Investigation Of Tweet Accuracy According To News Source

Merve

AŞIKUZUNOĞLU

30716014

INTRODUCTION

As is known, we live in the information age. Access to information has become very easy all over the world, especially with the introduction of the internet. This flow of information is generally provided through "news". There is also a social media accompanying the news and not every news on social media is true.

Therefore in this study, it was wanted to determine whether the news on social media is true or false.

LITERATURE REVIEW

When I do research to realize this project , I learned that I have to work with NLP. Natural language processing is called Natural Language Processing (NLP) in English literature. Natural language processing is a subcategory of artificial intelligence. There are two different languages in the computer world, one of which is machine languages, namely programming languages, and the other is natural languages. What is meant by natural languages are natural languages spoken by humans. Languages such as Turkish, English and Korean can be given as examples. The process of receiving and processing the language spoken by humans by machines is called natural language processing. 5 different projects can be given as examples of application areas, these

- Text Classification and Categorization
- Named Entity Recognition (NER)
- Part-of-Speech Tagging
- Paraphrase Detection
- Machine Translation

STRUCTURE OF THE SOLUTION PROPOSED

Firstly I chose a subject. My subject is **"milli savunma-savunma sanayi"**.Then I researched on news sites about this topic and I chose a reliable news site for label as a true. I did the same for the wrong news. I got the news on the subject using the web-scraping site method. I scrape tweet about subject.

I followed the text preprocessing steps. I only used 3 of them.

After teaching the machine the data set , I use count vectors for numerical processing of texts. I used naive bayes for supervised learning.Then I got confusion matrix, precision score, recall score, f score and accuracy values .

HOW TO USE THE SOFTWARE ?
WHAT ARE THE REQUIREMENTS TO RUN ?

```
pip install Scrapy
```

for web scraping

```
pip install snscrape
```

for twitter scraping

```
pip install nltk
```

for using nlp

```
from nltk.tokenize import word_tokenize
import nltk
import pandas as pd
import numpy as np
import re
import string
from nltk.corpus import stopwords
from nltk.tokenize import sent_tokenize
import textblob
from textblob import TextBlob
from textblob import Word
from sklearn.model_selection import train_test_split,cross_val_score
from sklearn import preprocessing
from sklearn.feature_extraction.text import CountVectorizer
from sklearn import model_selection, preprocessing, linear_model, naive_bayes, metrics
from sklearn.metrics import precision_score,recall_score,fbeta_score,confusion_matrix
```

import required libraries

RUNTIME EXAMPLES

c: > Users > merve > Desktop > tutorial > datamining > datamining > spiders > savunma.py

```python
import scrapy
from scrapy import Request
from ..items import ExampleItem
from urllib.parse import urljoin
from urllib.parse import urlparse
class SavunmaSpider(scrapy.Spider):
    name = 'savunma'
    allowed_domains = ['www.millisavunma.com']
    start_urls = ['http://www.millisavunma.com/savunma-sanayi-haberleri/page/13/']

    def parse(self, response):
        haberler=response.xpath("//div[@class='news-content']")
        for haber in haberler:
            haber_adi=haber.xpath(".//div[@class='post-content-text']/h3/a/text()").get()
            haber_tarih=haber.xpath(".//div[@class='grid-date-post']/text()").get()
            haber_icerik=haber.xpath(".//div[@class='news-short-content']/text()").get()
            yield{
                'haber basligi':haber_adi,
                'haber tarih':haber_tarih,
                'haber icerik':haber_icerik

            }

```

*For true news

```python
import scrapy
from scrapy import Request
from ..items import ExampleItem
from urllib.parse import urljoin
from urllib.parse import urlparse
class SavunmaSpider(scrapy.Spider):
    name = 'fake_savunma'
    allowed_domains = ['https://www.zaytung.com/']
    start_urls = ['http://zaytung.com/aramasonuc.asp?cx=010830566949380726139%3Awh6pqlu77_k&cof=FORID%3A9&ie=UTF-8&q=milli+savunma&sa=Ara']

    def parse(self, response):
        haberler=response.xpath("//div[@class='news-content']")
        for haber in haberler:
            haber_adi=haber.xpath(".//div[@class='post-content-text']/h3/a/text()").get()
            haber_tarih=haber.xpath(".//div[@class='grid-date-post']/text()").get()
            haber_icerik=haber.xpath(".//div[@class='news-short-content']/text()").get()
            yield{
                'haber basligi':haber_adi,
                'haber tarih':haber_tarih,
                'haber icerik':haber_icerik
```

*For false news

During web scraping I transferred what I found (haber_adi, haber_tarih, haber_içerik) to the excel file and I gave a label (true/false). Then I scrape tweet with this code.

In [57]:
```
import snscrape.modules.twitter as sntwitter
import pandas as pd
import numpy as np
```

In [76]:
```
maxTweets = 50
for i,tweet in enumerate(sntwitter.TwitterSearchScraper('türk savunma sanayi + since:2019-12-31 until:2020-01-16').get_items()):
    if i > maxTweets :
        break
    print(tweet.content)
    print(",")
    print(tweet.username)
    print(",")
    print(tweet.date)
    print(",")
```

I have taken the following into consideration while doing this;

- ❖ Dates of the news I have scraped
- ❖ I wrote the words on the news topic (savunma sanayi, milli uydu,roketsan etc.)

Then I transferred what I found (tweet.content, tweet.username, tweet.date) to the excel file.

[163]:
```
tweet=pd.read_excel("tumtwetler.xlsx") # tweetleri dataframe yaptık
tweet_df=pd.DataFrame(tweet)
print(tweet_df)
```

```
                                                tweet
0                                  Can You play Chess?
1    Bayraktar TB3 Ve Akıncı Yerli Uçak Motoru PD17...
2    aselsan Bence tb2 gibi daha ufak sihalarda yer...
3    PD-170 Akıncı TİHA'da ve yeni geliştirilecek B...
4                    Teşekkürler Sayın SelçukBayraktar
..                                                 ...
145   Tam da Ukrayna'nın ve Türkiye'nin MI6 bşk ziy...
146  Zona Positive (Ermenistan): Türkiye, Ukrayna'n...
147  Rusya Türkiye ilişkilerini takip eden biri Rus...
148  Türkiye saman ve hayvan ithal etmek yerine Rus...
149  Türkiye'den müthiş başarı: Rusya, İngiltere, U...

[150 rows x 1 columns]
```

```
In [165]: haber=pd.read_excel("rrhaberr.xlsx") # haberleri dataframe yaptık
          haber_df=pd.DataFrame(haber)
          print(haber_df)
```

```
                                                haber  dogru_yanlis
0   Cumhurbaşkanımız Sayın Recep Tayyip Erdoğan aç...          True
1   Cumhurbaşkanı Recep Tayyip Erdoğan, ""Savunma ...          True
2   Cumhurbaşkanı Recep Tayyip Erdoğan'ın Vahdetti...          True
3   Milli Savunma Bakanlığı ve Kara Kuvvetleri Kom...          True
4   HAVELSAN, Yeni Tip Denizaltı Projesi'ndeki 6 d...          True
5   Cumhurbaşkanı Recep Tayyip Erdoğan, ASELSAN Gö...          True
6   Milli Teknoloji Hamlesi seferberliğinde önemli...          True
7   T3 Vakfı Mütevelli Heyeti Başkanı ve Baykar Te...          True
8   Cumhurbaşkanımız Recep Tayyip Erdoğan ile Ukra...          True
9   Milli Savunma Bakanlığı (MSB), Deniz Kuvvetler...          True
10  Cumhurbaşkanlığı Savunma Sanayii Başkanı İsmai...          True
11  Roketsan Genel Müdürü Murat İkinci, Türkiye'ni...          True
12  MİLLİ Savunma Bakanı Hulusi Akar ile Sanayi ve...          True
13  Türk Havacılık ve Uzay Sanayii (TUSAŞ) tarafın...          True
14  Üstlendiği görevleri başarıyla yerine getiren ...          True
15  Cumhurbaşkanlığı Savunma Sanayii Başkanı İsmai...          True
16  ROKETSAN tarafından geliştirilen yerli füzeler...          True
17  Cumhurbaşkanı Recep Tayyip Erdoğan, Roketsan L...          True
18  Dünyanın en prestijli savunma sanayi listesi o...          True
19  Türkiye'nin havacılık motorlarında lider şirke...          True
20  Cumhurbaşkanlığı Savunma Sanayii Başkanı Prof....          True
21  ASELSAN ve Katmerciler firmaları arasında imza...          True
22  SAVUNMA SANAYİİ BAŞKANI PROF. DR. İSMAİL DEMİR...          True
23  Türkiye Uzay Ajansı Başkanı Serdar Hüseyin Yıl...          True
24  Türkiye'nin beşinci nesil bir muharip uçak üre...          True
25  Türk savunma sanayisinin tüm kesimlerinin katı...          True
```

Then I did text processing. For the text processing , I applied the following operations

1. Lowercase

```
In [166]: # tweetler için yapıldı
          tweet_df['tweet'] = tweet_df['tweet'].str.lower()
          print(tweet_df['tweet'])
```

```
0                                    can you play chess?
1      bayraktar tb3 ve akıncı yerli uçak motoru pd17...
2      aselsan bence tb2 gibi daha ufak sihalarda yer...
3      pd-170 akıncı tiha'da ve yeni geliştirilecek ...
4                        teşekkürler sayın selçukbayraktar
                             ...
145     tam da ukrayna'nın ve türkiye'nin mi6 bşk ziy...
146    zona positive (ermenistan): türkiye, ukrayna'n...
147    rusya türkiye ilişkilerini takip eden biri rus...
148    türkiye saman ve hayvan ithal etmek yerine rus...
149    türkiye'den müthiş başarı: rusya, ingiltere, ...
Name: tweet, Length: 150, dtype: object
```

```
In [167]: # haberler için yapıldı
          haber_df['haber'] = haber_df['haber'].str.lower()
          print(haber_df['haber'])
```

```
0      cumhurbaşkanımız sayın recep tayyip erdoğan aç...
1      cumhurbaşkanı recep tayyip erdoğan, ""savunma ...
2      cumhurbaşkanı recep tayyip erdoğan'ın vahdetti...
3      milli savunma bakanlığı ve kara kuvvetleri kom...
4      havelsan, yeni tip denizaltı projesi'ndeki 6 d...
5      cumhurbaşkanı recep tayyip erdoğan, aselsan gö...
6      milli teknoloji hamlesi seferberliğinde önemli...
```

2. Punctuation marks, deletion of special characters

```
In [168]: # tweetler için noktalama işaretleri, özel karakterler silindi
          tweet_df['tweet'] = tweet_df['tweet'].str.translate(str.maketrans('', '', string.punctuation))
          tweet_df['tweet']
```

```
Out[168]: 0                                can you play chess
          1      bayraktar tb3 ve akıncı yerli uçak motoru pd17...
          2      aselsan bence tb2 gibi daha ufak sihalarda yer...
          3      pd170 akıncı tİhada ve yeni geliştirilecek ba...
          4                    teşekkürler sayın selçukbayraktar
                                    ...
          145     tam da ukraynanın ve türkiyenin mi6 bşk ziyar...
          146    zona positive ermenistan türkiye ukraynanın kı...
          147    rusya türkiye ilişkilerini takip eden biri rus...
          148    türkiye saman ve hayvan ithal etmek yerine rus...
          149    türkiyeden müthiş başarı rusya İngiltere ukra...
          Name: tweet, Length: 150, dtype: object
```

```
In [169]: #haberler için noktalama işaretleri, özel karakterler silindi
          haber_df['haber']  = haber_df['haber'] .str.translate(str.maketrans('', '', string.punctuation))
          print(haber_df['haber'] )
```

```
0    cumhurbaşkanımız sayın recep tayyip erdoğan aç...
1    cumhurbaşkanı recep tayyip erdoğan savunma san...
2    cumhurbaşkanı recep tayyip erdoğan'ın vahdetti...
3    milli savunma bakanlığı ve kara kuvvetleri kom...
4    havelsan yeni tip denizaltı projesindeki 6 den...
5    cumhurbaşkanı recep tayyip erdoğan aselsan göl...
6    milli teknoloji hamlesi seferberliğinde önemli...
7    t3 vakfı mütevelli heyeti başkanı ve baykar te...
8    cumhurbaşkanımız recep tayyip erdoğan ile ukra...
```

3. Deletion of special numbers

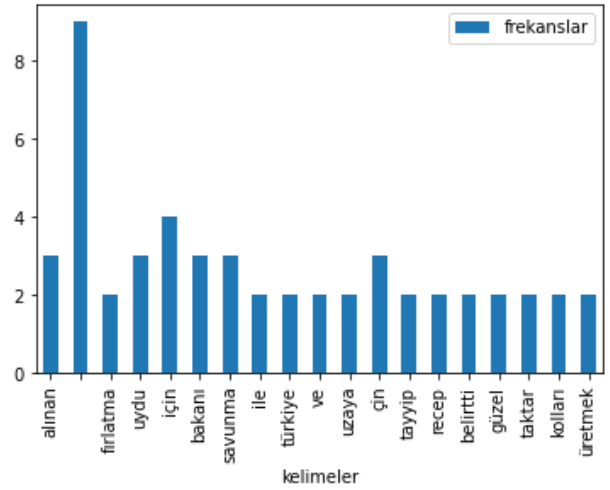
```
In [170]: #sayılar silindi tweetler için
          tweet_df['tweet'] = tweet_df['tweet'].str.replace('\d','')
          print(tweet_df['tweet'])
```

```
0                                  can you play chess
1       bayraktar tb ve akıncı yerli uçak motoru pd İ...
2       aselsan bence tb gibi daha ufak sihalarda yerl...
3       pd akıncı tİhada ve yeni geliştirilecek bayra...
4                       teşekkürler sayın selçukbayraktar
                              ...
145      tam da ukraynanın ve türkiyenin mi bşk ziyare...
146     zona positive ermenistan türkiye ukraynanın kı...
147     rusya türkiye ilişkilerini takip eden biri rus...
148     türkiye saman ve hayvan ithal etmek yerine rus...
149     türkiyeden müthiş başarı rusya İngiltere ukra...
Name: tweet, Length: 150, dtype: object
```

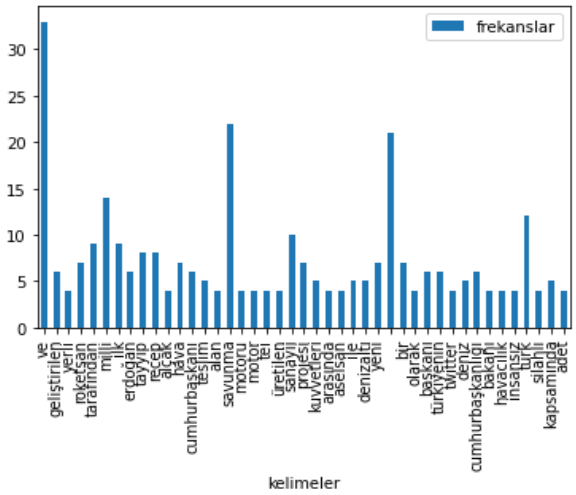
```
In [171]: #sayılar silindi haberler için
          haber_df['haber']  = haber_df['haber'] .str.replace('\d','')
          print(haber_df['haber'])
```

```
0      cumhurbaşkanımız sayın recep tayyip erdoğan aç...
1      cumhurbaşkanı recep tayyip erdoğan savunma san...
2      cumhurbaşkanı recep tayyip erdoğan'ın vahdetti...
3      milli savunma bakanlığı ve kara kuvvetleri kom...
4      havelsan yeni tip denizaltı projesindeki  deni...
5      cumhurbaşkanı recep tayyip erdoğan aselsan göl...
6      milli teknoloji hamlesi seferberliğinde önemli...
7      t vakfı mütevelli heyeti başkanı ve baykar tek...
8      cumhurbaşkanımız recep tayyip erdoğan ile ukra...
9      milli savunma bakanlığı msb deniz kuvvetleri k...
```

If we show the words in the true and fake news graphically to see frequency of words



*fake news



*true news

I did a modeling to teach the dataset to the machine.

```
In [206]: #true-false haberlerim için bir model oluşturma işlemi
          train_x, test_x, train_y, test_y = model_selection.train_test_split(haber_df['haber'],haber_df['dogru_yanlis'],random_state = 1)
```

```
In [207]: # 'dogru_yanlis' sütununun altında yer alan true-false ifadelerini numerik değere dönüştürme işlemleri için
          #encoder dönüştürücüsü tanımladım
          encoder = preprocessing.LabelEncoder()
```

```
In [208]: #modelimde ki train_y, test_y de var olan 'dogru_yanlis' sutünları için encoder işlemi gerçekleşti
          train_y = encoder.fit_transform(train_y)
          test_y = encoder.fit_transform(test_y)
```

There are 3 different approaches to numerical processing of texts:

- ❖ Count Vectors
- ❖ TF-IDF
- ❖ Word Embedding

I chose Count Vectors. According to my dataset, each row represents news. We will see the frequency of the words in these news with count vector. The process will proceed like this ; in each column in the data set there are unique words then these words will be evaluated together with the news on each line. In these news, numbering will be

made according to the frequency of occurrence depending on the words in the column and this numeric value will be written under the related column.

```
[219]: vectorizer = CountVectorizer()
       vectorizer.fit(train_x)

:[219]: CountVectorizer()

[226]: x_train_count = vectorizer.transform(train_x)
       x_test_count = vectorizer.transform(test_x)
       vectorizer.get_feature_names()[0:11]

:[226]: ['acil',
         'adayı',
         'adet',
         'ajansı',
         'ak',
         'akar',
         'akinci',
         'aksungur',
         'alan',
         'alanda',
         'alanında']

[227]: x_train_count.toarray()

:[227]: array([[0, 0, 0, ..., 0, 0, 0],
               [0, 0, 0, ..., 0, 0, 0],
               [0, 0, 1, ..., 0, 0, 0],
               ...,
               [0, 0, 0, ..., 0, 0, 0],
               [0, 1, 0, ..., 0, 0, 0],
               [0, 0, 0, ..., 0, 0, 0]], dtype=int64)
```

Finally I used a machine learning algorithm. I chose to use naive bayes. Naive Bayes is a probabilistic machine learning algorithm that can be used in a wide variety of classification tasks. It is also preferred because of its fast training. Naive bayes algorithms are mostly used in sentiment analysis, spam filtering and suggestion systems. Therefore, I found it appropriate to use this algorithm for social media analysis.

```
In [55]: nb = naive_bayes.MultinomialNB()
         nb_model = nb.fit(x_train_count,train_y)
```

Then I found accuracy, confusion matrix, precision, recall and f-measure.

Accuracy

Accuracy is one of the simplest criteria to understand and interpret. Frequently used to test machine learning classification algorithms. Accuracy score is between 0 and 1, and the model is considered successful for scores approaching 1.

```
[59]: accuracy = model_selection.cross_val_score(nb_model, x_test_count, test_y, cv = 5).mean()
      print("Accuracy:", accuracy)

      Accuracy: 0.7999999999999999
```

Confusion Matrix

A confusion matrix is used to interpret the results of an established classification model and cross-examine errors in the relationship between actual and predicted values.

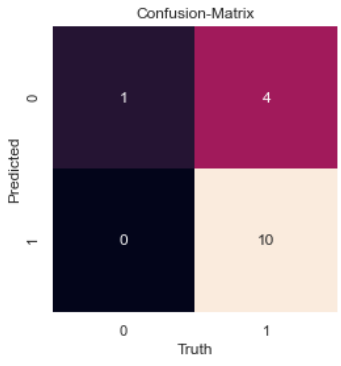
```
73]: test_data_predict=nb_model.predict(x_test_count)
     conf_matrix=confusion_matrix(test_y,test_data_predict)
     conf_matrix

73]: array([[ 1,  4],
            [ 0, 10]], dtype=int64)
```

There are 4 incorrect classifications

If we visualize the confusion matrix,

```
[54]: import seaborn as sns; sns.set()
      import matplotlib.pyplot as plt
      names = np.unique(test_y)
      sns.heatmap(con_m, square=True, annot=True, fmt='d', cbar=False,xticklabels=names, yticklabels=names)
      plt.xlabel('Truth')
      plt.ylabel('Predicted')
      plt.title('Confusion-Matrix')
```



in true values 14 out of 15 values are true, 1 is false, in predict values 10 out of 15 values are correct and 5 are false.

Precision

The precision shows how many of the values we estimate as true are actually true.

```
In [75]: precision =precision_score(test_y,test_data_predict)
         print("Precision:", precision)

         Precision: 0.7142857142857143
```

Recall

Recall is a metric that shows how much of the transactions we need to predict as Positive.

```
In [80]: recall=recall_score(test_y,test_data_predict)
         print("Recall:", recall)

         Recall: 1.0
```

F-measure

The F-measure value shows us the harmonic mean of the Precision and Recall values. The reason why it is a harmonic mean instead of a simple average is that we should not ignore extreme cases.

```
In [85]: f_measure=fbeta_score(test_y,test_data_predict,beta=1) #beta precison etkisini belirler
         print("F-measure:", f_measure)

         F-measure: 0.7575757575757576
```

Testing tweets

```
In [114]: tweet_df=pd.Series(tweet_df["tweet"])

In [115]: tweet_df_c=vectorizer.transform(tweet_df)

In [116]: matris=nb_model.predict(tweet_df_c)

In [117]: matris

Out[117]: array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1,
                 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1], dtype=int64)
```

1: True news

0: False news

I gave the tweets in the second dataset to the model I created and had the machine guess whether my tweets were true or false.

According to the results I got ; 1 show the right news , 0 show the false news.

RESULT AND INTERPRETATION

The Naive Bayes algorithm model, which we trained with the news dataset we collected, predicted the accuracy and falsity of the new data.

An accuracy ratio of **0,79.9** was obtained. This gives us the information that the predictions to be made will be **79,9%** correct.This value must be between 0 and 1. And the closer the value is to 1 the better the model will predict.

FUTURE WORK

For the development of the project , an increase in the number of news can be achieved by collecting news from many reliable and unreliable sources on the subject. Accuracy values can be tested using different machine learning algorithms. If these operations are repeated with the algorithm with higher accuracy, more meaningful results can be obtained.

REFERENCES

- *reliable news site: http://www.millisavunma.com/*
- *fake news site:https: www.zaytung.com/*
- *https://www.youtube.com/watch?v=C5cfpY7Gedk&t=56os*
- *https://medium.com/@arzuyldz_26994/scrapy-ile-web-scraping-web-kaz%C4%B1ma-nas%C4%B1l-yap%C4%B1l%C4%B1r-%EF%B8%8F-841cf7645c10*
- *https://www.sezerbozkir.com/2018/05/saf-python-ile-veri-kazima-web-scraping/*
- *https://www.youtube.com/watch?v=Fp4AnPVDRMk*
- *https://ichi.pro/tr/nlp-metin-on-isleme-teknikleri-257512615464420*
- *https://www.veribilimiokulu.com/blog/natural-language-toolkitnltk/*
- *https://www.veribilimiokulu.com/blog/sahte-haberlerin-belirlenmesi/*
- *https://www.kaggle.com/onurakkse/veri-bilimi-notlar*
- *https://ichi.pro/tr/nlp-metin-on-isleme-teknikleri-257512615464420*
- *https://devhunteryz.wordpress.com/2019/12/02/naive-bayes-siniflandirici/*
- *https://www.veribilimiokulu.com/blog/naive-bayes-yontemiyle-siniflandirma-classification-with-naive-bayes-python-ile-uygulama/*
- *https://www.machinelearningplus.com/predictive-modeling/how-naive-bayes-algorithm-works-with-example-and-full-code/*
- *https://medium.com/@gulcanogundur/do%C4%9Fruluk-accuracy-kesinlik-precision-duyarl%C4%B1l%C4%B1k-recall-ya-da-f1-score-300c925feb38*
- *https://yigitsener.medium.com/veri-bilimi-s%C4%B1n%C4%B1fland%C4%B1rma-model-%C3%A7%C4%B1kt%C4%B1lar%C4%B1n%C4%B1-de%C4%9Ferlendiren-metrikler-confusion-matrix-accuracy-437f5633c82b*
- *https://www.veribilimiokulu.com/blog/dogal-dil-isleme-nedir-ve-uygulama-alanlari-nelerdir/*