**1-** In the assignment, first, it was requested to create a Dataset. It was planned to collect a certain number of text data from the topics "Brain Injury", "Type 2 Diabetics" and "Kidney Failure" and teach it to the machines.

The file named "dataset\_maker.py" aimed to create a dataset with data retrieved from the internet.

General structure of Dataset (In Spyder IPython Console):

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

During the homework, the number of elements of the dataset was manipulated (to show its effect on accuracy). But as a constant, we ended up with a dataset creation model that pulls around 500-600 pieces of data.

metin, ekran görüntüsü, yazılım, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure 1 Example :Dataset that has 579 elements with 4 keys

**FULL CODE AND EXPLANATION:**

from paperscraper.pubmed import get\_pubmed\_papers

import pandas as pd

import time

The "paperscraper" module, which the teacher offered as an alternative, was used in the assignment. If it is not installed on your computer, don’t forget to download it with: “pip install paperscraper” command.

The aim was to build a DataFrame structure with the pandas module and to provide time between requests with the time module.

def collect\_data(keywords):

Inside our Python file, a function called collect\_data(keywords) has been created. Our aim with this function will be to extract the desired titles from the desired number of articles by using the keywords in the received keyword list.

    data = []

    for keyword in keywords:

        try:

            papers = get\_pubmed\_papers(keyword, max\_results=200)

        except Exception as e:

            print(f"Error fetching papers for {keyword}: {e}")

            continue

First of all, we create a list called "data" in the function where we will keep the paper values. Then, we start browsing through our keyword list. We start the operations we will use for each keyword by using one of the functions of the paperscaper module. The first is get\_pubmed\_papers(keyword, max\_results=200). The "max\_result" parameter is important here, we will determine the number of elements in the dataset accordingly. This function will return us a DataFrame data structure.

        for a in range(len(papers)):

            try:

                if papers['title'][a] == None or papers['authors'][a] == None or papers['abstract'][a] == None:

                    continue

                paper\_data = {

                    'Title': papers['title'][a],

                    'Authors': papers['authors'][a],

                    'Abstract': papers['abstract'][a],

                    "Label for HW": keyword

                }

                data.append(paper\_data)

            except Exception as e:

                print(

                    f"Error processing paper data for {keyword} (Index: {a}): {e}")

            time.sleep(0.0005)  # Add a delay of milisecond between requests

By browsing through the "papers" structure, which we returned as a DataFrame structure above, we get the values we want in a second for loop: Title, authors, Abstract and the Class Label we added ourselves. After obtaining the paper\_data we want, we add this data to the list we created called data.

    # Convert data to DataFrame

    df = pd.DataFrame(data)

    df.to\_csv('dataset.csv', index=False)

search\_query = ["brain injury", "type 2 diabetics", "kidney disease"]

collect\_datas(search\_query)

At the end of the function, the .csv file is appended.

At the end of the document, the list containing the last keywords is put into the function as a parameter and executed.

**k-neighbors**

YAŞANANLAR:

2- normal classification:

Number of elements in dataset: 58:

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Number of dataset is 580:

3- n sayısını değiştirince:

N sayısını değiştirmek ne ne yapar ne eder vb.

metin, ekran görüntüsü, yazı tipi, menü içeren bir resim

Açıklama otomatik olarak oluşturuldu metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Naïve Bayes**

**2-**

**Element sayısı 58 iken:**

**metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**metin, yazı tipi, ekran görüntüsü, tasarım içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**3-**

Alpha parametresi ne yapar ne eder vb. Açıklama

Ilki 58 ikinci 580 iken

**metin, yazı tipi, ekran görüntüsü, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Figure 2ss

**metin, yazı tipi, ekran görüntüsü, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**Random Forest**

**2-**

**metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**3-**

**metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**SVM**

**Element sayısı 58 iken:**

**2-** **metin, yazı tipi, ekran görüntüsü, tipografi içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**3- C parametresi arttıkça ne olur tanımı vb.**

**metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu**

**metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu**

* Element sayısı 58 iken, Datasetteki eleman sayısı az olduğu için, overfitting olmaya meyili bir modelimiz olur. Tüm algoritmalarda train oldukça yüksek(hatta %100) çıkarken, test değerlerinde durum değişmiştir gibi bişiler diyelim bence
* K-neihgbor algoritması Datasetin boyutu arttıkça kötüleşmiş gibi geldi eğer böyleyse bakalım slayta bide chatgpt
* Element sayısı arttığında(580) özellikle SVM ‘in verimi çok fazla artış göstermiştir. Yeterli dataset verildiğinde iyi çalışır(slaytta da diyordu bunu diye hatırladım şuan)