**T.C.**

**İZMİR EKONOMİ ÜNİVERSİTESİ**

**MBP 200**

**GÖRSEL PROGRAMLAMA III**

**YAPAY ZEKA İLE GELECEK YIL GERÇEKLEŞECEK SİBER OLAYLARIN TAHMİNİ**

**20212425023**

**Mustafa Cüneyt Kafes**

**Öğretim Üyesi**

**Burak Evrentuğ**

**İZMİR - 2022**

1. **ÖZET**
2. **GİRİŞ**
3. **BÖLÜM 1: MEVCUT VERİLERİN TANIMI VE İNCELENMESİ**
4. **BÖLÜM 2: MEVCUT VERİLERE YÖNELİK İSTATİSTİKLER**
   * **KISIM 1: VERİ FİLTRELEME**
5. **BÖLÜM 3: GELECEK YILA YÖNELİK VERİ TAHMİNİ**
6. **SONUÇ VE ÖNERİLER**
7. **REFERANSLAR**

**ÖZET**

Gündelik hayatımızda gerçekleştirdiğimiz davranışlardan veya karşılaştığımız olaylardan yola çıkarak gelecekte başımıza gelebilecek durumların öngörülmesi, Makine öğrenmesi ile geliştirilen yapay zeka modelleri sayesinde daha da kolaylaşmış ve bu modeller günümüz sektörlerinde sıkça kullanmasıyla yaşamımızda oldukça yer edinmiştir.

Buna bağlı olarak, bahsi geçen yapay zeka sistemlerinin Siber Güvenlik alanında nasıl kullanılabileceği düşünülmüş ve gerçekleşme potansiyeli olan Siber Güvenlik olaylarının tespit edilmesi yönünde işbu proje geliştirilmiştir. Projenin amacı 2005-2020 tarihleri arasında Siber Güvenlik alanında gerçekleşmiş olan olayların analizini yaparak makine öğrenmesi ile geliştirilen yapay zeka modelinin geleceğe yönelik tahminlerde bulunmasını sağlamaktır.

Proje kapsamında python yazılım diliyle birlikte, “kaggle.com” isimli internet sitesi üzerinden temin edilen veri seti kullanılmıştır. Veri işleme ve görselleştirme kısımlarında matlab, pandas, numpy, seaborn, wordcloud kütüphaneleri, veri tahmin aşamasında yapay zekayı desteklemek için sklearn ve csv kütüphaneleri kullanılmıştır. İşlemler, colab online derleyicisi ve local (yerel) bilgisayar üzerinde bulunan PyCharm isimli IDE ile gerçekleştirilmiş ve test edilmiştir. Veri setleri beslenerek gelecek yıllarda Siber Güvenlik alanındaki olaylara yönelik tahmin sisteminin sürekli olarak gerçekleşmesi ve buna yönelik olarak tedbirlerin uygulanması sektörde bulunan tüzel kişiler tarafından beklenmektedir.

Anahtar kelimeler: Siber Güvenlik, Tahmin, Gelecek Yıl Siber Olaylar, Python

**GİRİŞ**

Gündelik hayatımızda gerçekleştirdiğimiz davranışlardan veya karşılaştığımız olaylardan yola çıkarak gelecekte başımıza gelebilecek durumların öngörülmesi, Makine öğrenmesi ile geliştirilen yapay zeka modelleri sayesinde daha da kolaylaşmış ve bu modeller günümüz sektörlerinde sıkça kullanmasıyla yaşamımızda oldukça yer edinmiştir.

Neredeyse her sektörde kullanılmaya başlayan ve popülerliği gittikçe artan Yapay Zeka modellerinin insanları korumaya yönelik alanlarda da kullanılması gerektiği düşünülmektedir. Teknolojinin sürekli geliştiği ve elektronik cihazların gündelik hayatımızda oldukça yer edindiği günümüz dünyasında ise internetin korunmaya ihtiyaç duyduğu en önemli hususlardan biri de sanal mecralardır. Sanal mecralarda insanları korumak öncelik olarak bilgi toplayan kurum ve kuruluşların görevidir. Siber Güvenlik alanında yeni sistemler sürekli olarak geliştirilmekte, kurum ve kuruluşların ise buna bağlı olarak kendi güvenlik sistemlerini güncel tutarak bilgi güvenliğini hizmet verdikleri insanlara sağlamaları ve mağduriyet yaşatmamaları gerekmektedir.

İşbu proje sayesinde, geçmiş yıllarda Siber Güvenlik alanında yaşanan olaylar incelenerek gelecek yılda yeni gelişmelerin ne yönde olacağı makine öğrenmesi ile geliştirilen yapay zeka modeli tarafından tahmin edilmektedir. Projedeki modelin kullanılmasıyla bahsi geçen şirketler, üretilen bu tahmin değerlerine göre belirli önlemler alabilecek ve mağduriyet potansiyelini en aza indirebileceklerdir. Projede kullanılan veri setinin beslenmesi ve gelecek yıllar için kullanılabilir hale getirilmesi sektörde aktif olarak yer alan tüzel kişilerden beklenmektedir.

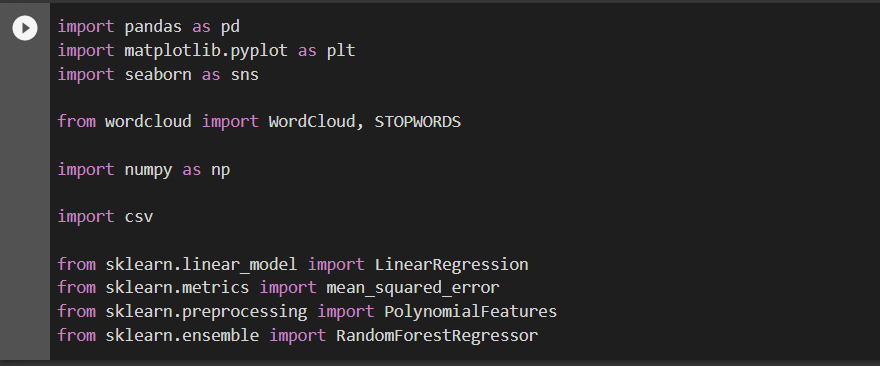
Proje, Python programlama dili ile geliştirilmiş olup çeşitli 3. Taraf kaynaklı kütüphanelerden yararlanılmıştır.

Proje Siber Güvenlik alanında yer alan her gerçek ve tüzel kişi tarafından, iyi niyetli geliştirilmek üzere kullanılabilir

**BÖLÜM 1**

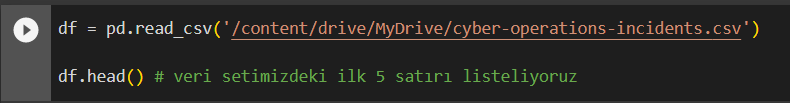
**MEVCUT VERİLERİN TANIMI VE İNCELENMESİ**

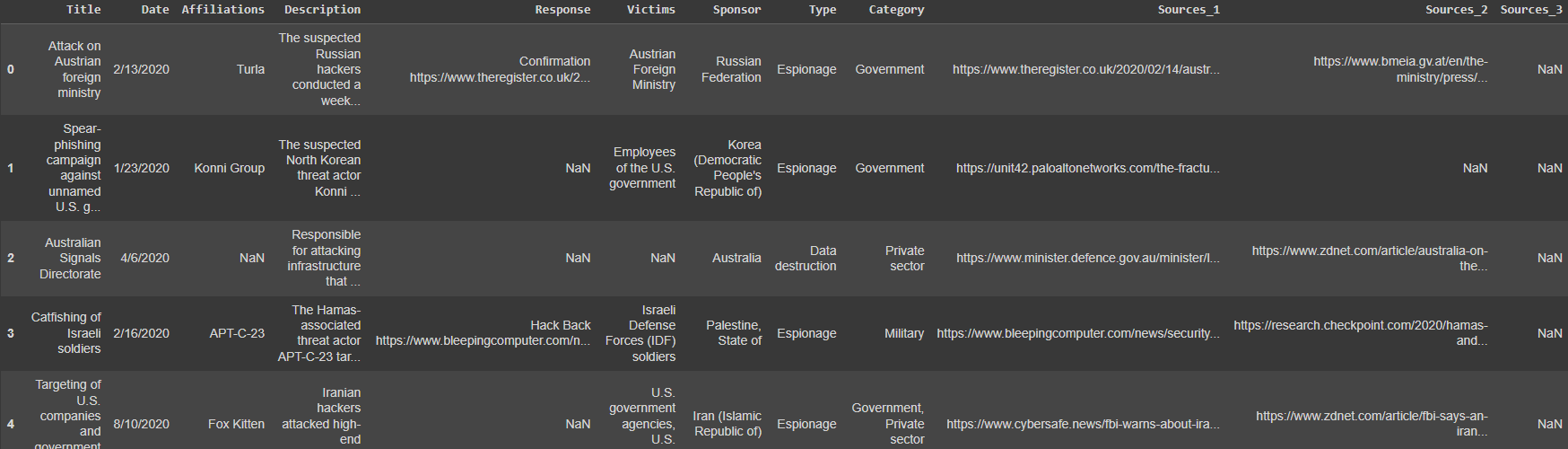
2005-2020 yılları arasında gerçekleşmiş olan Siber Güvenlik olaylarının bir kısmını içeren veri seti “kaggle.com” isimli internet sitesi üzerinden indirilmiş [0] ve projede kullanılacak olan kütüphaneler tanımlanmıştır. Yapılan işlemler aşağıdaki görselde yer almaktadır.



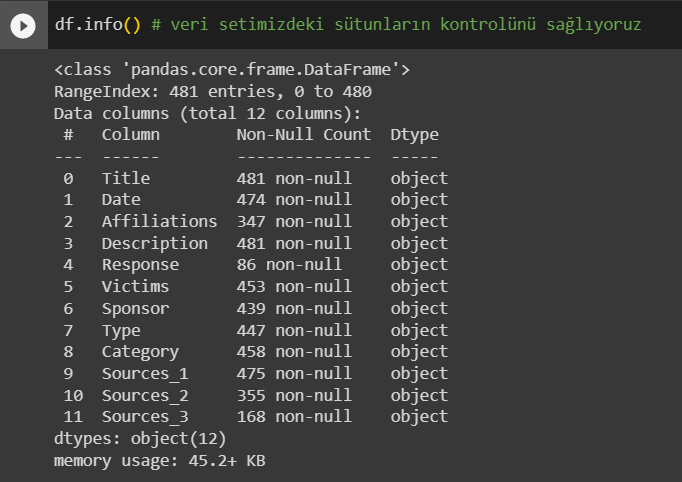
Pandas ve numpy kütüphanesi veri setinde bulunan veriler üzerinde işlemler yapmak, matplotlib ve seaborn kütüphaneleri veri görselleştirmeleri sağlamak, wordcloud veri setinde bulunan verilerin içeriğini görselleştirerek veriler hakkında genel bilgi sahibi olmak, csv kütüphanesi regresyon işlemi için yeni bir veri seti oluşturarak farklı dosyaya kaydetmek, sklearn kütüphanesi yapay zekayı eğitmek ve tahmin üretmek amaçlarıyla kullanılmıştır.

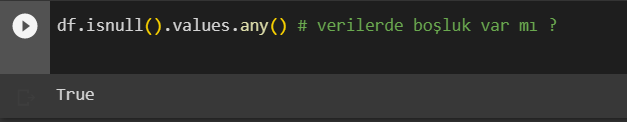
“kaggle.com” isimli internet sitesi üzerinden indirilen CSV uzantılı veri seti dosyası pandas kütüphanesi aracılığıyla okunarak, veri setinde bulunan ilk 5 satır kontrol amacıyla listelenmiştir. Aşağıda yapılan işlemlere ait görseller bulunmaktadır.

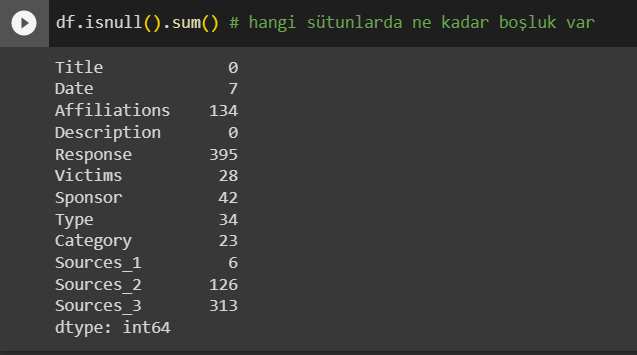




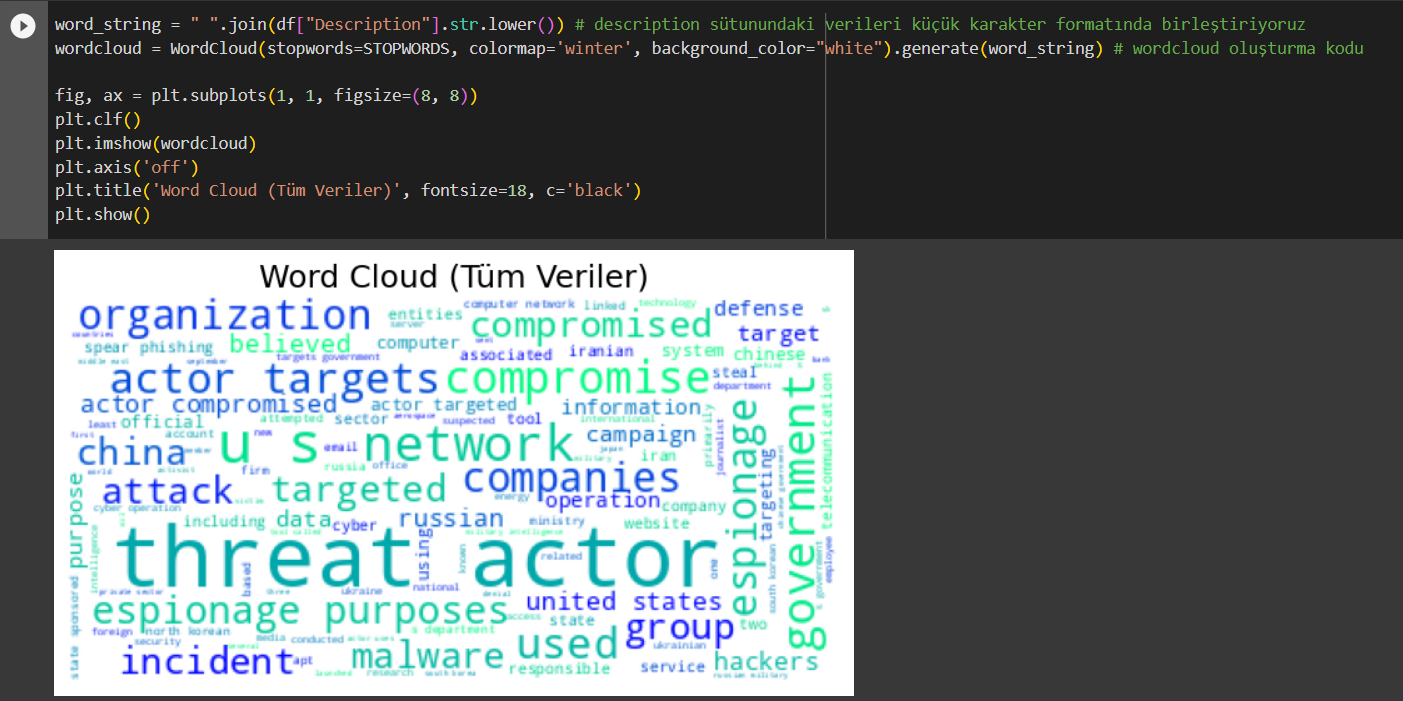
Veri setindeki sütunların kontrolünü sağlanarak veri setinde boş değerlerin olup olmadığı hakkında bilgi sağlayabilmek için `df.isnull().values.any()` komutu, boş değerler bulunan veri sayısına ulaşmak için ise `df.isnull().sum()` komutu kullanılmıştır.







WordCloud yapısı kullanılarak sahip olunan tüm veriler arasında en çok bulunan kelimeler görüntülenmiş, Verilerin içeriği hakkında genel bir bilginin tablo olarak sunulması amaçlanmıştır.

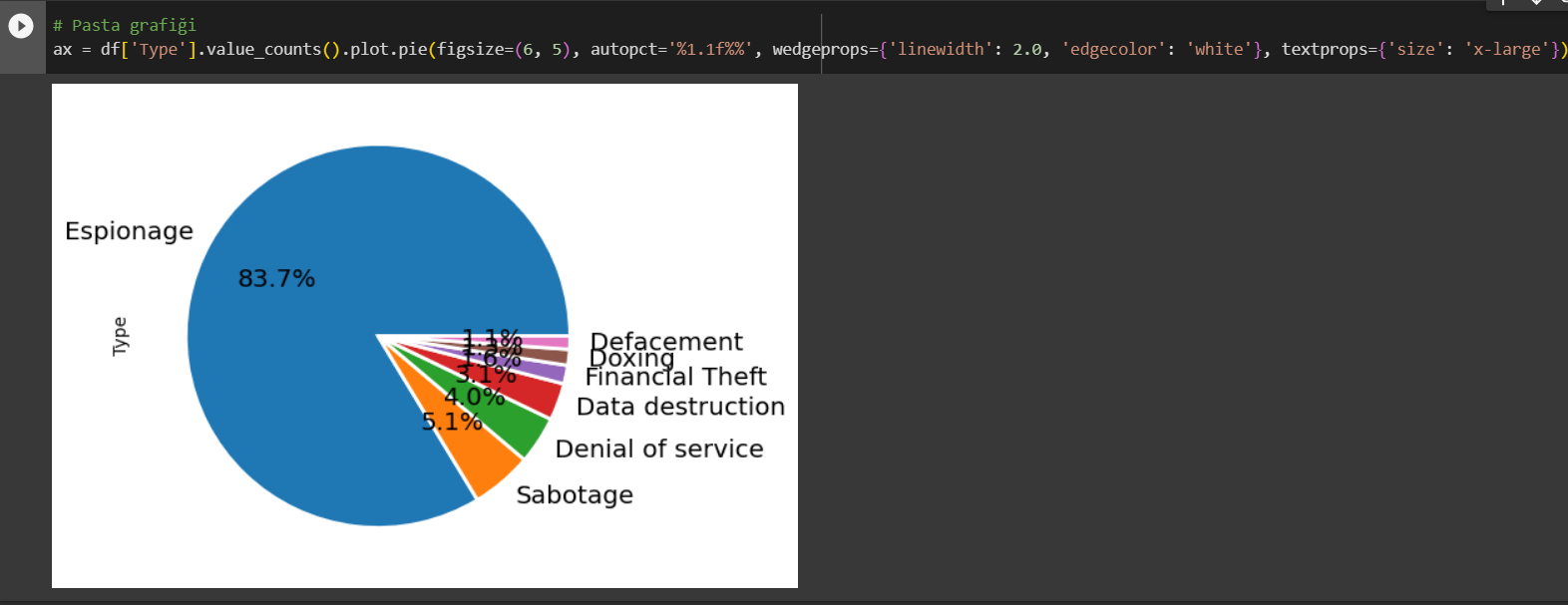


**BÖLÜM 2**

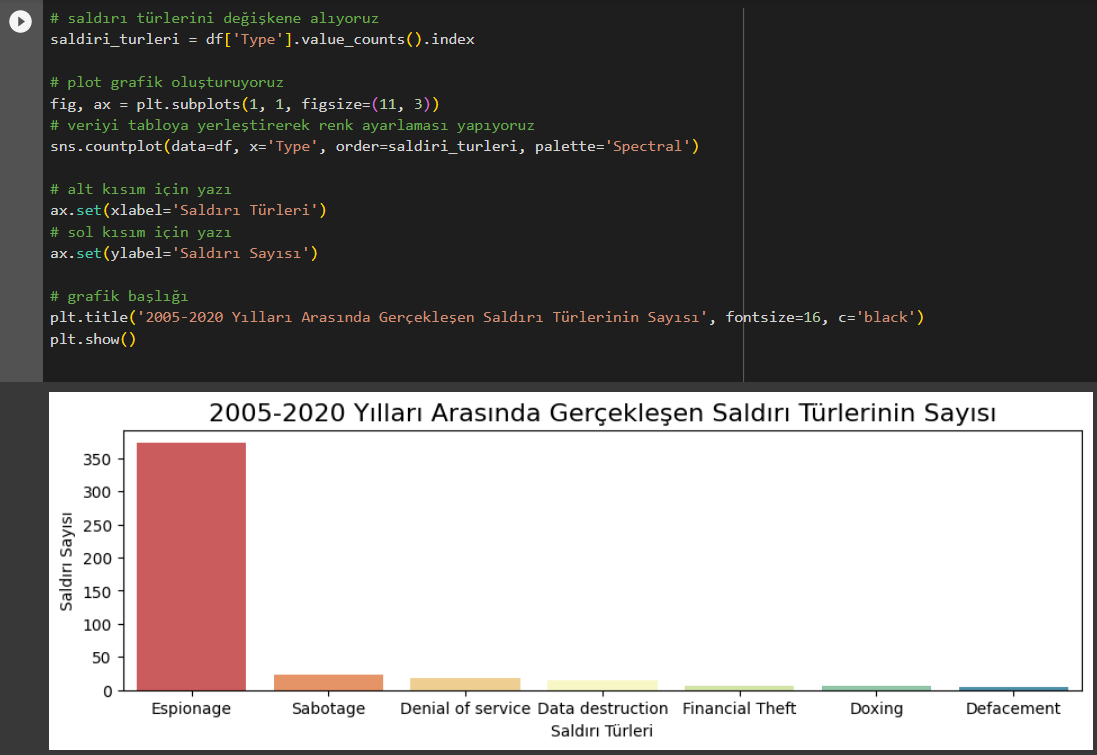
**MEVCUT VERİLERE YÖNELİK İSTATİSTİKLER**

Bu aşamada veriler görselleştirilerek, genelden özele doğru bir inceleme yapılmıştır. Bilgilerin görselleştirilme kısmında pasta grafiği (pie), sütun grafiği (barh ve bar plot), çizgi grafiği (line plot), saçılım grafiği (scatter plot) kullanılmıştır.

Aşağıdaki görselde 2005-2020 tarihleri arasında gerçekleşen siber saldırıların türlerine ait pasta grafiği yer almaktadır.



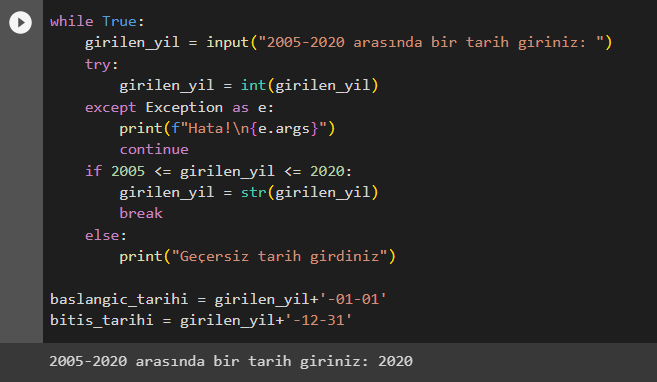
Aşağıdaki görselde 2005-2020 yılları arasında gerçekleşen saldırıların türlere göre sayısı sütun grafiği şeklinde yer almaktadır



**KISIM 1: VERİ FİLTRELEME**

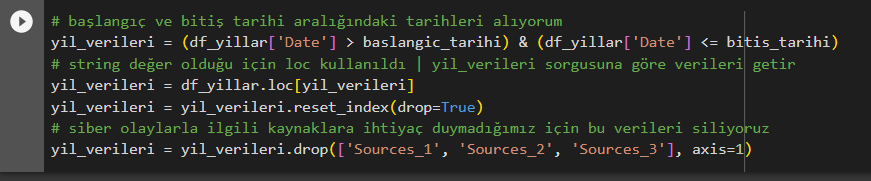
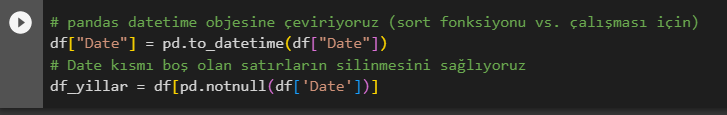
Bu aşamada veriler tarih, ülke ismi gibi spesifik kriterlere göre incelenerek görselleştirilmiştir.

Aşağıdaki görselde kullanıcının istediği yıla yönelik sistemde araştırma yapabilmesine olanak sağlayan basit bir arayüz tasarlanmıştır.

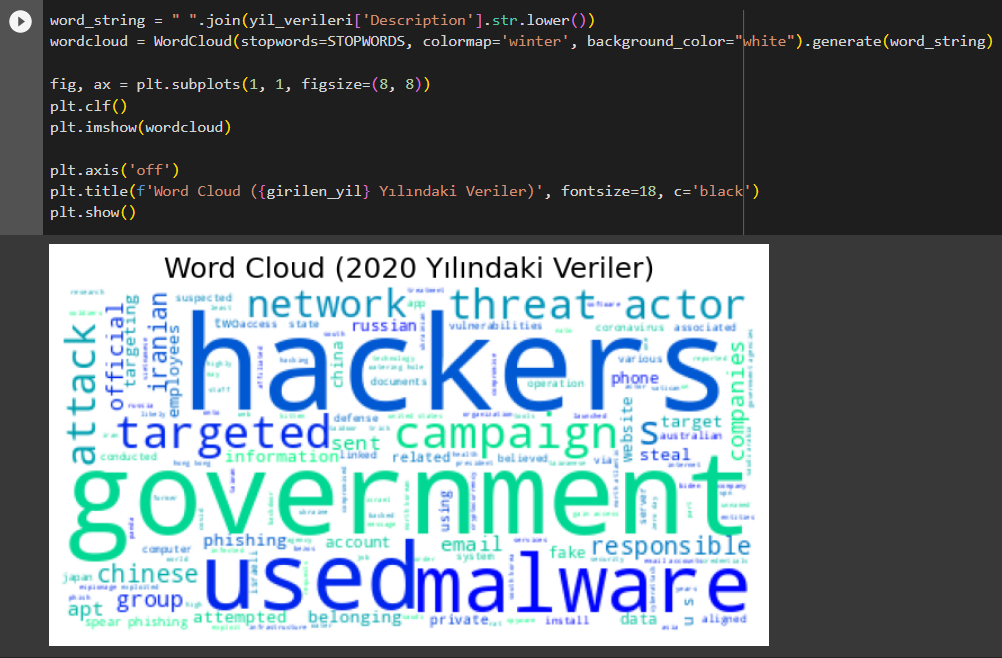


Date sütununda yer alan tarih verilerinin pandas kütüphanesi tarafından "tarih" olarak görülebilmesi ve bazı özel fonksiyonların kullanılabilmesi için pandas kütüphanesinde tanımlı "datetime" objesine dönüştürülmesi gerekmektedir.

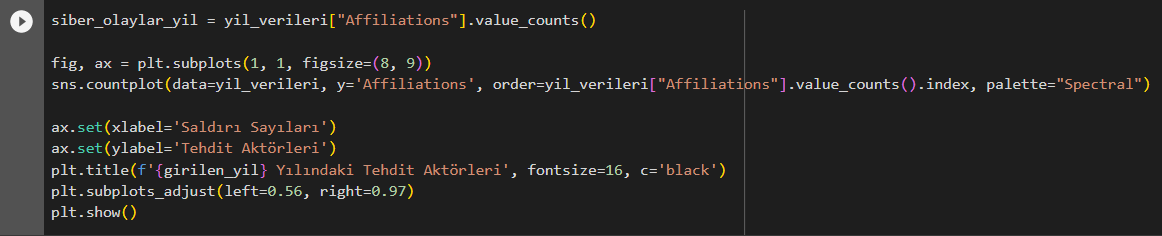
Aşağıda bulunan görsellerde Date sütununda yer alan veriler pandas kütüphanesinde bulunan "datetime" objesine çevrilmiş, Date kısmı boş olan satırlar veri setinden silinmiş ve kullanıcının seçmiş olduğu yıl ile ilgili veri setinde filtreleme işlemi yapılmıştır.

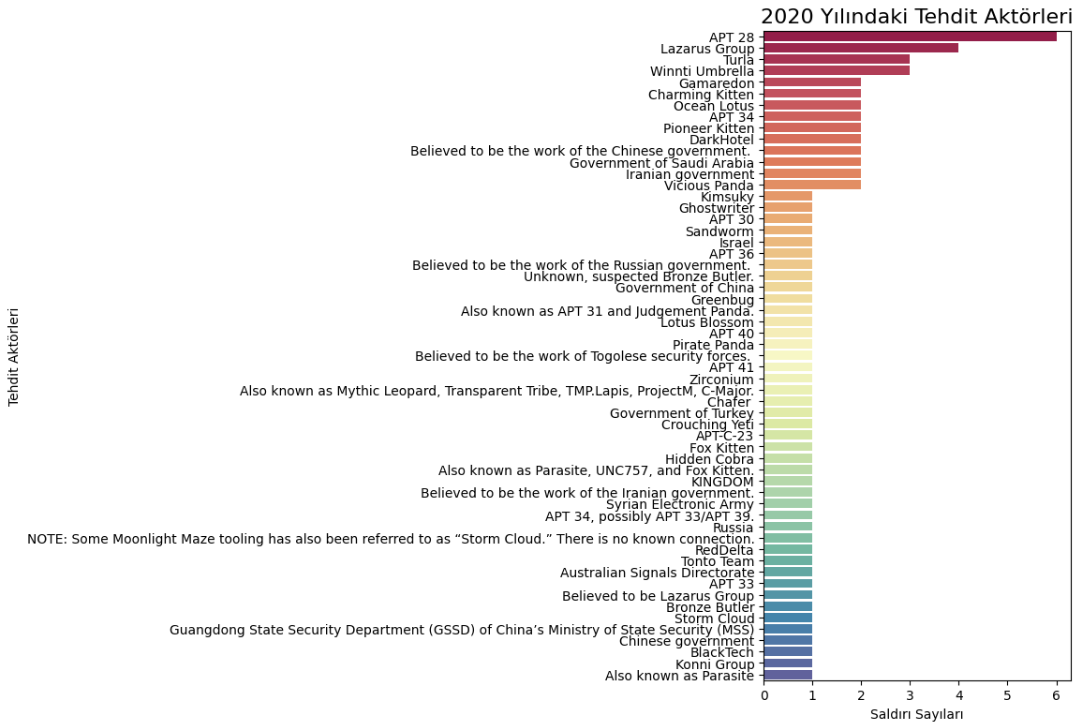


Veri setinde, kullanıcın seçmiş olduğu yıl ile eşleşen verilerin içeriği hakkında genel bilgi sahibi olmak üzere WordCloud yapısı uygulanmıştır. (2020 Yılı Siber Olaylar İçeriği)

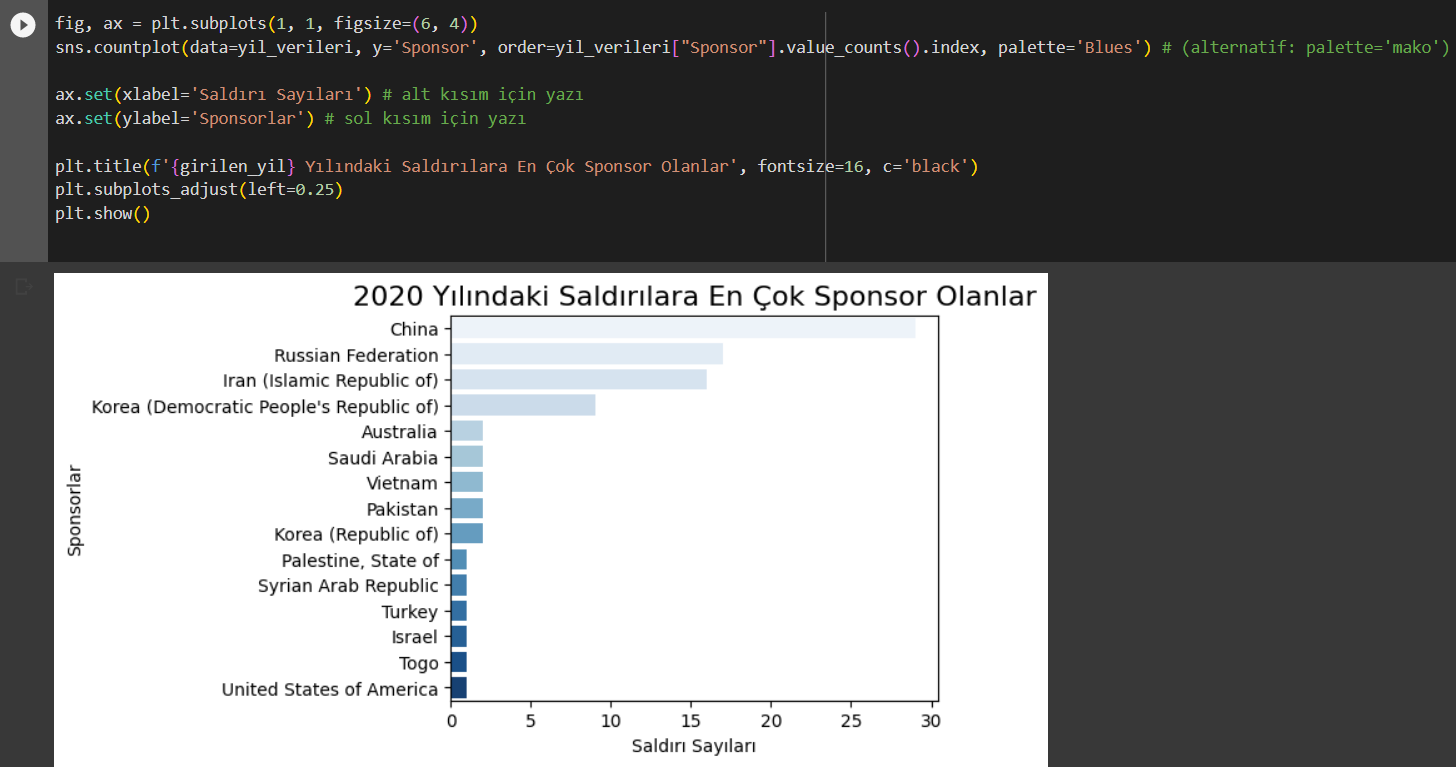


Kullanıcının seçmiş olduğu yıl ile ilgili tehdit aktörlerinin aktiviteleri incelenerek veriler sütun grafiğine yansıtılmıştır. (2020 Yılı Siber Güvenlik Tehdit Aktörleri Tablosu)

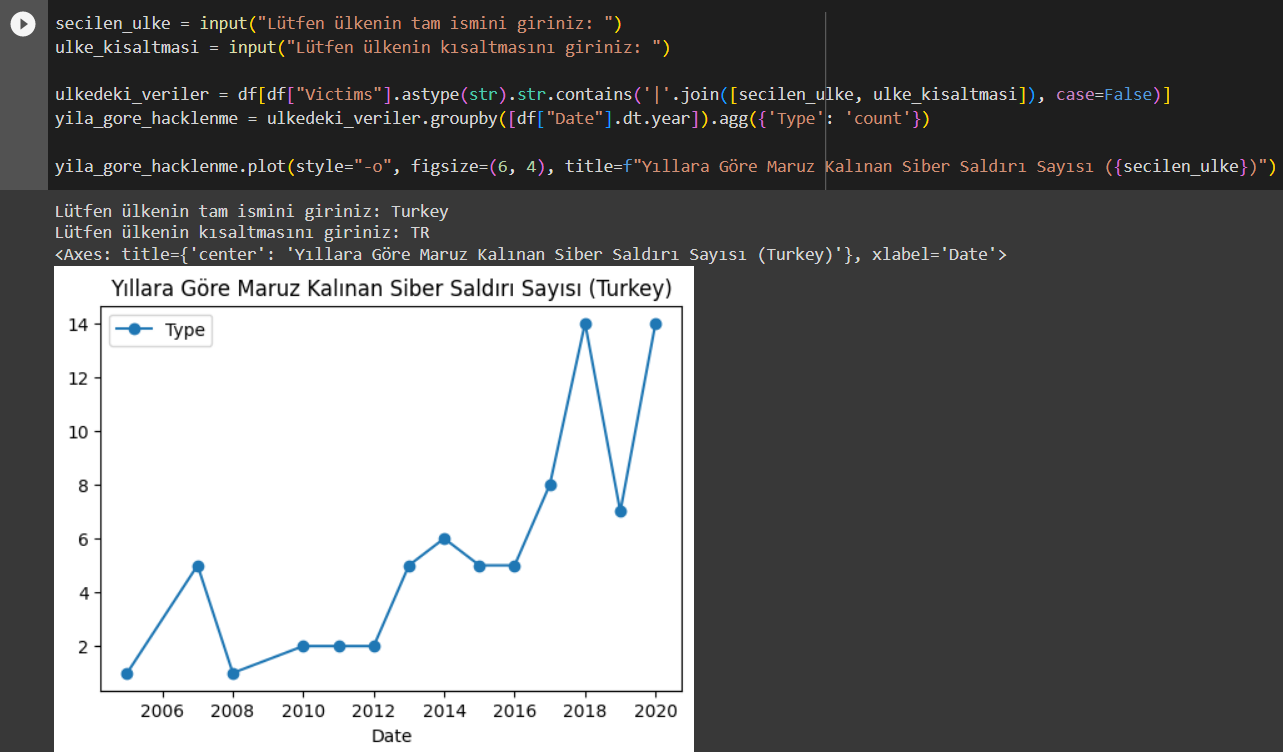




Aşağıdaki görselde kullanıcının seçtiği tarihte tehdit aktörerlerine en çok sponsor olan ülkelerin listelenmesine yönelik işlemler yapılmıştır. (2020 Yılı Siber Saldırılara / Tehdit Aktörlerine En Çok Sponsor Olan Ülkeler)



Kullanıcının seçmiş olduğu ülkeye göre, hükümetin maruz kaldığı siber saldırı sayısı listelenmiştir. (Türkiye’nin Yılara Göre Maruz Kaldığı Siber Saldırı Sayısı)



**BÖLÜM 3**

**GELECEK YILA YÖNELİK VERİ TAHMİNİ**

Bu aşamada veri tahminini gerçekleştirmek üzere yapay zekanın performansı artırmak istenmiş, buna bağlı olarak regresyon işlemleri uygulanmıştır.

**Regresyon:** Regresyon denklemi birden fazla değişken arasındaki ilişkiyi analiz etmeye ve buna bağlı olarak yeni tahminler üretilebilmesine olanak sağlar.

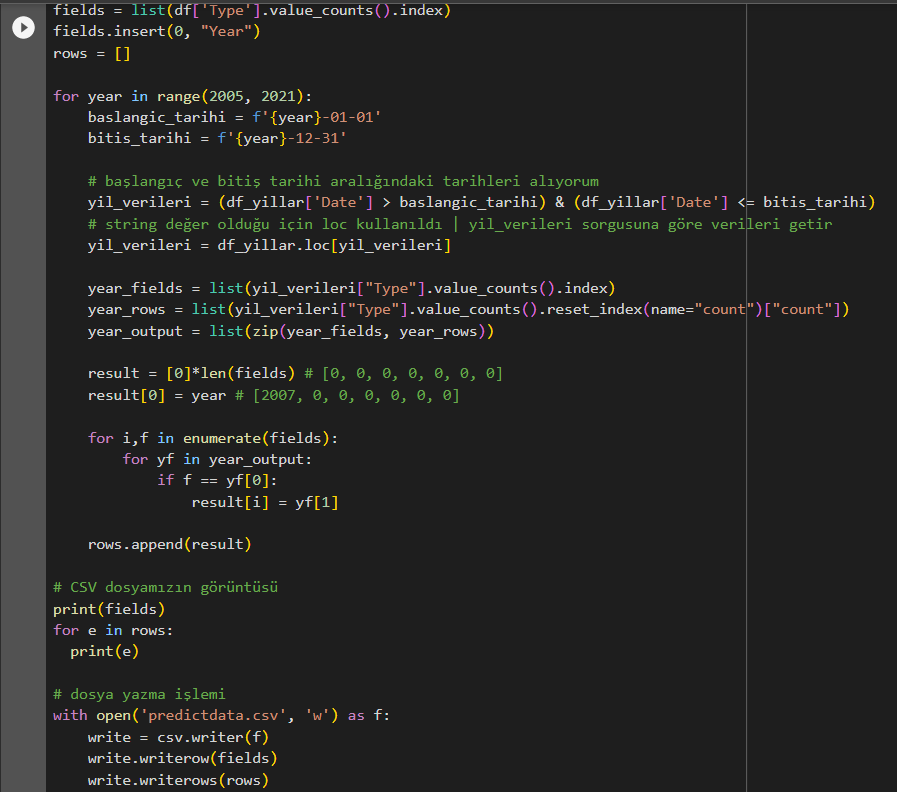
2020 yılındaki "Defacement" (arayüz görünümünü değiştirme) saldırısına yönelik verileri hesaplamak için kalan saldırı türleri bağımsız değişken olarak kabul edilir ve bu veriler kullanılarak bir regresyon denklemi oluşturulur. Bu sayede "Defacement" sayısı 2020 yılı için tahmin edilir. Ardından 2020 yılındaki gerçek değer ile arasındaki fark incelenir ve MSE, RMSE değerleri bulunur. Bu işlem bağımsız değişkenlere olduğu sürece her yılda tahmin yapmak için kullanılabilir.

**MSE (Mean Squared Error):** ortalama hatanın karesi

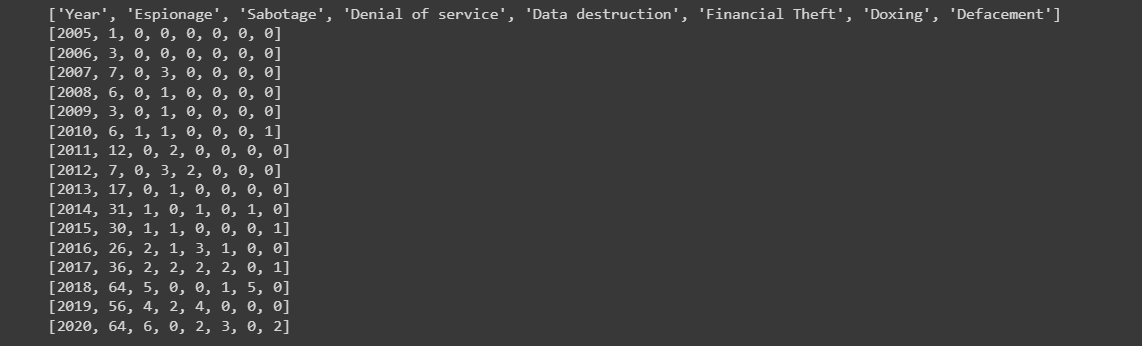
**RMSE (Root-Mean-Square Deviation):** ortalama karekök sapması

2020 yılındaki verilerin tahmin edilip karşılaştırılması ve 2021 yılındaki verilerin tahmin edilmesi aşamasında multiple linear regression (çoklu doğrusal regresyon) yöntemi kullanılmıştır. Tahmin ediciler ve yanıt arasında doğrusal bir ilişki olduğu ve basit yorumlanabilir bir model olduğu için bu regresyonda çoklu doğrusal yöntem tercih edilmiştir.

Gelecek yıl "Defacement" arayüz saldırısının ne kadar olacağını tahmin edebilmek için kullanacağımız regresyon denkleminde ihtiyaç duyduğumuz şey; hangi yıl hangi saldırıların ne kadar olduğudur. Bu yüzden aşağıdaki kod bloğunda sahip olunan verilerle yeni bir CSV dosyası yaratılmıştır.

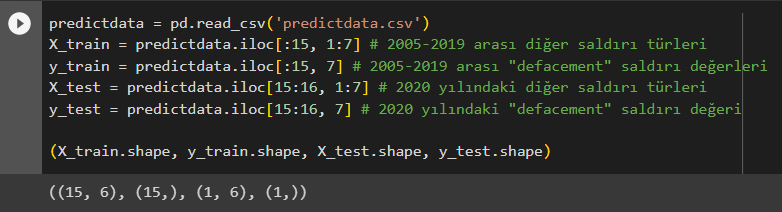


Yukarıda yazılmış olan kodların çıktısı aşağıdaki şekilde olacaktır. Bahsi geçen çoklu doğrusal regresyon denklemi bu tablo üzerinde uygulanacaktır.

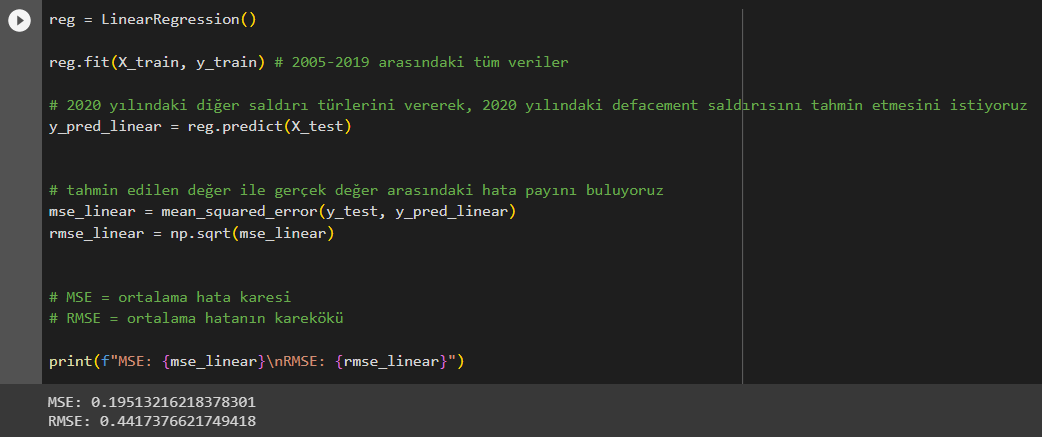


Oluşturulan CSV uzantılı dosyanda işlemlerin gerçekleştirilebilmesi için pandas kütüphanesiyle okuma işlemi sağlanmıştır. Ek olarak makine öğrenmesi modelimizi eğitme aşamasında kullanacağımız X\_train, y\_train, X\_test ve y\_test isimli değişkenler tanımlanmıştır. X\_train 2005-2019 yılları arasındaki defacement hariç saldırı türlerinin değerlerini, y\_train 2005-2019 yılları arasındaki defacement saldırı türünün değerlerini içermektedir. Bu değişkenler modelimizi eğitmek için kullanılacaktır. X\_test 2020 yılındaki defacement hariç saldırı türlerinin değerlerini, y\_test ise 2020 yılındaki defacement saldırı değerini içermektedir.

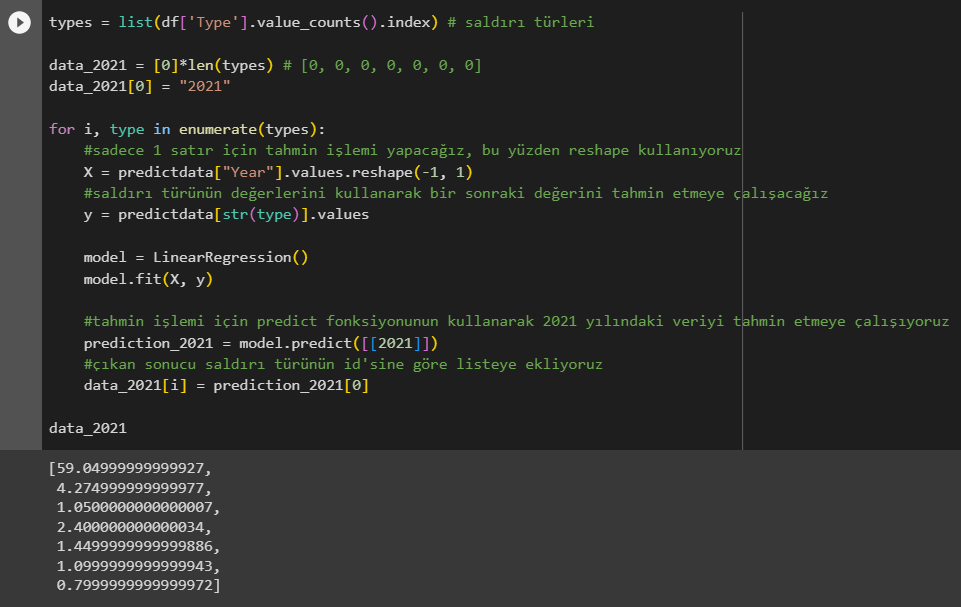
Kullanılacak olan regresyon yönteminin doğruluk payını ölçebilmek adına, modelin önce var olan veriler üzerinde kullanılması amaçlanmaktadır. Model 2005-2019 yılları arasındaki veriler ile eğitildikten sonra, 2020 yılındaki defacement saldırı verisini tahmin etmesi beklenecektir.



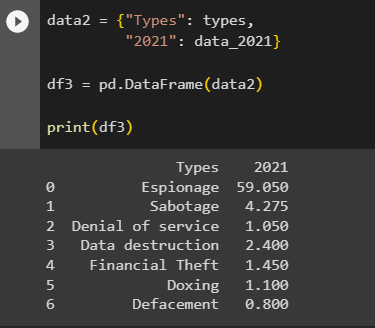
Aşağıdaki görselde regresyon işlemi kullanılmak üzere class(sınıf) tanımlaması yapılmış ve fit() fonksiyonu 2005-2019 arasındaki veriler ile modelimizi eğitmek amacıyla kullanılmıştır. Ardından X\_test verisi verilerek defacement verilerini tahmin etmesi beklenmiştir. Gerçek değerler (y\_test) ile tahmin edilen değerler (y\_predict\_linear) arasındaki MSE ve RMSE değerleri incelenerek ekrana yazdırılmıştır. 0.1 ve 0.4 değerleriyle modelin gerçeğe yakın tahminler üreterek, doğruluk payının yüksek olduğu kanıtlanmıştır.



Aynı yöntemler modelin 2021 tahmini oluşturmasını sağlamak için kullanılmıştır. Aşağıdaki görselde yapılan işlemler yer almaktadır.

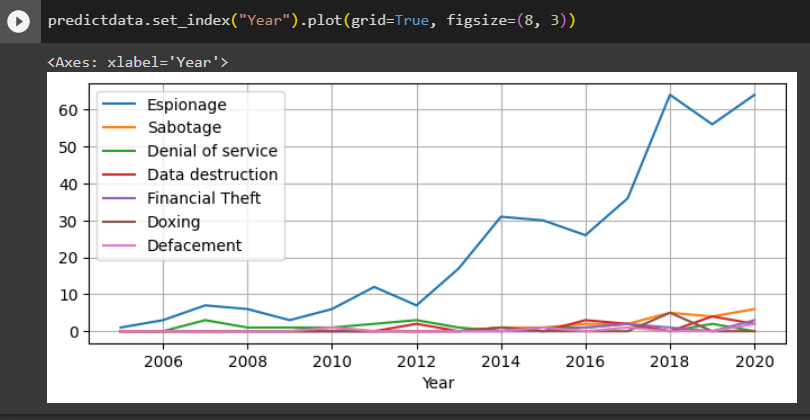


Yukarıdaki kodlarda yapılmış işlemlerin çıktısı daha anlamlı bir şekilde aşağıdaki görselde yer almaktadır.

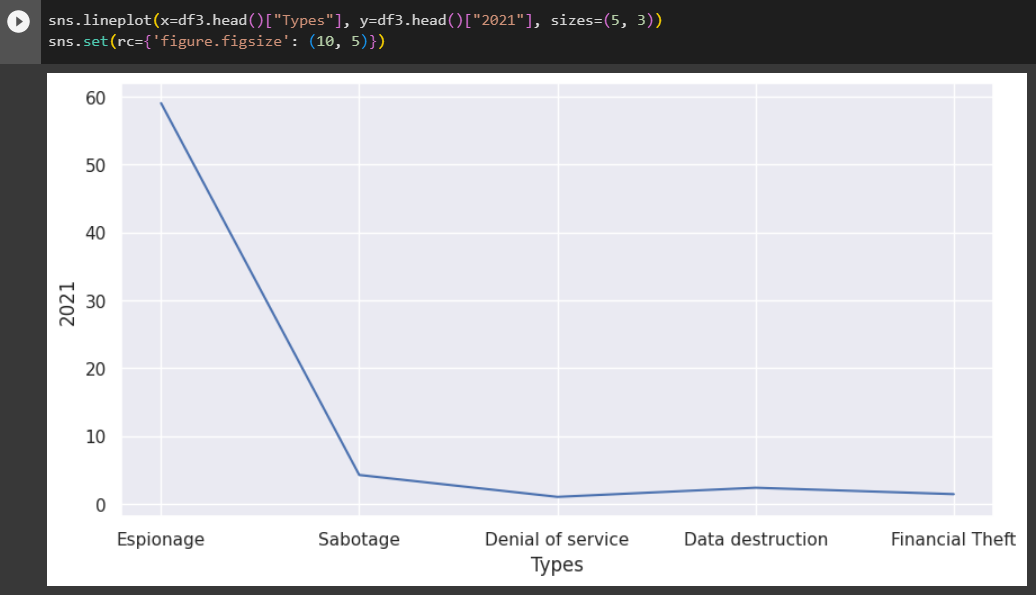


Aşağıdaki görselde 2005-2020 yılları arasında gerçekleşen siber olayların türlerine

göre sayısını gösteren çizgi grafiği yer almaktadır. (2005-2020 Yılları Arası Gerçekleşen Siber Olaylar)



Aşağıda bulunan grafikte ise Makine Öğrenmesi ile geliştirilen Yapay Zeka modeli tarafından tahmin edilmiş olan 2021 yılı verilerine ait siber olayların sayısı türlerine göre belirtilmiştir. (2021 Yılı Gerçekleşmesi Öngörülen Siber Olayların Sayısı)



**SONUÇ VE ÖNERİLER**

Proje kapsamında Siber Güvenlik alanında gerçekleşen olayların belirli bir kısmı kullanılarak Makine Öğrenmesi ile Yapay Zeka modeli geliştirilmiştir. Geliştirilen yapay zeka modelinin gelecek yıldaki verilerin nasıl olacağına dair tahmin yapması sağlanmıştır. Yapay Zeka modeli yapılan testlere göre %86.5 doğruluk oranı vermektedir.

Doğruluk oranı hesaplama yöntemi:

( 1.96 x RMSE ) x 100

Projedeki veri setlerinin incelenmesi sonucu yapay zeka modelinin yaptığı tahminlere göre göre; 2021 yılı espionage (casusluk) türündeki siber saldırıların sayısı 59, sabotage (sabotaj) türündeki siber saldırıların sayısı 4, denial of service (sistem reddi) türündeki saldırıların sayısı 1, data destruction (veri imhası) türündeki saldırıların sayısı 2, financial theft (hırsızlık) türündeki saldırıların sayısı 1, doxing (gizli bilgileri yayma) türündeki saldırıların sayısı 1, defacement (arayüz görünümü değiştirme) türündeki saldırıların sayısı 0 olacaktır.

Siber Güvenlik sektöründe aktif olarak rol alan tüzel ve gerçek kişilerin işbu projede kullanılan veri setini gerçek bilgilerle besleyerek gelecek yıllara yönelik gerçek veri tahminleri yapılmasını sağlamaları beklenmektedir.

**REFERANSLAR**

[0]: <https://www.kaggle.com/datasets/fireballbyedimyrnmom/cyber-incidents-up-to-2020> (21 Mayıs 2023)

[1]: <https://github.com/ml6973/Course> (21 Mayıs 2023)

[2]: [https://realpython.com/linear-regression-in- python/#:~:text=You%20can%20predict%20the%20output,the%20intercept%20to%20the%20sum.&text=That's%20the%20prediction%20using%20a%20linear%20regression%20model](https://realpython.com/linear-regression-in-%20python/#:~:text=You%20can%20predict%20the%20output,the%20intercept%20to%20the%20sum.&text=That's%20the%20prediction%20using%20a%20linear%20regression%20model) (21 Mayıs 2023)

[3]: <https://www.youtube.com/watch?v=A6jKo7OjAKw> (21 Mayıs 2023)

[4]: <https://dergipark.org.tr/tr/download/article-file/550950> (21 Mayıs 2023)

Mustafa Cüneyt Kafes