaşılabilir. Fakat ikinci tablodan bakılacağı üzere, gelirin düşüklüğü yaşam memnuniyetini çok etkilememiştir. İki özellik arasında korelasyon olmadığı görülmektedir. Burdan anlaşılacağı üzere kullanıcılar gelirlerinden çok memnun olmasalar da yaşam memnuniyetlerine yüksek verme eğilimdelerdir.

6- Beklenti



Son grup olan beklenti grubu da incelendiğinde , verilerin çoğunlukla 4 e yakın değerlerde toplandığı görülebilir. Dağılım diğer birkaç gruba göre daha fazladır. İkinci tablo incelendiğinde korelasyonun çok net görülebildiği en yüksek etkileşim olan özelliğin beklenti olduğu anlaşılmaktadır. Beklenti için puanlar yükseldikçe yaşam memnuniyeti puanları da aynı oranda yükselmiştir.

Modellerin Uygulanması:

Veriler model eğitimine uygun hale getirildikten sonra farklı modeller ile eğitime sokularak sonuçlar değerlendirilmiştir.

Burada kullanılan modeller şunlardır:

Linear Regression (Doğrusal Regresyon)

Random Forest Regressor (Rastgele Orman Regresörü)

Support Vector Regression (Destek Vektör Regresyonu)

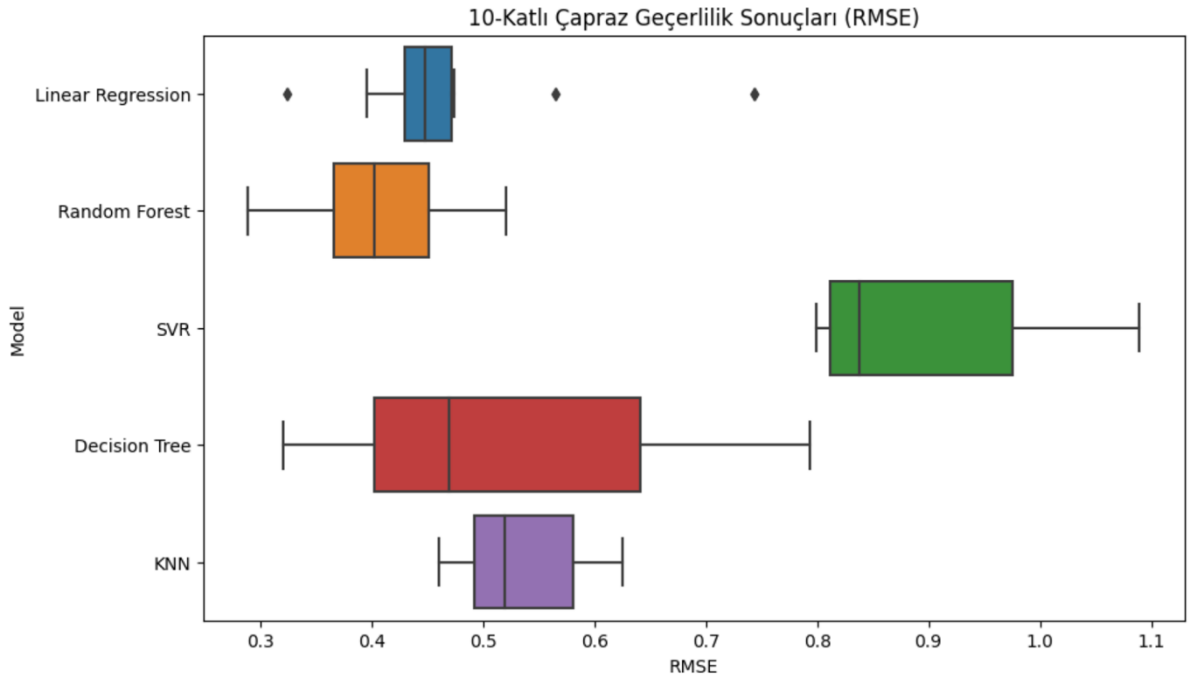
Decision Tree Regressor (Karar Ağacı Regresörü)

K-Nearest Neighbors Regressor (K-En Yakın Komşu Regresörü)

Bu modellerin her biri, 10-katlı çapraz geçerlilik kullanılarak eğitilmiştir. 10-katlı çapraz geçerlilik, modelin genelleme yeteneğini değerlendirmek için kullanılan bir tekniktir. Veri seti 10 eşit büyüklükteki parçaya bölünür. Model, 9 parça üzerinde eğitilir ve kalan parça üzerinde test edilir. Bu işlem 10 kez tekrarlanır, her seferinde farklı bir parça test seti olarak kullanılır.

Modelin performansı, negatif ortalama karesel hata (neg_mean_squared_error) ile ölçülür. Ancak bu metrik doğrudan yorumlanamaz, çünkü hataların kareleri alındığından, hata birimleri de karesel hale gelir. Bu nedenle, sonuçların daha anlaşılır olması için hataların karekökü alınır. Bu değerlere RMSE (Root Mean Squared Error) denir ve hata ölçütünün orijinal birimlerine dönüşmesini sağlar.

Son olarak, her modelin çapraz geçerlilik sonuçları bir kutu grafiği (boxplot) ile görselleştirilmiştir. Kutu grafiği, bir veri dağılımının özetini görselleştirmenin bir yoludur. Çizgi grafikte her "kutu", modelin 10 farklı çapraz geçerlilik sonucunu gösterir. Her kutunun orta çizgisi, medyan sonucu gösterir; kutunun üst ve alt çizgileri, çeyreklikler (IQR, yani ilk ve üçüncü çeyrekler) arasındaki sonuçları gösterir; ve "bıyıklar", sonuçların genel aralığını gösterir. Bu, modellerin genel performansını karşılaştırmak için kullanılabilir.



	Linear Regression	Random Forest	SVR	Decision Tree	KNN
0	0.323910	0.288773	0.809513	0.458984	0.490877
1	0.565563	0.520855	1.027533	0.793305	0.600622
2	0.396066	0.314925	0.860155	0.357771	0.460724
3	0.442401	0.434366	0.827096	0.396412	0.625916
4	0.743603	0.470906	0.820432	0.665475	0.498455
5	0.429153	0.360989	0.848413	0.320713	0.517356
6	0.473746	0.457211	0.799126	0.570714	0.479404
7	0.453255	0.387404	1.089600	0.420883	0.567501
8	0.434078	0.381764	1.013557	0.481070	0.585540
9	0.465754	0.418132	0.802728	0.667618	0.521974

Yukarıdaki verilen tablo ve listeler incelenip aşağıdaki gibi yorumlanmıştır.

Linear Regression (Doğrusal Regresyon): Bu modelin hata değerleri genellikle 0.32 ile 0.74 arasında değişmektedir. Doğrusal regresyon genellikle daha basit bir model olduğu için diğerlerine göre daha yüksek bir hata değeri vermiştir.

Random Forest (Rastgele Orman): Bu modelin hata değerleri genellikle 0.28 ile 0.52 arasında değişmektedir. Bu, Random Forest modelinin genellikle daha iyi genelleme yeteneğine sahip olduğunu gösterir. Bu model genellikle karmaşık ilişkileri ve özellik etkileşimlerini daha iyi yakalayabilir.

SVR (Support Vector Regression): Bu modelin hata değerleri genellikle 0.80 ile 1.08 arasında değişmektedir. SVR, genellikle verilerdeki karmaşık ilişkileri yakalamak için kullanılan bir başka modeldir ancak bu durumda daha yüksek hata değerleri üretmiştir.

Decision Tree (Karar Ağacı): Bu modelin hata değerleri genellikle 0.32 ile 0.79 arasında değişmektedir. Karar ağacı modelleri genellikle hızlı ve yorumlanabilir olsalar da, genellikle daha karmaşık modeller kadar iyi performans göstermezler.

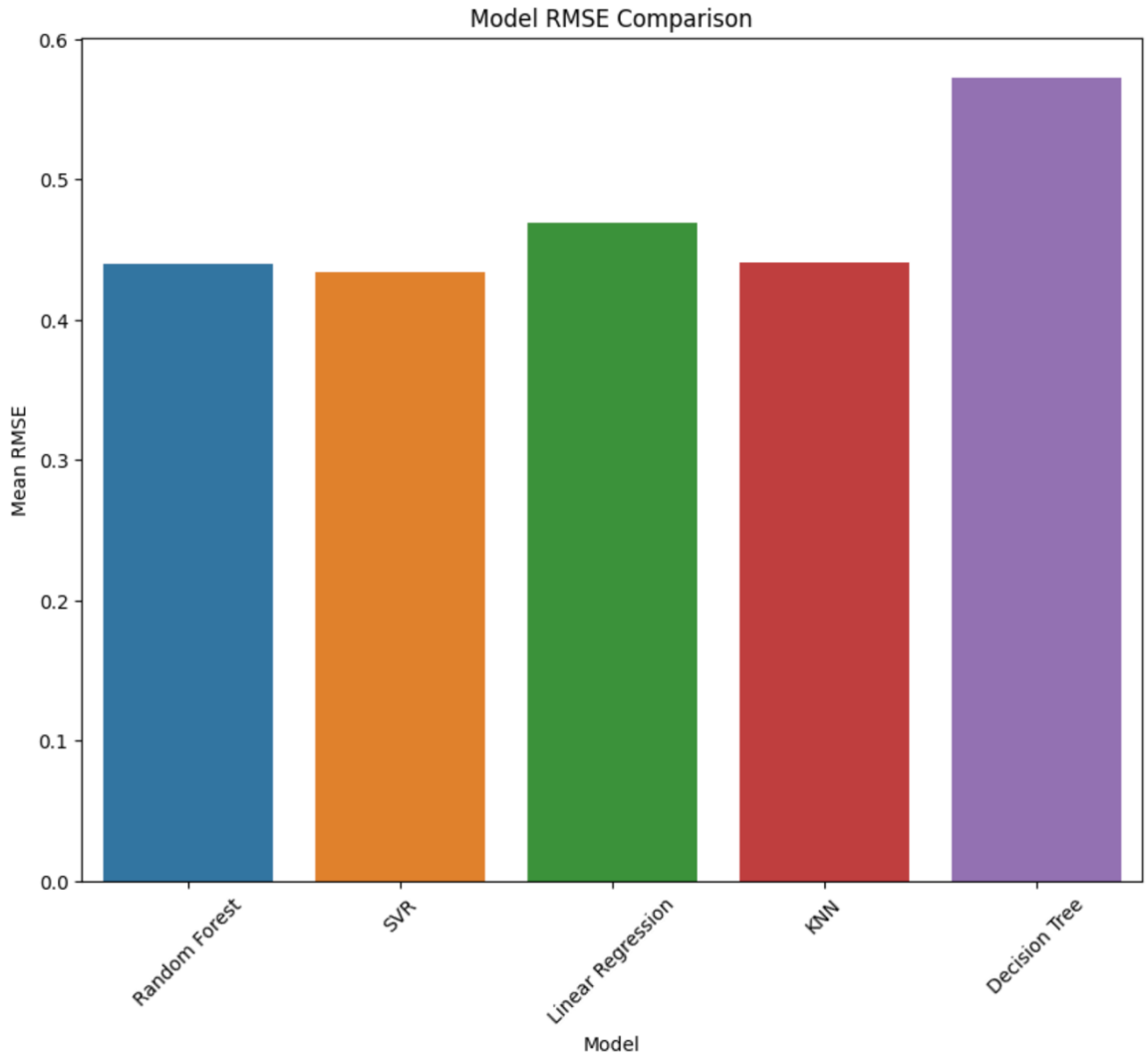
KNN (K-Nearest Neighbors): Bu modelin hata değerleri genellikle 0.46 ile 0.62 arasında değişmektedir. KNN, her örneği en yakın komşularına göre tahmin etmeye çalışır. Bu durumda, hata değerleri diğer modellere kıyasla biraz daha düşüktür.

Tablodan çıkan genel sonuç, Random Forest modelinin bu veri seti için en iyi performansı sergilediği üzerinedir. Bununla birlikte, hangi modelin kullanılacağına karar verirken, yalnızca modelin tahmin performansını değil, aynı zamanda modelin yorumlanabilirliğini ve işlem süresini de dikkate almak önemlidir.

Modellerin Farklı Yöntemlerle Geliştirilmesi:

Bu kısımda yapılan tahminlerin ve modellerin daha iyi sonuç vermesi için farklı modellerin özellik seçme ve boyut azaltma teknikleri kullanılarak denenmiştir. Bu, model performansını iyileştirmek ve daha anlaşılır hale getirmek için sıklıkla kullanılan bir yöntemdir. Seçilen modeller için özellik seçme (SelectKBest), normalleştirme (StandardScaler), ana bileşen analizi (PCA), teknikleri uygulanmıştır.

Her bir model için çapraz geçerlilik (cross-validation) uygulanarak bu süreçteki hataların karekökünün (RMSE) ortalamasını ve standart sapmasını içeren tablosu aşağıya eklenmiştir.



Model	Mean RMSE	Std Dev
Random Forest	0.440023	0.085628
SVR	0.433708	0.100835
Linear Regression	0.469102	0.098283
KNN	0.441047	0.069462
Decision Tree	0.572490	0.108028

Sonuçlar:

Verilen değerler ve tablolar incelendiğinde, uygulanan tekniklerin modellerin eğitimi konusunda daha iyi bir sonuca ulaştırdığı gözlemlenebilir. Uygulanan teknikler sonucunda her bir modelin performansına bakılırsa:

Random Forest, 0.440023'lik bir ortalama RMSE ile makul bir performans göstermiştir. Standart sapma 0.085628'dir, bu da modelin tutarlı tahminler yaptığını gösterir.

SVR (Support Vector Regression), 0.433708'lik bir ortalama RMSE ile en düşük ortalama RMSE değerine sahiptir, bu da tahminlerinin genellikle daha doğru olduğunu gösterir. Ancak standart sapması

0.100835 ile diğer modellerden biraz daha yüksektir, bu da tahminlerinin biraz daha az tutarlı olabileceğini gösterir.

Linear Regression modeli, 0.469102'lik bir ortalama RMSE ile diğer iki modelden daha yüksek bir hata oranına sahiptir. Standart sapması da 0.098283 ile diğer modellerle benzerdir.

KNN (K-Nearest Neighbors), 0.441047'lik bir ortalama RMSE ile makul bir performans göstermiştir. Standart sapması 0.069462 ile en düşük değere sahiptir, bu da modelin tahminlerinin oldukça tutarlı olduğunu gösterir.

Decision Tree, 0.572490'lık en yüksek ortalama RMSE değeri ile en düşük performansı sergiler. Standart sapması da 0.108028 ile en yüksektir, bu da modelin tahminlerinin diğerlerine göre daha az tutarlı olduğunu gösterir.

Bu sonuçlara dayanarak, SVR modeli genellikle en doğru tahminleri yaparken, KNN modeli en tutarlı tahminleri yapmıştır. Decision Tree modeli ise genellikle en düşük performansı sergilemiştir.