### ABSTRACT

Bundan 20 yil once hayatimizda olmayan bir cihaz olan telefonlar suanda hemen hemen hepimizle her an beraberler ve bu telefonlari aslinad icindeki uygulamalar icin kullaniyoruz. Madem bizim icin bu kadar onemli bizde bu nedenden oturu bu uygulamalari indirdigimiz tek alan olan (iphonelar icin) bu uygulamalarin ne kadara satilmasi gerektigini oneren yapay zeka sistemi ve o zamanin trendlerine gore bize uygulama cesidi oneren bir yapay zeka urettik.Bu calismamizi Makine ogrenimi algoritmalari ile gerceklestirdik.

Makine Ogrenimi algoritmalarimizi secerken dikkat ettigimiz 1.unsur verisetimizin az veriye sahip olmasi ve buna bagli olarak cok fazla kategorik degiskeni olmas idi (genre).2 unsur ise price predict yaparken cok fazla verimizin ucretsiz olmasindan oturu verimizin overfitting veya underfitting olma durumuydu. Bu kistaslari goz onune alarak sirasi ile Classification icin:Decision Tree , Random Forest , LinearRegression Ridge,Lasso Algoritmalarini kullandik.Bu algoritmalarin en iyisinin accuracy sonuclari 99%(Decision Tree) ve 1.0MSE (Lasso)gibi cok iyi bir sistem oldugunu kendi icerisinde tutarladik.

### INTRODUCTION

“Mobil cihazlarin dunya capinda son 40 yilda 500.000’den 5,000,000,000 adete kadar artmasi ve gelisen donanimlar sayesinde ozellikle mobil oyun sektoru olmak uzere uygulama magazasindaki sektorun hacminin devasa buyumesi kacinilmaz olmustur”.[3]

Insanlar artik gercek hayatta yasamaktan cok sosyal medyada yasar zaman gecirir hale gelmistir. Ustelik kendisini eglendirmek ve zaman gecirmek icin yaptigi bir aktivite olan “oyun” oynama ihtiyacini da akilli telefonlar araciligi ile saglamaya baslamistir.

Sektorun bu onlenemez yukselisi elimizde daha once veri olmamasi ve insanlarin para yatirma egilimlerinin bu sektore kayacagini ongorememesinden dolayi uygulama marketler anarsik bir durumda atil sekilde kalmislardir.

Bu verisetinde appstore’den aldigimiz verilerimizin aldigi puanlardan;version numarasi,cesiti,fiyat … ve bunlarla iliskili 18 features ile veri analizi yaptik ve gorsellestirdik.

Ve sonrasinda bize bir oneri sistemi sunacak olan ,sistemimizi “akilli” hale cevirecek olan makine ogrenmesi modellerini egitmek icin 5000’den fazla verimizi belirli parcalara bolerek Buradan hangi uygulamalarin aslinda daha cok kullanildigini ve fiyatinin yaklasik olarak ne kadar olmasi gerektigini matematiksel olarak group oneri sistemimize egittik.Bu sonuclarin her ne kadar yuksek dogruluk degerine sahip olsa da aslinda veri sayimizin yeterli olmadigini gosteriyor.

## VERI ONISLEME

Elimizdeki verileri makine ogrenmesi modellerine sokmadan once bunlara data cleaning data reducancy data transform ve data add islemleri uygulamamiz gerekiyordu.

AsagidakI tabloda (1) kirmizi ile yazili olan features’I kendimiz kesifci veri analizi yaptiktan sonra olmasi gerektigine karar verdik , bu sayede tahmin sistemi gelistirebildik.

Bazi tekrarlayan ve gereksiz kolonlari dropladik (tabloda hepsi bulunmakta)

Ve cok fazla verimiz string formatinda iken bunlari int veya float formatina cevirdik

|  |  |
| --- | --- |
| **Track Name** | Oyun ismi |
| **Size (bytes)** | Boyut |
| **Currency** | Para birimi |
| **Price** | Fiyat |
| **Rating Count Tot** | Toplam degerlendirme sayisi |
| **Rating Count Ver** | Toplam degerlendirme sayisi hakkinda |
| **User Rating** | Kullanici degerlendirme puani |
| **User Rating Ver** | Kullanici degerlendirme puani hakkinda |
| **Ver** | Version |
| **Cont Rating** : | Puanlama |
| **Prime Genre** : | Tur |
| **Sup Devices Num** : | Destekleyen Cihaz numarasi |
| **ipadSC urls num** : | **XX (Bilinmiyor)** |
| **lang num** : | **XX (Bilinmiyor)** |
| **vpp lic** : | **XX (Bilinmiyor)** |
| **isFREE** |  |

## MAKINE OGRENMESI ONCESI VERI ISLEME

Verilerimizi makine ogrenimi algoritmalarina sokmadan once sayisal veriler oldugundan emin olmaliyiz.string degerler birakmamaliyiz

Ve isFree & price kolonlarinin kendi icerisindeki korelasyon gibi cakisma durumlarini gozden gecirip birini etkisiz hale getirmeliyiz ki modelimiz dogru sonuc verebilsin.

Ve 0.in sagindaki sayilarin hassaisyeti gibi temel veri isleme hareketlerini yaptik

## THEORY

“Denetimli ogrenme classification ve siniflandirma olarak ikiye ayrilmaktadir Classification surekli olmayan yani kategorik verilerin cevaplarini dondururken continuous verileri ise regression kullanarak bir yaklasim yapariz.Bizim de elimizdeki veriye bakarak bu denetimli ogrenme siniflarinin ikisine de ihtiyacimiz oldugunu :Decision Tree , Random Forest Linear Regression gibi modelleri secmemiz gerektigini fikrini benimsedik [1]

”

## Machine Learning Algoritmalari

### “A1-Decision Tree

Bir karar ağacı, kararların bir ağacını oluşturan basit ve faydalı bir algoritmadır. Hem sınıflandırma hem de regresyon analizinde kullanılır. Bu yöntem, bir veri kümesini kökten yaprak düğümlere doğru büyüyen bir ağaç yapısına bölmeyi içerir. Her yaprak düğümünün durumu kök düğümüne bağlıdır ve yaprak düğüm koşulu sağlandığında büyür. Bu algoritmanın amacı, verilen veri kümesinden karar kurallarını öğrenerek gelecekteki değerleri tahmin etmektir. Sınıflandırmada, bir karar ağacı entropi ve bilgi kazancı kullanılarak oluşturulur. Entropi, ayrıca Shannon entropisi olarak da bilinir ve sonlu bir S kümesi için H(S) ile gösterilir. Verinin homojenliğini (belirsizlik veya rastgelelik) ölçer. Entropi çıkış değeri 0 ile 1 arasında değişir, burada 0 tamamen homojen veriyi temsil eder ve 1 eşit olarak bölünmüş veriyi temsil eder.

### A2-Random forest,

birçok karar ağacının bir araya gelerek bir model oluşturduğu bir ensemble (bir araya getirme) öğrenme tekniğidir. Her bir karar ağacı, rastgele özelliklerin bir alt kümesini kullanarak eğitilir ve ardından bu ağaçlar bir araya gelerek tahminler yapar. Bu yöntem, overfitting'i azaltır ve genel olarak daha iyi bir tahmin yapma yeteneğine sahiptir. Random forest, sınıflandırma ve regresyon problemleri için yaygın olarak kullanılır

### B1-Linear Regression

Doğrusal regresyon, bağımlı ve bağımsız değişkenler arasındaki istatistiksel ilişkiyi bulmak için kullanılır. Bağımsız değişken tahmin edici, bağımlı değişken ise yanıt değişkenidir ve hata, düzlemden uzaklık olarak ifade edilir. Amacı, verileri en iyi şekilde temsil eden bir çizgi çizmektir. Bu, bir değişkenin diğer bir değişkenle tam olarak ifade edilemediği durumlarda kullanılır, örneğin boy ve sıcaklık arasındaki ilişki gibi.

### B2-Ridge Regression

Ridge regresyon, lineer regresyonun bir türevidir ve genellikle aşırı uyum (overfitting) problemlerini ele almak için kullanılır. Temel amaç, katsayıların büyüklüğünü sınırlamak ve bu şekilde modelin karmaşıklığını azaltmaktır. Ridge regresyonu, toplam kare hatasına bir ceza terimi ekleyerek çalışır. Bu ceza terimi, katsayıların karelerinin toplamıdır. Bu sayede, model aşırı uyum yapma eğiliminde olan katsayıları kontrol altına alır.

### B3-Lasso Regression

Lasso regresyonu, ridge regresyonuna benzer bir şekilde lineer regresyonun bir türevidir ve aşırı uyumu azaltmak için kullanılır. Ancak, lasso regresyonu, katsayıların toplam mutlak değerini sınırlayarak çalışır. Bu, gereksiz özellikleri belirleme ve modelin karmaşıklığını azaltma eğilimindedir. Lasso regresyonu, toplam kare hatasına bir ceza terimi ekleyerek çalışır, ancak bu ceza terimi katsayıların mutlak değerlerinin toplamıdır.

“[1]

.

### Makine ogrenmesi Modelleri ile Tahmin

Olusturdugumuz isFree kolonunu target olarak belirledik ve geri kalan tum feature’lari x degiskeni olarak belirledik yaklasik P sutun L satirlik matris elde ettik

Ve bu veri setimizi train\_test\_split methodu kullanarak parcalayarak karistirdik [scikit]

A blue and yellow squares with black text

Description automatically generated

Daha sonrasinda bize lazim olan algoritmayi yani Decision Tree(4) yi uygulamak istedik.Decision Tree Teknik olarak basitce tablodaki gibi calisir (3)

A diagram of a triangle

Description automatically generated

Modeli uyguladiktan sonra bize verdigi accuracy rate =76.66 idi

## DOGRULUK GELISTIRME

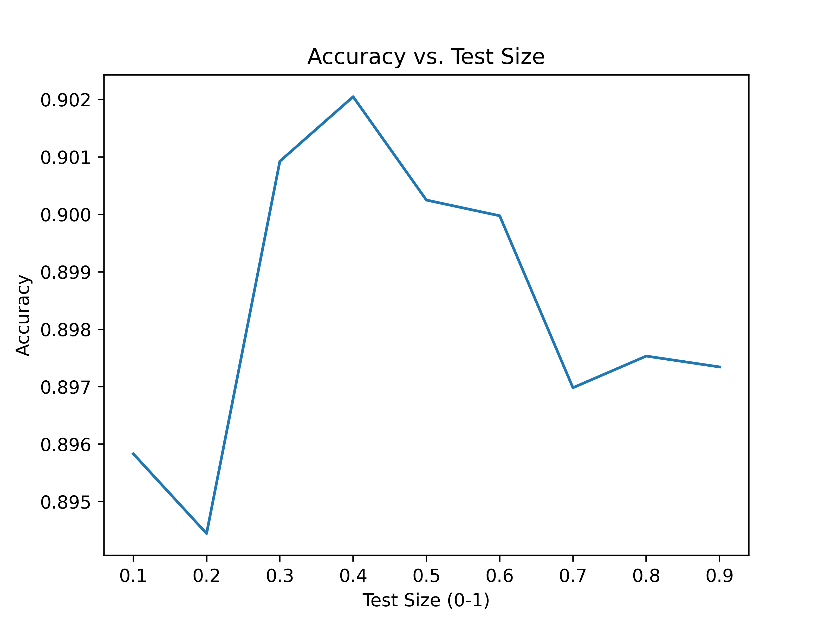
Kronolojik olarak gittigimizde accuracy rate decision tree icin =76.6 gibi gorece dusuk bir degerdi. Biz de train\_test\_split fonksiyonunun icindeki parametreler olan test size ve random state icin en optimal degerleri bulursak accuracy artirabilecegimizi dusunduk ve decision tree dahil bundan sonraki tum makine ogrenmesi algoritmalari icin bu ozelligi uyguladik.aradaki accuracy farkini yolladigimiz kaynak kod icinde inceleyebilirsiniz.

### SONUC

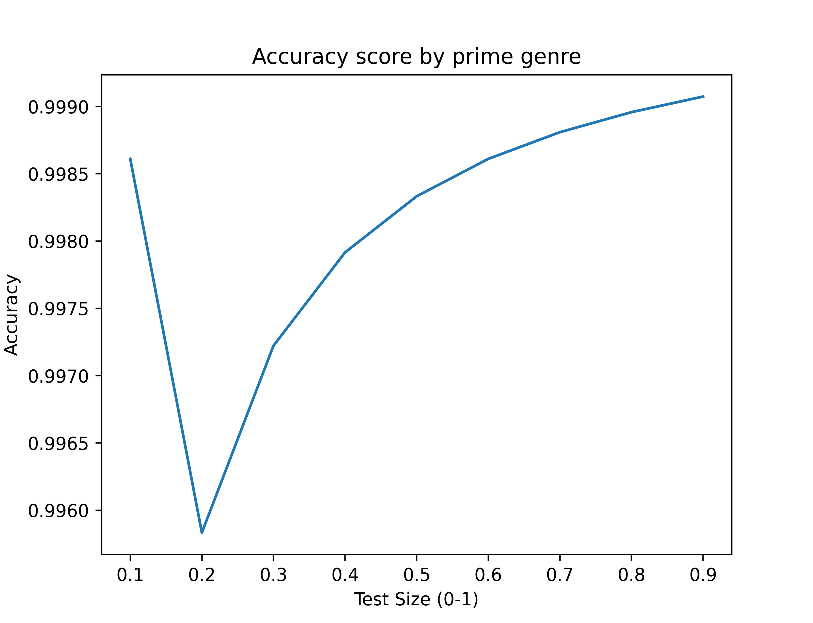
Modellerin tahminine gecmeden once hangi algoritmalari kullandigimizi soyledik,fakat verimizde target degiskenimizi degistirdigimiz senaryolar oldugu icin decision tree(2) kere kullandik.

|  |  |
| --- | --- |
| **ALGORITHM** | **Y\_TARGET** |
| Decision Tree | isFree |
| Random Forest | isFree |
|  |  |
| Decision Tree | Genre |
|  |  |
| Linear Regression | price |
| Ridge Regression | price |

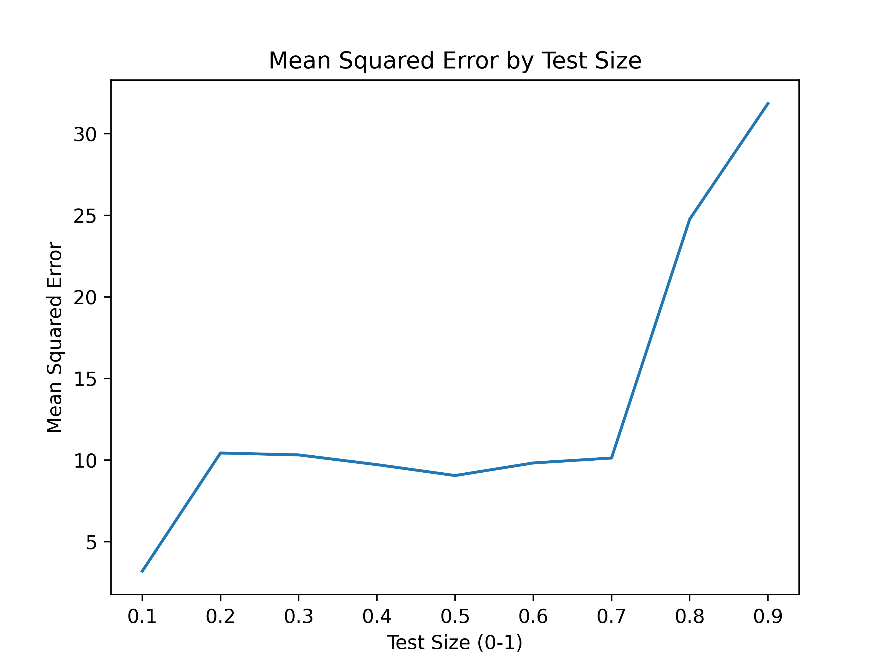
### 1-Decision Tree



### 2-Random Forest



### 3- Linear Regression

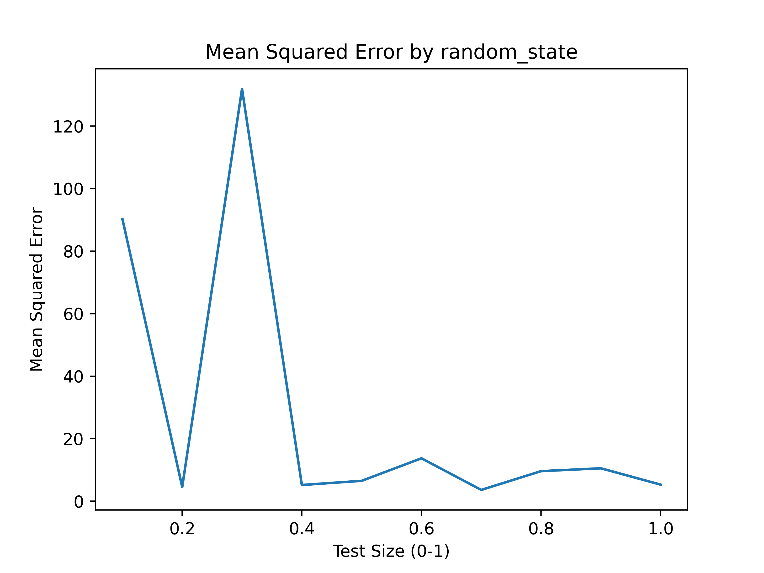


4-Random Forest (II-Genre)

A graph with blue lines

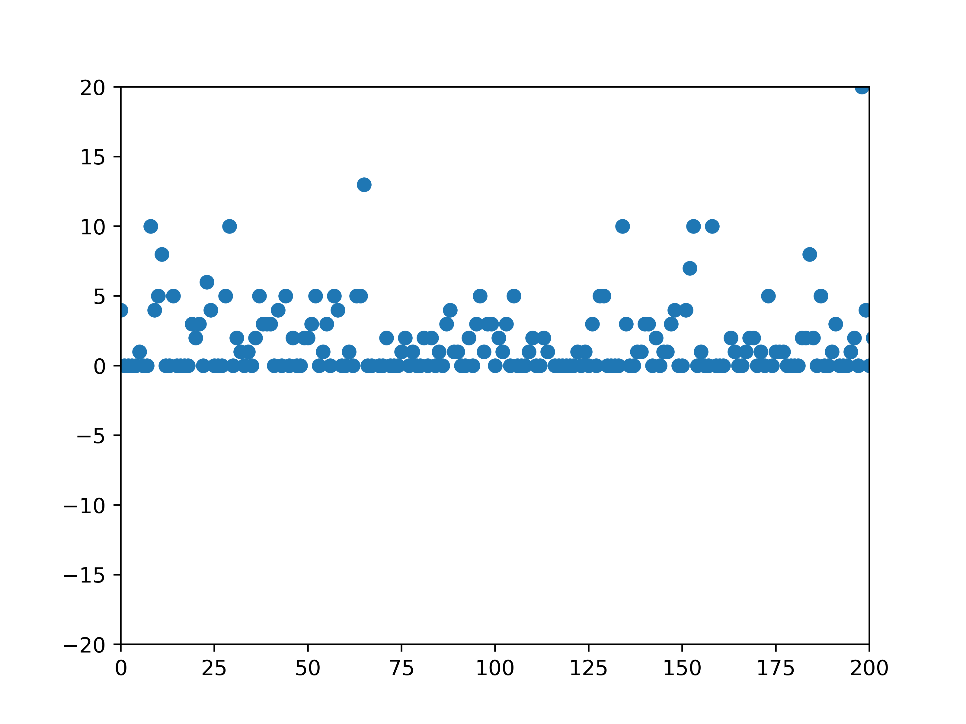
Description automatically generated

### 5-Ridge Regression



### DATA CHECKING

Verilerimizin dogruluk degerlerini bulduk. IsFree bagimli degiskenine bagli 2 algoritmamizda 98+ % dogruluk oranina ulasti. Overfitting ihtimaline karsi verilerimizi gorsellestirip bazi istatistiki degerlerine ulastik



Buradaki verilerin tamamina yakinin 0 ve cevresinde oldugunu ve bu durumun gercek degerleri etkileyip etkilemedigini anlamak istedik bunun icin boxplot kutularindan ve gercek hayatta da halihazirda kullandigimiz telefonumuzunun(iphone) app marketine girip ilk 50 veri icin baktik

Soldan saga olmak uzere sirasiyla boxplot grafiklerini

Plt1 : (0-1 USD)

Plt2 : (0-3 USD)

Plt3 : (0-7 USD)

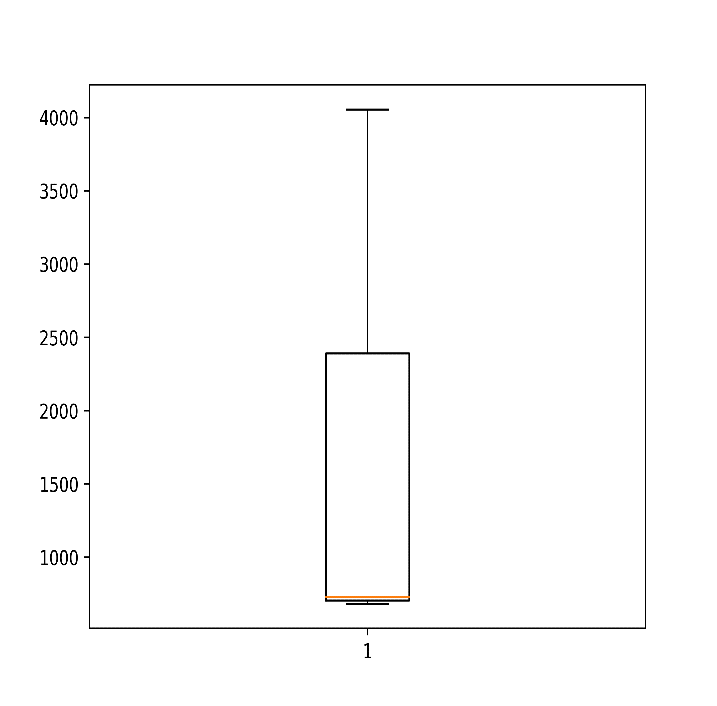
Plt4 : (all)

Bu degerleri alirken df[“price].sort\_values().value\_counts() methodunu kullandik

A diagram of a line

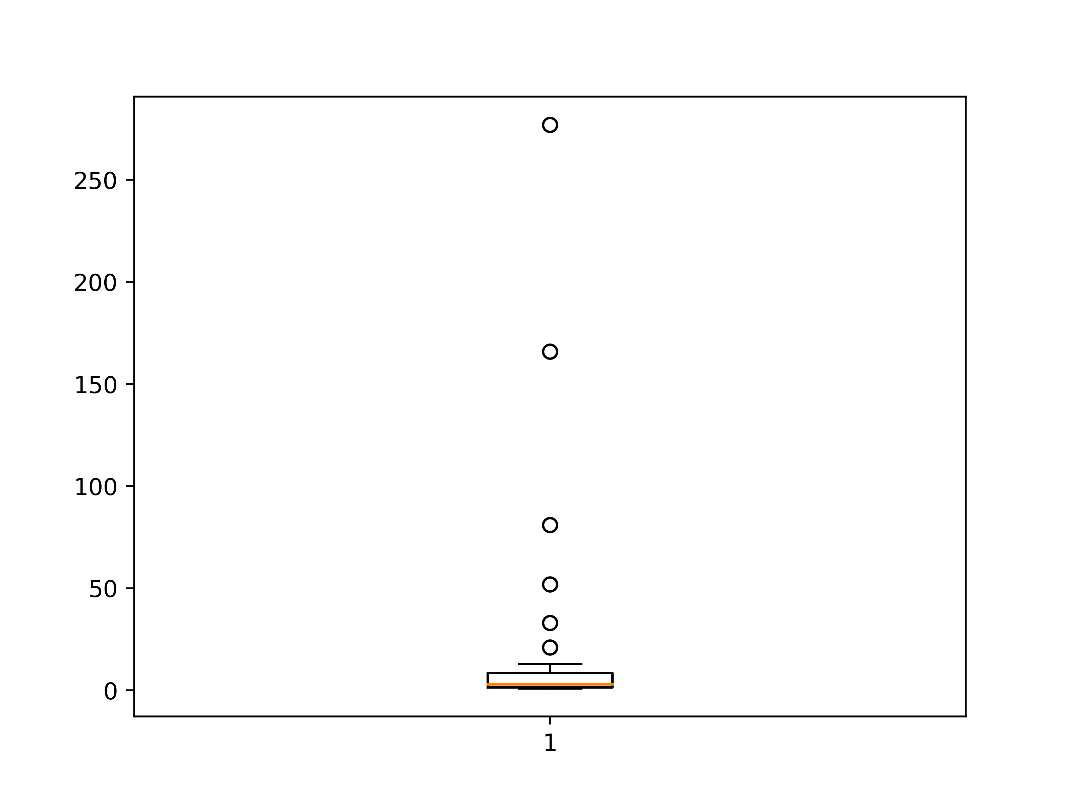
Description automatically generatedA screen shot of a computer

Description automatically generated

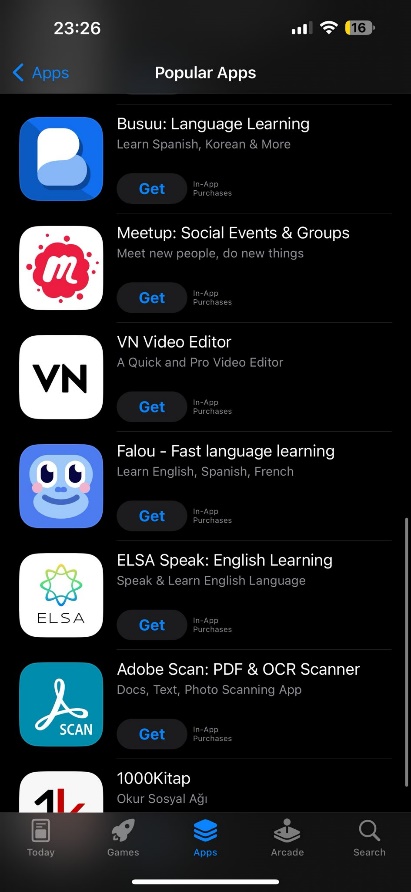
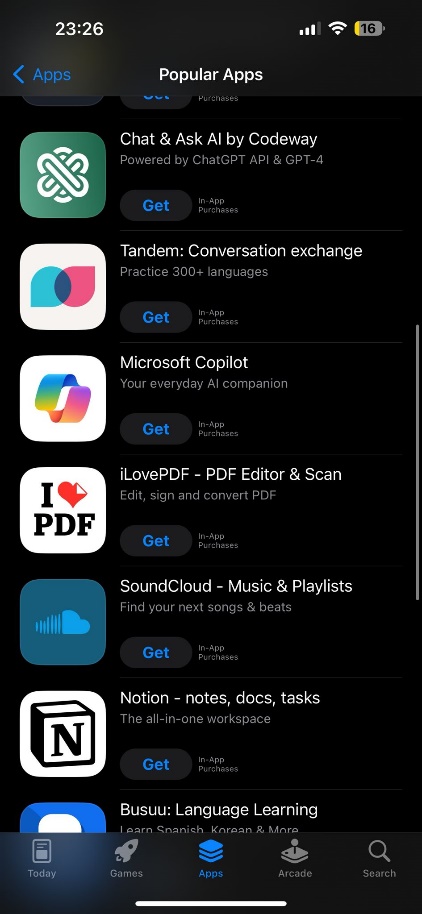
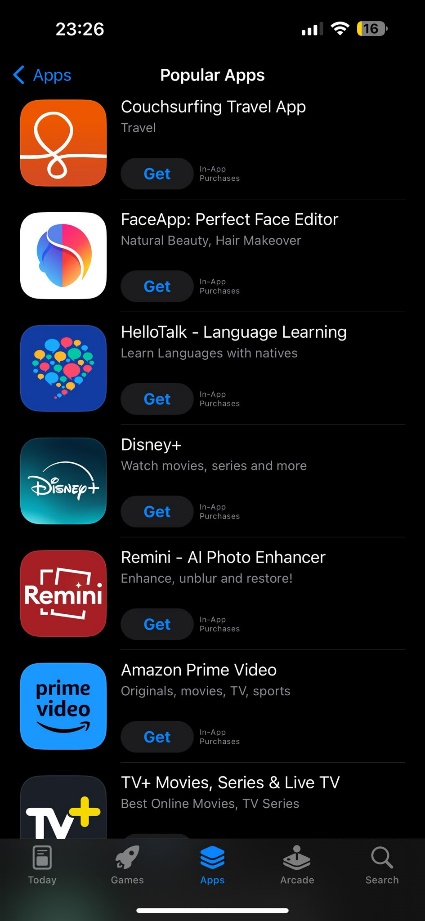
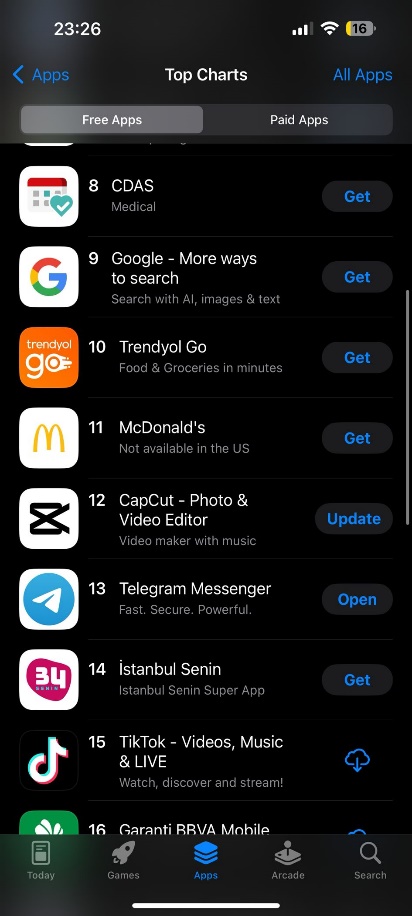
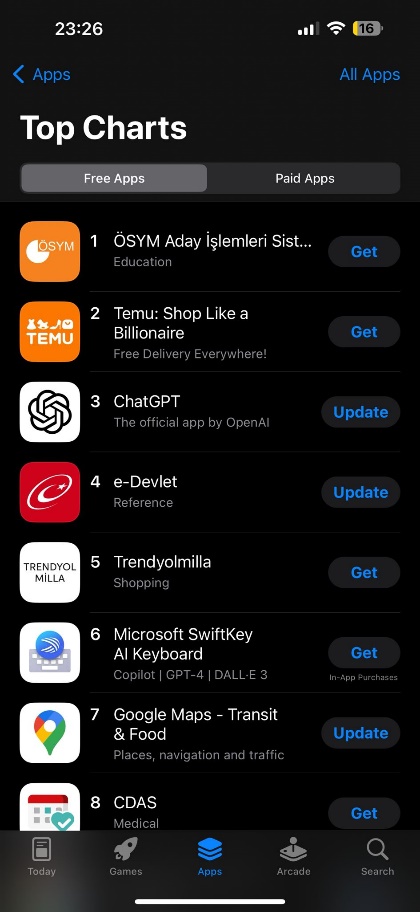
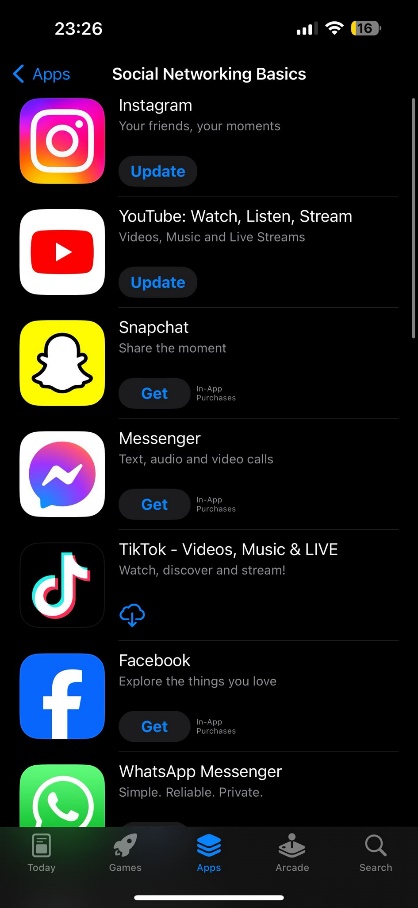
A screenshot of a computer

Description automatically generated

Karsilastigimiz sonuclara baktigimizda buldugumuz sonuclarin dogrulugunu gostermis olduk.bazi onerilere kulak verdikten sonra (0-2)USD araligi disindakileri kendi icerisinde bir boxplot haline getirdik.



2.kez verimizin dogrulugundan emin olduk.son adim olarak kendi kisisel telefonlarimizdan en popular top uygulamalarin ekran goruntusunu aldik.

Aldigimiz ekran goruntuleri her biri kendi kategorisinde en cok indirilen,genelde en popular uygulamalar. Buradaki hicbir uygulama ucretli degil, bu bizim verisetimiz ve scatterplot gorselimiz ile tamamen uyumlu.

### KAPANIS

Herkesin kimlik karti gibi telefon sahibi oldugu bir donemde telefonlar ve bunlarin icinde onlari kullanmamiza neden olan uygulamalarin analizini yapmak istedik. Bu oneri sistemini yapmamizdaki amac . Reklam sektorunun hangi uygulama sektorlerine yatirim yapmasi,Oyun gelistiricilerin ve start-up larin o anki moda olan uygulamalari takip edebilmesi caga uyum saglamasini hedefledik .Yapay zeka destekli bu modelimiz bize verdigi sonuclar ise tablo ile basarili oldugunu gostermektedir

|  |  |
| --- | --- |
| Algorithm Name - Target | Accuracy Score (%) |
| Decision Tree -isFree | **90,17** |
| Random Forest -isFree | 75,7 |
| Random Forest – prime\_Genre | **99,9** |

|  |  |
| --- | --- |
| Algorithm Name - Target | MSE |
| Linear Regression | **3.19** |
| Ridge Regression | 3.61 |

### Referances:

1-Javed, Noman & Hanif, Ambreen & Abdullah, Ali & Ullah, Shakir. (2019). Stock Price Forecast Using Recurrent Neural Network. ER -

2-POJON, Murat. *Using machine learning to predict student performance*. 2017. Master's Thesis.

3-Nayebi, M., Cho, H. & Ruhe, G. App store mining is not enough for app improvement. Empir Software Eng 23, 2764–2794 (2018). <https://doi.org/10.1007/s10664-018-9601-1>

### Links

[train\_test\_split — scikit-learn 1.5.0 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html)

[ppcic-ml-modelselection.ipynb - Colab (google.com)](https://colab.research.google.com/github/MLRG-CEFET-RJ/ml-class/blob/master/ppcic_ml_modelselection.ipynb#scrollTo=p_Wx4vawBl9n)

[Evergreen\_16x9\_Python\_UGC\_20s\_la.EN (youtube.com)](https://www.youtube.com/watch?v=k1liepnnSjA&ab_channel=TirendazAkademi)

[ChatGPT](https://chatgpt.com/)

# UYGULAMALAR ILE YAPAY ZEKA FINAL PROJE TESLIM RAPORU

### APPLE STORE UYGULAMA MARKET ANALIZI VE ONERI SISTEMLERI

# Bilgilendirmeler:

# 1-Inceledigimiz makalelerde basliklarin turkce halini tam olarak bulamadik abstract=oz gibi

2- Ayni sekilde kullandigimiz makine ogrenmesi ve yazilim terimlerini anlamlarini kaybetmemesi icin her defasinda turkceye cevirmedik process=islem , properties,features farki gibi anlasilmasi gereken konular icin.

3-Bu sunumu hazirladigimiz bilgisayar US klavyeye sahip oldugu icin turkce kelimeler kullanilamamistir.

# MURAT FERHAT DERYA

2020280014

# BARISCAN TOSYALI

2021280050