Yapay Zeka Ödev 1
Genetik Algoritma İle Metin Sınıflandırma

Kürşat Kömürcü 18014038

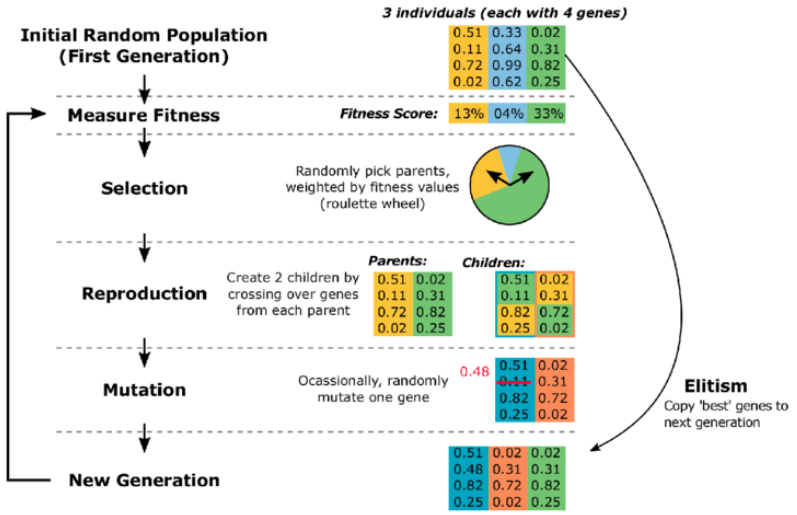
Video Link:

https://www.youtube.com/watch?v=egLfjYHu69M

Genetik Algoritma Nedir?

Genetik algoritma evrim mekanizmasına (en iyinin yaşaması) dayanan bir algoritmadır. Amaç, uygunluk fonksiyonunun (fitness function) maksimizasyon / minimizasyon optimizasyonudur. K adet rastgele üretilmiş durum ile başlar. Durumlara kromozom, durumlar kümesine ise popülasyon denir. Yeni kromozom ve durumların üretilmesinde seçme, yeniden kopyalama ve değiştirme operatörleri kullanılır.

Seçme (Selection): Bir sonraki nesli üretecek kromozomların seçilmesidir.
Yeniden kopyalama (Reproduce) : Yeni çözümler üretmek için çaprazlama (crossover) işlemi yapılır.
Değiştirme (Mutation) : Kromozom'un bazı değerlerini rasgele değiştirir.



Genetik Algoritmanın Performansını Etkileyen Faktörler:

Popülasyon büyüklüğü / Kromozom sayısı: Kromozom sayısını arttırmak çalışma zamanını arttırırken, azaltmak da kromozom çeşitliliğini yok eder.
Mutasyon Oranı: Kromozomlar birbirine benzemeye başladığında hala çözüm noktalarının uzağında bulunuyorsa mutasyon işlemi GA'nın sıkıştığı yerden (tüm kromozomlar aynı platoda) kurtulmak için tek yoludur. Ancak yüksek bir değer vermek GA'ın kararlı bir noktaya ulaşmasını engelleyecektir.
Kaç Noktalı Çaprazlama Yapılacağı: Normal olarak çaprazlama tek noktada gerçekleştirilmekle beraber yapılan araştırmalar bazı problemlerde çok noktalı çaprazlamanın çok yararlı olduğunu göstermiştir.
Çaprazlamanın sonucu elde edilen bireylerin nasıl değerlendirileceği: Elde edilen iki bireyin birden kullanılıp kullanılamayacağı.
Durum kodlanmasının nasıl yapıldığı: Bir parametrenin doğrusal ya da logaritmik kodlanması GA'nın performansında önemli bir farka yol açabilir.

Başarı değerlendirmesinin nasıl yapıldığı: Akıllıca yazılmamış bir değerlendirme işlevi, çalışma zamanını uzatabileceği gibi çözüme hiçbir zaman ulaşmamasına da neden olabilir.

Veri Ön İşleme:

Öncelikle veri setindeki cümleler kelimelerine ayrılmıştır. Daha sonra ise stopwords denilen gereksiz kelimeler ve noktalama işaretleri çıkarılarak kelime havuzu oluşturulmuştur. Bu kelimeler kullanılarak popülasyonu oluşturacak bireylerin oluşturulması amaçlanmaktadır.

```
from nltk.corpus import stopwords
"""
Cümleler kelimelere ayrildiktan sonra stopwords word denilen gereksiz kelimeler cümlelerden atilmiştir.
Daha sonra ise her kelime tek bir listede toplanarak veri setinin kelime havuzu oluşturulmuştur.
"""
stop_words = set(stopwords.words('english'))
word_pool = x
word_pool = word_pool.ravel()

word_pool_tokens = []
for i in word_pool:
    token = nltk.word_tokenize(i)
    for word in token:
        if (word not in stop_words) and (word.isalnum()):
            word_pool_tokens.append(word)

# print(word_pool_tokens) # kelime havuzu
word_pool_tokens
```

✓ 0.8s

```
Output exceeds the size limit. Open the full output data in a text editor
['Go',
 'jurong',
 'point',
 'crazy',
 'Available',
 'bugis',
 'n',
 'great',
 'world',
 'la',
 'e',
 'buffet',
 'Cine',
 'got',
 'amore',
 'wat',
 'Ok',
 'lar',
 'Joking',
 'wif',
 'u',
 'oni',
```

Genetik Algoritma İle Metin Sınıflandırma:

Bu çalışmada Genetik Algoritma kullanılarak beş farklı veri seti üzerinde binary metin sınıflandırma yapılmıştır ve hiperparametreler incelenmiştir. Sınıflandırma yapılırken ek bir sınıflandırma algoritması kullanılmamıştır. Kullanılan yöntem şu sekildedir:

1- Oluşturulan bireye rastgele kelimeler atanır.
2- Bireyin ilk yarısının olumlu veriler için kelimeleri tuttuğu, ikinci yarısının ise olumsuz veriler için kelimeleri tuttuğu varsayılır. Örneğin 100 gen içeren bir bireyin ilk yarısının “güzel, harika, müthiş” gibi kelimeler içermesi beklenirken son yarısının ise “berbat, kötü, beğenmedim” gibi kelimeler içermesi beklenir.
3- Her adımda genetik algoritmanın seçme, yeniden kopyalama, değiştirme özellikleri kullanılarak hedeflenen bireye ulaşılmaya çalışılır.

İyilik (Fitness) Fonksiyonu:

İyilik fonksiyonu oluşturulurken öncelikle birey ikiye bölünmüştür. Daha sonra veri setindeki cümleler kelimelere ayrılmıştır. Ayrılan kelimeler sadece bireyin ilk yarısında ise olumlu, sadece bireyin ikinci yarısında ise olumsuz sayılmıştır. Daha sonra ise olumlu ve olumsuz sayılarından fazla olanı toplam değişkenine eklenerek fonksiyon toplam değişkenini döndürmüştür.

```
def fitness_function(individual, x):
    toplam = 0
    pos_arr = individual[:len(individual) // 2]
    neg_arr = individual[len(individual) // 2:]

    for i in range(len(x)):
        count_pos = 0
        count_neg = 0
        for j in x[i]:
            tokens = nltk.word_tokenize(j)

            for token in tokens:
                if (token in pos_arr) and (token not in neg_arr):
                    count_pos += 1
                if (token in neg_arr) and (token not in pos_arr):
                    count_neg += 1

            if count_pos > count_neg:
                toplam += count_pos

            elif count_neg > count_pos:
                toplam += count_neg

    return toplam
```

Veri Setleri Ve Elde Edilen Sonuçlar

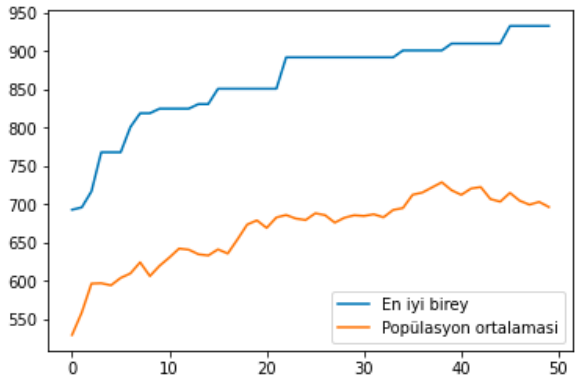
Algoritma eğitilirken iterasyon sayısı 50, populasyon büyüklüğü 100, populasyondaki her bireyin uzunluğu (N) , crossover olasılığı 0.7, mutasyon olasılığı ise 0.1 seçilmiştir. Her verisetinin rastgele 500 satırı alınarak algoritma eğitilmiştir.

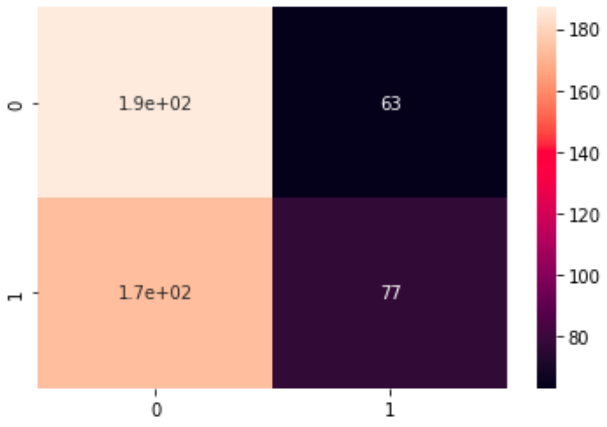
Veriseti 1 – Sms Spam Collection Dataset:

Veriseti spam olan ve olmayan smsleri içermektedir. Spam smsler 1 ile spam olmayan smsler ise 0 ile etiketlenmiştir.

Üretilen en iyi kelime listesi aşağıdaki gibidir:

['16' 'yr' 'eyes' 'CaRE' 'week' 'reassuring' 'NTT' 'anything' 'Mins' 'txt'
'Shijutta' 'U' 'sis' 'Hottest' 'operator' 'dis' 'Txt' 'She' 'dearly'
'Claim' 'maximize' 'receive' 'End' 'Evr' 'town' '09066364349' 'backdoor'
'abt' 'Call' 'Call' 'Actually' 'contact' 'number' 'everyones' 'home'
'comp' 'sir' 'PICS' 'MSG' 'SMS' 'Call' '08719181503' 'order' 'Hmmm' 'K'
'2003' 'send' 'Well' 'C' 'amp' 'anybody' 'go' 'I' 'Cup' 'js' 'want'
'REWARD' 'one' 'customer' '0870' 'Chance' 'REMINDER' '12hrs'
'08714712394' 'No' 'word' 'gt' '2' 'claim' 'Double' 'AGE' 'Holiday'
'like' 'mobile' 'answering' 'know' 'mins' '4' 'someone' 'Gr8' 'Text'
'coming' 'Sony' 'Logon' 'mates' 'u' 'call' '2' 'get' '1' 'ur' 'Nokia'
'phone' '2' 'enjoy' 'da' 'cause' 'luv' 'FREE' 'please']

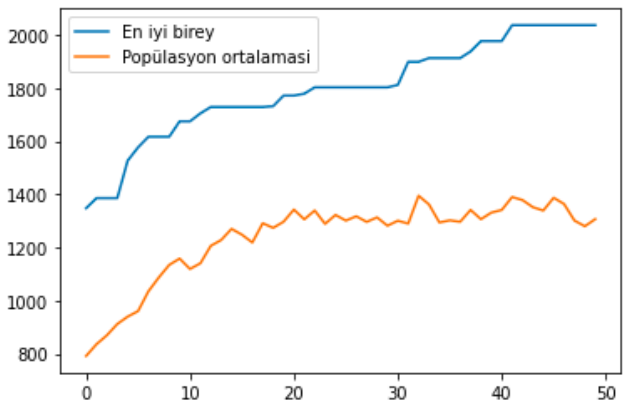


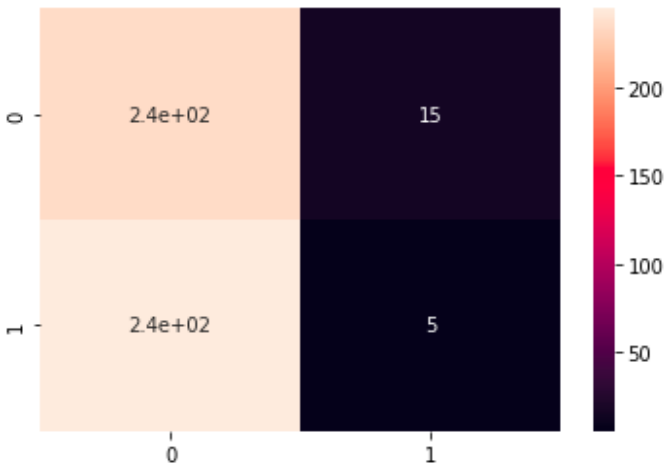


Veriseti 2 – Beyazperde Dataset:
Veriseti Türkçe film yorumları içermektedir. Olumlu yorumlar 1 ile olumsuz yorumlar ise 0 ile etiketlenmiştir.

Üretilen en iyi kelime listesi aşağıdaki gibidir:

['basariyla' 'c1kti' 'vasat' 'vermeleri' 'batiya' 'guzel' 'gidilmemesi'
 'sinemada' 'islam' 'olmasi' 'soundtracklar' 'sinemada' 'izleyenlerdenim'
 'olacak' 'ne' 'kademe' 'eden' 'deil' 'yasattigi' 'sinir' 'kalkmasinlar'
 'zamanki' 'aksiyonla' 'imaji' 'bolumde' 'hayal' 'geldi' 'kendimden'
 'izlenebielecek' 'altin' 'arkadaslar' 'filmden' 'ilerleyen' 'oscar'
 'derece' 'herzaman' 'almasina' 'seyler' 'toparlayamadi' 'yanlis'
 'heralde' 'diyemicem' 'dozunda' 'bulamiyorsaniz' 'muzikleri' 'tanimi'
 'yazan' 'yutana' 'izlediyseniz' 'degildir' 'kadar' 'insan' 'bu'
 'oldugunu' 'ya' 'kadar' 'bir' 'izlenmesi' 'filme' 'ama' 'devam' 'olarak'
 'durumu' 'film' 'sadece' 'sifir' 'konusu' 'olarak' 'sadece' 'bu' 'dogru'
 'bir' 'olabilirdi' 'filmin' 'iyi' 'var' 'filmi' 'mi' 'filmin' 'de' 'oldu'
 'filmlerden' 'foster' 'kesin' 'de' 'bitmesini' 'film' 'filmdi' 'bi' 'iyi'
 'bir' 'bunu' 'sonu' 'kelimeyle' 'kadar' 'tek' 'ben' 'en' 'bence' 'ne']

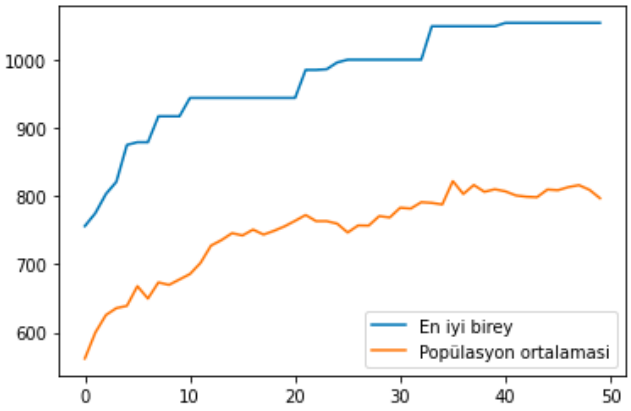


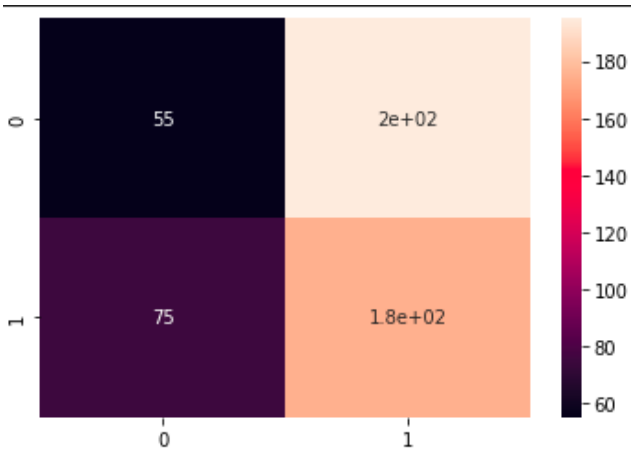


Veriseti 3 – Financial Sentiment Analysis Dataset:
Veriseti finans haberleri içermektedir. Olumlu haberler 1 ile olumsuz haberler ise 0 ile etiketlenmiştir. Nötr haberler çıkarılmıştır.

Üretilen en iyi kelime listesi aşağıdaki gibidir:

['mln' 'hope' 'second' 'China' 'group' 'period' 'mln' 'The' 'lower' 'year'
 'group' 'share' 'loss' 'period' 'FB' 'upgrade' 'The' 'consolidating'
 'EUR' '2008' 'Mongolia' 'Cramo' 'https' '2010' 'complements' 'compared'
 'million' '2009' 'million' 'grew' 'Finnish' 'circuit' 'Dutch' 'models'
 'said' 'period' 'AAPL' 'pretax' 'quarter' 'In' 'net' 'Oyj' 'mn'
 'received' 'sales' 'negotiations' 'largest' 'Lee' 'steel' 'second'
 'acquisitions' 'still' 'AAPL' 'EUR46m' 'first' 'Kone' 'EUR0' 'Tough'
 'euro' 'declined' 'plan' 'money' 'machinery' 'euros' 'stop' 'Ad'
 'Metsaliitto' 'prospects' 'said' 'Results' 'looking' 'USD' 'drug' 'GLD'
 'make' 'positive' 'Department' 'Diageo' 'costs' 'investors' 'third' '6'
 'LONG' 'maker' 'half' 'Agricultural' 'information' 'America' 'love' 'get'
 'EPS' 'administration' '45' 'XLF' 'rivals' 'Baltic' 'orders' 'Music'
 'HLF' '2008']

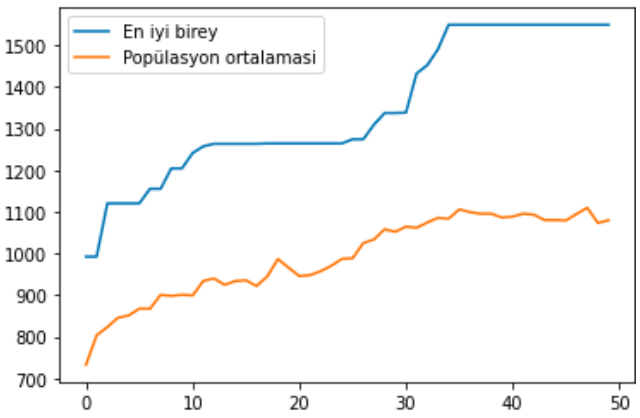


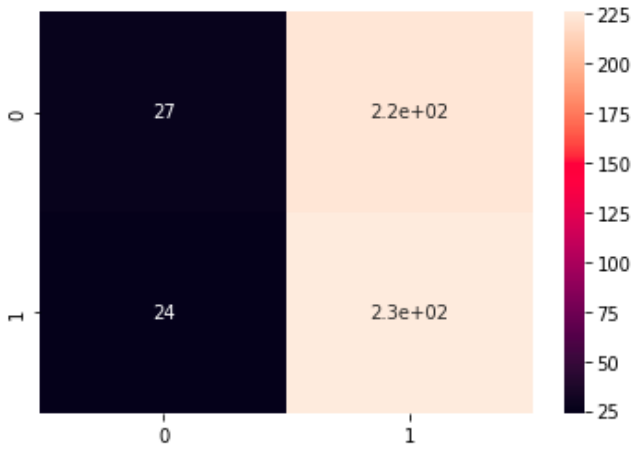


Veriseti 4 – Mağaza Yorumları Duygu Analizi Dataset:
Veriseti mağaza yorumları içermektedir. Olumlu yorumlar 1 ile olumsuz yorumlar ise 0 ile etiketlenmiştir. Tarafsız yorumlar çıkarılmıştır.

Üretilen en iyi kelime listesi aşağıdaki gibidir:

['antenli' 'halıda' 'aldım' 'yok' 'bana' 'gönderdiler' 'işe' 'sözde'
 'için' 'çıkmaya' 'ne' 'gün' 'yorumları' 'Harika' 'Ürün' 'TÜM' 'biraz'
 'yönleri' 'daha' 'tam' 'Fiyat' 'çok' 'gücü' 'ucu' 'guzel' 'ürün'
 'kesinlikle' 'TAVUĞUN' 'iyi' 'üzerinde' 'da' 'memnun' 'var' 'bu' 'geldi'
 'bir' 'değil' 'kontrol' 'olduğu' 'tavsiye' 'cezvede' 'çok' 'telefonda'
 'hallettiler' '8' 'güzel' 'tuşuna' 'diğer' 'seçeneği' 'Kısmı' 'ederim'
 'DNS' 'isteyenler' 'kimseye' 'tiz' 'giriyor' 'ay' 'paketlenmis' 'orda'
 'Sorularımız' '2' 'nisan' 'gerek' 'yılın' 'ÇÜNKÜ' 'aparatida'
 'götürdügum' 'geliyor' 'Ürünümün' 'gordugumde' 'sorunsuz' 'almıştım' 'HD'
 'kablosunun' 'yenisi' 'kullanıyorum' 'özellikle' 'falan' 'kulaklıktaki'
 'kaldırmamalısınız' 'şekilde' 'hızlı' 'ben' 'tutacak' 'İndirime'
 'alırken' 'ÜRÜNÜ' 'Aşırı' 'düşünüyorum' 'küçük' 'degistirdiler' 'ise'
 'üretim' 'aletlerine' 'bildiğimiz' 'aynı' 'dakika' 'kullanmaya'
 'alabilirsiniz' 'torbasız']

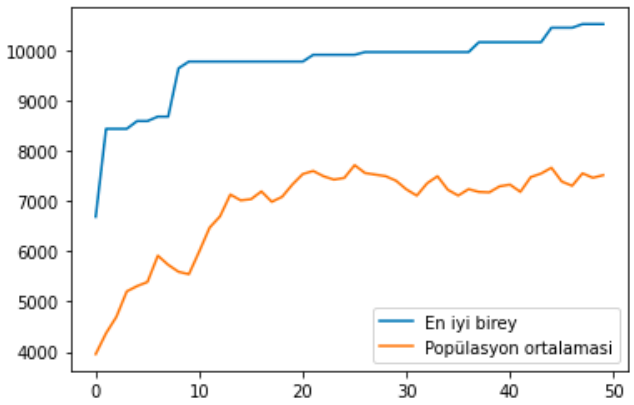


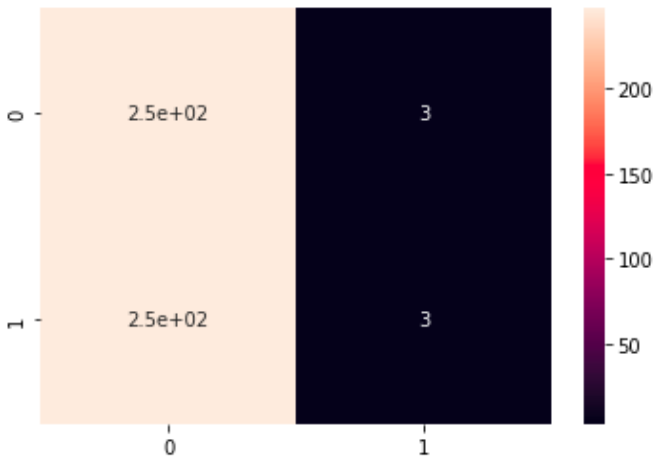


Veriseti 5 – Movie Dataset:
Veriseti film yorumları içermektedir. Olumlu yorumlar 1 ile olumsuz yorumlar ise 0 ile etiketlenmiştir.

Üretilen en iyi kelime listesi aşağıdaki gibidir:

['stopping' 'long' 'watched' 'Stormare' 'And' 'really' 'great' 'sort'
 'Show' 'Barrymore' 'pretty' 'adversaries' 'predictable' 'real' 'clear'
 'Minus' 'mantis' 'Richard' 'hell' 'single' 'often' 'done' 'films'
 'leaves' 'Coonhound' 'resolution' 'films' 'producing' 'opens' 'directors'
 'What' 'Johnny' 'nearly' 'strange' 'fugitive' 'trip' 'Amy' 'expensive'
 'assignments' 'NOT' 'heard' 'prove' 'without' 'english' 'right' 'mother'
 'science' 'explore' 'conveniently' 'takes' 'Woman' 'live' 'way' 'br'
 'film' 'film' 'The' 'makes' 'one' 'It' 'dealing' 'supporting' 'end'
 'film' 'routine' 'atrocious' 'delivery' 'She' 'girl' 'good' 'think'
 'probably' 'scenes' 'gore' 'tale' 'I' 'brownstone' 'nerd' 'fatally'
 'Balsam' 'found' 'good' 'movie' 'many' 'characters' 'STEP' 'watch' 'I'
 'MANTIS' 'br' 'fall' 'director' 'graphic' 'like' 'school' 'something'
 'But' 'This' 'would' 'make']



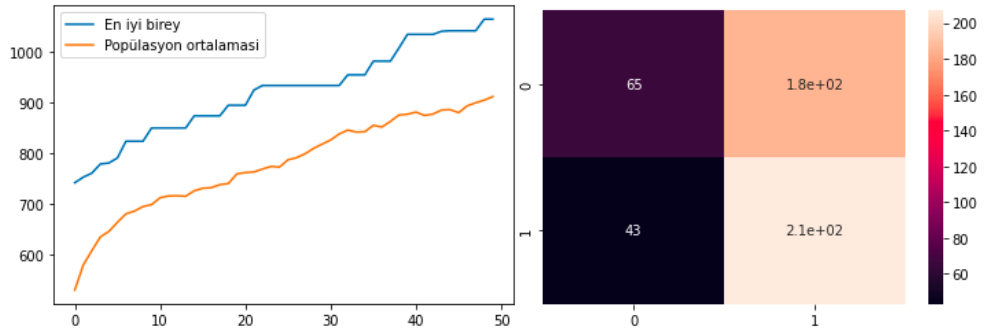


Populasyon Büyüklüğü Ve Mutasyon Oranının Fitness Fonksiyonu İle İlişkisi

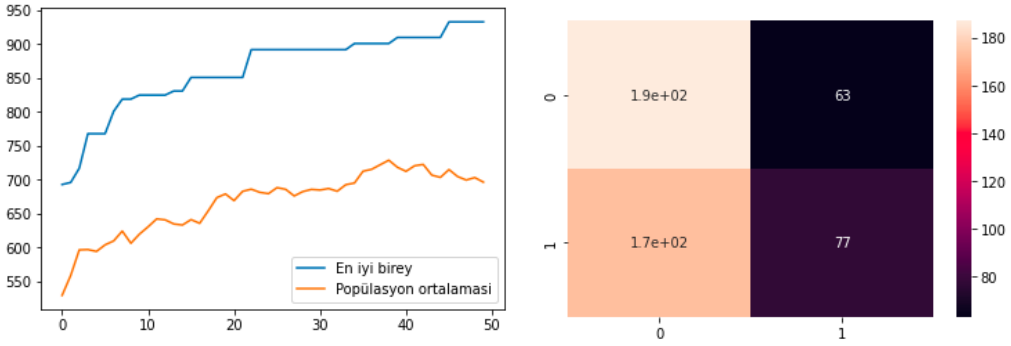
Sms Spam Collection Veriseti ile denemeler yapılmıştır.

Populasyon büyüklüğü 100 iken:

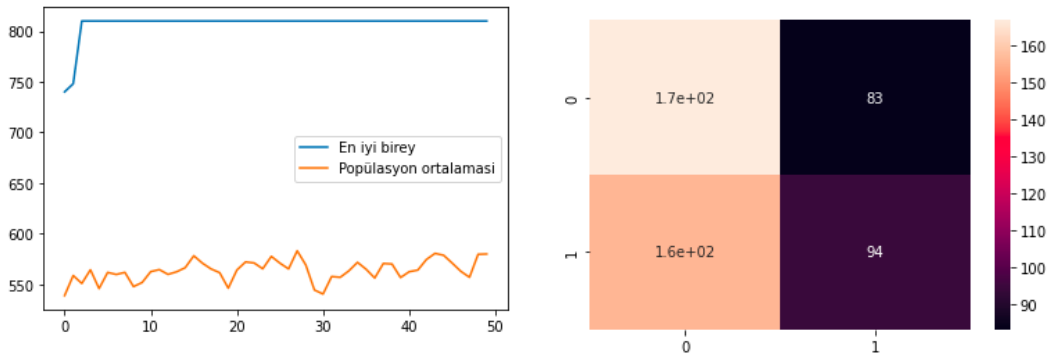
0.05 mutasyon oranı için;



0.1 mutasyon oranı için;

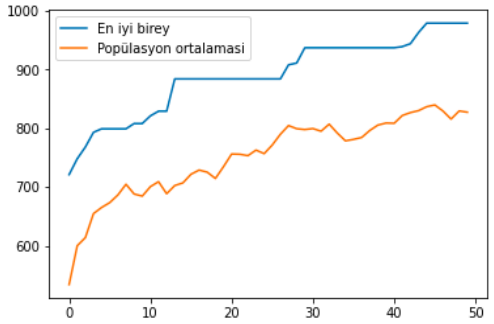


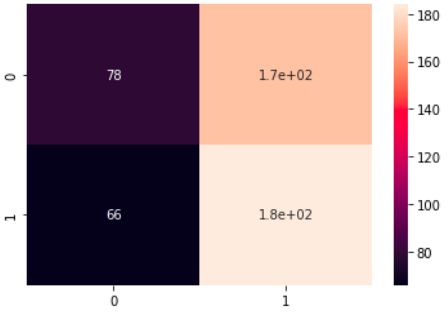
0.3 mutasyon oranı için;



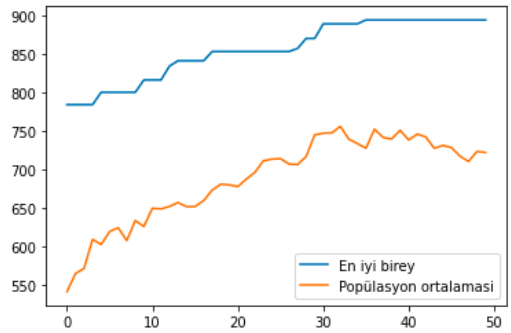
Populasyon büyüklüğü 50 iken:

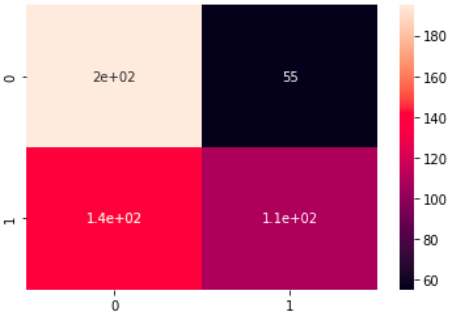
0.05 mutasyon oranı için;



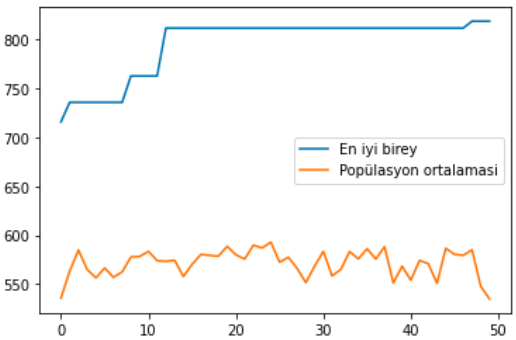


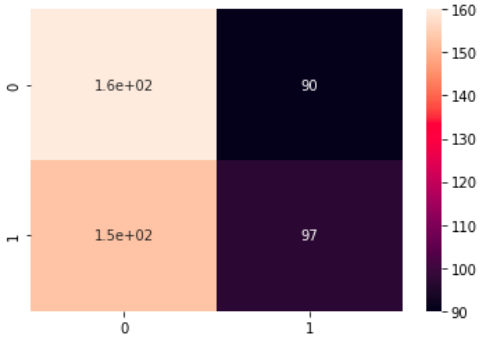
0.1 mutasyon oranı için;





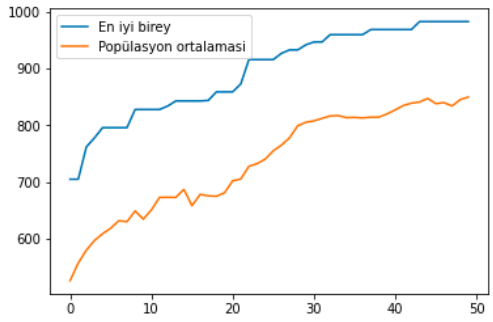
0.3 mutasyon oranı için;

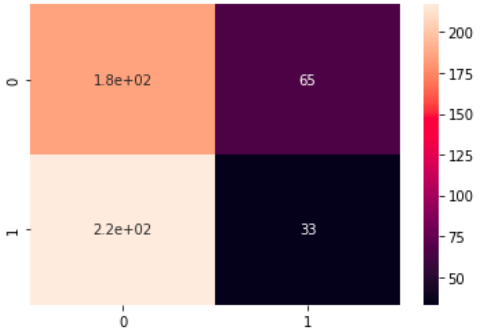




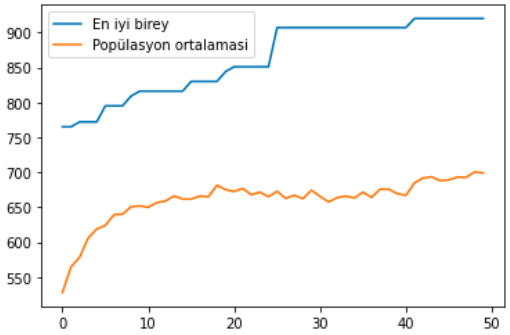
Populasyon büyüklüğü 200 iken:

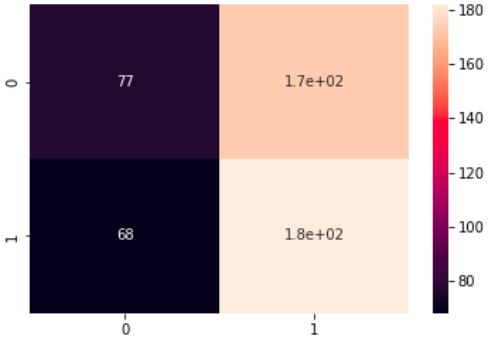
0.05 mutasyon oranı için;



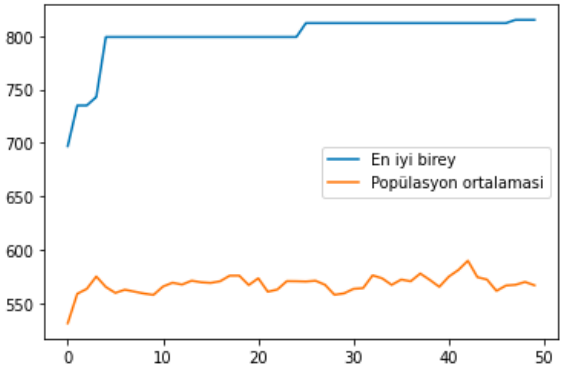


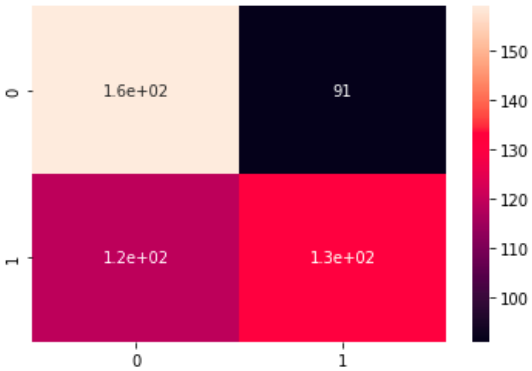
0.1 mutasyon oranı için;





0.3 mutasyon oranı için;





	0.05	0.1	0.3
50	979	894	819
100	1064	933	810
200	983	920	815

Değerlendirme Ve Yorumlar:

Mutasyon seviyesi yükseldikçe fitness fonksiyonunun değeri düşmektedir. Çünkü ilgili bireyde çok fazla kelime değiştiği zaman crossover işleminin bir önemi kalmamaktadır. Mutasyon çok düşük olursa da fitness fonksiyonu belli bir seviyede kaldığı zaman kelimeler çok az değişeceği için aynı seviyede kalmaya devam edebilir. Yukarıdaki tabloda birinci verisetimiz olan Sms Spam Collection Veriseti ile denemeler yapılmıştır ve en yüksek fitness fonksiyonunun 0.05 mutasyon oranında olduğu görülmüştür.

Denemelerde popülasyon sayısı 100 olduğu zaman en yüksek fitness fonksiyonu değerine ulaşılmıştır. Popülasyon sayısı az olduğunda çeşitlilikte az olacağı için başarı oranı düşebilir.

Ayrıca başarı oranını arttırmak için iterasyon sayısı arttırılabilir, bireylerin içerdiği listenin uzunluğu arttırılabilir, daha iyi veri ön işleme yöntemleri kullanılabilir ya da fitness fonksiyonu bir makine öğrenmesi algoritması ile yazılıp genetik algoritma ile bu makine öğrenmesi algoritmasının doğruluk oranı iyileştirilebilir.

Yararlanılan Kaynaklar Ve Verisetleri:

1- https://sites.google.com/view/mfatihamasyali/yapay-zeka
2- https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset
3- https://www.kaggle.com/datasets/sbhatti/financial-sentiment-analysis
4- https://www.kaggle.com/datasets/shub99/sentiment-analysis-data
5- https://www.kaggle.com/datasets/ozcan15/turkish-sentiment-analysis-data-beyazperdecom?select=train.csv
6- https://www.kaggle.com/datasets/burhanbilenn/duygu-analizi-icin-urun-yorumlari
7- https://medium.com/@kocelifk/genetik-algoritma-nedir-a79414e96e22