BLM 4510 YAPAY ZEKA 2. ÖDEV RAPORU

Dersin Yürütücüsü : Doç.Dr.Mehmet Fatih AMASYALI

Öğrenci Ad/Soyad : Alaaddin Göktuğ AYAR

Öğrenci No : 19011603

Ödev Konusu : Bir veri kümesi oluşturup makine öğrenmesi algoritmalarını çalıştırmak.

Ödev konusu:

Ödevde dataset Google formlar üzerinden açılan bir anket üzerinden oluşturulmuştur. Formun amacı kişinin verdiği bilgiler üzerinden en sevdiği alışveriş tipinin bulunmasıdır. Bu bağlamda kullanıcıya 13 soru sorulmuştur. Bu sorular sonucunda bir sınıflandırma problemi olan alışveriş tipinin tahmin edilmesi 5 farklı algoritma ile gerçekleştirilecektir.

Dataset için oluşturduğum form'un linki:

https://docs.google.com/forms/d/e/1FAIpQLSfJF6APKAIbx2qyCiFo8ARHxgWEE2JVvFvyQPBvvgyCL3Zkag/viewform?usp=sf_link

Bu algoritmalar;

- Logistic Regression
- Random Forest Classifier
- Support Vector Machine
- K Neighbors
- Decision Tree

Son soru target variable belirttiği için algoritmalarda toplamda 12 parametre bulunacaktır.

Toplamda 219 kişi ankete katılmıştır ve algoritmanin eğitiminde bu forma verilen cevaplar kullanılmıştır.

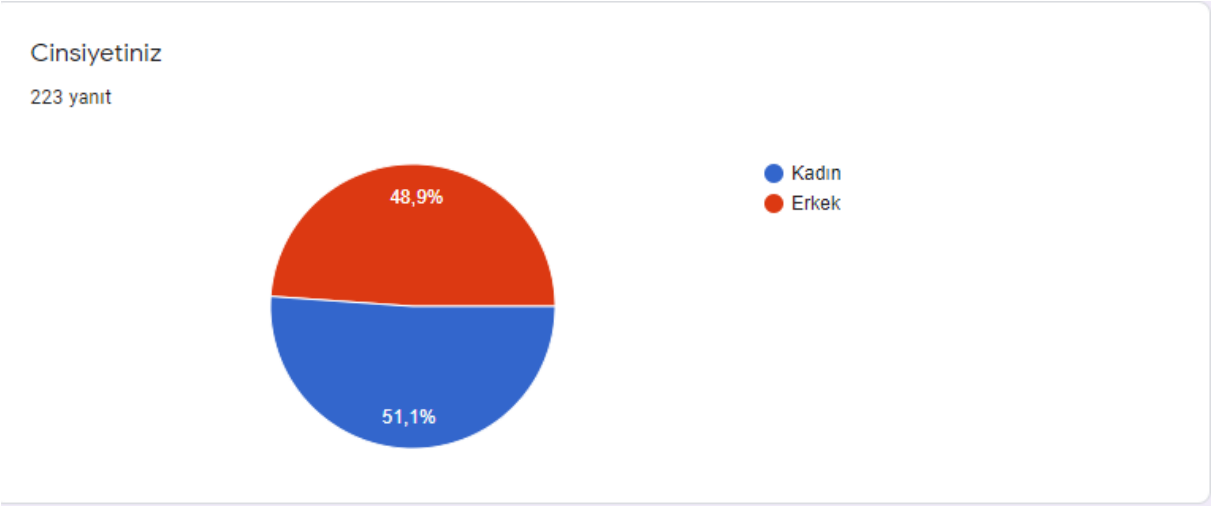
Tahmin edilecek hedef değişkenleri ;

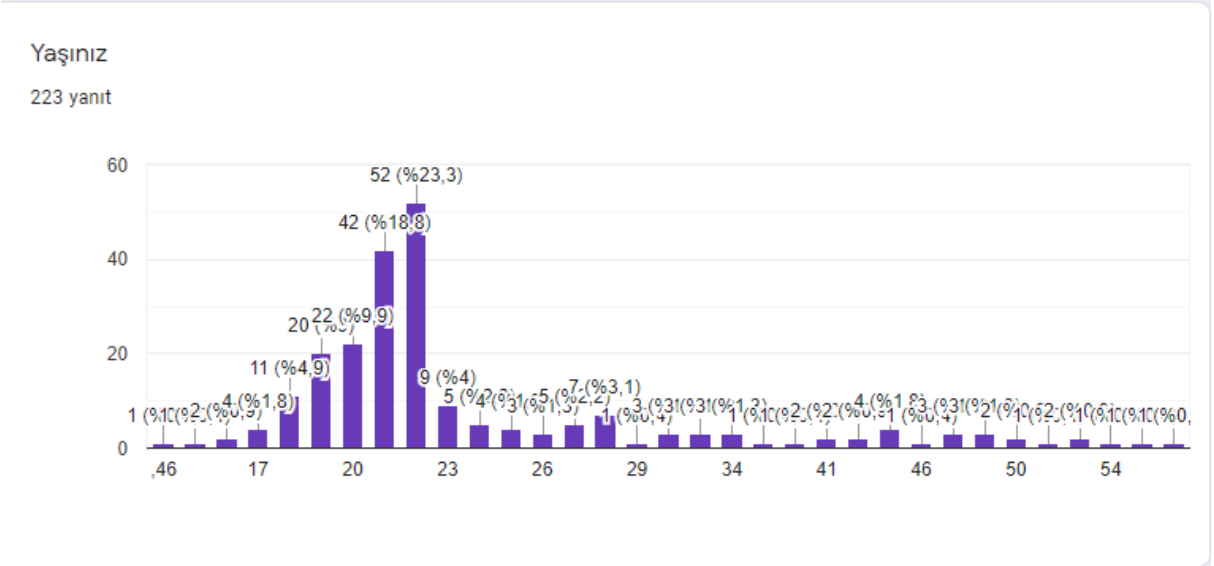
- Giyim - Teknoloji - Ev/Yaşam - Kitap/Dergi

Tahminin yapılması için sorulan sorular aşağıdaki gibidir

- Cinsiyetiniz - Yaşınız
- Hangi mağazaya gitmeyi tercih edersiniz?
- Hangi mağazaya gitmeyi tercih edersiniz?
- Hangi mağazaya gitmeyi tercih edersiniz?
- En sevdiğiniz mevsim nedir?
- Alışverişlerinizde dolar kurunun sizin için önem derecesi nedir?
- Alışverişe ayırdığınız bütçenizden memnunluk dereceniz nedir?
- Sosyal yaşantınızı nasıl değerlendirirsiniz?
- Online alışveriş sitelerinden hangisini tercih edersiniz?
- Ne sıklıkla alışverişe çıkarsınız?
- Günlük uyku süreniz ortalama kaç saattir?
- En sevdiğiniz alışveriş türü nedir? // target

Verilen Cevapların Yorumlanması ve Algoritmaya Olası Etkileri



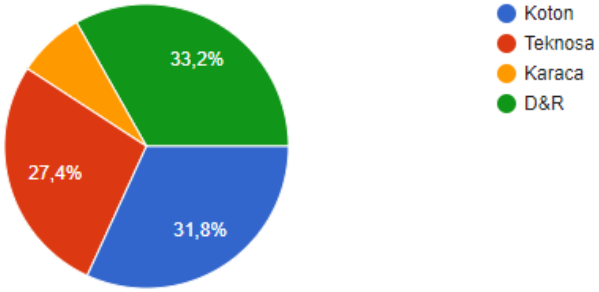


Cinsiyet neredeyse yarı yarıya bir oranla dağılım gösteriyor bu durum dışarıdan gelecek yeni bir inputun kadın veya erkek olma olasılığının eşit olduğundan iki durum için de yorumlamaya açık bir sonuç.

Genel ankete katılanların yaş aralığı 18-22 yaş aralığı kişiler bu durum belirli bir alışveriş tipinin baskın çıkmasını etkileyecek bir durum. Yaş aralığı daha geniş ve her yaştan insanın fazlasıyla doldurduğu bir dataset sonuca daha olumlu etki edecektir.

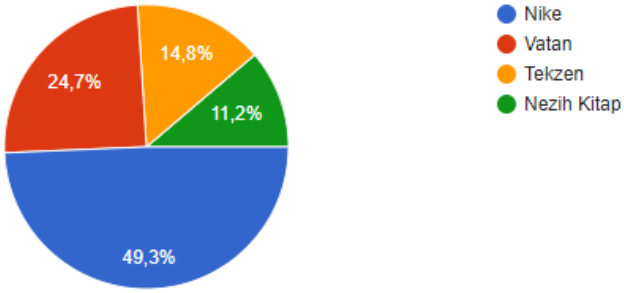
Hangi mağazaya gitmeyi tercih edersiniz?

223 yanıt



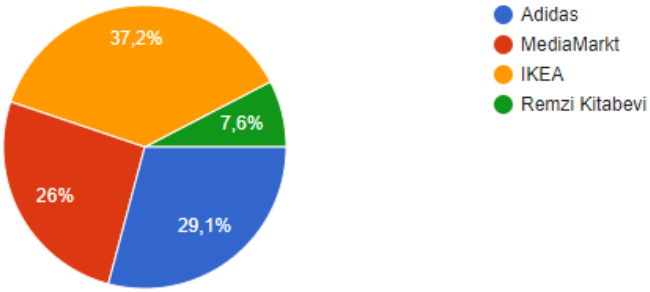
Hangi mağazaya gitmeyi tercih edersiniz?

223 yanıt



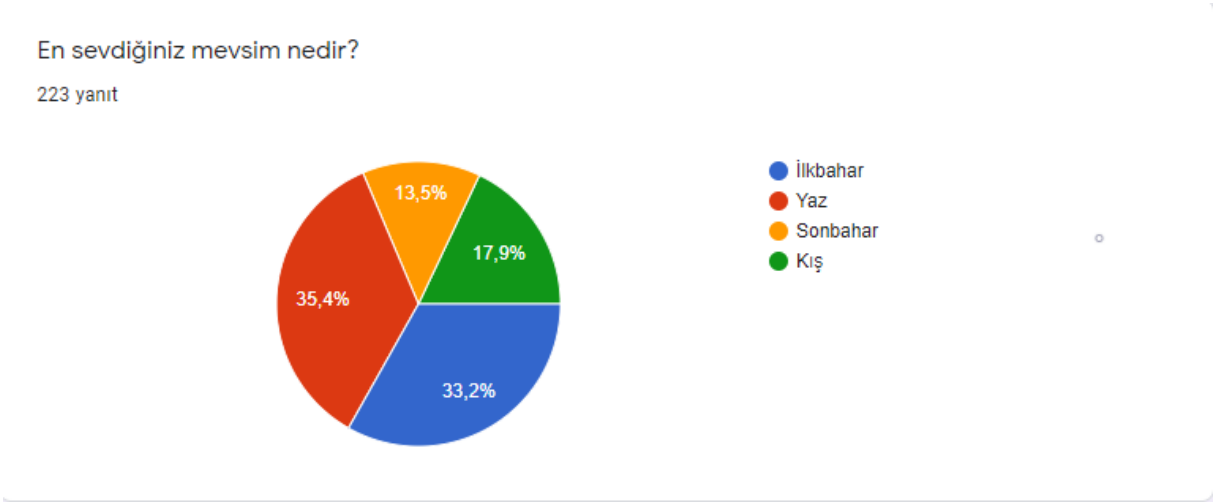
Hangi mağazaya gitmeyi tercih edersiniz?

223 yanıt

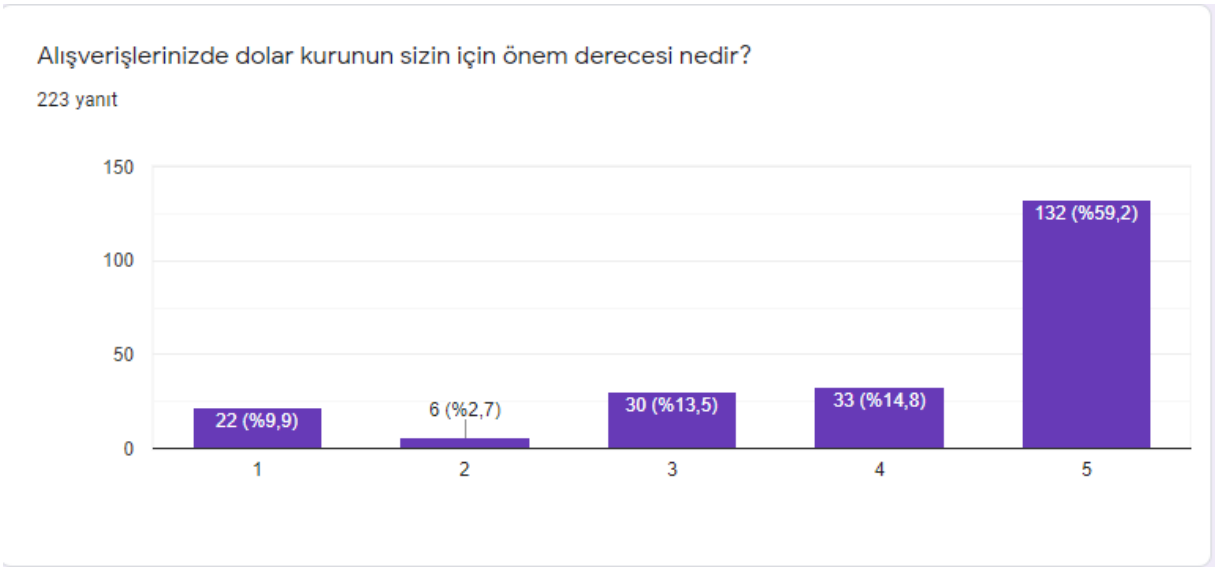


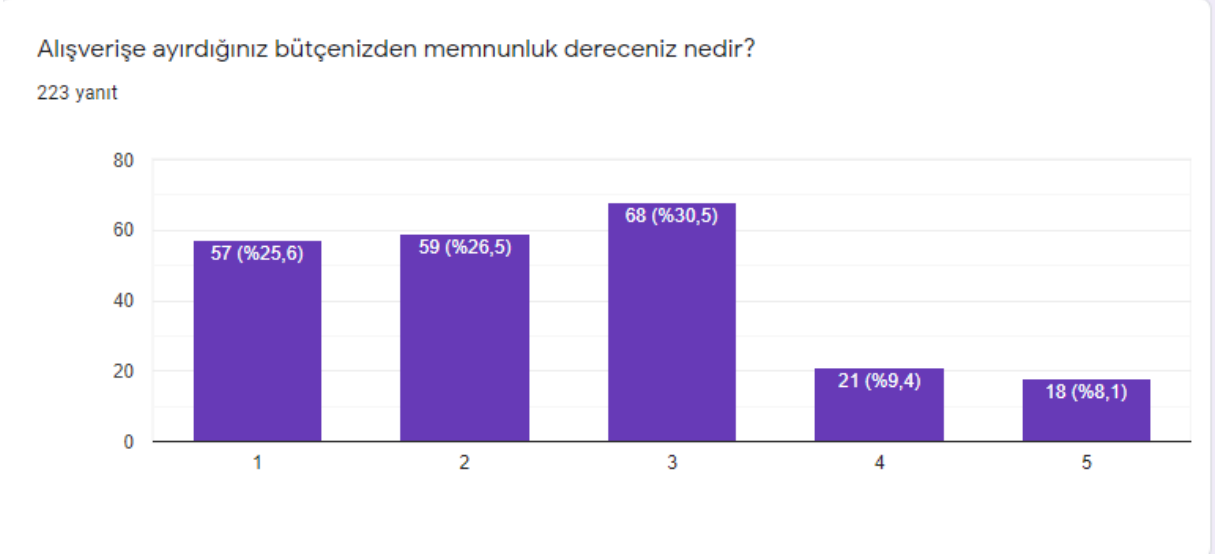
3 tane aynı sorudan sorduğum bu soruda verilen cevaplar algoritmaya doğrudan etki edebilecek cevaplar. Bu sorulardaki amacım her sorudaki mağazalardan birinin hedef değişkenleri işaret edecek değişkenler olmasıydı.

Kullanıcıların genel anlamda giyim mağazalarını tercih ettiği daha sonra teknoloji mağazalarını tercih ettiği görülüyor. Bu soruların cevabına göre sevdikleri alışveriş türünün daha net belli olacağı, sonucun da buna benzer sonuçlar olma ihtimali yüksek.

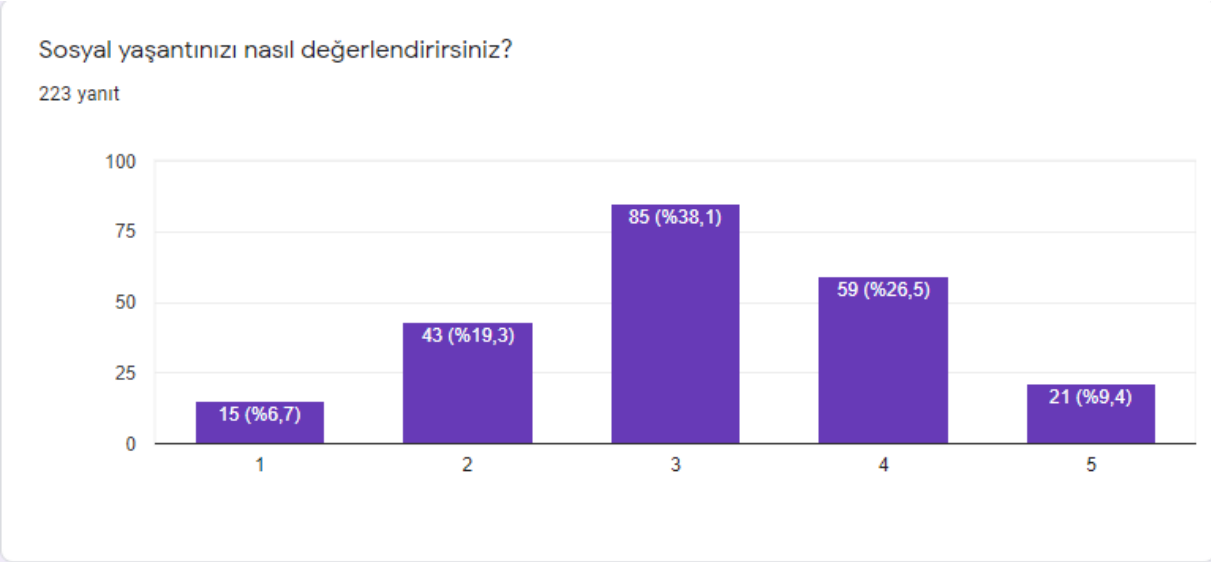


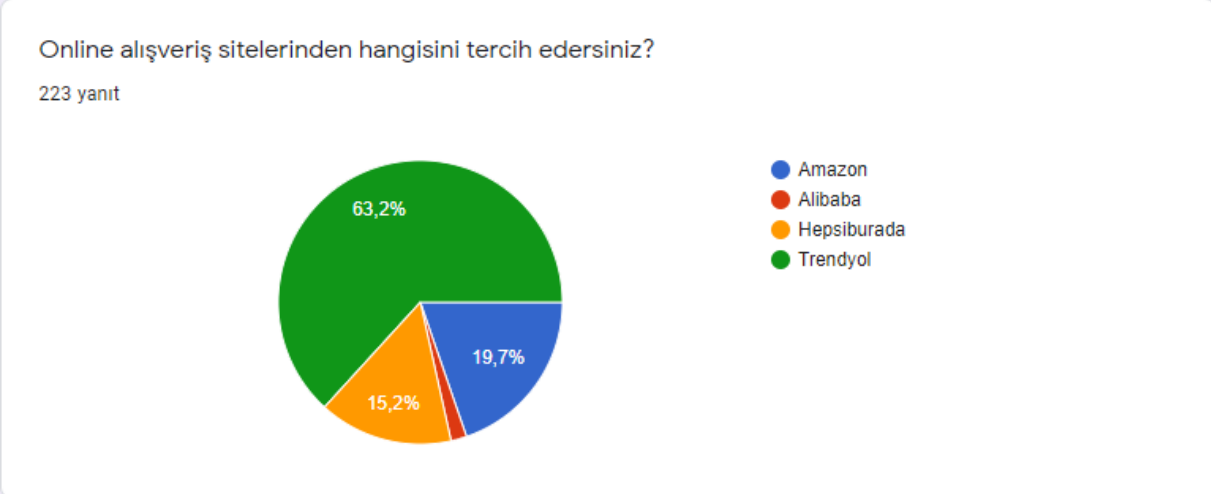
Bu soru biraz daha cevabın belirlenmesinde çok üst düzey bir etkisi olmayacak bir şekilde sadece kullanıcıların belirli bir örüntü oluşturmasını beklediğim bir soru.





Dolar kurunun önem derecesi biraz daha teknoloji ve giyim alanını tercih edenlerin önem duyacağı bir alan olacağını düşünüp koyduğum bir soruydu ancak. Günümüzde her alışveriş tipinin dolara yüksek bir etkisi olduğundan bu parametreye verilen cevaplar büyük bir çoğunlukla 5 olduğundan algoritmaya etkisi en az olan parametrelerden biri olacaktır. Sonuca herhangi bir belirleyici etkisi diğer parametrelere göre daha az olacaktır.

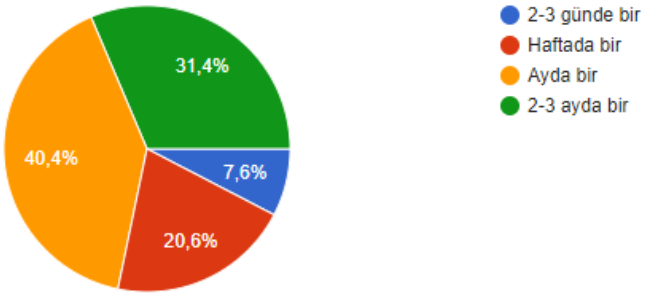




Sosyal yaşantı ve online site tercihlerinin de alışveriş tipine belirli etkileri olabilir. Online alışveriş sitelerinden tercih edilme olasılığı trendyol için fazla çıktığından biraz daha bu özellik seçiciliğini kaybetmiştir.

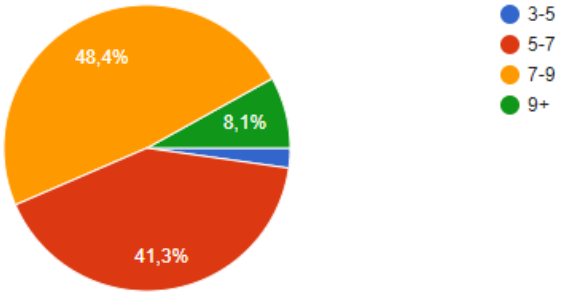
Ne sıklıkla alışverişe çıkarsınız?

223 yanıt



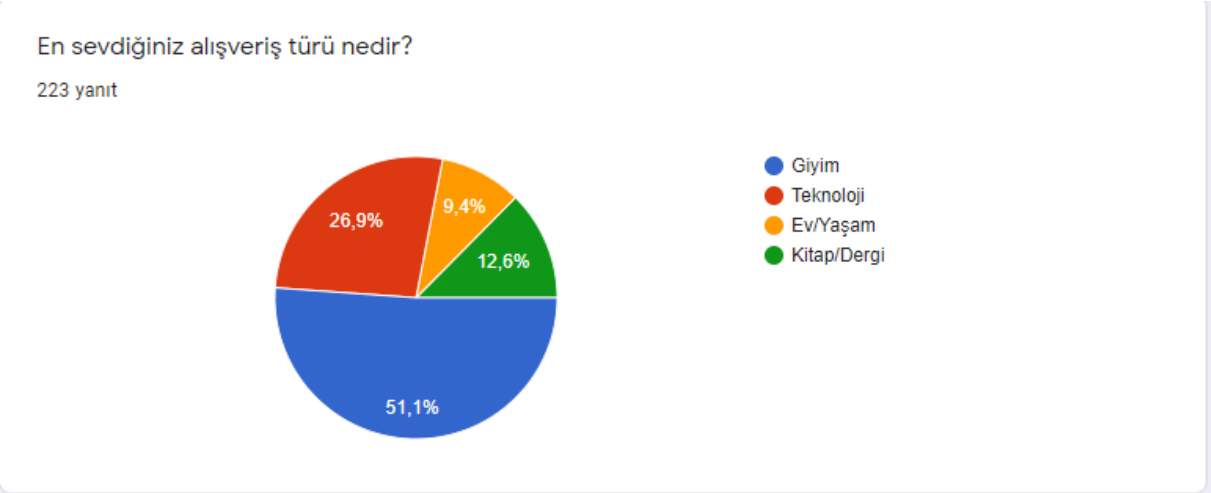
Günlük uyku süreniz ortalama kaç saattir?

223 yanıt



Hangi sıklıkla alışverişe çıktığımız sorusunda cevaplar biraz daha dağılmış ve alışverişe çıkma sıklıklarının insanların ne tarz bir alışveriş yapmayı sevdiğine önemli bir miktar etkisi olacaktır.

Günlük uyku süresi herhangi bir beklentim olmadan, alışveriş tipine önemli bir etki yapmayacağına inandığım bir soruydu ancak kendi içinde cevaplar için bir düzen oluşturma ihtimalini düşündüm ancak verilen cevaplar çok ortada ve birbirine 5-7 7-9 yakın olduğu için etkisi tartışılacak bir soru.



Önceki sayfalarda önemi yüksek dediğim ve biraz daha standart sapması yüksek olan soruların etkisinin giyim yönünde alışveriş tercihinin daha çok olacağı yöndeydi. Sonuçları incelediğimizde giyim tercihinin bir hayli yüksek olduğun görüyoruz.

Kodun Anlatımı

Csv dosyası şeklinde bulunan dataset , sisteme dataframe şeklinde okunur. Ve algoritmamiz için bir anlam ifade etmeyen kullanıcının formu hangi saat ve dakikada doldurduğu bilgisi sütunu çıkarılır.

PCA işlemi yapılmadan önce dataset içinde bulunan sütunlardaki sayılar diğerlerine göre daha büyük olduğundan PCA işlemi sırasında bu sütunların birleşmesi sonucu etki etmesi gerekenden daha büyük bir etkiye sahip olmaması için sütunlara standardization işlemi uygulanır.

Datasetin algoritmalara verilmesi sırasında kullanılacak features ve target columnları belirlenir.

Oluşan algoritmalari fit etmek için sıra ile datasetin ilk hali , normalize edilmiş hali ve pca uygulanmış halleri ayrı ayrı tutulur. Oluşturulan datalar train = 0.8 ve test = 0.2 olmak üzere parçalara ayrılır. Cross Validation işlemi 0.8 train datası üzerine uygulanacaktır.

Kullandığım 5 algoritmayi dataseti vermeden önce, dataset içinde bulunan yazı ile yazılmış cevaplar ve diğer sorularda bulunan yazı ile cevapların encode işlemi yapılır ve dataset sayılara dönüştürülür.

Kullandığım 5 algoritma sklearn kütüphanesinden alınmış olan fonksiyonlardır. Cross Validation işlemi Logistic Regression algoritmasi hariç GridSearchCV() fonksiyonu kullanarak yapılmıştır. Logistic regression algoritmasinda ise algoritma için kullandığım fonksiyon ile Cross Validation yapma imkanı olduğundan o fonksiyon kullanılarak yapılmıştır.

GridSearchCV() fonksiyon için verdiğim parametleri deneyerek K-Fold Cross Validation uygular, bizim ödevimiz için K sayısı 10'dur. Cross validation işlemi parametreler ve verdiğimiz train datası bölünerek en iyi sonucu hangi değerlerde alabileceğimiz belirlenir.

Belirlenen parametreler kullanılarak algoritma oluşturulur ve algoritma train datası ile fit edilecek test datası ile test edilir.

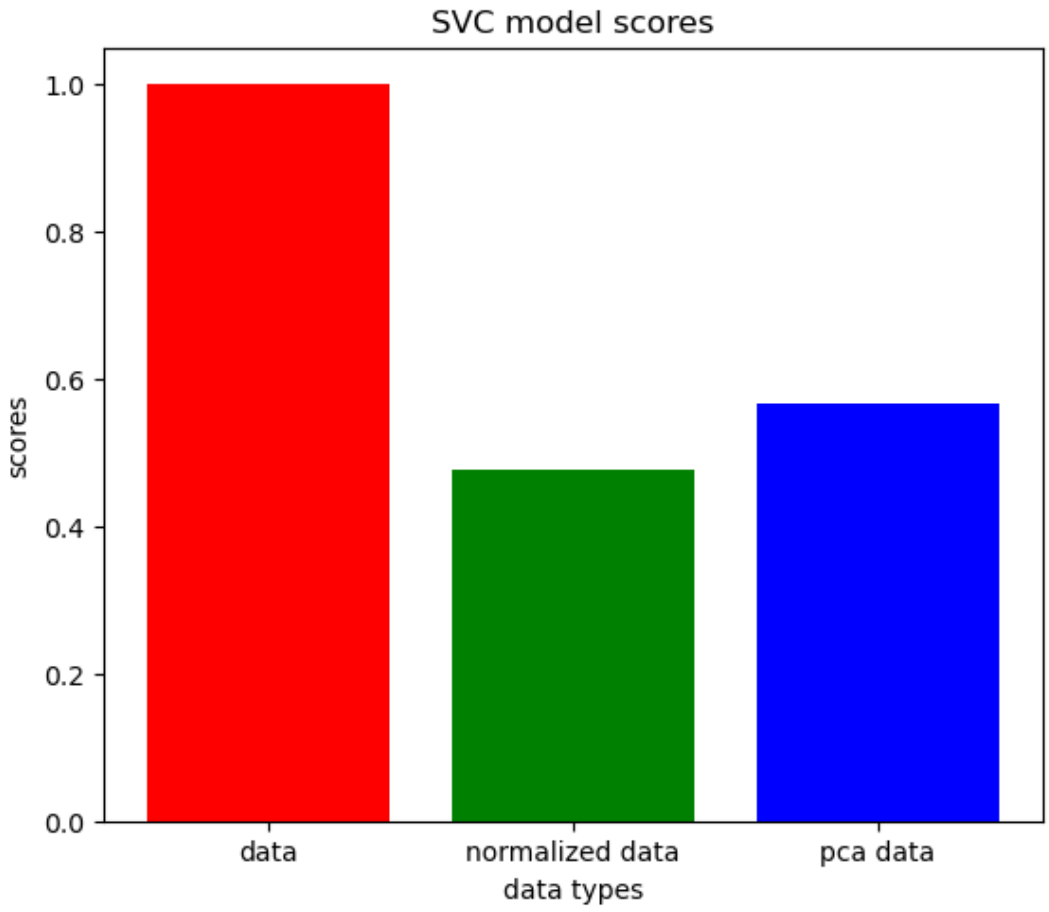
Data , Normalized Data , PCA uygulanmış Data karşılaştırmaları

Logistic Regression


```
Accuracy for LogisticRegression on data:  0.9090909090909091
Accuracy for LogisticRegression on normalized data:  0.8863636363636364
Accuracy for LogisticRegression on pca data:  0.5909090909090909
```

Logistic Regression kullanımında sınıflar arası düz bir çizgi çekilmesindense daha kıvrımlı bir çizgi ile boundary oluşturulmaya çalışılır. Bizim datamızda bazı sorulan soruların net bir şekilde sonuca etki etme özelliği bulunduğundan o sorulara verilen cevaplar ile alışveriş tipine verilen cevap uyumlu bir şekilde ise yani kullanıcı Nike ve Adidas mağazalarını işaretleyip Teknoloji seçmiyor ise logistic regression kullanılarak gayet sınıflandırılabilir bir durumda. Eğer kullanıcı böyle çelişkili bir işaretleme yapıyor ise bu durumda diğer sorulardan train aşamasında ortaya çıkardığımız düzen bize yardımcı olacak. Data Normalize edilirse de bu durumda herhangi bir farklılık olmayacak sadece bazı datalar birbirine fazla yaklaşacağından tahmin aşamasında küçük bir miktar düşüş yaşadık ancak onun dışında algoritma performansını koruyor. PCA algoritmasinda bazı birbiri ile alakası olmayan sutünlar birleşeceğinden ve standart sapmayı yüksek tutarak bu birleştirmeyi yapmaya çalıştığımızda algoritma train datası içinde boundary'ı istenilen yere tam olarak koyamayacaktır ve başarımız data ve normalized data hallerine göre önemli bir ölçüde düşecektir . PCA algoritma için önemli olan bazı dataları , bu datalara göre daha önemsiz olan datalar ile birleştirdiğinde başarıya olumsuz bir etkisi olmuştur.

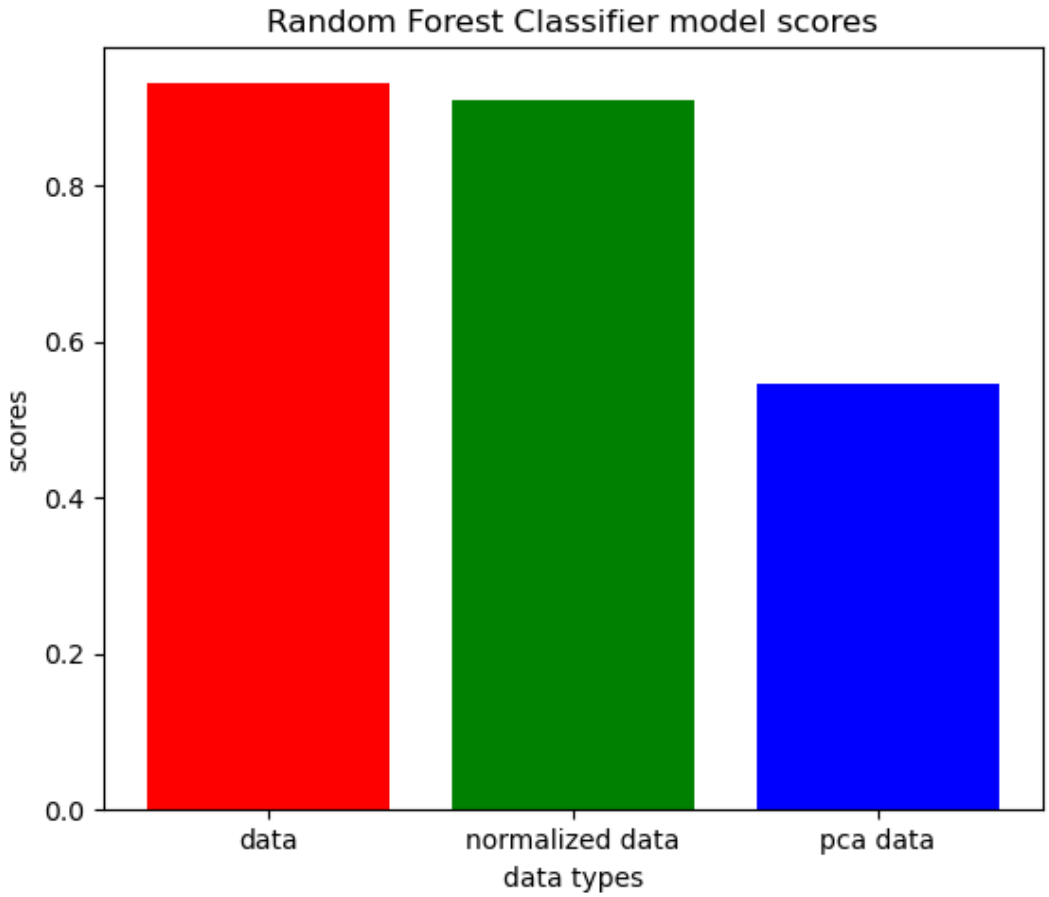
Support Vector Machine




```
Accuracy for SVC on data:  1.0
Accuracy for SVC on normalized data:  0.47727272727273
Accuracy for SVC on pca data:  0.5681818181818182
```

SVC algoritmasinda datanın ilk hali için 1.0 lık bir sonuç almışız. Bu durum biraz daha overfit olduğunu gösteriyor. Overfit olma ihtimalinin dışında bir başka ihtimal ise datasetimde bulunan bir sütunun perfect seperator olduğudur ancak bu kısım biraz daha datasetin oluşturulma aşamasında düşünmem gereken bir durumdu , algoritmalarin kullanımına yabancı olduğumdan bu durumu biraz geç farkettim. Dataset bazı cevapları açısından net yönlere yönlendirebilir ancak GridSearch ile en iyi parametreleri seçerken dataset keskin sınırlar ile ayrılabildiğinden ve SVC için overfite açık olduğundan 'Kernel ' : 'Linear' olarak seçilir. Linear olarak seçilen kernela uygulanan normalizasyon ve pca işlemleri datasetin biraz daha linear çizgiler ile ayrılabilir yapısına zarar vereceğinden score'larımızda bir düşme olacaktır.

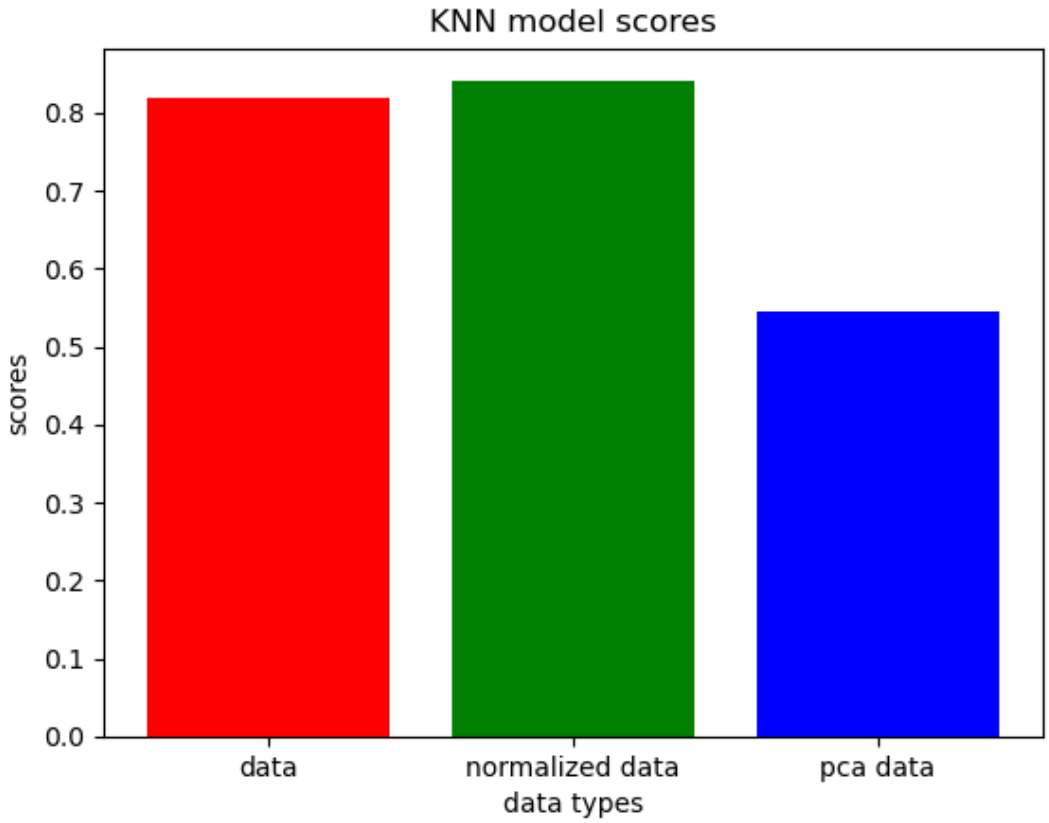
Random Forest Classifier




```
Accuracy for Random Forest on CV data:  0.9318181818181818
Accuracy for Random Forest on CV normalized data:  0.9090909090909091
Accuracy for Random Forest on CV pca data:  0.5454545454545454
```

Random Forest algoritmasi kullandığımız özelliklerin öneminin algılanmasının daha çok öne çıktığı bir algoritma olduğundan , normalize işleminin uygulanması datalarının öneminin algılanmasında katsayıların azalmasından dolayı küçük bir etki edecek olsada bu durum aldığımız skoru çok az bir miktar etkileyecektir. Random Forest normalizden de sonra hala özelliklerin önemini rahat bir şekilde kavrayabilecektir. Ancak PCA işlemi özelliklerin birleştirilmesini kullanacağından bazı önemli özelliklerin önemliler ile karışması ve değerlerinin kaybolması durumu Random Forest tarafından yakalanmalarını zorlaştıracak ve skorumuzu önemli bir ölçüde düşürecektir. PCA'nin Random Forest için olumlu tek yanı algoritmain çalışma hızını biraz daha arttıracak olmasıdır. Ancak önemli özelliklerin öneminin algılanmasını hedeflediğimiz Random Forest algoritmasinda hızlandırılmış ancak doğruluğunu önemli bir ölçüde kaybetmiş bir algoritma bizim için çok olumlu olmayacaktır.

K-Nearest Neighbors (KNN)




```
Accuracy for KNN on  data:  0.8181818181818182
Accuracy for KNN on normalized data:  0.8409090909090909
Accuracy for KNN on pca data:  0.5454545454545454
```

En yakin n komşunun bulunduğu algoritmaya normalization işlemi uygulanması demek yakın olan dataların daha da yakına geleceği anlamına geliyor. Bu durumda iki nokta arası mesafeyi hesaplayarak en yakın komşularını seçen algoritma , normalize edilmiş data için yanlış sınıflandırma ihtimalini daha da azaltacaktır. Oluşturduğum datasette özellikler arası fark zaten çok büyük değerler olmadığından skor artmış olsa bile çok yüksek bir artım olmamıştır ancak normalize etme işlemi knn algoritması için önemli ve performansı arttıran bir uygulamadır. PCA uygulanmış datada hem sütün sayısı yani özellik sayımız daha da azalacak ve standart sapmayı yükseltmeyi hedefleyen PCA uygulanmış datada mesafeler artacaktır. Bu durum özelliklerin öneminin azaltılması ve iki nokta arası uzaklıkların artması sonucu knn algoritmasını olumsuz bir şekilde etkilemiştir.

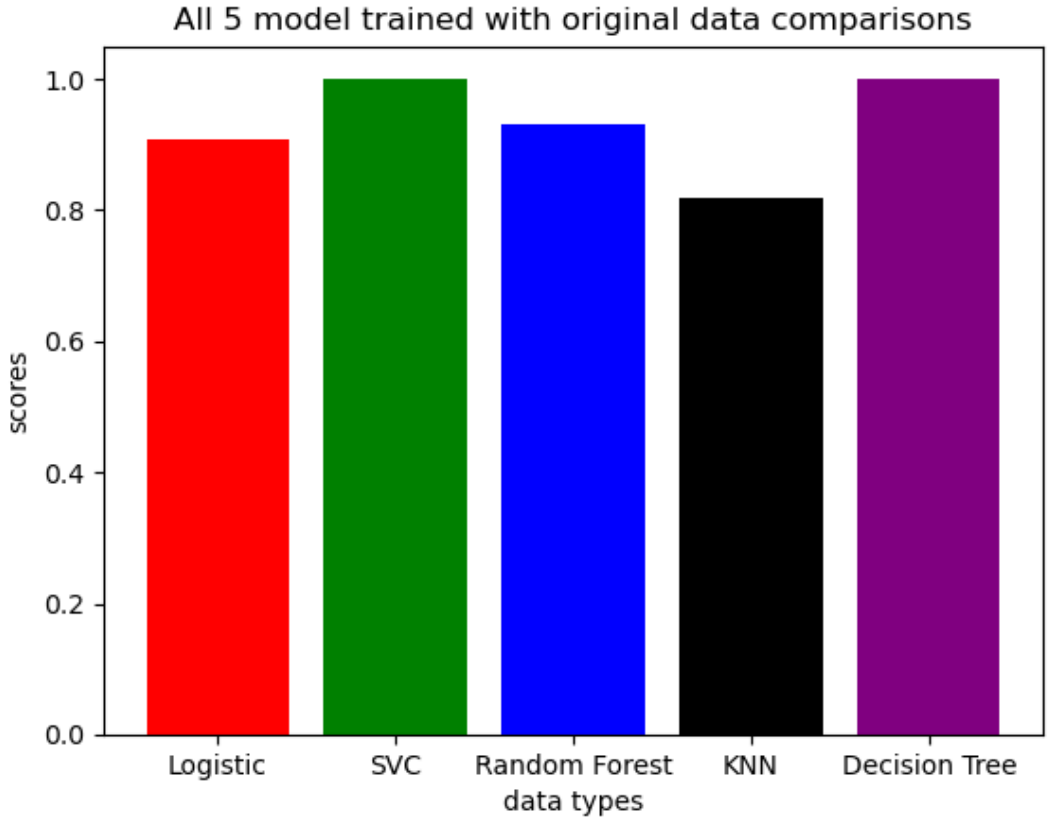
Decision Tree


```
Accuracy for decision tree on data:  1.0
Accuracy for decision tree on normalized data:  0.8863636363636364
Accuracy for decision tree on pca data:  0.5681818181818182
```

Temelde Decision Tree kullanan Random Forest algoritması için yazdığım bütün şeyler bir bakıma bu algoritma için de geçerlidir. Perfect Seperator yaratmış olmam tree algoritmalı için çok net cevaplara ulaşmalarına neden olmuştur. PCA uygulanması ise özelliklerin değerini daha iyi anlayan Decision Tree algoritması için özellikleri azaltmak ve önemli özellikleri önemsizler ile harmanlayarak ulaştığımız özelliklerin , önemli özellik değerlerini yitirmesine neden olur. Bu sebeplerden data , normalized data ve pca uygulanmış datada böyle sonuçlar elde ederiz.

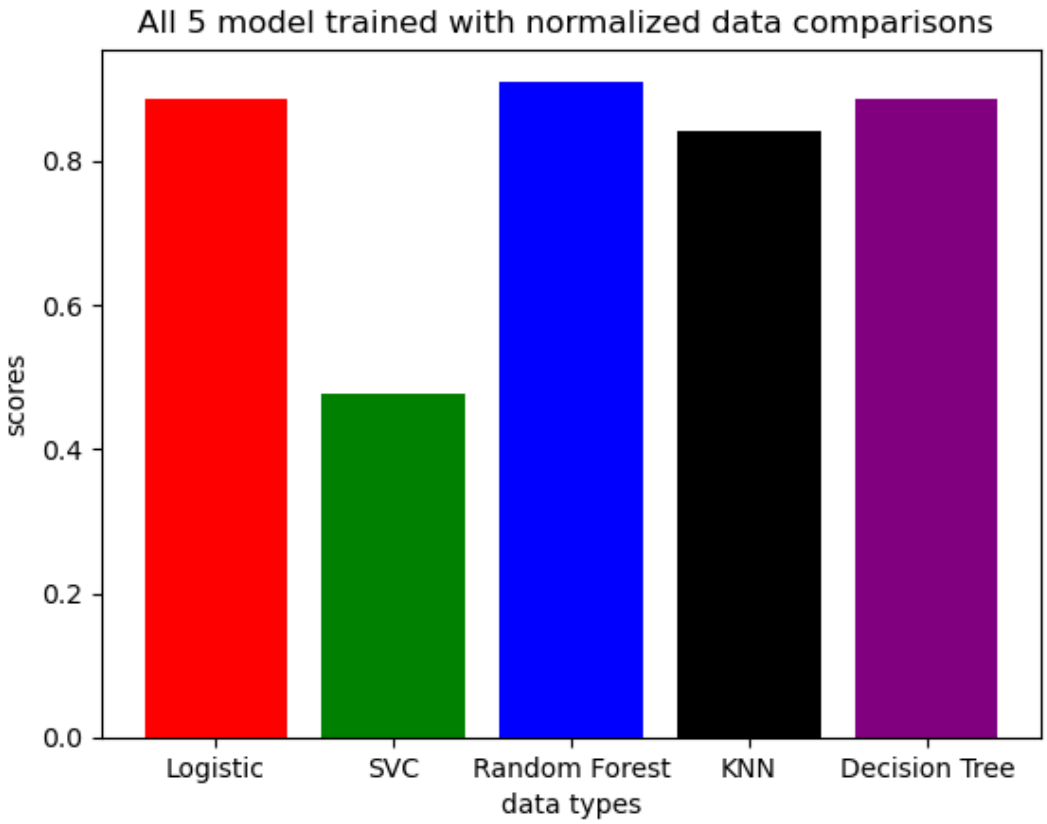
Tahminleyicilerin Performansları Arasında İstatistiksel Farklar

Original Data , Normalized Data ve PCA uygulanmış Data için modellerin aldığı skorlar;

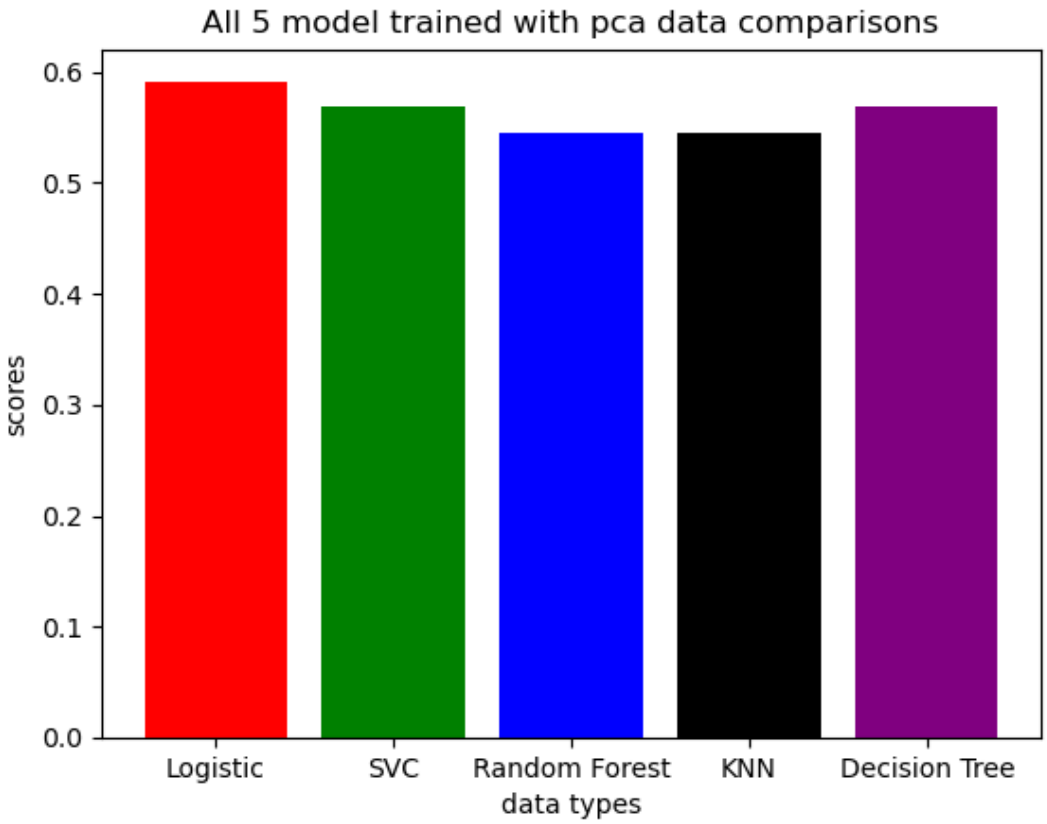


Datamız üzerinde yaptığımız karşılaştırmada linear modellerin , datasetin biraz daha linear şekilde ayrılabilir yapısı olduğundan diğer algoritmalara göre fazla geride kalmadığı görülüyor. Decision Tree için perfect seperator durumu öne çıkıyor. KNN algoritmasının diğer algoritmalara göre biraz daha geride kaldığı gözüküyor, bu durum hem normalizyon uygulanmamış olmasının etkisi, hem de data sayımızın az ve daha birlikte bir yapısı olduğundan knn algoritması ne kadar iyi bir skor elde etse bile, bu tip datasetler için diğer algoritmalar kadar başarı gösteremediği görülüyor.

Logistic regressiondan daha iyi bir boundary çıkartan SVC algoritması için, logistic regression için yüksek bir skor alan dataset. SVC'nin eğitiminde kullanıldığında çok daha iyi bir boundary çıkacak ve skoru üst seviyelere taşıyacaktır.



Normalized uygulanmış datasette, knn performansının arttığı görülürken, komşular arası mesafenin azalması decision tree, random forest için çok önemli bir etkiye sahip değilken knn algoritması için diğer modellere göre daha fazla iyileşme göstermesini sağlamıştır.



PCA uygulanması tree algoritmalarında ve knn algoritmasında önemli düşüşlere yol açmıştır. Özellik sayısının ve önemlerinin azalması ayrıca standart sapmanın artması ve komşular arası mesafelerin artması bu algoritmaları olumsuz yönde diğer iki algoritmaya göre daha çok etkilemiştir. Regression ise dataset biraz daha cluster yapısından ayrılma ve standart sapması arttığından dolayı performansında düşüş yaşamıştır ancak bu düşüş tree ve knn algoritmaları kadar sert olmamıştır.