

**DERİN ÖĞRENME VE MAKİNE ÖĞRENMESİ TEKNİKLERİ
KULLANILARAK GÜMRÜK İTHALATINDAKİ USULSÜZ
İŞLEMLERİN TAHMİN EDİLMESİ**

YÜKSEK LİSANS TEZİ

İSMAİL ERGÜN

**MERSİN ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ**

**İŞLETME BİLGİ YÖNETİMİ
ANABİLİM DALI**

**MERSİN
EKİM 2021**

**DERİN ÖĞRENME VE MAKİNE ÖĞRENMESİ TEKNİKLERİ
KULLANILARAK GÜMRÜK İTHALATINDAKİ USULSÜZ
İŞLEMLERİN TAHMİN EDİLMESİ**

YÜKSEK LİSANS TEZİ

İSMAİL ERGÜN

**MERSİN ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ**

**İŞLETME BİLGİ YÖNETİMİ
ANABİLİM DALI**

**Danışman
Dr. Öğr. Üyesi Evrim Ersin KANGAL**

**MERSİN
EKİM 2021**

ÖZET

Dış ticaret şirketlerinin vergi kaçırma eğilimlerinin hızla artması hem ülkelerin gelirlerini azaltmakta hem de rekabet ortamının bozulmasına neden olmaktadır. Bu sorunla başa çıkmak için makine öğrenmesi ve derin öğrenme modelleriyle usulsüzlüğün önceden tahmin edilerek vergi kaybının önüne geçilmesi planlanmıştır. Bu çalışma; 2013-2014 yılları arasında Nijerya'da faaliyet gösteren gümrük merkezlerinin 200.000 kayıttan oluşan gümrük ithalat verileri üzerinden yapılmıştır. Bu tezin temel amacı, açık kaynaklı Python kütüphaneleri ve yöntemleri kullanılarak şirketler tarafından değeri düşük veya yanlış beyan edilen yasadışı ithalat işlemlerini tahmin etmektir. Çalışma sonuçlarına göre; F-puan değerlendirmesinde K-EYK algoritması 0,93 ile en iyi performansı gösterirken, İKE-EAA sonuçlarına göre, Derin Öğrenme modeli %83,88 ile en başarılı sonucu vermiştir.

Anahtar Kelimeler: İthalat, Vergi, Makine Öğrenmesi, Derin Öğrenme, Gümrük Dolandırıcılığı Tespiti

Danışman: Dr. Öğr. Üyesi Evrim Ersin KANGAL, İşletme Bilgi Yönetimi Anabilim Dalı, Mersin Üniversitesi, Mersin

ABSTRACT

The rapid increase in the tax evasion tendencies of foreign trade companies both reduces the incomes of the countries and causes the deterioration of the competitive environment. To cope with this problem, it is planned to prevent tax loss by predicting fraud with machine learning and deep learning models. This job; Between 2013 and 2014, more than 200,000 customs import data of customs centers operating in Nigeria were made. The main purpose of this thesis is to predict illegal import transactions that are undervalued or misrepresented by companies using open source Python libraries and methods. According to the research results; While K-NN algorithm showed the best performance with 0.93 in F-score evaluation, Deep Learning model gave the most successful result with 83.88% according to ROC-AUC results.

Keywords: Imports, Tax, Machine Learning, Deep Learning, Customs Fraud Detection

Advisor: Asistant Prof.Dr. Evrim Ersin KANGAL, Department of Business Information Management, Mersin University, Mersin



TEŞEKKÜR

Tezin hazırlanması aşamasında, nerede olursa olsun, her gün her saat iletişim kurabildiğim, gece gündüz yardım ve katkılarını esirgemeyen danışman hocam Dr. Öğr.Üyesi Evrim Ersin KANGAL hocama, bilgi ve tecrübesinden olabildiğince faydalandığım Anabilim Dalı Başkanım Prof.Dr. Cemile ÇELİK hocama, öneri ve görüşlerinden dolayı jüri üyesi Dr. Öğr. Üyesi Ali Kemal HAVARE hocama ve yüksek lisans programında ders gördüğüm tüm hocalarıma sonsuz teşekkürlerimi sunarım. Bu çalışmada, yeteri kadar ilgilenemediğim eşim ve çocuklarıma da ayrıca teşekkürlerimi sunuyorum.

İsmail ERGÜN



İÇİNDEKİLER

	Sayfa
İÇ KAPAK	i
ONAY	ii
ETİK BEYAN	iii
ÖZET	iv
ABSTRACT	v
TEŞEKKÜR	vi
İÇİNDEKİLER	vii
TABLOLAR DİZİNİ	viii
ŞEKİLLER DİZİNİ	ix
KISALTMALAR ve SİMGELER	x
1. GİRİŞ	1
2. KAYNAK ARAŞTIRMALARI	3
2.1. Literatür Araştırması	3
2.2. Gümrük Usulsüzlüğü	4
2.3. CIF Kıymet	6
2.4. Armonize Sistem	6
2.5. Gümrük Tarife İstatistik Pozisyonu (GTİP)	7
3. MATERYAL ve YÖNTEM	8
3.1. Veri Setinin İncelemesi	8
3.2. Verideki Özniteliklerin Görselleştirmesi ve Açıklaması	13
3.3. TensorFlow ve Keras	22
3.4. Yapay Sinir Ağları	22
3.5. Makine Öğrenmesi	23
3.6. Derin Öğrenme	23
3.7. Denetimli Öğrenme	24
3.8. Denetimsiz Öğrenme	24
3.9. K-En Yakın Komşu (K-EYK)	25
3.10. Destek Vektör Makineleri (DVM)	26
3.10.1. Sigmoid Çekirdek	27
3.10.2. Polinom Çekirdek	28
3.10.3. Doğrusal Çekirdek	28
3.10.4. Radyal Tabanlı Fonksiyon Çekirdeği (RTF)	29
3.11. Sınıflandırma	30
3.12. Veri Setinin Eğitim ve Test Olmak Üzere İkiye Bölünme Aşaması	30
3.13. Sentetik Azınlık Yüksek Hızda Örnekleme Tekniği (SAYHÖ)	30
3.14. Performans Değerlendirmesi	31
3.15. Karışıklık Matrisi	31
3.16. Doğruluk	32
3.17. Kesinlik	32
3.18. Hassasiyet	32
3.19. F-1 Skoru	33
3.20. İşlem Karakteristik Eğrisi (İKE)	33
4. BULGULAR ve TARTIŞMA	35
4.1. Araştırmanın Amacı	35
4.2. Araştırmanın Sınırlılıkları	35
4.3. Araştırmanın Önemi	36
4.4. Veri Ön İşleme	36
4.5. Verilere Modellerin Uygulanması	37
4.5.1. K-EYK Algoritmasının Uygulanması	38
4.5.2. Destek Vektör Makinesi Algoritmasının Uygulanması	39

4.5.3. Derin Öğrenme Modelinin Uygulanması	40
4.6. Algoritmaların Performans Değerlendirmesi	41
5. SONUÇ ve ÖNERİLER	44
KAYNAKLAR	46
ÖZGEÇMİŞ	52



TABLolar DİZİNİ

	Sayfa
Tablo 2.1. Elmanın GTİP sistem kodları	7
Tablo 3.1. Veri setinin ilk 5 satırı	8
Tablo 3.2. Veriye ait özniteliklerin türleri	9
Tablo 3.3. Gümrük işlemlerine ait veri seti açıklaması	9
Tablo 3.4. Türçe adlandırılan başlıklar	10
Tablo 3.5. Özniteliklerin değişken sayıları	11
Tablo 3.6. Karışıklık matrisi	31
Tablo 4.1. SAYHÖ yöntemi uygulama sonucu	37
Tablo 4.2. K-EYK algoritması karışıklık matrisi sonuçları	38
Tablo 4.3. K-EYK algoritması performans değerlendirmesi	38
Tablo 4.4. DVM algoritması performans sonuçları	39
Tablo 4.5. Derin öğrenme performans değerlendirmesi sonucu	41
Tablo 4.6. K-EYK, DVM ve Derin öğrenme performans değerlendirmesi sonuçları	42

ŞEKİLLER DİZİNİ

	Sayfa
Şekil 3.1. Yıllara göre veri sayıları	12
Şekil 3.2. Isı haritası	12
Şekil 3.3. En çok yasadışı işlem yapılan aylar	13
Şekil 3.4. En çok yasadışı işlem yapılan günler	14
Şekil 3.5. En çok yasadışı işlemi yapan gümrük idareleri	14
Şekil 3.6. En çok yasadışı işlemi yapan firmalar	15
Şekil 3.7. En çok yasadışı işlemin yapıldığı ürün	15
Şekil 3.8. En çok yasadışı işlem yapılan ülkeler	16
Şekil 3.9. En çok yasadışı işlem yapılan CIF değeri	16
Şekil 3.10. En çok yasadışı işlem yapılan ürün sayısı	17
Şekil 3.11. En çok yasadışı işlem yapılan ürün ağırlıkları	17
Şekil 3.12. Toplam vergi tutarı bazında en çok yasadışı işlem yapılan vergi miktarları	18
Şekil 3.13. En çok yasadışı işlem yapılan vergi cezası tutarları	18
Şekil 3.14. En çok yasadışı işlem yapan gümrük idaresi ve firma eşleşmesi	19
Şekil 3.15. CIF değeri histogram grafiği	20
Şekil 3.16. Toplam vergi tutarı histogram grafiği	20
Şekil 3.17. Ürün sayısı histogram grafiği	20
Şekil 3.18. Ürün fiziksel ağırlık histogram grafiği	21
Şekil 3.19. Vergi cezası tutarı histogram grafiği	21
Şekil 3.20. Yapay sinir ağı yapısı	22
Şekil 3.21. Makine öğrenmesinin çalışma diyagramı	23
Şekil 3.22. Denetimli makine öğrenmesine ait diyagram	24
Şekil 3.23. Denetimsiz makine öğrenmesine ait diyagram	25
Şekil 3.24. K-EYK Algoritması Örneği	25
Şekil 3.25. K Değeri Örneği	26
Şekil 3.26. DVM algoritması yapısı	27
Şekil 3.27. Sigmoid çekirdek diyagramı	27
Şekil 3.28. Polinom çekirdek diyagramı	28
Şekil 3.29. Doğrusal çekirdek diyagramı	29
Şekil 3.30. RTF çekirdek diyagramı	30
Şekil 3.31. Örnek İKE eğrisi	34
Şekil 4.1. Yasal ve yasadışı olan işlemler	37
Şekil 4.2. K-EYK algoritması İKE eğrisi	38
Şekil 4.3. DVM çekirdekleri İKE-EAA eğrileri	40
Şekil 4.4. Derin öğrenme modeli karışıklık matrisi	40
Şekil 4.5. Derin Öğrenme İKE eğrisi	41
Şekil 4.6. K-EYK, DVM ve derin öğrenme İKE-EAA grafiği	42

KISALTMALAR ve SİMGELER

Kısaltma/Simge	Tanım
ABD	Amerika Birleşik Devletleri
API	Application Programming Interface (Uygulama Programlama Ara yüzü)
BACUDA	Band of Customs Data Analysts (Gümrük Veri Analistleri Grubu)
CIF	Cost, Insurance and Freight (Mal bedeli, Sigorta ve Taşıma)
CTGAN	GAN Tabanlı Veri Sentezleyici
CPU	Central Processing Unit (Merkezi İşlem Birimi)
DP	Doğru Pozitif
DN	Doğru Negatif
DVM	Destek Vektör Makineleri
EAA	Eğri Altındaki Alan
YP	Yanlış Pozitif
YN	Yanlış Negatif
FOB	Free On Board (Gemi Bordasında Teslim)
GTİP	Gümrük Tarife İstatistik Pozisyonu
GPU	Graphics Processing Unit (Grafik İşlemci Ünitesi)
HS	Harmonized System (Armonize Sistem)
İKE	İşlem Karakteristik Eğrisi
K-NN	K-Nearest Neighbors (k-En Yakın Komşular)
RBF	Radial Basis Function (Radyal Taban Fonksiyonu)
ROC	Receiver Operating Characteristic Curve (İşlem Karakteristik Eğrisi)
SAYHÖ	Sentetik Azınlık Yüksek Hızla Örnekleme Tekniği
WCO	World Customs Organization (Dünya Gümrük Örgütü)

1. GİRİŞ

Verilerin dijital ortamda tutulmasına ilk olarak 1986 yılında başlanmıştır (Hilbert & Lopez, 2011). Günümüzdeyse hemen hemen tüm veriler dijital ortamdadır. Veri kaydedebilen elektronik cihazların kullanımının artması, teknolojinin ve bilgisayar sistemlerinin gelişmesiyle birlikte büyük veriler gün geçtikçe daha da büyümektedir. Öyle ki, dünyadaki tüm kitapların toplamı, dijital ortamdaki tüm insan verilerinin toplamının en fazla %6'sını oluşturmaktadır (Price, 2020). Güncel verilere göre günde en az 2,5 kentilyon veri üretildiği tahmin edilmektedir (Domo, 2020).

Data (Veri); ham gerçeklerdir. Veri, bilginin yapı taşlarını oluşturan kelimeler, sayılar, görüntüler, videolar veya diyagramları içerir (Richmond, 2006). Information (Enformasyon); tüm bu verilerden bir anlam çıkararak karar verme sürecinde problemleri çözmeye veya bir fırsatı gerçekleştirmeye yardımcı olmaktır. Her mesajda olduğu gibi enformasyonda da bir verici bir de alıcı bulunmaktadır, enformasyonun amacı alıcının bir konudaki düşüncelerini değiştirmektir (Yılmaz, 2013), değerlendirmesi ya da davranışı üzerinde bir etki yaratmaktır (Devenport & Laurence, 2001). Knowledge (Bilgi); zihinde bir biliş oluşturma ve farkındalık oluşturarak bu verilerle ve enformasyon sonucunda artık karar vermedir. Yani kısaca bilgi, değer yaratmaktır (Liew, 2007). Bu nedenlerle; verilerden bir tahmin veya çıkarım elde etmek için bilgiye dönüştürülmesi gerekir. Makine öğrenmesinde, verilerin bilgiye dönüştürülmesi için de verilerin madenciliği yapılarak veri analiz yöntemleri uygulanmalıdır. Veri analizi; toplanan verilerin kütlesine düzen, yapı ve anlam getirme sürecidir (Marshall & Rossman, 1989).

Bilginin elde edilmesinden sonra işlenmesi, yorumlanması, denenmesi, paylaşılması, kaydedilmesi, değerlendirilmesi, saklanması gibi bilgi yönetim süreci aşamalarından da geçirilmesi gerekmektedir. Bilginin yapı taşı olan verilerin analiz edilmesi, büyük verilerde makine öğrenmesi ve derin öğrenme yöntemleriyle yapılmaktadır. Veri analizi 1999 yılında tanıtılmıştır (Webber, 2013).

Bu çalışmada da Nijerya ülkesine ait gümrük ithalat verilerinin olduğu bir veri seti üzerinde tahminlerde bulunmak için klasik makine öğrenmesi ve derin öğrenme metotları kullanılarak veri analizi yapılmıştır. Bu analiz, gerekli olan paketleri ve ortamları sunan Anaconda yazılımı ve Python programlama dili kullanılarak yapılmıştır. Buradaki veri analizi ikili sınıflandırma yöntemidir. Uygulanacak yöntemler sonucunda bilgisayardan, yapılan ithalat işleminin yasal veya yasadışı olup olmadığını tahmin etmesi istenir. Yasal veya yasadışı işlemi tahmin etmek için iki ayrı makine öğrenmesi ve derin öğrenme modeli kullanılmış ve sonuçları karşılaştırılarak en iyi performansı gösteren modeller tezin sonunda belirtilmiştir.

Tez, beş bölümden oluşmaktadır. Çalışmanın birinci bölümü giriş bölümüdür. İkinci bölümde; literatür araştırması yapılmış, tezin konusuyla ilgili bilgiler ve tanımlara yer verilmiştir.

Üçüncü bölüm olan materyal ve yöntem bölümünde; analizi yapılacak olan Nijerya ülkesine ait gümrük ithalat verilerinin bulunduğu bir veri setiyle ilgili bilgiler paylaşılmıştır. Ayrıca çalışmada kullanılan makine öğrenmesi ve derin öğrenme algoritmalarının neler olduğu, hangi yöntemlerin kullanıldığı ve bu yöntemlerin detaylı bilgilerine yer verilmiştir. Dördüncü bölüm olan bulgular ve tartışma bölümünde; araştırmanın amacından, öneminden ve sınırlılıklarından bahsedilmiştir. Öte yandan, verilere makine öğrenmesi ve derin öğrenme modellerinin uygulanması, bu modellerin performans sonuçlarının değerlendirilmesi ve tartışma konuları belirtilmiştir. Beşinci ve son bölüm olan sonuçlar ve öneriler bölümünde; araştırmanın sonucu yazılarak, kurumlara ve araştırmacılara birtakım önerilerde bulunulmuştur.



2. KAYNAK ARAŞTIRMALARI

Bu bölümde, literatür araştırması yapılmış, gümrük usulsüzlüklerinin neler olduğu, nasıl yapıldığı ve hukuksal boyutları anlatılmış, gümrük terminolojisinde yer alan bazı tanımlara yer verilmiştir.

2.1. Literatür Araştırması

Gümrük ithalatlarındaki yasadışı işlemlerin tespiti için makine öğrenimiyle ilgili 2020 yılı temmuz ayında ABD'nin New York eyaletinde düzenlenen 26. ACM SIGKDD Uluslararası Bilgi Keşfi ve Veri Madenciliği Konferansında gümrük ithalatındaki yasadışılığın tahmin edilmesine yönelik makine öğrenmesi çalışmaları yayınlamıştır. BACUDA (Band of Customs Data Analysts, Gümrük Veri Analistleri Grubu) ve Dünya Gümrük Örgütü üyeleriyle veri bilimcilerin ortaklaşa yürüttükleri bu çalışmada; açık kaynak kodlu yayınlanan bu algoritmalarından, üye ülkeler tarafından fayda sağlaması açısından kendi verilerini kullanmaları ve algoritmanın geliştirilmesi hedeflenmiştir.

Konferansta, çalışmalarına yer veren ve bu konuda makale yayınlayan (Kim, ve diğerleri, 2020) kendi çalışmalarında klasik makine öğrenmesi modelleri olan Karar Ağacı (Decision Tree), Aşırı Gradyan Arttırma (XGBoost), Lojistik Regrasyon (Logistic Regression), Rastgele Orman (Random Forest), Uyarlamalı Güçlendirme (AdaBoost), Hafif Gradyan Güçlendirme Makineleri (LGBM), Kategori Yükseltme (CatBoost) ve derin öğrenme modeli uygulamışlardır.

Kaçakçılık ve diğer gümrük sahtekârlığı türleriyle mücadele çabalarını güçlendirmek, gümrük sahtekârlığının önlenmesi ve tespiti için yasal düzenlemelerin ve pratik önlemlerin sayısını analiz ederek, gümrük dolandırıcılığı yaptığından şüphelenilen kişilere yöneltilen cezai suçlamalarla ilgili (Vasileska, 2012), istatistiksel verileri çalışmasında analiz etmiştir.

Gümrük dolandırıcılığının önüne geçilebilmesi için (Roman, ve diğerleri, 2009) yapılandırılmamış veri tabanlarının tutarsızlıkları ve belirsizliklerinden muzdarip olan gümrük idarelerine yönelik yeni bir veri tabanı sistemi önerisinde bulunmuşlardır.

Gümrük sahtekârlıklarının tespiti için (Vanhoeyveld, Martens, & Peeters , 2020) EasyEnsemble yöntemi adında bir yöntem geliştirdiklerini ve bu sayede gümrük vergi kayıplarının önüne geçerek caydırıcı tedbirler sağladıklarına yönelik yaptıkları çalışmalarında yüksek EAA değeri elde ettiklerini söylemişlerdir.

Gümrük ihracat ve ithalatındaki yasadışı işlemlerin önlenmesine yönelik manuel denetimlerin en aza indirilmesi ve risk yönetimine ilişkin hesaplamalı istihbarat çözümlerinin geliştirilmesi için (Digiampietri, ve diğerleri, 2007) bazı kurum ve kuruluşlarla ortaklaşa yaptıkları HARPIA projesinde çalışmalarını sunmuşlardır.

Seçim simülasyonu yardımıyla, ithal edilen ürünlerin gümrük memurları tarafından kontrol edilmeden önce ne tür ürünler oldukları hakkında bir takım ön bilgiler elde ederek (Kim, ve diğerleri, 2020) stratejik çalışmalar yapmışlardır.

Bu tez çalışmasında; (Kim, ve diğerleri, 2020) dışında klasik makine öğrenmesi modelleri olan k-En Yakın Komşu (k-Nearest Neighbors), Destek Vektör Makineleri (Support Vector Machine) ve Derin Öğrenme (Deep Learning) modeli uygulanmıştır.

Uygulama aşamasında derin öğrenme modeli için farklı yol ve yöntemler denenmiştir. Modelimizde 5 gizli katman kullanılmış ve “dense” katmanı nöronlara bağlılığı farklılaştırılarak 1000, 500, 100, 50, 10 şeklinde seçilmiştir. Derin öğrenmede (Kim, ve diğerleri, 2020) %80 doğruluk oranı elde etmişken, yapılan ince ayarlar sonucunda doğruluk %83’e çıkarılmıştır.

2.2. Gümrük Usulsüzlüğü

Türkiye Cumhuriyeti kanunlarında, kişilerin 4458 sayılı Gümrük Kanunu’na aykırı davranışları “kabahat” olarak hükmedilmiş ve idari para cezası ile cezalandırılmaları gerektiği şeklinde düzenlenmiştir. 5607 sayılı Kaçakçılık Kanunu kapsamında yapılan davranışlar ise “suç” olarak tanımlanmış ve hapis cezasını gerektiren yaptırımlar olarak uygulanmaktadır.

Çalışmamızda; 5607 sayılı Kaçakçılık Kanununda belirtilen hususların dışında, sadece 4458 sayılı Gümrük Kanunu’nun kapsamında yer alan hususlara değinilecektir.

Gümrük usulsüzlüğünü açıklamak maksadıyla; Türkiye Cumhuriyeti kanunlarında yer alan suç ve kabahat ayırımından söz edilmiş olsa da gümrüklerde yaşanan usulsüzlükler, özellikle ithalat aşamasındaki vergi kaçırma olayları tüm dünya ülkelerinde var olan hususlar olduğundan, çalışmamızın bu kapsamda evrensel niteliği bulunmaktadır.

Yabancı bir ülkeden mal veya hizmet alınmasına ithalat, kişi veya kuruluşlarca üretilen mal veya hizmetlerin yurtdışına satılmasına ise ihracat denmektedir (MEB, 2011).

4458 sayılı Gümrük Kanunu’na bakıldığında, “Cezalar” başlığı altında; 231 ve 233’üncü maddelerde, Vergi Kaybına Neden Olan İşlemlere Uygulanacak Cezalar başlığı altında 234 ve 238’inci maddelerde, Usulsüzlüklere İlişkin Cezalar başlığı altındaki 239 ila 241’inci maddelerde, gümrük ithalatındaki usulsüzlüklere ilişkin ek vergiler ve cezalarla ilgili açıklamalar yapıldığı görülmektedir (Gümrük Kanunu, 1999). Bu kanun maddelerinin gümrük usulsüzlükleri konusundaki hükümleri incelendiğinde; Gümrük Kanunu’nun 231’nci maddesi; vergi cezalarının zamanaşımı sürelerinden, 232’inci maddesi vergi cezalarıyla birlikte alınması gereken cezaların aynı anda tebliğ edilmesinden, 233’üncü maddesi belirtilen vergi cezalarının uygulanmasından önce ihbar verilmesi durumunda muhbire ödeme yapılmasına ilişkin hususlardan bahsedildiği görülmektedir.

Vergi kaybına neden olan işlemlere bakıldığında özetle;

- Gümrük Kanunu'nun 234'üncü maddesi, gümrük tarifesine esas unsurlarda veya vergilendirmeye ilgili olan işlemlerde, ithal edilen mal veya eşyaların sayı, ağırlık gibi ölçülerindeki hatalı bildirimlerde veya beyan edilen gümrük vergisi ile muayene sonucu ortaya çıkan vergi miktarı arasında, %5 den daha büyük bir aykırılık söz konusu olduğunda, yaptırım uygulanacak vergi cezalarıdır (Büyükgöral & Türkoğlu, 2015).
- Gümrük kanununun 235'inci maddesi; ithali yasak, lisansa, şarta, izne, kısıntıya veya belli kuruluşların vereceği uygunluk veya yeterlilik belgesine tabi eşyalarla ilgili aykırılıkların gümrük idaresince tespit edilmesi sonucunda uygulanacak olan vergi cezalarıdır (Uzun, 2016),
- Gümrük Kanunu'nun 236'nci maddesi, antrepolarda veya gümrük idareleri tarafından müsaade edilen yerlere konulan eşyalarda yapılan kontrollerde eksiklik veya fazlalık çıkması, bu yerlerden izin alınmadan eşya çıkarılması ve gümrüksüz satış mağazalarından (freeshop vb.) hak sahibi olmayanlara eşya satışı gibi hususlarda verilecek olan vergi cezalarıdır.
- Gümrük Kanunu'nun 237'nci maddesi, özet beyan verilmemesi, özet beyanla kontrol sonucunda tespit edilen eşyanın fazla veya eksik olması veya beyanların kaybolduğu veya çalındığının ispat edilememesi durumlarında uygulanacak olan vergi cezalarıdır.
- Gümrük Kanunu'nun 238'inci maddesi, dâhilde işleme, gümrük kontrolü altında işleme veya geçici ithalat rejimiyle ilgili ihlallerin meydana gelmesi sonucunda uygulanacak olan vergi cezalarıdır (Uzun, www.nazaligundem.com, 2019).
- Gümrük Kanunu'nun 239'uncu maddesi, Vergilerden muaf olan ithal veya ihraç edilecek olan bir eşyanın gümrük işlemlerine tabi tutulmaksızın gümrük kapıları dışındaki yerlerden izinsiz olarak geçirilmesiyle ilgili uygulanacak olan vergi cezalarıdır (Yılmaz Ö. , 2018).
- Gümrük kanununun 241'inci maddesi, Gümrük Kanunu'nda belirtilmeyen ancak Gümrük Kanunundaki yetkilere dayanan diğer usulsüzlüklere uygulanacak vergi cezalarıdır.

Yukarıda maddeler halinde tanımlanan kabahatler ve yasadışı işlemler, Türkiye Cumhuriyeti'ne ait gümrük usulsüzlüklerine ilişkin kanun maddeleri olup, devletin vergi kaybını önlemek için düzenlenmiştir. Diğer ülkelerin de vergi kayıplarını önlemek için kendilerine özgü yasal düzenlemeleri ve yaptırımları bulunmaktadır.

Gümrük ithalat işlemlerinde, ithalatçı firmaların özellikle devlete daha az vergi ödemek maksadıyla, ithal ettikleri ürünlerin veya hizmetlerin CIF değerini düşük göstermek, ithal eşyanın

miktar veya ağırlıklarını farklı beyan etmek gibi durumlar ortaya çıkmaktadır. Bu tip yasadışı işlemler gümrük idareleri tarafından gümrük kontrolünden geçerek, eşyalar gümrük muayene memurları tarafından muayene edilmektedir. Kontrol amacıyla, görevli gümrük memurları haricinde müfettişler, müfettiş yardımcıları, kontrolörler ve gümrük idare amirleri tarafından ikinci muayeneleri her zaman yapılabilmektedir (Gümrük yönetmeliği, 2009).

Gümrük idarelerinin kendi yapısal sistemlerinde elektronik ortamda bir takım risk analizleri de yapılmaktadır. Daha önce yasadışı işlem yapan bir firmanın yeni bir beyanname açması durumunda, bu firmanın ürünleri özellikle denetlenmektedir (TB. Yönerge, 2020).

2.3. CIF Kıymet

CIF İngilizce Cost (Maliyet), Insurance (Sigorta), Freight (Navlun) kelimelerin baş harflerinden oluşan uluslararası bir terimdir (Çoban, 2020). Navlun ürüne ait gemi yük taşıma ücretidir. CIF bedel; mal veya ürünün imalatçıdan alındığı fiyatın üzerine sigorta ve taşıma ücretinin de eklenmesinden sonra oluşan toplam fiyata denilmektedir. Dış ticarete ihracattaki değer FOB ile belirlenirken, ithalatta ise CIF değeriyle belirlenir (Özdamar, 2013). Bununla birlikte ithal edilen ürünlere uygulanacak vergi cezaları CIF değeri üzerinden uygulanmaktadır (Adak, 2010).

2.4. Armonize Sistem

HS (Harmonized System, Harmonize Sistem) başlangıç uzunluğu 6 haneden oluşan ve ülkelere göre uzunluğu değişmekte olan uluslararası bir koddur. Bu kod uluslararası ticareti kolaylaştırmak amacıyla, Dünya Gümrük Örgütü (WCO) tarafından bir mal sınıflandırma sistemi olarak geliştirilmiştir. 14 Haziran 1983 tarihinde Gümrük İş Birliği Konseyi tarafından Brüksel’de yapılan sözleşmeyle Türkiye’de bu sınıflandırma sistemine taraf olmuştur (3501 Sayılı Kanun, 1988). Ayrıca, 23 Mart 2021 tarihi itibarıyla, 183 ülke ve Gümrük birlikleri bu sözleşmeye taraf olmuş ve gönüllü olarak uygulamaktadırlar (WCO, 2019) (Wind, 2008). Bu sınıflandırmanın ilk bölümlerinde çoğunlukla eşyanın ham veya yarı mamul hali yer alır. Bölümler ilerledikçe bu ham maddenin işlenmiş halinin işlem düzeyi de artar.

HS kodun ilk 6 rakamı Dünya Gümrük Örgütü’ne üye tüm ülkelerce kullanılan Armonize Sistem (Harmonized System-HS) Nomanklatürü (adlar dizgisi) kodunu, 7-8’inci rakamları Avrupa Birliği ülkeleri tarafından kullanılan Kombine Nomanklatür (Combined Nomenclature-CN) kodunu, 9-10’uncu rakamları farklı vergi uygulamaları nedeniyle açılan pozisyonları gösteren kodlarını, 11-12 inci rakamlar ise istatistik kodlarını göstermek için kullanılmaktadır (www.muhasabenews.com, 2019). Ayrıca Armonize Sistemin, kullanıcı ülkelerin kendi ulusal

ihtiyaçlarına göre alt açılımlar yapmasına imkân veren bir mimari yapısı vardır (www.tarifemevzuati.com, 2021). Bu alt açılımlar Türkiye’de Gümrük Tarife İstatistik Pozisyonu (GTİP) olarak adlandırılır.

2.5. Gümrük Tarife İstatistik Pozisyonu (GTİP)

Avrupa birliği ülkeleri ve Türkiye, Armonize koda ait 6 basamaklı sayıya, kendi aralarında oluşturdukları 2 basamak daha ekleyerek, Kombine Nomanklatür’ü (Commission Implementing Regulation (EU) 2019/1811, 2019) adı verilen ve Avrupa Birliği ülkeleri arasındaki mal ve ürün ithalatında kullandıkları 8 basamaklı bir kod üretmişlerdir.

Türkiye’de farklı vergi uygulamalarını yansıtmak amacıyla 2 basamak daha ekleyerek 10 basamaklı bir kodlama sistemi elde etmiştir. İstatistik kodlarını göstermek amacıyla 2 basamak daha eklenerek sonuçta 12 basamaklı ulusal bir kod sistemi oluşturulmuş (Altıntaş, 2018) ve bu 12 basamaklı ulusal kod sistemine de Gümrük Tarife İstatistik Pozisyonu denilmektedir. Örneğin Elma’nın Ticaret Bakanlığının resmî sitesindeki GTİP kodu aşağıdaki tabloda verilmiştir.

Tablo 2.1. Elmanın GTİP sistem kodları

GTİP Sistem Kodu	Ürün
080810	Elma
081330000000	Elma (kurutulmuş)
080810800011	Elma; Golden (taze)
080810800013	Elma; Starking (taze)
080810800019	Diğer elmalar (taze)

Kaynak: Ticaret Bakanlığı, 2021

3. MATERYAL ve YÖNTEM

Bu bölümde; veri setinin içeriği anlatılmış, veri ön işleme yapılmış ve veri setinin görselleştirmesi yapılarak veriye ait öznitelikler hakkında ayrıntılı bilgiler verilmiştir. Daha sonra makine öğrenme ve derin öğrenme modelleri açıklanmış, bu modellerin farklı fonksiyonları, sınıflandırmaları ve doğruluk sonuçlarının nasıl yapıldığı hususları incelenmiştir.

3.1. Veri Setinin İncelemesi

Bu çalışmada kullanılan veri, Nijerya ülkesine ait olan anonimleşmiş gümrük veri setidir (Choi, 2020). Bu veri içerisinde; 200.000 satır ve 14 öznitelikten oluşan, 130 ülkeden yapılan 2013 ve 2014 yıllarına ait gümrük ithalat işlemlerinin gerçek bilgileri yer almaktadır. Veri bilimcileri tarafından firmaların bilgi güvenliğini sağlamak adına CTGAN modeli kullanılarak orijinal veriler sentetik verilere dönüştürülmüştür. CTGAN, yüksek doğrulukta sentetik tablo verileri oluşturabilen GAN tabanlı bir veri sentezleyicidir (Xu, Skoularidou, Infante, & Veeramachaneni, 2019). Veri setinin ilk beş satırına ait içerik Tablo 3.1’de gösterilmiştir.

Tablo 3.1. Veri setinin ilk 5 satırı

Unnamed: 0	year	month	day	OFFICE	IMPORTER.TIN	TARIFF.CODE	ORIGIN.CODE	CIF_USD_EQUIVALENT	QUANTITY	GROSS.WEIGHT	TOTAL.TAXES.USD
0	2013	1	1	OFFICE60	IMP545640	8703232225	CNTRY680	172710.0	1.0	28487.0	3492.0
1	2013	1	1	OFFICE24	IMP252323	8703321220	CNTRY376	7450.0	1.0	185994.0	503.0
2	2013	1	1	OFFICE59	IMP194834	8704212025	CNTRY680	2684.0	1.0	452843.0	5706.0
3	2013	1	1	OFFICE60	IMP668188	3901100000	CNTRY277	7412.0	2018.0	26546.0	3182.0
4	2013	1	1	OFFICE59	IMP395619	8517180000	CNTRY759	2786.0	7044.0	1958.0	587.0

Tabloya bakıldığında OFFICE, IMPORTER.TIN ve ORIGIN.CODE’un orijinal verilerinin yerine, bilgi güvenliği nedeniyle dönüştürme işlemi yapılarak, sayısal veri haline getirildiği görülmektedir.

Veri içerisindeki özniteliklerin türlerine bakıldığında 3 tanesi kategorik değişken, 11 tanesi de numerik değişken olduğu görülmüştür. Tablo 3.2’de kategorik ve numerik türlerin neler olduğu açıklanmıştır.

Tablo 3.2. Veriye ait özniteliklerin türleri

Unnamed: 0	Numerik
year	Numerik
month	Numerik
day	Numerik
OFFICE	Kategorik
IMPORTER.TIN	Kategorik
TARIFF.CODE	Numerik
ORIGIN.CODE	Kategorik
CIF_USD_EQUIVALENT	Numerik
QUANTITY	Numerik
GROSS.WEIGHT	Numerik
TOTAL.TAXES.USD	Numerik
RAISED_TAX_AMOUNT_USD	Numerik
illicit	Numerik

Veri setindeki her bir öznitelige karşılık gelen Türkçe açıklamalar ve örnek veriler Tablo 3.3’de gösterilmiştir.

Tablo 3.3. Gümrük işlemlerine ait veri seti açıklaması

Başlıklar	Açıklama	Örnek veri
Unnamed:0	İsimsiz (dizide veri bulunmamaktadır)	1
year	İthalatın yılı	2013
month	İthalatın ayı	7
day	İthalatın günü	16
OFFICE	Gümrük idaresi	OFFICE33
IMPORTER.TIN	İthalat işlemini yapan firma	IMP194834
TARIFF.CODE	Tarife kodu (Armonize sistem kodu)	8704212025
ORIGIN.CODE	İthalat işleminin yapıldığı ülke	CNTRY680
CIF_USD_EQUIVALENT	Ürünlere verilen benzersiz kod (CIF veya GTİP kodu)	2684.0
QUANTITY	Ürünün miktarı	1.0
GROSS.WEIGHT	Ürünün brüt ağırlığı	452843
TOTAL.TAXES.USD	ABD doları cinsinden ürünün toplam vergi tutarı	5706.0
RAISED_TAX_AMOUNT_USD	ABD doları cinsinden yükseltilmiş vergi tutarı	0.0
illicit	“0” yasal “1” yasadışı	1

Verilerin daha iyi anlaşılması ve öznitelik başlıklarının ne anlama geldiğinin açıklanabilmesi için, İngilizce olan başlıklar Türkçe’ye çevrilmiştir. Türkçe yapılan başlıklar Tablo 3.4’de gösterilmiştir.

Tablo 3.4. Türkçe adlandırılan başlıklar

Orijinal Başlıklar	Türkçeleştirilen Başlıklar
Unnamed:0	ADSIZ
year	YIL
month	AY
day	GÜN
OFFICE	GÜMRÜK İDARESİ ADI
IMPORTER.TIN	İTHALATÇI FİRMA
TARIFF.CODE	ÜRÜN KODU
ORIGIN.CODE	GÜMRÜK ÜLKE KODU
CIF_USD_EQUIVALENT	CIF DEĞERİ
QUANTITY	ÜRÜN MİKTARI
GROSS.WEIGHT	ÜRÜN BRÜT AĞIRLIK
TOTAL.TAXES.USD	TOPLAM VERGİ TUTARI
RAISED_TAX_AMOUNT_USD	YÜKSELTİLMİŞ VERGİ TUTARI
illicit	YASADIŞI

- ADSIZ isimli başlık hiçbir anlam ifade etmemektedir.
- YIL, veri setindeki işlemlerin hangi yıla ait olduğunu göstermektedir.
- AY, işlemin yılın hangi ayında gerçekleştiğini göstermektedir.
- GÜN, işlemin ayın kaçınıcı gününde gerçekleştiğidir.
- GÜMRÜK İDARESİ ADI, işlemin hangi gümrük idaresinde yapıldığını belirtir.
- İTHALATÇI FİRMA, ithalat işleminin hangi ithalat firması tarafından gerçekleştirildiği bilgisini vermektedir.
- ÜRÜN KODU, ithalatı yapılan ürüne ait Armonize Sistem kodunu belirtmektedir.
- GÜMRÜK ÜLKE KODU, ithalatın yapıldığı ülkenin kodunu belirtmektedir. Burada veri setini üretenler tarafından gizlilik nedeniyle ülke isimleri kodlanmıştır.
- CIF DEĞERİ, işleme ait ürünün navlun, sigorta ve fatura bedeli toplamından oluşan CIF değerinin Amerikan doları cinsinden gösterilmiş değeridir.
- ÜRÜN MİKTARI, ithal edilen ürünün sayısal miktarıdır.
- ÜRÜN BRÜT AĞIRLIK, ürünlerin her birinin ağırlığını ifade etmektedir.
- TOPLAM VERGİ TUTARI, ithal edilen ürünün CIF değeri üzerinden hesaplanan ve devlete ödenmesi gereken vergi tutarını göstermektedir.
- YÜKSELTİLMİŞ VERGİ TUTARI, ithal edilen ürünlerde gerçeğe aykırı beyanda bulunduğu tespit edilmesi halinde, gümrük idaresi tarafından ithalatçı firmaya verilen gümrük vergi cezası bilgisidir.
- YASADIŞI, işlemin yasal veya yasadışı olduğunu belirtmektedir. İşlemin yasal olması halinde 0 (sıfır) ibaresi bulunmakta, yasadışı olması halinde 1 ibaresi bulunmaktadır.

YÜKSELTİLMİŞ VERGİ TUTARI Amerikan doları cinsinden ifade edilmekte ve yapılan işlemin yasal ya da yasadışı olduğunu göstermektedir. Bununla paralel olan diğer öznitelik ise illicit (Yasadışı) kolonudur. Ne zaman ki; yükseltilmiş vergi tutarı özniteliğinde değer sıfırdan büyükse illicit kolonu 1, eşit sıfır ise 0 olmaktadır.

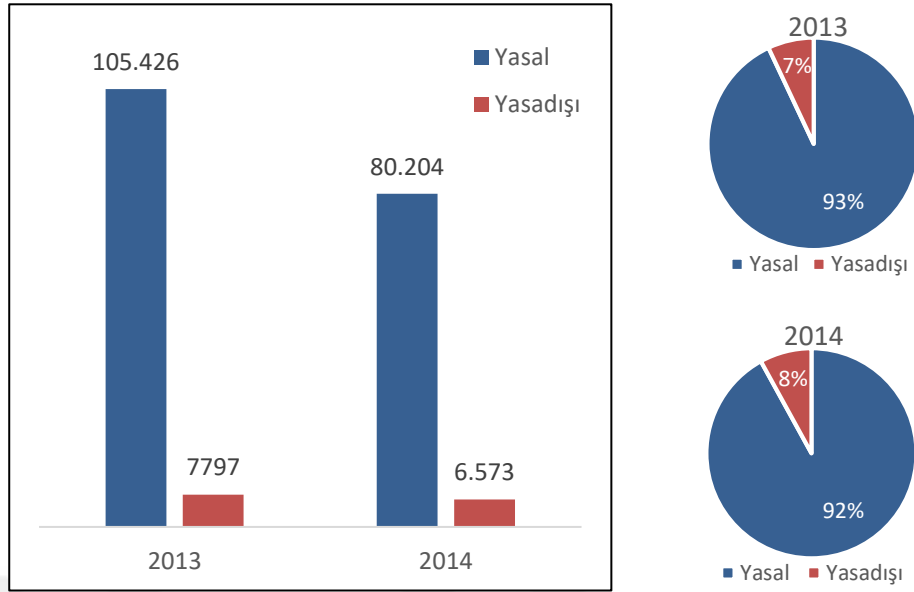
Tüm veri setindeki özniteliklerin alacağı olası değişken sayıları Tablo 3.5’de gösterilmiştir.

Tablo 3.5. Özniteliklerin değişken sayıları

Değişken türü	Değişken sayıları
YIL	2
AY	12
GÜN	31
GÜMRÜK İDARESİ ADI	73
İTHALATÇI FİRMA	15348
ÜRÜN KODU	6333
GÜMRÜK ÜLKE KODU	130
CIF DEĞERİ	68374
ÜRÜN SAYISI	18723
ÜRÜN FİZİKSEL AĞIRLIK	65775
TOPLAM VERGİ TUTARI	20395
YÜKSELTİLMİŞ VERGİ TUTARI	2305
YASADIŞI	2

Tabloya bakıldığında, YIL 2013 ve 2014 yıllarına ait verileri kapsadığı için 2 değişken, AY yılın 12 ayının tamamı, GÜN 31 günün tamamına ait verileri, GÜMRÜK İDARESİ ADI 73 adet gümrük idaresi olduğunu, İTHALATÇI FİRMA 15348 farklı firmanın işlem yaptığını, ÜRÜN KODU 6333 adet ürüne ait işlemin olduğunu, GÜMRÜK ÜLKE KODU 130 ülkeye ait ülke isimlerinin olduğu, CIF_DEĞERİ 68374 adet CIF değerinin yer aldığını, ÜRÜN SAYISI ürünlere ait 18723 adet farklı sayılarda ürün sayısı bilgisinin olduğunu, ÜRÜN FİZİKSEL AĞIRLIK 65775 adet farklı ürüne ait ağırlık bilgisinin yer aldığını, TOPLAM VERGİ TUTARI 20395 adet farklı vergi tutarı, YÜKSELTİLMİŞ VERGİ TUTARI 2305 adet farklı vergi cezası tutarı bilgisinin yer aldığını, YASADIŞI işlemlerin yasal veya yasadışı olduğunu belirttiğinden 2 değişkenin olduğunu göstermektedir.

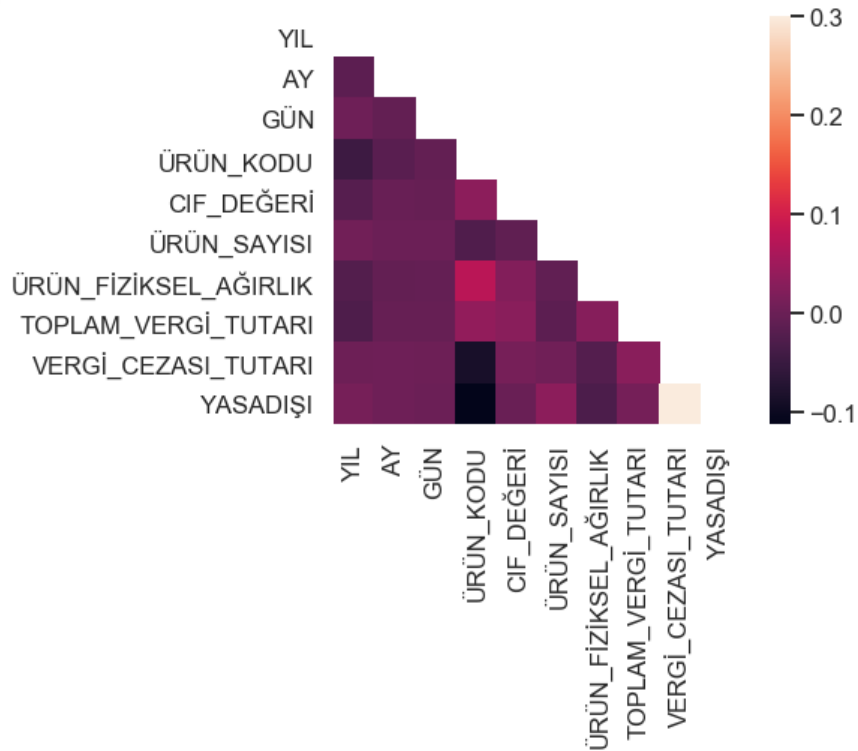
Yıllara göre veri durumuna bakıldığında; 2013 ve 2014 yıllarına ait yasal ve yasadışı işlemlere ait veri sayıları Şekil 3.1’de gösterilmiştir.



Şekil 3.1. Yıllara göre veri sayıları

Yasal ve yasadışı işlemler incelendiğinde, 2013 yılına ait verilerin %93'ü normal, %7'sinin yasadışı, 2014 yılına ait verilerin %92'si normal, %8'inin yasadışı işlemler olduğu görülmektedir.

Bu çok değişkenli veriler, ısı haritası adı verilen şekil ile çapraz incelenerek hücreler renklendirilmiş ve aralarındaki ilişki incelenmiştir. Isı haritası Şekil 3.2'de gösterilmiştir.

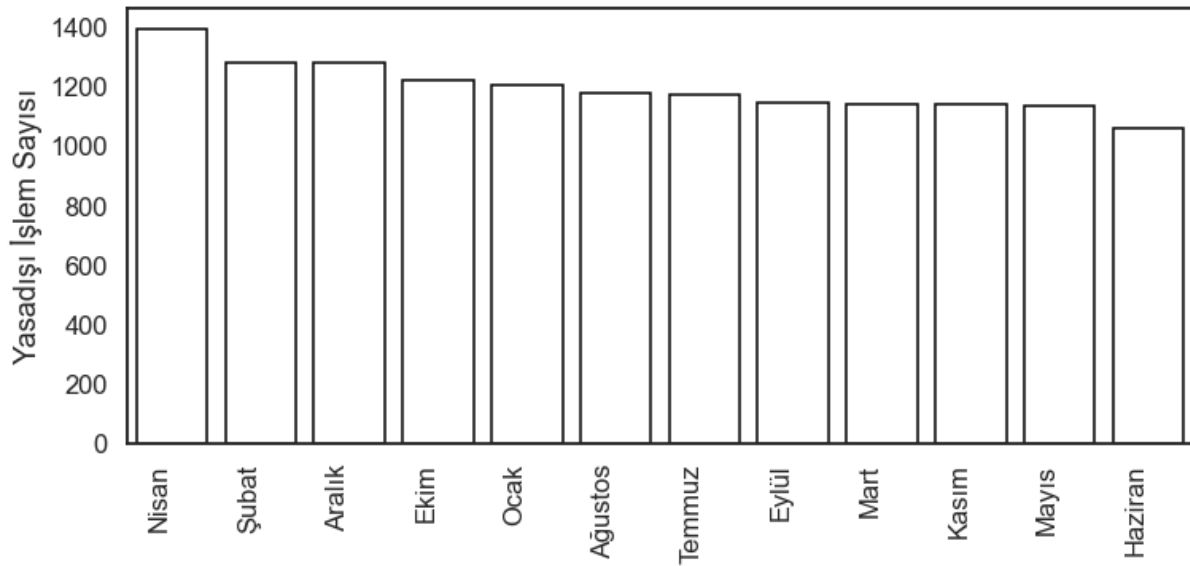


Şekil 3.2. Isı haritası

Isı haritası incelendiğinde VERGİ CEZASI TUTARI ve YASADIŞI öznitelikleri arasında açık renk (krem) yani yoğun bir ilişki olduğu görülmüştür. Bunun nedeni gümrük idareleri tarafından yasadışı işlem yapılan işlemlere ek vergi uygulanmasıdır. İthalatçı firmaya ceza olarak uygulanan bu vergiler VERGİ CEZASI TUTARI'na işlenmekte ve işlem yasadışı olduğu için de YASADIŞI bölümüne 1 rakamının eklenmesinden kaynaklanmaktadır.

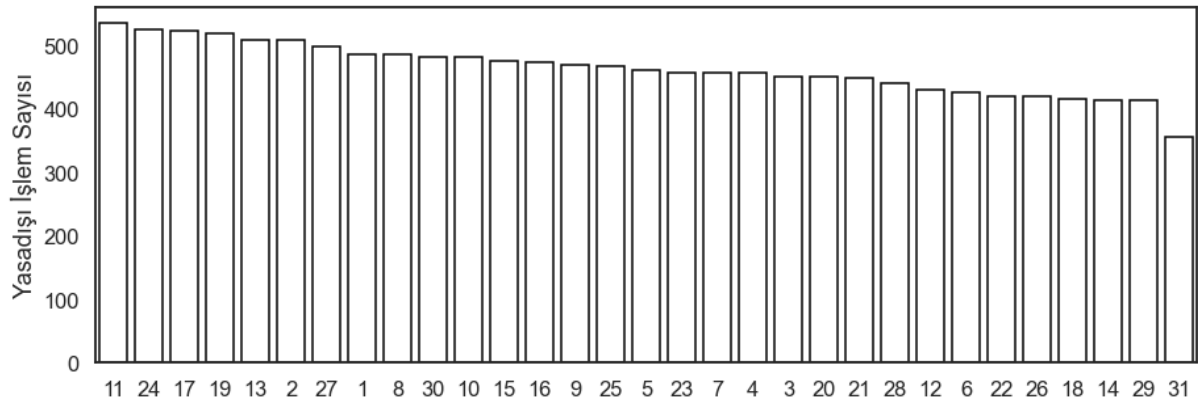
3.2. Verideki Özniteliklerin Görselleştirmesi ve Açıklaması

Veri seti, seaborn ve matplotlib kütüphaneleri kullanılarak görselleştirilmiş ve en çok yasadışı yapılan işlemlerin içerisinde her özniteliğe ait ilk 10 işlem sırasıyla görselleştirilmiştir. En çok işlem yapılan Ay'a bakıldığında 26.250 adet ile en çok Nisan ayında işlem yapıldığı, en çok yasadışı işlemin yapıldığı Ay'a bakıldığında ise, 1395 işlem ile en çok Nisan ayında yasadışı işlem yapıldığı görülmüştür. Yasadışı işlemlerin yapıldığı aylar sırasıyla Şekil 3.3'de görselleştirilmiştir.



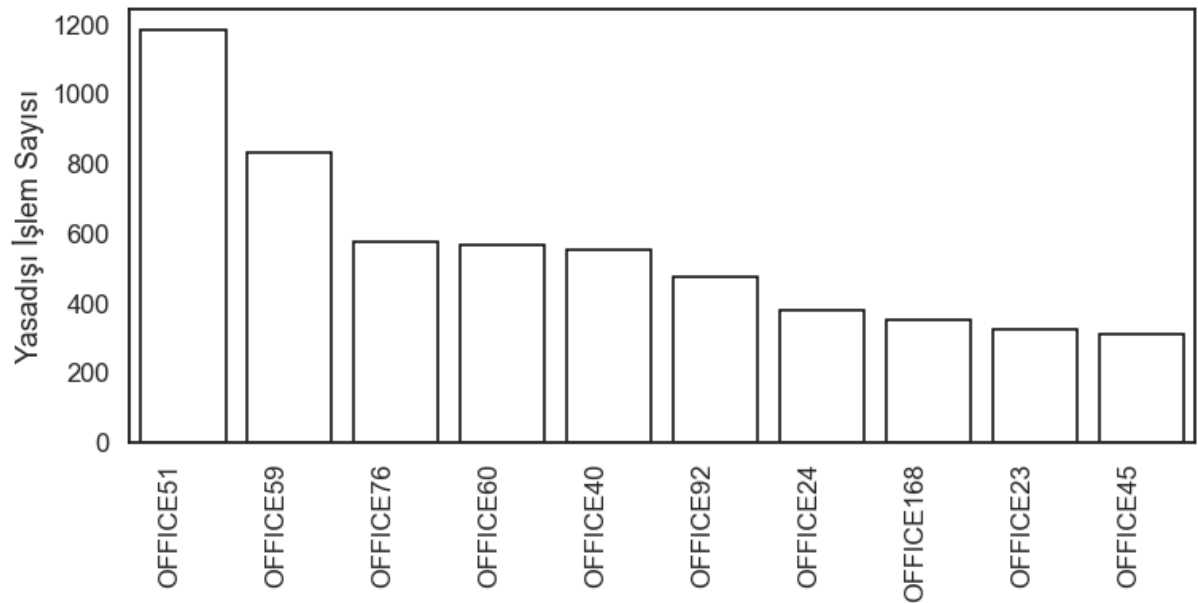
Şekil 3.3. En çok yasadışı işlem yapılan aylar

En çok işlem yapılan günlere bakıldığında; 8.344 adet ile ayın 11'inci gününde işlem yapıldığı görülmüştür. En çok yasadışı işlemin yapıldığı gün ise aşağıdaki Şekil 3.4'de gösterildiği gibi ayın 11'inci günü olup, bugünde yapılan toplam yasadışı işlem sayısı 535'dir.



Şekil 3.4. En çok yasadışı işlem yapılan günler

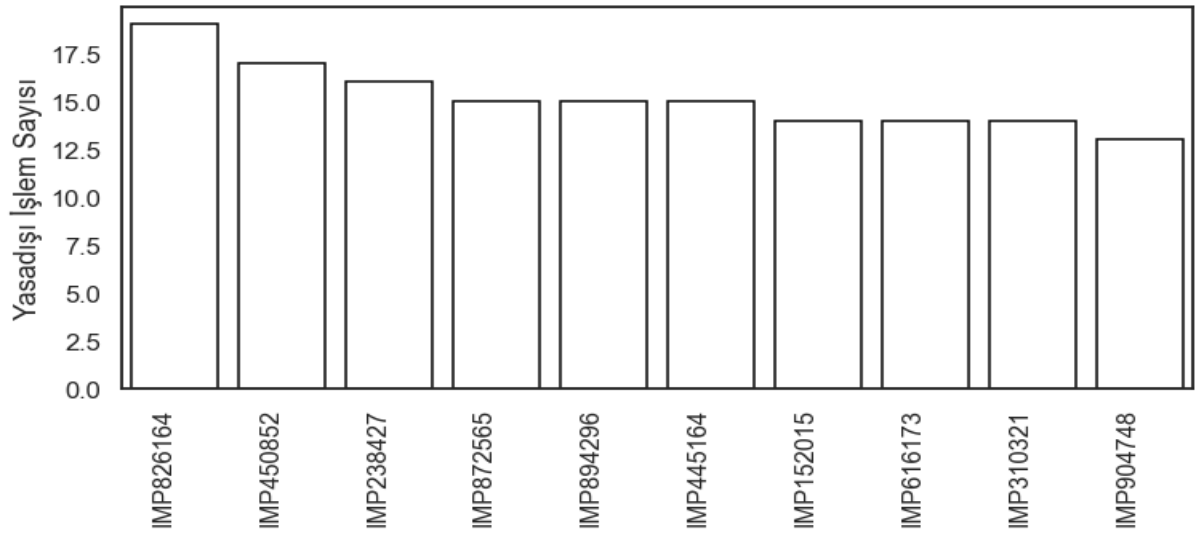
En çok işlem yapılan gümrük idarelerine bakıldığında; 37.425 adet ile en çok OFFICE51 isimli gümrük idaresinin işlem yaptığı görülmüştür. Benzer şekilde en çok yasadışı işlemin yapıldığı gümrük idaresine bakıldığında, 1184 işlem ile OFFICE51'dir. Bu durum aşağıdaki şekilde açık olarak görselleştirilmiştir.



Şekil 3.5. En çok yasadışı işlemi yapan gümrük idareleri

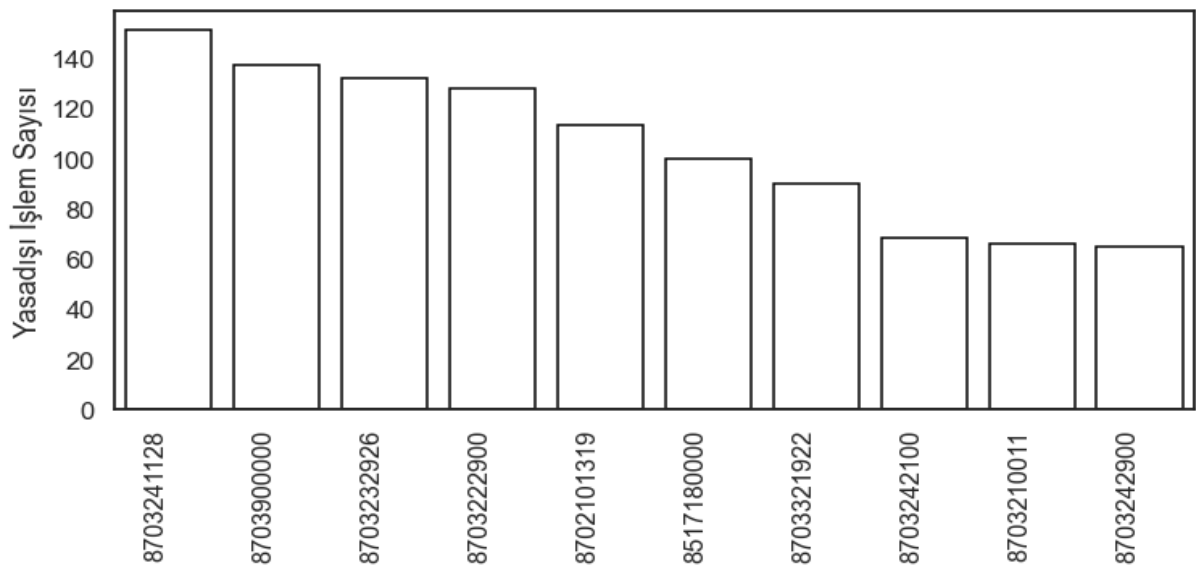
Gümrükteki yasadışı işlemlerin sebeplerinden biri olan İTHALATÇI FİRMA başlığı altında yer alan, ithalatı gerçekleştiren gümrük firmalarına bakıldığında, 1272 adet ile en çok IMP019420 isimli gümrük firmasının işlem yaptığı görülmüştür. En çok yasadışı işlemin yapıldığı ithalatçı firmaya bakıldığında, 1.087 toplam işlemi olan IMP826164 isimli firmanın 19 işlem ile en çok

yasadışı işlem yapan gümrük firması olduğu tespit edilmiştir. Yasadışı işlemlerin yapıldığı ilk 10 gümrük firması Şekil 3.6'da görselleştirilmiştir.



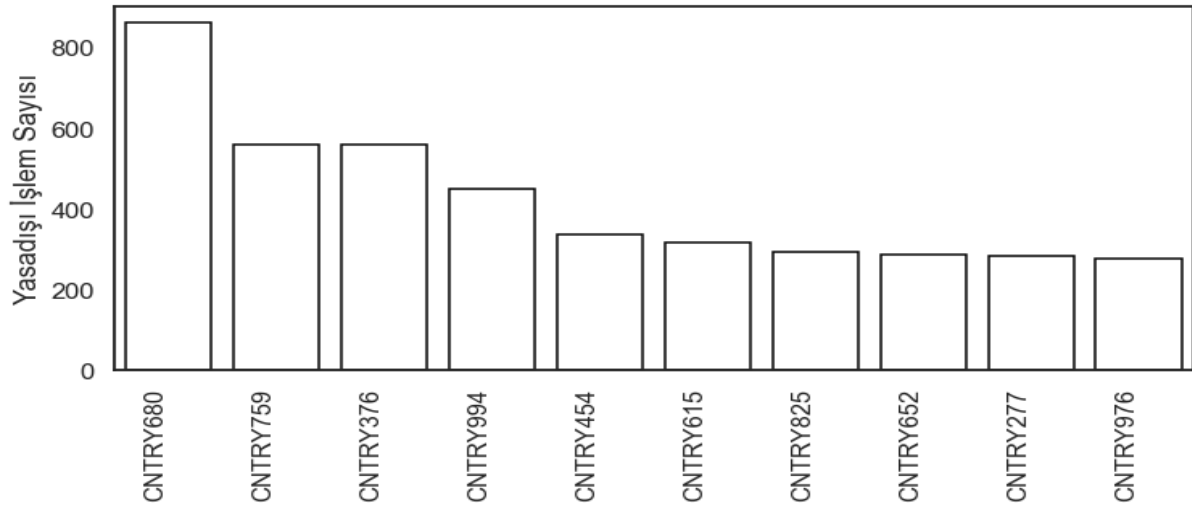
Şekil 3.6. En çok yasadışı işlemi yapan firmalar

İthalat işlemi esnasında beyan edilen ürünlere bakıldığında; 10513 işlem ile en çok 8703241128 HS kod numaralı ürünün ithalatı yapılmış olup, en çok yasadışı işlem açısından bakıldığında 151 adet ile yine 8703241128 HS kod numaralı ürün olduğu anlaşılmıştır. Bu kod araştırıldığında, silindir hacmi 3000 cc olan binek otomobile ait bir HS kodu olduğu öğrenilmiştir (www.nairametrics.com, 2011). Yasadışı işlemlerin yapıldığı ilk 10 ürün sırasıyla Şekil 3.7'de görselleştirilmiştir.



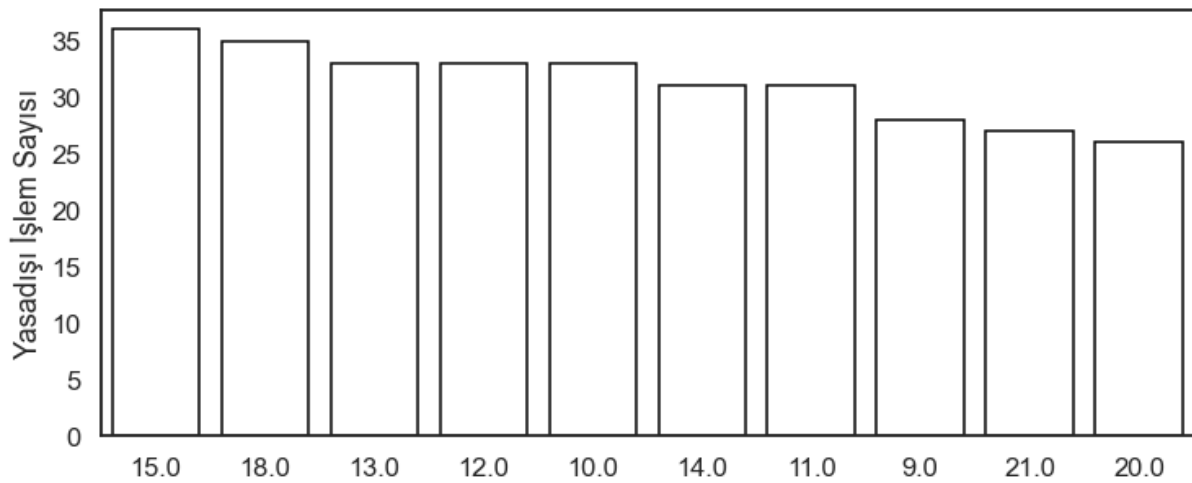
Şekil 3.7. En çok yasadışı işlemin yapıldığı ürün

En çok ithalatın yapıldığı ülkeye bakıldığında; 34461 işlem ile CNTRY680 isimli ülkeden en çok mal veya ürün alındığı analiz edilmiştir. En çok yasadışı işlemin yapıldığı ülkeye bakıldığında ise; 861 işlem ile en çok CNTRY680 isimli ülke olduğu tespit edilmiştir. Gizlilik gerekçesiyle ülkelerin ismi açık bir şekilde yazılmayarak, sayısal değer verilerek suretiyle kodlandığı değerlendirilmektedir. Yasadışı işlemlerin ülkelere göre dağılımı Şekil 3.8'de açık olarak belirtilmiştir.



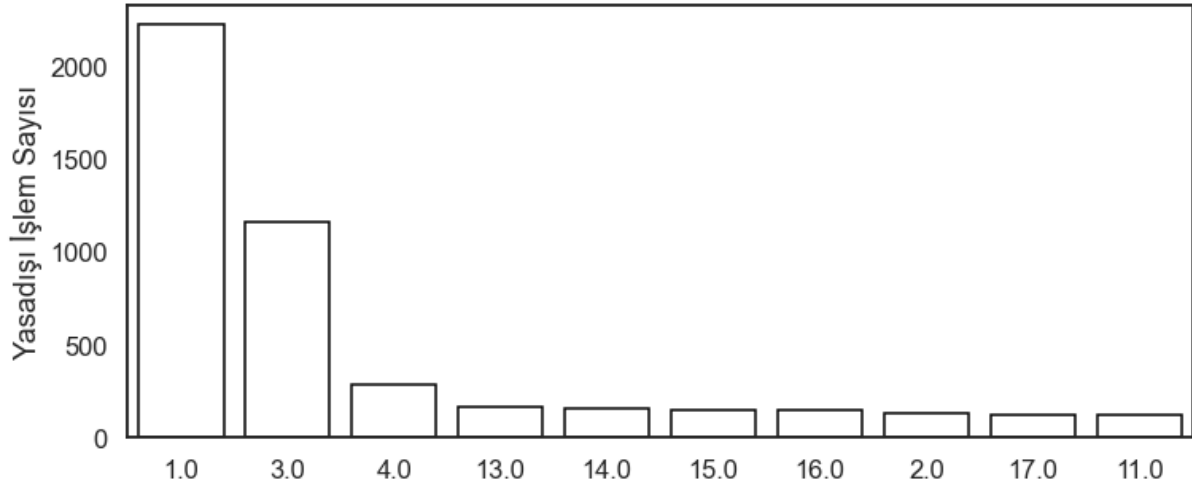
Şekil 3.8. En çok yasadışı işlem yapılan ülkeler

Verilerdeki yasadışılığı ortaya çıkaran etkenlerden biride CIF değeridir. En çok işlemin yapıldığı CIF değeri 14.0 dolar olan ürünlerin işlem sayısı yaklaşık olarak 184'dür. Diğer taraftan CIF değeri 15 dolar olan ürünlerin toplam sayısı 166'dır ve bunların 36 tanesi ise yasadışı işleme karşılık gelmektedir. İlk 10 CIF değerini görselleştirdiğimizde aşağıdaki şekli elde ederiz.



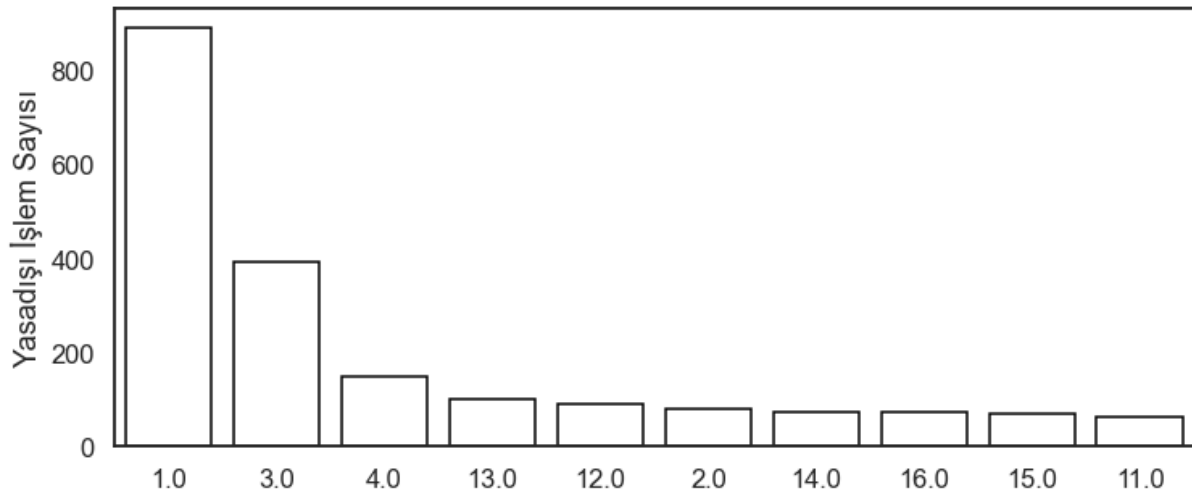
Şekil 3.9. En çok yasadışı işlem yapılan CIF değeri

İthalatı yapılan ürünlerden, miktar olarak işlemi en çok olan ürünün 48.843 adetle 1.0 miktartlı ürünler olduğu ve bu miktarın yaklaşık olarak 2.218'inin yasadışı olduğu Şekil 3.10'da gösterilmiştir.



Şekil 3.10. En çok yasadışı işlem yapılan ürün sayısı

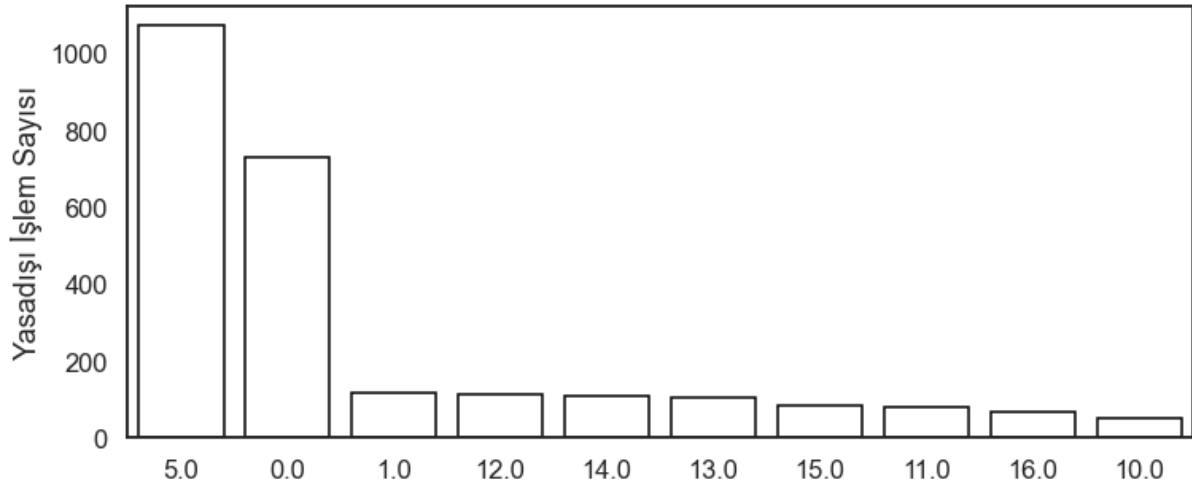
Veriye ürün ağırlıkları açısından bakıldığında en çok beyanın 24.538 adet ile 1.0 olduğu ve bu ağırlıktaki ürünlerin 889 defa yasadışı beyan edildiği görülmüştür. Ağırlık değerleri yasadışı işlem açısından görselleştirildiğinde aşağıdaki grafik elde edilir.



Şekil 3.11. En çok yasadışı işlem yapılan ürün ağırlıkları

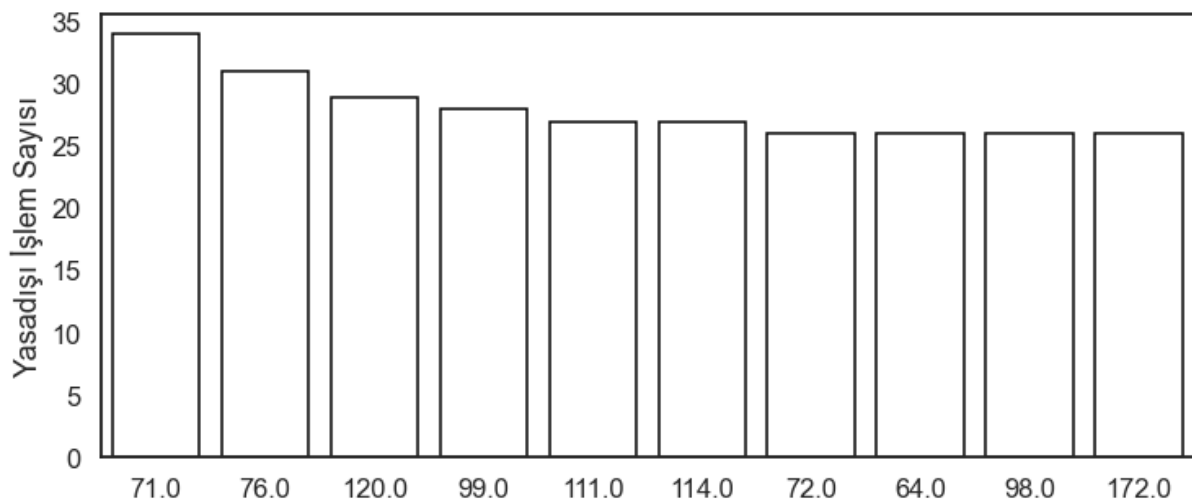
Devletler ithal ettikleri ürünlerden almış oldukları vergi miktarlarını CIF değeri üzerinden hesaplamaktadırlar. Bu değerlerde, ithalat işlemlerinde uluslararası para birimi olan Amerikan doları şeklinde belirtilmiştir. Bu ürünlere ödenecek vergi tutarlarıda, Amerikan doları cinsinden

hesaplanarak gümrük idaresine beyan edilmektedir. İthal edilen ürünlerde, 51.737 adet işlemle en çok vergi tutarı olarak beyan edilen tutarın 2.0 Amerikan doları olduğu görülmeye rağmen en çok yasadışı işlemin yapıldığı vergi tutarının 5.0 Amerikan doları olduğunu ve bunun toplam sayısının 1.071 işlem olduğu veriden elde edilmiştir.



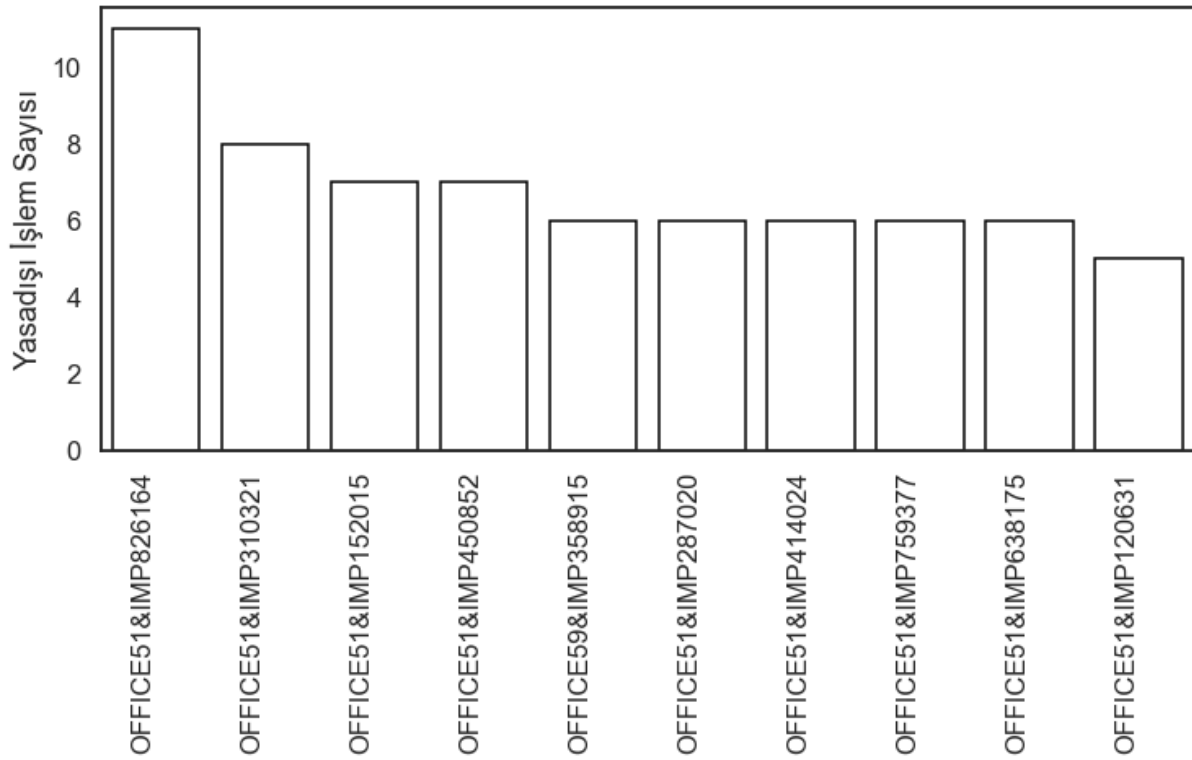
Şekil 3.12. Toplam vergi tutarı bazında en çok yasadışı işlem yapılan vergi miktarları

Gümrük idareleri tarafından yasadışı işlem yapılan firmalara ek olarak bir vergi tutarı ceza olarak yansıtılmaktadır. Veri setinde böyle bir etiketleme yasadışı işlemin varlığı anlamına gelmektedir. Bu vergi cezaları gümrük idareleri tarafından girilmiş ve Amerikan doları cinsinden belirtilmiştir. Ceza açısından bakıldığında aşağıda grafikte belirtildiği gibi 34 defa ile 71.0 Amerikan dolarının en çok vergi cezası olarak uygulandığı görülmüştür.



Şekil 3.13. En çok yasadışı işlem yapılan vergi cezası tutarları

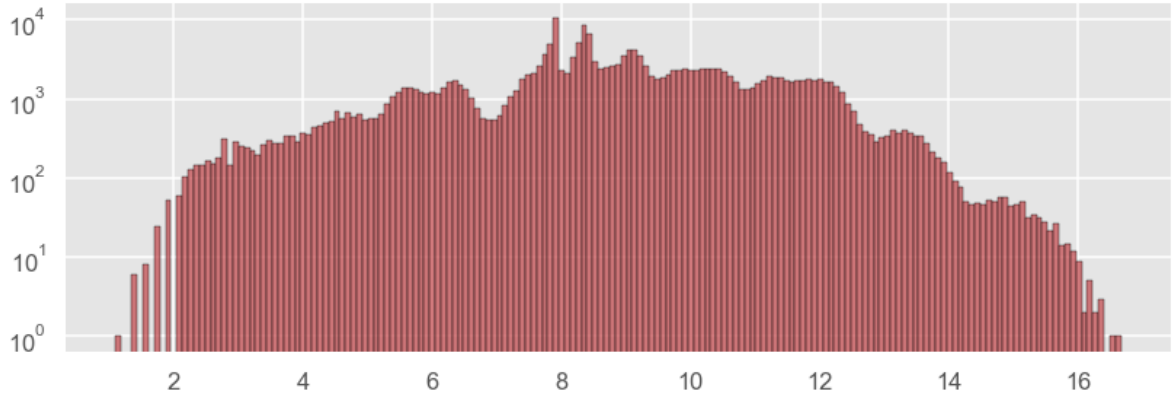
Gümrüklerde meydana gelen yasadışı işlemin sebepleri arasında olan, gümrük idarelerinde çalışan kötü niyetli memurlar ile bu işlemleri yapan ithalatçı firmaların yasadışı işlemleri sonucunda bu işlemlerin gerçekleşmesi değerlendirildiğinde, GÜMRÜK İDARESİ ADI ile İTHALATÇI FİRMA birleştirilerek GÜMRÜK İDARESİ ADI&İTHALATÇI FİRMA isimli yeni bir öznitelik elde edilmiştir. Bunun oluşturulmasının amacı, kötü niyetli olan gümrük idaresi ile aralarında bir anlaşma olduğu düşünülen ithalatçı firma eşleşmesini tahmin etmektir. Veri bu açıdan görselleştirildiğinde;



Şekil 3.14. En çok yasadışı işlem yapan gümrük idaresi ve firma eşleşmesi

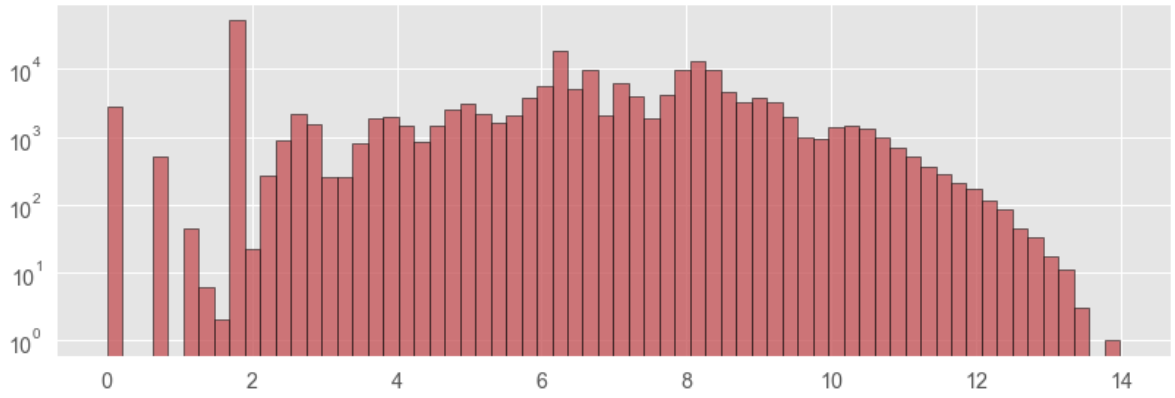
OFFICE51 isimli gümrük idaresi ile IMP826164 isimli ithalatçı firmanın 508 defa işlem yaptıkları, bu işlemlerden 11 tanesinin yasadışı olduğu açık olarak görülmektedir.

Veri seti içerisinde yer alan CIF değerinin histogram grafiği Şekil 3.15’de verilmiş olup, 7 ile 9 arasında bu değere ait maksimum seviyeler gözlemlenmektedir.



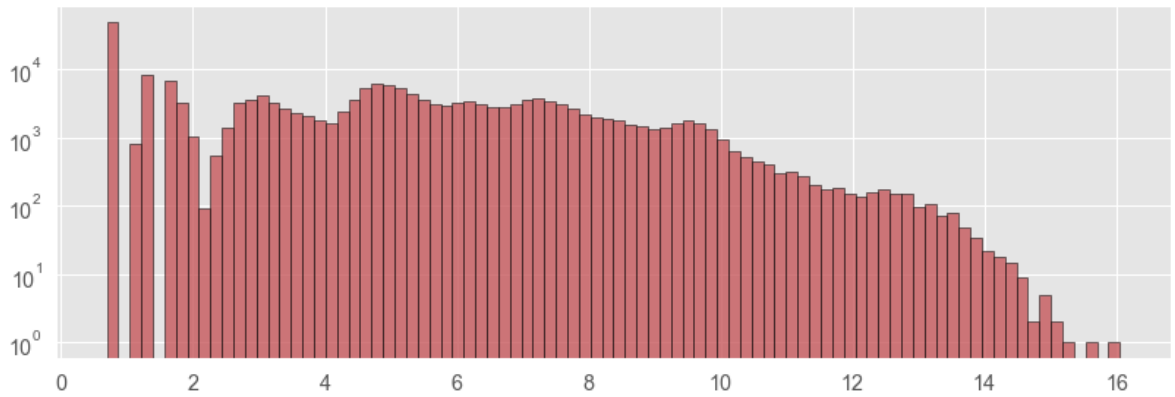
Şekil 3.15. CIF değeri histogram grafiği

Beyan edilen vergiler görselleştirildiğinde ürün başına en çok beyan edilen verginin 2 Amerikan doları olduğu Şekil 3.16’da açık olarak görülmektedir.



Şekil 3.16. Toplam vergi tutarı histogram grafiği

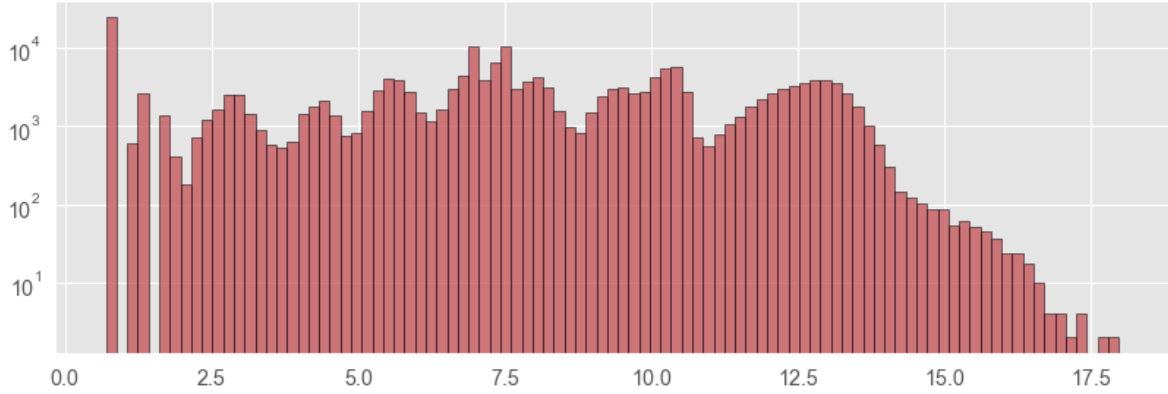
Ürünlerin birim başına adetlerin gösterildiği ürün sayısı histogram grafiği Şekil 3.17’de gösterilmiştir.



Şekil 3.17. Ürün sayısı histogram grafiği

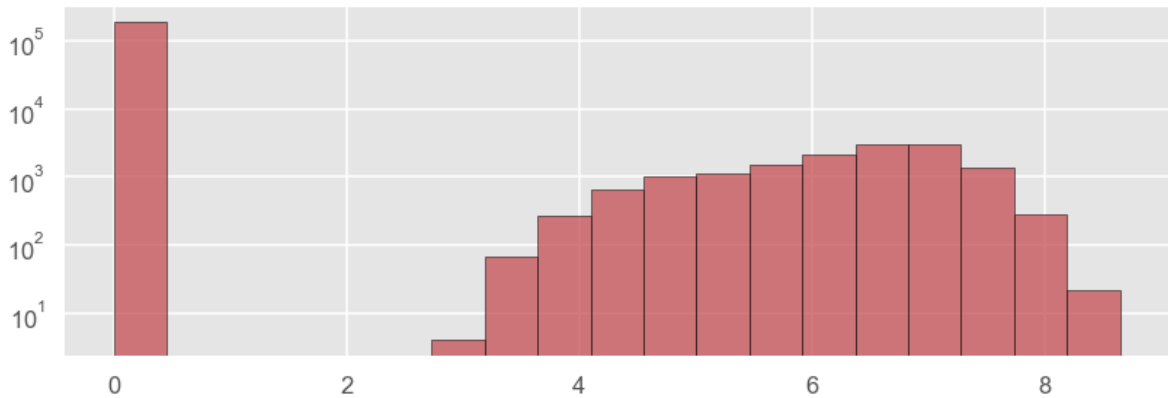
Histogram incelendiğinde, ürün sayısı en çok 1 olarak beyan edilen ürünlerde yığılma olduğu görülmüştür. Bunun nedeni, ürünlerin sayısal olarak her ürün başına 1 adet olarak belirtilmesinden kaynaklandığı tahmin edilmektedir.

Ürünlere fiziksel ağırlıkları açısından bakıldığında; ithalatçı firmaların çoğunlukla 1 kg. ağırlıklı ürünleri beyan ettikleri Şekil 3.18’de net olarak görülmektedir.



Şekil 3.18. Ürün fiziksel ağırlık histogram grafiği

İthal edilen ürünlerde ürünün bedeli, ağırlık, sayı, ürünün vergi tutarı gibi ürünün özelliklerinin beyan edildiği beyannamelerde, gerçeğe aykırı bir şekilde beyan verilmesi ve bunun da tespit edilmesi halinde, gümrük idareleri tarafından ithalatçı firmalara, ürün başına ceza olarak bir vergi tutarı yansıtılmaktadır. Veri setindeki vergi cezası tutarına ait Şekil 3.19’da gösterilen histogram incelendiğinde, yasal işlemlere ait olan miktar fazla olduğu için 0 oranının belirgin bir şekilde fazla olduğu, yasadışı işlemlerde 6 ve 7 dolar civarında vergi cezası uygulandığı görülmüştür.



Şekil 3.19. Vergi cezası tutarı histogram grafiği

3.3. TensorFlow ve Keras

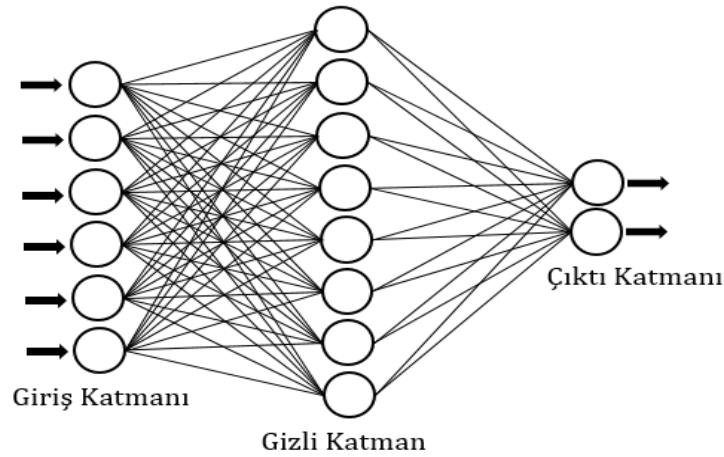
TensorFlow ücretsiz ve açık kaynaklı, genellikle sinir ağları gibi makine öğrenimi uygulamaları için kullanılan bir matematiksel yazılım kütüphanesidir. Google Brain ekibi tarafından 09 Kasım 2015 yılında yayınlanmıştır. Google kullanımı için geliştirilmiş fakat daha sonra anonimleşerek her alanda kullanılmaya başlanmıştır. Tensor (değişken sayıda eksene sahip normal bir ızgara üzerinde düzenlenmiş bir sayı dizisidir) ve flow (akış) kelimelerinin bir araya getirilmesiyle meydana gelmiştir. (Google Brain Team, 2-4 November 2016).

Tensorflow genel anlamda girdi olarak verilen verileri, kendine özgü bir takım matematiksel işlemler yardımıyla sınıflandıran veya tahminde bulunmaya yardımcı olan bir derin öğrenme kütüphanesidir.

Keras ise; TensorFlow üzerine inşa edilmiş ve Python diline API sağlayan üst düzey bir kütüphanedir (www.keras.io, 2021).

3.4. Yapay Sinir Ağları

Prensip açısından bakıldığında; yapay sinir ağları, insan beynine benzer bir sistem ile çalışır. Yapay sinir ağları öğrenme, hassasiyet ve genelleme gibi beynin yaptığı sıralamayı taklit ederek tahminde bulunur. Bir yapay zekânın öğrenme sürecinde, giriş ve çıkış verileri doğrultusunda edinilen bilgiyle tahminde bulunması beklenir (Walczak & Cerpa, 2003). Şekil 3.20’de yapay sinir ağlarının yapısı gösterilmiştir.



Şekil 3.20. Yapay sinir ağı yapısı

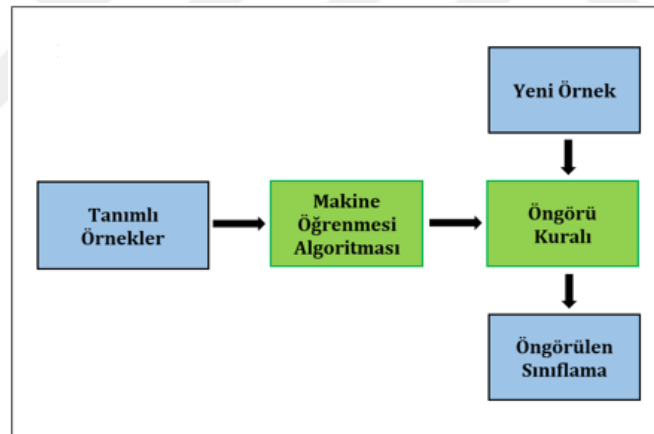
Şekil incelendiğinde nöronların giriş, gizli katman ve çıktı olmak üzere üç farklı ana katmandan meydana geldiği görülmektedir. Sinir ağına girecek olan her sütun ve satırdaki veriler ağın eğitimi için girdi verisi olarak kullanılır (LI, LI, & LI, 2017). Daha sonra çıkış katmanı ile giriş katmanı arasındaki öğrenme performansına bağlı olarak gizli katman sayısı belirlenir.

Aktivasyon fonksiyonu olarak sigmoid, hiper bolik tanjant, sinüs, step, eşik değer fonksiyonları kullanılmaktadır (Doğan & Türkoğlu, 2019). Bu çalışmamız ikili sınıflandırma olduğu için sigmoid katmanı tercih edilmiştir.

3.5. Makine Öğrenmesi

Makine öğrenmesi ilk olarak (Samuel, 1959) tarafından 1959 yılında geliştirilmiştir. Arthur Samuel bu çalışmasında bir bilgisayarın, programı yazan kişiden daha iyi dama oyunu oynadığını doğrulamak için makine öğrenmesi çalışması yapmıştır.

Makine öğrenimi, verilerden öğrenen ve zaman içinde doğruluğunu arttıran uygulamalar oluşturmaya odaklanan bir yapay zekâ dalıdır (Şahan, 2021). Makine öğreniminde algoritmalar, yeni verilere dayalı kararlar ve tahminler yapmak için büyük miktarda verideki kalıpları ve özellikleri bulmak üzere eğitilir (Demir, 2018). Algoritma ne kadar iyi olursa, daha fazla veri işledikçe kararlar ve tahminler o kadar doğru olur (IBM Cloud Education, Machine Learning, 2020). Şekil 3.21’de makine öğrenmesinin çalışma diyagramı gösterilmiştir.



Şekil 3.21. Makine öğrenmesinin çalışma diyagramı

3.6. Derin Öğrenme

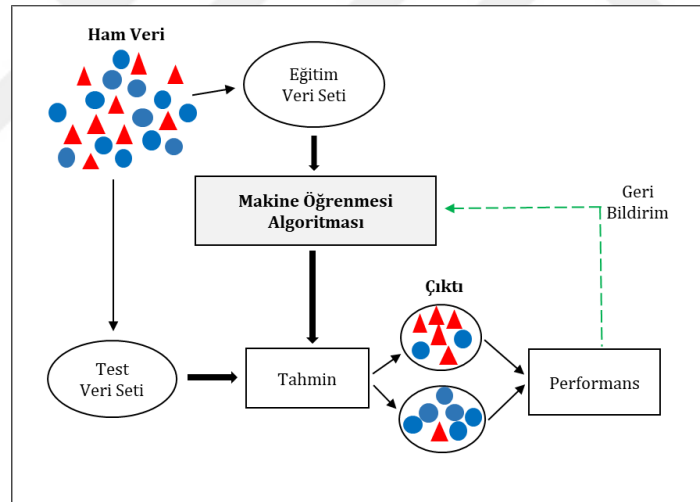
Derin öğrenme analizlerinde CPU (Merkezi İşlem Birimi) veya GPU (Grafik İşlemci Birimi) kullanılır. Carnegie Üniversitesinde dil işleme araştırma grubuna liderlik yapan Dr. Lan Lane ile yapılan bir röportajda, NVIDIA firması tarafından geliştirilen GPU’nun 33 kata kadar CPU’ya göre daha hızlı tanılama yaptığını ortaya koymuştur (Nemire, 2014).

Derin öğrenme, girdi olarak yüklenen verilerin çıktılarını tahmin eden bir yapay zekâ yöntemidir (Khanna, 2019). Yapay zekâyı eğitmek için hem denetimli hem de denetimsiz öğrenme yöntemleri kullanılabilir (Hinton, 2013).

3.7. Denetimli Öğrenme

Denetimli öğrenme, makine öğrenimi ve yapay zekânın bir alt kategorisidir. Verileri sınıflandırmak veya sonuçları doğru bir şekilde tahmin etmek için etiketli veri kümelerini kullanmak şartıyla, belirlenen algoritmanın eğitilmesiyle yapılan bir öğrenme biçimidir. Girdi verileri modele beslendikçe, modelin uygun şekilde yerleştirilmesini sağlayan bir takviye öğrenme süreciyle ağırlıklarını ayarlar (IBM Cloud Education, Supervised Learning, 2020)

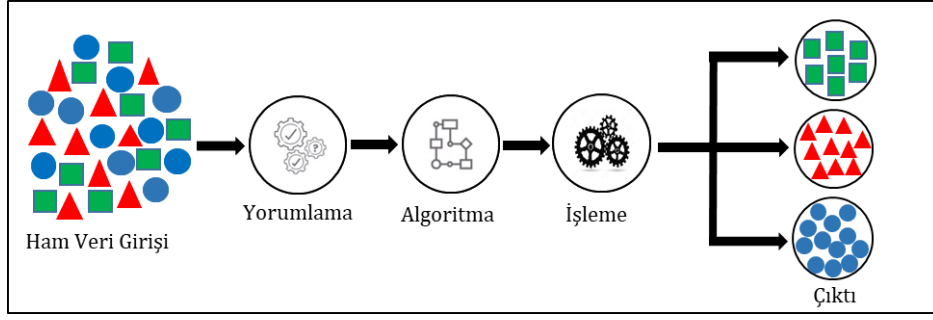
Sınıflandırma ve Regresyon şeklinde iki türlü denetimli öğrenme modeli vardır. Gümrük yasadışı işlemleri örnek alınırsa, veri setimizde yasadışı yapılan işlemler ve yasal yapılan işlemler bulunmaktadır. Bu işlemlerden hangisinin normal, hangisinin yasadışı olduğunu makineye denetimli bir şekilde öğrettikten sonra, makineden bundan sonraki verileri kendisinin değerlendirip, hangisinin yasal veya hangisinin yasadışı olduğunu tahmin etmesini beklememiz denetimli öğrenmeye güzel bir örnek olarak değerlendirilebilir. Bu çalışmada da sınıflandırma algoritmaları kullanılmış ve denetimli öğrenme modeli uygulanmıştır. Denetimli öğrenmeye ait diyagram Şekil 3.22’de gösterilmiştir.



Şekil 3.22. Denetimli makine öğrenmesine ait diyagram

3.8. Denetimsiz Öğrenme

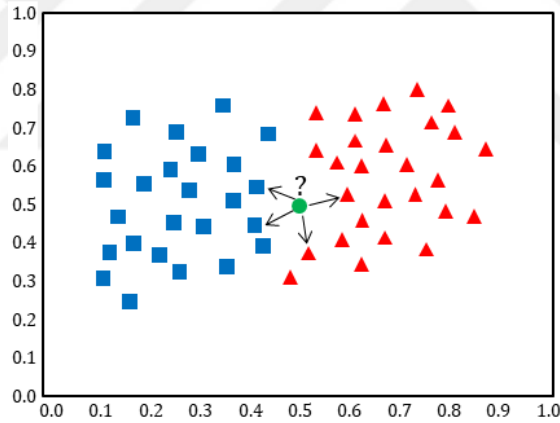
Denetimsiz öğrenme, eldeki verilerin sonuçları hakkında hiçbir bilginin olmadığı durumlarda kullanılmaktadır. Bu algoritmanın temelinde, verileri kendisinin kümelendirmesi beklenir ve hangi verinin hangi sınıfa ait olduğu tahmin edilmesi amaçlanır (Castelli , 2001). Denetimsiz öğrenmeye ait diyagram Şekil 3.23’de gösterilmiştir.



Şekil 3.23. Denetimsiz makine öğrenmesine ait diyagram

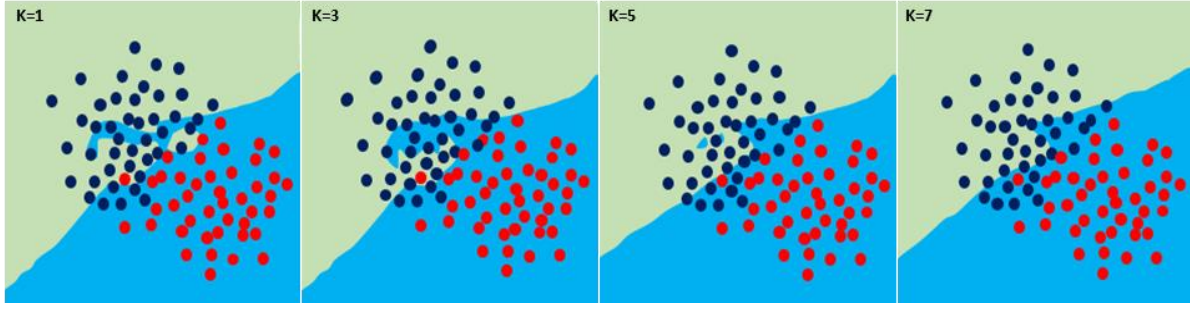
3.9. K-En Yakın Komşu (K-EYK)

En yaygın kullanılan sınıflandırma algoritmalarından bir tanesi k en yakın komşuluk algoritmasıdır. Mantığı Şekil 3.24’de gösterildiği gibi, iki farklı kümenin olduğu bir kümeye yeni bir nesne katıldığında, bu üyenin hangi kümeye ait olacağı mesafeye (örneğin oklit mesafesi) göre hesaplanır. Tabiatı gereği bu hesaplamalar yakın komşuluk sayısı olan k ile doğrudan ilişkilidir (Cover & Hart, January 1967).



Şekil 3.24. K-EYK Algoritması Örneği

Bu algoritmada en yakın komşu sayısının azaltılması veya çoğaltılması işlemi k değerinin sayısına bağlıdır (Agrawal, 2014). Örneğin düzlemdeki k değeri 1 olduğunda örnek sayısını sadece 1 olarak almakta ve en yakın komşu 1 olarak hesaplanmakta, k değeri 3 olduğunda örnek sayısı 3 olmakta ve en yakın komşu sayısı buna bağlı olarak 3 olmaktadır. Böylelikle k değeri yükseldikçe en yakın komşu sayısı da çoğalmaktadır. Eşitliğin bozulması için genelde k değeri tek sayılardan meydana gelmektedir. Şekil 3.25’de verilen örnekte görüldüğü gibi, k değeri arttırıldıkça mavi alanın sınırlarının genişlediği açık olarak görülmektedir.



Şekil 3.25. K Değeri Örneği

Burada k 'nın değerinin verildiği komşu sayısı ile sınıfa yeni katılan üyenin komşuları arasındaki mesafeye, Öklid mesafesi denir (Chomboon, Chujai, Teerarassamee, Keerdprasop, & Kerdprasop, 2015). Öklid mesafesi, bir öznitelik uzayında iki nokta arasındaki uzaklığı ölçmek için kullanılır (Hu, Huang, Ke, & Tsai, 2016). Hesaplaması Denklem 1'de açık olarak verilmiştir (Jose, 2018).

$$d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2}$$

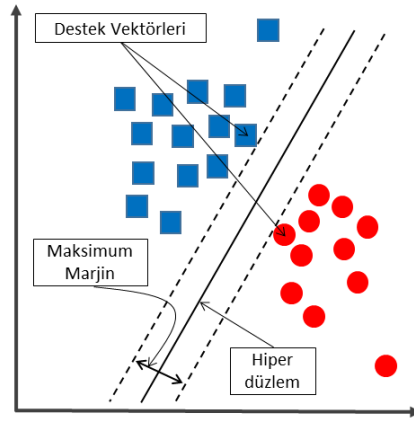
$$= \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (1)$$

Literatürde örneklemeler arasındaki mesafeyi belirleyen matematiksel yaklaşımlar sırasıyla Euclidian, Manhattan, Minkowski, Weighted vs. gibi yöntemlerdir. Ancak bu tez çalışmasında Öklid mesafesi kullanılmıştır.

3.10. Destek Vektör Makineleri (DVM)

Destek Vektör Makinesi hem sınıflandırma hem de regresyon zorlukları için kullanılabilen denetimli bir makine öğrenme algoritmasıdır (www.ichi.pro.tr, 2021). Ancak daha çok sınıflandırma problemlerinde kullanılmaktadır. Bu algoritmanın kalbinde, iki sınıfa ayıran bir hiper düzlemin belirlenmesi yatmaktadır. Fakat bu belirleme noktasında dikkat edilmesi gereken durum, uygun çekirdeğin tespit edilmesidir (Salomon, King, & Osborne, 2002). Literatürde birçok çekirdek işlevci bulunmaktadır. Ancak bu çalışmada; Sigmoid, Polinom, Doğrusal ve RTF çekirdek işlevcileri kullanılmaktadır. Tüm çekirdekler için C hiper parametresine tam sayılı bir değer atanması gerekmektedir. C hiper parametresi, daha düşük bir hata oranı elde etmek için veri noktalarının yanlış sınıflandırılmasına izin veren bir parametredir yani doğrusal olmayan veriler üzerinde verilere olan toleransı ayarlar. C değeri büyüdükçe verileri aşırı böler ve doğruluğu azaltır (Lundell, 2019). Aynı şekilde C 'nin değeri düşük verildiğinde de verilere uzak bir bölme yapacağından doğruluk yine azalacaktır. Bu nedenle C değerinin genellikle 0.1 ve 10.0 aralığında

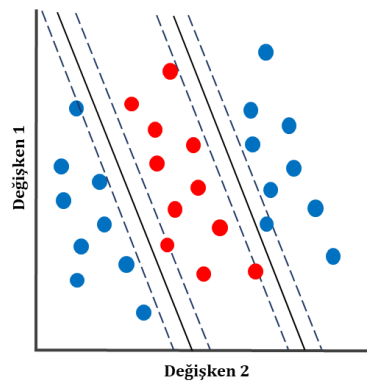
daha iyi performans gösterdiği bilinmektedir. DVM algoritmasının yapısıyla ilgili basit bir temsiliyet Şekil 3.26'da verilmiştir.



Şekil 3.26. DVM algoritması yapısı

3.10.1. Sigmoid Çekirdek

Sigmoid çekirdeği, makine öğrenmesi işlemlerinde özellikle DVM algoritmalarında iki katmanlı bir algılayıcı olarak sıkça kullanılmaktadır. Ayrıca yapay sinir ağlarında aktivasyon fonksiyonu olarak da tercih edilmektedir (Wang, Li, Song, & Rong, 2020). Yapısı; veri seti üzerinde iki ayrı marjın boyutu oluşturup, verilerin marjine olan uzaklıklarını hesaplayarak, ikili veya çoklu sınıflandırmada verilerin hangi sınıfa ait olduğunu ayırt etmek için düzenlenmiştir. Sigmoid çekirdeği doğrusal olmayan bir fonksiyona sahiptir (Asraf, Nooritawati, & Shah Rizam, 2012). Sigmoid çekirdek yapısına ait diyagram Şekil 3.27'de gösterilmiştir.



Şekil 3.27. Sigmoid çekirdek diyagramı

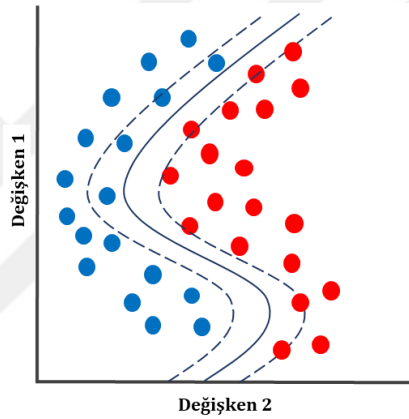
Sigmoid çekirdek yapısının fonksiyonu denklem 2'de verilmiştir.

$$k(x, y) = \tanh(ax^T y + c) \quad (2)$$

Burada c : kesme sabit olup, ayarlanabilir bir parametre ve a : hiper düzlemin yüzey normalidir (Güldoğan, 2017).

3.10.2. Polinom Çekirdek

Görüntü işleme konusunda başarılı performans göstermesi nedeniyle, makine öğrenmesi çalışmalarına bakıldığında, Polinom çekirdeğin sıkça kullanıldığı görülmektedir. Fakat diğer doğrusal olmayan verilerde de kullanıldığında, iyi performans gösterdiği görülmüştür (Wang, Leung, Lee, & Wang, 2012). Genellikle Polinom çekirdekler, verilerin belirli bir kararlılıkla dağılmadığı problemler için tercih edilir. Polinom çekirdek fonksiyonuna ait diyagram Şekil 3.28’da gösterilmiştir.



Şekil 3.28. Polinom çekirdek diyagramı

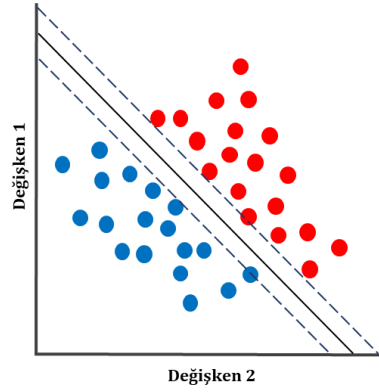
Polinom çekirdeğin denklemi şu şekildedir.

$$k(x_i, x_j) = (x_i \cdot x_j + 1)^d \quad (3)$$

Burada d Polinomun derecesidir (Ichi.pro, 2021).

3.10.3. Doğrusal Çekirdek

Destek Vektör Makineleri için tanımlanan en basit çekirdek yapısı doğrusal çekirdektir (www.crsouza.com, 2010). Buradaki temel amaç sınıflar arasında ayrımı yapacak olan en uygun doğrusal düzlemin belirlenmesidir. Fakat bazı durumlarda bu ayrım net bir şekilde hiper düzlem ile tayin edilemeyebilir. Bu noktalar algoritmanın tahmin etmiş olduğu hata olarak değerlendirilir. Bu hata algoritmanın performansını negatif yönde etkiler. Doğrusal çekirdek fonksiyonu ile ayırt edilen veri dağılımı Şekil 3.29’da gösterilmiştir.



Şekil 3.29. Doğrusal çekirdek diyagramı

Kesikli çizgiler üzerinde bulunan noktalar destek vektör noktalarıdır ve iki kesik çizgi arasındaki mesafe ayırım uzaklığı diye tanımlanır. Ayrıca doğrusal çekirdeğe ait denklem

$$k(x, y) = x^T y + c \quad (4)$$

ile verilir (Güldoğan, 2017).

3.10.4. Radyal Tabanlı Fonksiyon Çekirdeği (RTF)

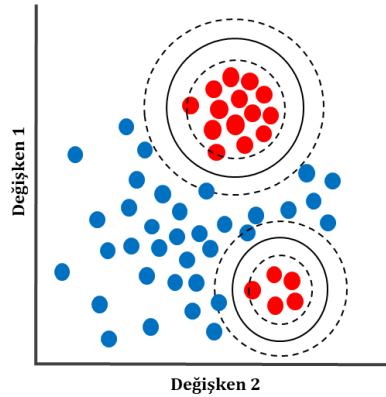
Radyal tabanlı fonksiyon çekirdeği, doğrusal olmayan veri setleri üzerinde kullanılan ve bazı parametreleri sigmoid çekirdeğe benzer olan bir çekirdek modelidir. Sigmoid, Polinom ve Doğrusal çekirdekten farklı olarak, RTF çekirdekte gamma hiper parametreleri kullanılmaktadır. Gamma hiper parametresi hem verilerin karar sınırlarını hem de tek bir eğitim örneğinin etkisinin ne kadar uzağa sirayet edeceğini belirler (Kumar, 2020). Gamma değeri düşük oldukça verilere olan mesafeleri uzak, gamma değeri yükseldikçe de verilere olan mesafe kısalmaktadır. Bu değerler çok düşük veya çok yüksek olduklarında algoritmanın doğruluk performansını azaltacaktır. Genellikle C hiper parametresinde olduğu gibi gamma değeri de 0.01, 0.1, 1.0, 10.0, 100.0, 1000.0 olarak tercih edilir. Fakat kullanıcılar gamma başlangıç değerini, deneme yanılma yoluyla algoritmanın performansını göz önüne alarak belirlerler (Lu, 2018).

RTF çekirdeğin denklemi şöyledir,

$$k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0 \quad (5)$$

buradaki γ parametresi, tek bir eğitim örneğinin ne kadar etkiye sahip olduğunu tanımlar (Apostolidis-Afentoulis & -Ina Lioufi, 2010).

RTF doğrusal olmayan veriler üzerinde daha etkilidir çünkü marjin uzaklıklarını dairesel bir biçimde ele almaktadır. RTF çekirdeğe ait diyagram Şekil 3.30'da gösterilmiştir.



Şekil 3.30. RTF çekirdek diyagramı

3.11. Sınıflandırma

Denetimli öğrenme kategorisine ait olan sınıflandırma, genel olarak ikili veya çoklu sınıflandırma şeklinde ikiye ayrılmaktadır. İkili sınıf 0 veya 1 olarak tanımlanırken, çoklu sınıflar ise ikiden fazla durumları içeren sınıfları tanımlar. Örneğin; bir kişinin kanser olup olmaması ikili sınıflara karşılık gelirken, iris çiçeğinin sahip olmuş olduğu tür ise çoklu sınıflara karşılık gelir.

Bu çalışmada, işlemin yasadışı veya yasadışı değil şeklinde ikili sınıflandırma sonucu tahmin edileceğinden, ikili sınıflandırmaya ait algoritmalar ve yöntemler kullanılmaktadır.

3.12. Veri Setinin Eğitim ve Test Olmak Üzere İkiye Bölünme Aşaması

Makine öğrenmesinde özellikle denetimli öğrenmede, veri setinin büyük bir bölümünün örneğin %70'inin eğitim verisi olarak, diğer %30'unu ise test verisi olarak ayrılması gerekmektedir. Bu oran yüzdesi %90-%10, %80-%20 veya %77-%33 şeklinde de olabilir. Bu oranın belirlenmesi elde edilen tahminin doğruluk derecesine göre azaltılıp çoğaltılabilir. Makine eğitim verileriyle algoritmanın tahminlerini ortaya koymakta ve test verileriyle de bu tahminlerin ne kadar doğru olduğunu sınamaktadır.

3.13. Sentetik Azınlık Yüksek Hızda Örnekleme Tekniği (SAYHÖ)

SAYHÖ algoritması ilk olarak (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) tarafından tanıtılmıştır. Pozitif veri miktarını negatif veri miktarına eşitlemek için, pozitif verilerin ken yakın komşu hesaplamasıyla istenen yüzde kadar rastgele sentetik veri çoğaltarak pozitif verilerin

oranını negatif veriler kadar yükseltmekte kullanılır (Sain & Purnami, 2015). SAYHÖ'nün denklemi şu şekildedir,

$$X_{syn} = X_i + (X_{knn} - X_i) \times \delta \quad (6)$$

burada δ , 0 ile 1 arasında rastgele üretilen bir sayıyı ifade etmektedir.

3.14. Performans Değerlendirmesi

Makine öğrenmesi çalışmalarında, verilerin modellerde eğitildikten sonra mutlaka performansının değerlendirilmesi gerekir. Çeşitli makine öğrenme modelleri ve tekniklerle veri seti eğitilebilir ancak performans değerlendirilmesi yapılmadığı sürece modelin doğruluğu güven vermeyecektir.

Denetimli öğrenme tekniklerinin kullanıldığı sınıflandırma problemlerinde, model performansını değerlendirebilmek için en yaygın kullanılan yöntemlerden biri, hedef niteliğin sınıflarına ait gerçek ve tahmin değerlerinin bir arada gösterildikleri, karışıklık matrisinden hareketle hesaplanan değerlerdir (Balaban & Kartal, 2015).

3.15. Karışıklık Matrisi

Karışıklık matrisi uygulanan modelin sonucunun bir çıktısı olarak düşünülebilir. Her bir sütun elde edilen tahminlerin sonuçlarını, satırlar ise gerçekte olan sınıfları göstermektedir (Güneş, ve diğerleri, 2011). Yani veri setindeki gerçekte var olan durum ile algoritmanın doğru ve yanlış tahminlerinin sayısını tablo şeklinde vermesidir.

Tablo 3.6. Karışıklık matrisi

	Tahmin Pozitif	Tahmin Negatif
Gerçek Pozitif	Doğru Pozitif (DP)	Yanlış Negatif (YN)
Gerçek Negatif	Yanlış Pozitif (YP)	Doğru Negatif (DN)

Burada;

- DP doğru pozitiflerin sayısıdır. Modelin pozitif sınıfı doğru tahmin ettiği orandır.
- YP yanlış pozitiflerin sayısıdır. Modelin pozitif sınıfı yanlış tahmin ettiği orandır.
- DN doğru negatiflerin sayısıdır. Modelin negatif sınıfı doğru tahmin ettiği orandır.
- YN yanlış negatiflerin oranıdır. Modelin negatif sınıfı yanlış tahmin ettiği orandır.

3.16. Doğruluk

Eğitilen modelin başarı oranını görmek için kullanılan bir yöntem olup, matematiksel olarak

$$\text{Doğruluk} = \frac{(DP + YN)}{(DP + DN + YP + YN)} \quad (7)$$

Denklem 7’de verildiği gibi doğru sınıflandırılmış örnek sayının toplam örnek sayısına oranıyla hesaplanır (Taylor, 1997). Hesaplanan sonuç 1’e ne kadar yakın ise algoritmanın sınıflandırma performansı o kadar başarılı demektir.

Performans değerlendirmesi başlığı altında bahsedildiği gibi, doğruluk bir modelin başarısını ölçmek için kullanılabilir fakat tek başına yeterli olmayacaktır. Doğru ölçüt için Kesinlik veya Hassasiyet sonuçlarına ya da her ikisinin önemli olduğu durumlarda harmonik ortalamalarını veren F skorunun da hesaplanması gerekir.

3.17. Kesinlik

Kesinlik, algoritmanın ne kadar gerçek pozitif değeri tahmin ettiğinin bir ölçüsüdür. Başka bir deyişle, algoritmanın pozitif tahmin ettiği oranların gerçekte kaç tanesini doğru tahmin ettiğinin bir göstergesi olarak tanımlanabilir. Matematiksel olarak kesinlik denklemi aşağıda verilmiştir (Sokolova, Japkowicz, & Szpakowicz, January 2006).

$$\text{Kesinlik} = \frac{DP}{(DP + YP)} \quad (8)$$

Denklem 8’e bakıldığında, kesinlik doğru pozitiflerin tüm pozitif sonuçlara (hem doğru pozitif hem yanlış pozitifler) oranı olduğu açık olarak görülmektedir. Burada herhangi bir algoritma için elde edilen kesinlik değeri olabildiğince yüksek olmalıdır.

3.18. Hassasiyet

Hassasiyet, pozitif olarak tahmin edilen doğru pozitif verilerin gerçek doğruların sayısının, gerçek doğrularla sahte yanlışların sayısının toplamına bölümünden oluşur. Matematiksel olarak

$$Hassasiyet = \frac{DP}{(DP + YN)} \quad (9)$$

ile temsil edilir (Gupta, Rawal, Narasimhan, & Shiwani, 2013). Denklem 9'a bakıldığında, hassasiyetin hesaplanmasında yanlış negatiflerin önemi ön plana çıkmaktadır. Başka bir deyişle; hassasiyet, pozitif sınıftan gelen gözlemlerin doğru tahmin edilme oranı olarak tanımlanabilir (Döring, 2018).

3.19. F-1 Skoru

F ölçütü, kesinlik ve hassasiyet ölçütlerinin harmonik ortalamasıdır. Genellikle, kesinlik yüksek çıktığı durumlarda hassasiyet düşük veya tersi olmaktadır. Burada hem kesinliğin hem de hassasiyetin önemli olduğu durumlar için her ikisinin harmonik ortalaması alınır. Basit ortalama alınabilir ancak aykırı değerlere karşı daha tutarlı olduğu için harmonik ortalama kullanılmalıdır. Yani kesinlik ve hassasiyeti dengelemek için F skorunun sonuçlarına bakılmalıdır. F skoru matematiksel karşılığı;

$$F_1 = 2 \times \frac{Kesinlik \times Hassasiyet}{Kesinlik + Hassasiyet} \quad (10)$$

ile verilir (Vakili, Ghamsari, & Rezaei, 2019).

3.20. İşlem Karakteristik Eğrisi (İKE)

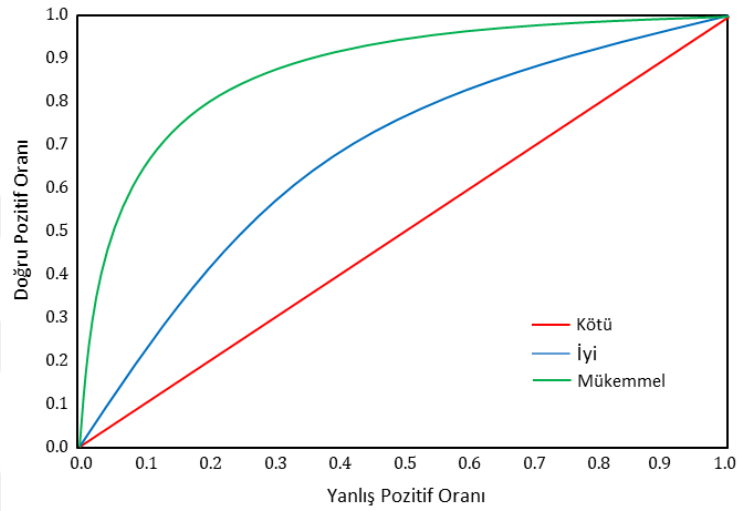
Hassasiyetin, Kesinliğe olan oranıyla yani doğru pozitif oran ile yanlış pozitif oran arasındaki ilişki İKE eğrisi olarak adlandırılır. Karışıklık matrisinden elde edilen sonuçlar aracılığıyla hesaplanan Kesinlik, Hassasiyet ve F skorunun haricinde, modelin performans değerlendirmesi İKE eğrilerinin sonuçlarına bakılarak da yapılması mümkündür.

Doğru Pozitif Oran (DPO) ile Yanlış Pozitif Oran (YPO) hesaplaması aşağıdaki eşitliklerle yapılır.

$$DPO = \frac{DP}{(DP + YN)} \quad (11)$$

$$YPO = \frac{YP}{(YP + DN)} \quad (12)$$

İKE eğrisi altındaki alan Eğri Altındaki Alan (EAA) olarak tanımlanır ve farklı İKE eğrilerinin karşılaştırılmasının bir göstergesidir (Hanley & McNeil, 1982). DPO'yu YPO'ya karşı çizen ve esas olarak sinyali gürültüden ayıran bir olasılık eğrisidir (Bhandari, 2020). Eğrinin çizilmesi (0, 0) ile başlar, algoritmanın doğru tahminleri oldukça eğri yukarı ve sağa doğru devam eder, yanlış tahminlerde aşağı iner, en sonunda (1, 1) de tamamlanır, böylelikle bir eğri ortaya çıkar (Majnik & Bosnic, 2013). Bu eğriye örnek Şekil 3.31'de gösterilmiştir. Eğri altındaki alan ne kadar fazla olursa başarı oranı o kadar yüksek olur.



Şekil 3.31. Örnek İKE eğrisi

4. BULGULAR ve TARTIŞMA

Bu bölümde, araştırmanın amacı, sınırlılıkları ve önemi anlatılmıştır. Öte yandan; veriler algoritmalar tarafından eğitilmeden önce birtakım ön işlemlere tabi tutulmuştur. Daha sonra algoritmaların eğitime uygun hale getirilen verileri tahmin etmesi sağlanmış ve en sonunda performans değerlendirmesi yapılmıştır.

4.1. Araştırmanın Amacı

Tezin genel amacı, verilerden faydalanarak çıkarımlar yapmak, veriyi yorumlamak ve nihayetinde faydalı olabilecek değerli bilgiye erişmektir.

Veri olarak, Nijerya ülkesine ait iki yıllık gümrük ithalat verileri kullanılmıştır. Bu veriler incelendiğinde, ithalatçı firmalar tarafından işlemler yapılırken, devlete daha az vergi vermek maksadıyla, vergi miktarı daha düşük olan bir ürünü beyan ettikleri ya da ithal edilen ürünün sayısal veya ağırlık miktarlarını gerçek dışı beyan ederek gümrük vergisini daha az ödedikleri görülmüştür.

Bu tespitler üzerine, klasik makine öğrenmesi algoritmaları ve derin öğrenme metotları kullanarak, ithalatı yapılan veya yapılacak olan ürünlerin cins, miktar, bedel, vergi tutarı gibi bilgilerinden faydalanılması suretiyle, işlemlerin yasal veya yasadışı işlem olup olmadığını tahmin etmek ve böylelikle devletin vergi kaybının önüne geçilmesi amaçlanmaktadır.

4.2. Araştırmanın Sınırlılıkları

Literatürde yapılan araştırmalar esnasında sadece Nijerya ülkesine ait gümrük ithalat verilerine rastlanmıştır. Bu nedenle örneklem olarak sadece Nijerya ülkesine ait ithalat verileri kullanılmıştır. Nijerya'nın ithalat durumuna bakıldığında, dünya bankası verilerine göre 2020 yılındaki kişi başı gayri safi milli hasılası 5.19 bin ABD Doları ve toplam 448.12 milyar ABD Doları gayri safi milli hasılaya sahip olduğu görülmüştür. 2019 yılında toplam 47.36 Milyar ABD Doları ithalat ve 53.61 Milyar ABD Doları ihracat gerçekleştirmiştir. Bu da dünya ekonomisinin 0,38'ini temsil etmektedir (www.tradingeconomics.com, 2021). Farklı ülkelerin gümrük ithalat verilerine erişme imkânı olmadığından karşılaştırma yapılması da mümkün olmamıştır. Öte yandan veri setinin orijinal hali 5 yılın verisini kapsamasına rağmen, bazı çekincelerle sadece 2 yıla ait olan veriler paylaşılmıştır. Verinin çok olması, modellerde performansı olumlu yönde etkileyecektir. Bu nedenle 5 yıla ait olan verilerle yapılacak olan modellemelerden daha iyi sonuçlar elde edileceği tahmin edilmektedir.

4.3. Araştırmanın Önemi

Gümrüklerdeki ithalat işlemlerinde yapılan usulsüz işlemlerle ilgili olarak; gümrük idarelerince henüz herhangi bir kontrol yapmadan, gümrük işlemlerine başlanılmadan, beyan aşamasında, bilgilerin sisteme girilmesiyle birlikte işlemin yasadışı olup olmadığı tahmin edilmiştir. Bu açıdan değerlendirildiğinde; gümrük idareleri ve gümrük personelleri için zamandan tasarruf, işgücünün azaltılması, maliyetlerin en aza indirilmesi ve vergi kaybının önüne geçilmesi öngörülmektedir. İnsan beyninin yapamayacağı bu işlemler makine öğrenmesi ve derin öğrenme modelleri kullanılarak yapılmıştır. Bu algoritmalar, mevcut veriler üzerinden hesaplamalar yaparak olayın nasıl gerçekleştiğini öğrenir ve bir sonraki periyotta ne olacağını tahmin ederler.

Gümrük ithalat işlemlerindeki usulsüzlüklerin tahmin edilmesi bağlamında literatürde birçok makine öğrenmesi algoritmaları ve derin öğrenme metotları kullanılmıştır. Fakat ilk defa bu çalışmada, K-EYK ve DVM algoritmaları kullanılmıştır. Ayrıca çalışmanın evrensel niteliği de bulunmaktadır. Herhangi bir ülke veya gümrük idaresi kendi ithalat verileriyle bu çalışmadaki yöntemleri kullanarak veya farklı makine öğrenmesi algoritmalarıyla usulsüzlükleri önceden tahmin edebilecektir.

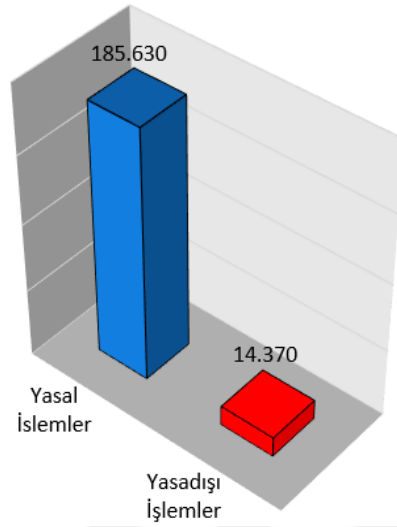
4.4. Veri Ön İşleme

Daha önce veri seti üzerinde birtakım ön işlemler yapılmıştı fakat bu işlemler görselleştirme aşamasında gerekli olan işlemlerdir. Modellerin eğitilmesi için birtakım ön işlemlerin de yapılması gerekmektedir. Uygulanacak olan 3 ayrı modelde de ortak yapılması gerekenler olduğundan veri seti tekrar ön işleme tabi tutulmuştur. İşlemler sırasıyla anlatılmaktadır.

Veri seti ilk önce içe aktarılmıştır. Daha sonra veride yer alan özelliklerin yapısını incelediğimizde 3 kategorik, 11 numerik olmak üzere 14 öznitelik olduğu görülmüştür. Eksik hücrelerin varlığına bakılmış ve herhangi bir boş hücre olmadığı, kategori olarak 5 adet kayan sayı, 6 adet tam sayı ve 3 adet nesne olduğu görülmüştür. Ayrıca veri seti içerisinde yer alan ve herhangi bir kimliği temsil etmeyen öznitelik veri setinden çıkarılmıştır.

Python'un veriler üzerinde işlem yapabilmesi için nesne karakterlerinin olmaması gerekmektedir. Bu nedenle nesne olan GÜMRÜK İDARESİ ADI, İTHALATÇI FİRMA ve GÜMRÜK ÜLKE_KODU isimli öznitelikler Labelencoder fonksiyonu ile sayılara dönüştürülmüştür. Daha sonra yine bu veri setini kullanacağımız için yeni bir isimle sadece sayılardan oluşan veri seti pandas kütüphanesine kaydedilmiştir.

Yasadışı yapılan işlemlerin sayısına bakıldığında Şekil 4.1’de gösterildiği gibi 185.630 yasal işlem ve 14.370 yasadışı yapılan işlem olduğu görülmüştür.



Şekil 4.1. Yasal ve yasadışı olan işlemler

Bu durum veri setinin dengesiz veri seti olduğunu göstermektedir. Bu dengesizliği ortadan kaldırabilmek için SAYHÖ yardımıyla veri seti dengeli hale getirilmiştir.

Tablo 4.1. SAYHÖ yöntemi uygulama sonucu

	Yasal işlemler	Yasadışı işlemler
Önceki Değerler	(0) :185630	(1) :14370
Sonraki Değerler	(0) :185630	(1) :185630

Böylece, pozitif durumlar negatif durumlar ile dengelenmiştir.

Veri seti p-değeri açısından irdelendiğinde; gümrük idaresi, gümrük idaresinin ülkesi, ithalatçı firma ve ithal edilen ürünün uluslararası kodu, yasadışılığa sebep olan en önemli öznitelikler olduğu tespit edilmiştir. Bu ön işlemler sonucunda, daha sonra kullanmak üzere veri seti farklı bir isimle yine pandas kütüphanesine kaydedilmiştir.

Analiz, 3 ayrı algoritmayla ayrı ayrı yapıldığı için veri dönüştürme, normalleştirme, boyut indirgeme, eğitim ve test verilerinin ayrılması gibi diğer ön işlemler algoritmaların eğitimleri öncesinde ve modelin özelliğine göre ayrı ayrı yapılmıştır. Algoritmaların uygulama aşamalarında bu hususlara değinilmiştir.

4.5. Verilere Modellerin Uygulanması

Bu bölümde bir takım ek ön işlemler mevcut veriye uygulanacak ve daha sonra hem modelin eğitimi tamamlanacak hem de performans değerlendirmesi yapılacaktır.

4.5.1. K-EYK Algoritmasının Uygulanması

Pandas ve numpy kütüphaneleri içe aktarıldıktan sonra analize uygun hale getirilen dengelenmiş veri seti içe aktarılmıştır. YASADIŞI kolonu 0 ve 1'den meydana gelmektedir. 0 yasal işlemi 1 ise yasadışı işlemi tanımlamaktadır. Geri kalanlar ise kararı etkileyen özniteliklerdir. Veri setini bölme aşamasına gelindiğinde, en iyi sonuç alındığı için veriler %80'i eğitim, %20'si test veri seti olarak ayrılmıştır.

K-EYK algoritmasına ait kod dizesi uygulanırken, k sayısı 3, 5 ve 7 olacak şekilde sırasıyla denenmiştir. En iyi sonucun k=7'de olduğu tespit edilmiş ve algoritmanın doğruluk oranı %92,26 olarak hesaplanmıştır. Doğruluk oranının tespitinden sonra, modelin 40.000 veri üzerinden yapmış olduğu işlemlere ait karışıklık matrisi Tablo 4.2'de gösterilmiştir.

Tablo 4.2. K-EYK algoritması karışıklık matrisi sonuçları

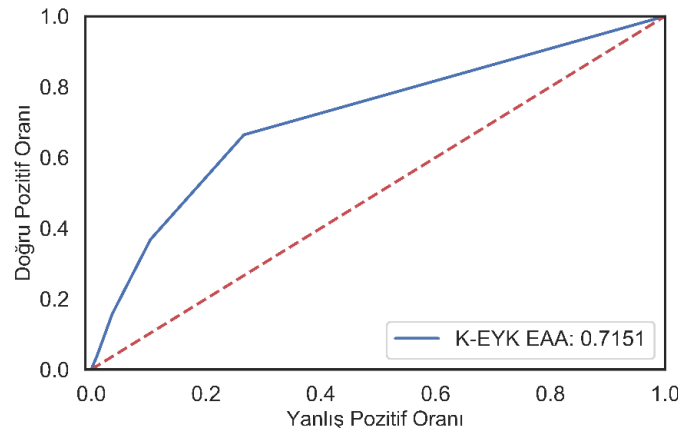
	Tahmin Pozitif	Tahmin Negatif
Gerçek Pozitif	36748	378
Gerçek Negatif	2762	112

Karışıklık matrisi sonuçlarına bakıldığında, değerlendirme metrikleriyle ilgili sonuçlar Tablo 4.3'de gösterilmiştir.

Tablo 4.3. K-EYK algoritması performans değerlendirmesi

Sonuçlar	Kesinlik	Hassasiyet	F-1 Skoru	Doğruluk
	0.93	0.99	0.96	%92,26

Ayrıca İKE eğrisini çizdirdiğimizde aşağıdaki grafik elde edilmiştir.



Şekil 4.2. K-EYK algoritması İKE eğrisi

Grafiğe bakıldığında EAA değerinin %71,51 olduğu görülmüştür.

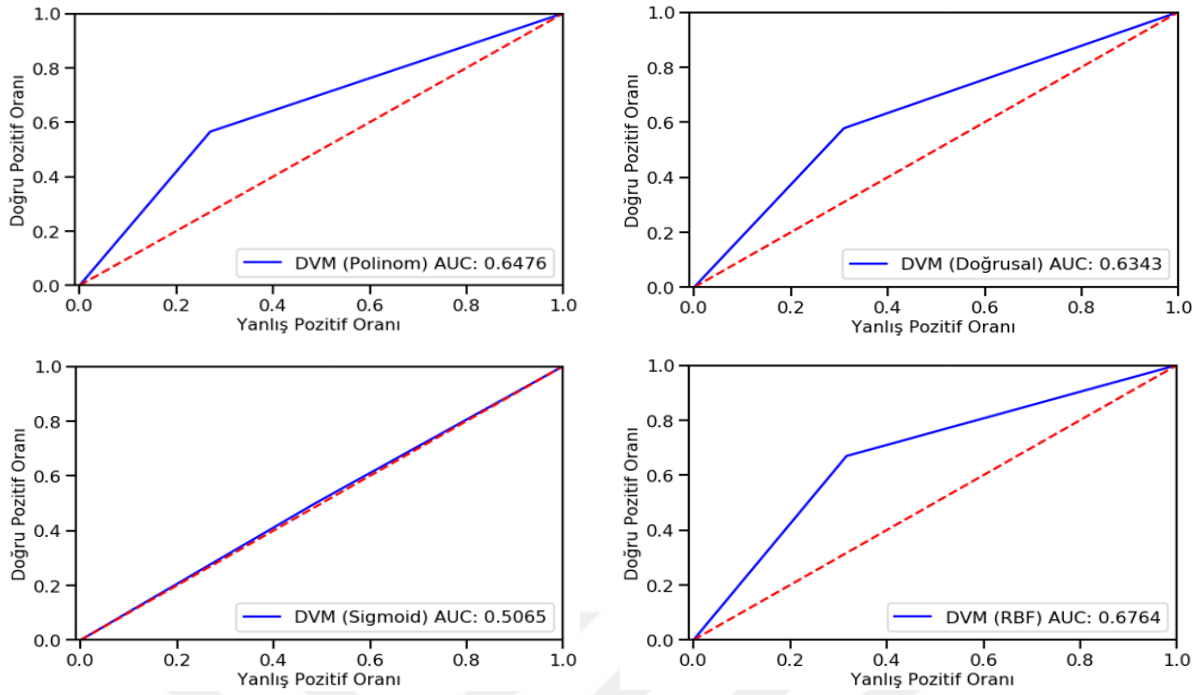
4.5.2. Destek Vektör Makinesi Algoritmasının Uygulanması

Bu bölümde, Destek Vektör Makinelerinin performans değerlendirmeleri Doğrusal, Polinom, Sigmoid ve RTF çekirdek olmak üzere dört farklı çekirdek üzerinden değerlendirilecektir. Her çekirdeğin veriler üzerinde oluşturduğu düzlem ve eğriler birbirlerinden farklı olacağından veri üzerine çekirdekler tek tek uygulanmıştır. Veri seti 2013 ve 2014 yıllarına ait verilerden oluşması nedeniyle ilk önce 2013 yılı eğitim, 2014 yılı test veri seti olarak ayarlanmıştır. Yıllara göre ayrılan veri seti dengeli olduğundan dengesizliği gidermek için herhangi bir işlem yapılmamıştır. Daha sonra yıllara ayırmadan, tüm veri seti analize tabi tutulmuş ve %80 eğitim, %20 test veri seti olarak ayarlanmıştır. Bahsi geçen bu bölünme oranı için en iyi performans sonuçları elde edilmiştir. Ayrıca, modelin bir takım ezberleme problemleri ile karşılaşmaması için veriler rastgele bir şekilde ayrılmıştır.

Her iki eğitim ve test veri seti ayırımından sonra, tüm verilerin aynı birime getirilmesi ve varyansı artan verileri eşitlemek için MinMaxScaler metodu ile veriler ölçeklendirilmiştir. Algoritmaların C hiper parametresi C=0.01, C=0.1, C=1.0 ve C=100 olacak şekilde ayrı ayrı test edilmiştir. C değeri ne kadar büyük olursa marjin o kadar dar olacağından işlem uzun sürmüştür. C değeri arttırıldıkça model aşırı uyumlu hale gelmekte, C değeri azaldıkça ise model yetersiz bir uyum göstermektedir. Hiper parametreye hangi değer verileceği bazı argümanlarla hesaplansa da manuel olarak denenmek suretiyle bu işlem yapılmıştır. Buraya kadar olan tüm işlemler mevcut çekirdekler için aynı yöntemle uygulanmıştır. Çekirdeklerden elde edilen sınıflandırma raporu aşağıda tablo halinde ve grafik olarak verilmiştir.

Tablo 4.4. DVM algoritması performans sonuçları

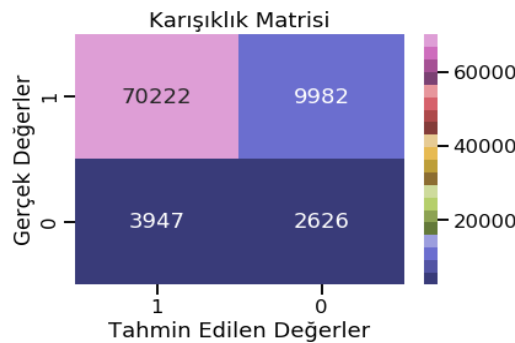
Çekirdek Tipi	C Hiper Parametresi	Gamma değeri	DP	YP	DN	YN	Doğruluk	Kesinlik	Hassasiyet	F-1 Skoru
Doğrusal	0.1	-	25507	1268	1666	11559	%68,25	0.95	0.69	0.80
Polinom	0.1	-	27112	1293	1612	9983	%71,81	0.95	0.73	0.83
Sigmoid	0.01	-	14349	1074	1056	13521	%51,35	0.93	0.51	0.66
RTF (RBF)	0.01	0.1	25380	954	1929	11737	%68,27	0.96	0.68	0.80



Şekil 4.3. DVM çekirdekleri İKE-EAA eğrileri

4.5.3. Derin Öğrenme Modelinin Uygulanması

Benzer olarak derin öğrenme modelinde 2013 yılı eğitim, 2014 yılı test olmak üzere ikiye ayrılmıştır. Derin öğrenme modeli 5 katmana ayrılmıştır ve katmanlar sırasıyla 1000, 500, 100, 50, 10 olacak şekilde nöronlar ile doldurulmuştur. Aktivasyon fonksiyonu olarak relu ve sınıflandırma için sigmoid tercih edilmiştir (Daşkın, Balık Şanlı, Özkan, & Kartal, 2018). Optimizasyon yöntemi olarak adam, kayıp fonksiyonları hesaplamak için ikili çapraz entropi eğitim sayısı (epoch) 20 olarak belirlenmiştir. Modelden elde edilen karışıklık matrisi sonuçları Şekil 4.4’de gösterilmiştir.



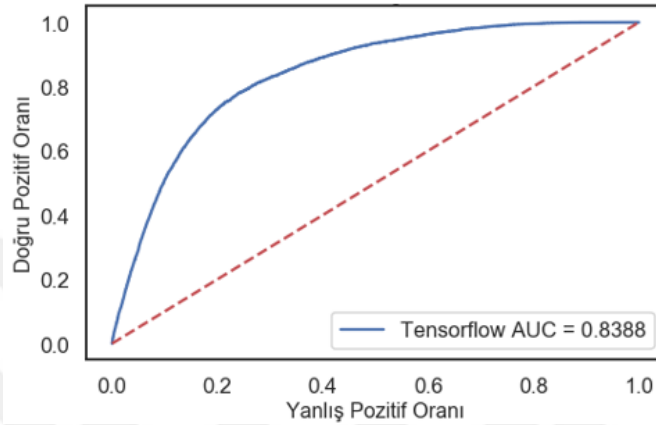
Şekil 4.4. Derin öğrenme modeli karışıklık matrisi

Karışıklık matrisi kullanılarak elde edilen performans sonuçları ise aşağıdaki tabloda verilmiştir.

Tablo 4.5. Derin öğrenme performans değerlendirmesi sonucu

Sonuçlar	Kesinlik	Hassasiyet	F-1 Skoru	Doğruluk
	0.95	0.88	0.91	%83,94

Karışıklık matrisi sonuçlarının performans değerlendirmesinden sonra, İKE-EAA oranı da hesaplanmış olup, sonuç Şekil 4.5’de gösterilmiştir.

**Şekil 4.5.** Derin öğrenme İKE eğrisi

Model %83,88 oranında İKE-EAA sonucu vermiştir. Gümrük dolandırıcılığı tespiti için derin öğrenme metodunu uygulayan (Kim, ve diğerleri, 2020)’nin çalışmasında %80,10 oranında bir oran elde edilmişken, modelde bazı ince ayarlar yapılarak bu oran %83,88’a çıkarılmıştır.

4.6. Algoritmaların Performans Değerlendirmesi

Yasadışı işlemlerin tahmin edilmesinde, ilk olarak uygulanan K-EYK algoritması için yakın komşuluk sayısı 3, 5, 7 olarak atanmış ve en iyi sonucun 7 olduğu tespit edilmiştir. Bu durumda F skorun yaklaşık olarak 0.96 EAA değerinin ise %71,51 olduğu raporlanmıştır. Diğer taraftan DVM’nin söz konusu çekirdeklerdeki hiper parametre değerleri için farklı performans sonuçları elde edilmiştir. Bu sonuçlar çerçevesinde; C hiper parametresi Sigmoid: C=0.01, Polinom: C=0.1, Doğrusal: C=0.1 ve RTF: C=0.01 ve gamma=0.1 değeri verildiğinde en iyi performans sonuçlarının elde edildiği gözlemlenmiştir.

C hiper parametresine daha yüksek bir değer verildiğinde Overfitting (aşırı uyum) daha düşük değerler atandığında Underfitting (yetersiz uyum) sorunu gözlemlenmiştir. Ayrıca C’ye daha yüksek bir değer verilmesinin işlem süresi açısından uzun sürdüğü görülmüştür.

Performans değerlendirmesi sonucunda, Doğrusal çekirdek F-1 skoru 0.80, İKE-EAA sonucu %63,43, Polinom çekirdek F-1 skoru 0.83, İKE-EAA sonucu %64,76, RTF çekirdek F-1

skoru 0.80, İKE-EAA sonucu %67,64 ve Sigmoid çekirdek F-1 skoru 0.66, İKE-EAA sonucu %50,56 olmuştur.

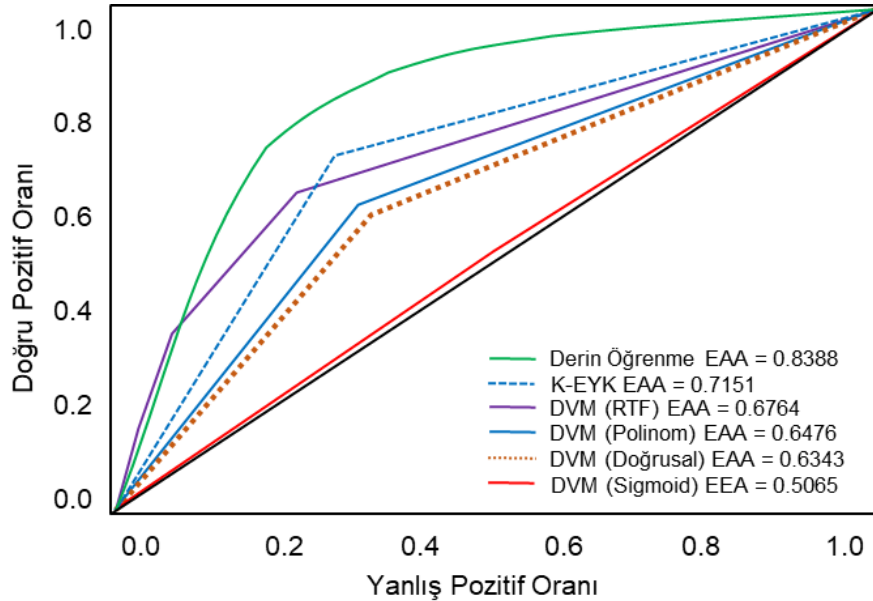
DVM çekirdeklerde en iyi F-1 skorunu 0.83 ile Polinom çekirdek göstermiştir. En iyi İKE-EAA sonucunu da %67,64 oranıyla RTF çekirdekte elde edilmiştir. En kötü performansı Sigmoid çekirdek uygulaması göstermiştir.

Derin öğrenme modeli için elde edilen performans değerlendirmesine bakıldığında F-1 skoru 0.91, İKE-EAA oranı %83,88 olarak tespit edilmiştir.

Tüm algoritmaların F skoru için performans değerlerini gösteren Doğruluk, Kesinlik, Hassasiyet ve F-1 skoru değerleri Tablo 4.6'da İKE-EAA sonuçlarına göre performans sonuçları Şekil 4.6'da gösterilmiştir.

Tablo 4.6. K-EYK, DVM ve derin öğrenme performans değerlendirmesi sonuçları

Algoritma	Doğruluk	Kesinlik	Hassasiyet	F-1 Skoru	İKE-EAA
K-EYK	%92,26	0.93	0.99	0.96	%71,51
DVM (Doğrusal)	%68,25	0.95	0.69	0.80	%63,43
DVM (Polinom)	%71,81	0.95	0.73	0.83	%64,76
DVM (Sigmoid)	%51,35	0.93	0.51	0.66	%50,65
DVM (RTF)	%68,27	0.96	0.68	0.80	%67,64
Derin Öğrenme	%83,94	0.95	0.40	0.91	%83,88



Şekil 4.6. K-EYK, DVM ve derin öğrenme İKE-EAA grafiği

Hassasiyet değeri, yasal olan işlemlerin yasadışı işlemler olarak görülmesi ihtimalinin olduğu sonuçlardır. Yani veri seti örneğinde yasadışı işlem sayısı 14.370 olarak gösterilmiş olup,

algoritma bu sayının üstünde yasadışı işlem ortaya çıkaracaktır. Avantajı; mümkün olduğunca yasadışı işlemin tahmin edilmesi, sakıncası ise; kontrol edilmesi gereken gümrük işlem sayısını arttırmasıdır.

Öte yandan; kesinlik oranı ölçüt olarak kabul edildiğinde, yasadışı işlemlerin, yasal işlemler gibi değerlendirileceği anlamına gelir. Bu da yasadışı işlemlerin büyük oranda gözden kaçırılması demektir. Avantajı; kontrol edilmesi gereken işlemin az oluşudur. Sakıncası ise; yasadışı olan işlemlerin az tahmin edilerek birçoğunun gözden kaçırılmasıdır. Bu sorunun çözümü için, Hassasiyet ve Kesinlik oranının harmonik bir ortalamasını alan F skorunun sonuçları temel alınmıştır. Diğer taraftan; Hassasiyetin Kesinliğe olan oranıyla hesaplanan İKE-EAA oranı bu çalışmada performans ölçütü olarak kabul edilmiştir.



5. SONUÇ ve ÖNERİLER

Sonuç: Modellerin performansları F skoru ve İKE-EAA sonuçları elde edilerek değerlendirilmiştir. Performans değerlendirmesinde sadece F skoru veya sadece İKE-EAA sonuçları üzerinden de değerlendirme yapmak mümkündür. F skoru, Hassasiyet, Kesinlik, İKE-EAA gibi sonuçlar verinin özelliğine ve verideki beklentiye göre göreceli tercihlerdir. Bunun kararını veriyi analiz ettirenler veya veri bilimcisinin kendisi karar vermelidir.

Çalışma sonucunda; F skoruna göre 0,96 oranında K-EYK algoritması, İKE-EAA değerlerine göre de %83,88 oranında Derin Öğrenme modeli en iyi performansı göstermiştir. Bu sonuçlar gösteriyor ki; bir ülkenin gümrük ithalatındaki usulsüzlükleriyle ilgili en başından hiçbir kontrol yapılmadan, işlemlere başlanılmadan, beyan aşamasında gerçeğe aykırı bir durumun olup olmadığı yaklaşık olarak %84 oranında doğru tahmin edilmektedir. Buna bağlı olarak özellikle, denetim ve kontrollerin azalacağı, ithalatçı firmaların evrak kontrollerinin asgari düzeye indirileceği ayrıca gözden kaçan işlemler nedeniyle devletin alamadığı vergi gelirlerinin de artacağı değerlendirilmektedir.

Genel olarak; devlet kurumları veya işletmeler, ellerinde bulundurdıkları yığın verileri yapay zekâ, makine öğrenmesi veya derin öğrenme yöntemleriyle analiz etmelidir. Bu yöntemler geleneksel analiz yöntemlerinden çok farklıdır. Burada, gelecek olaylar ve gelişmelerle ilgili birtakım öngörüler, tahminler ve çıkarımlar elde edilmektedir. Bu bağlamda, bu tür bilimsel çalışmalara daha fazla yer verilmesi, yenilikçilik ve yaratıcılık konusunda esnek kuralların olması gerekmektedir. Bu amaçla, varsa hukuki düzenlemelerin ve bürokratik sorunların çözüme kavuşturulması da önem arz etmektedir. Bu çalışmaların yapılabilmesi için de ihtiyaca uygun personel yetiştirilmesi, teknolojik ürün ve hizmetlerin temin edilmesi gerekir. Tüm bu anlatılanlar uygulandığında kurumsal açıdan fayda sağlanacağı gibi, nihayetinde devlete faydalı olacağı sonucuna varılmıştır.

Kurumlar için öneriler: Bu çalışmada kullanılan veri seti gerçek olmakla birlikte gizlilik gerektiren bölümler CTGAN yöntemiyle dönüştürülmek suretiyle herkese açık bir şekilde paylaşılmıştır. Benzer şekilde, devlet kurumlarının veya işletmelerin kendi verilerini, gizlilik gerektiren bölümlerini dönüştürmek suretiyle paylaşımları mümkündür. Bu sayede, veri bilimiyle uğraşanlar bu verileri analiz ederek bir tahminde bulunabileceklerdir. Devlet kurumlarının veya işletmelerin buna rağmen birtakım gizlilik gerekçeleri ve çekinceleri varsa kendi imkanlarıyla bu çalışmaları yapmaları önerilmektedir.

Araştırmacılar için öneriler: Bu çalışmadaki veriler, bazı algoritmalarda aşırı uyum veya yetersiz uyum göstermiştir. Sayı olmayan ifadelerin sayılara dönüştürülmesi gerekmiştir. Ayrıca veri setindeki dengesizliklerin üstesinden gelinmesi gibi zorluklarla karşılaşmıştır.

Özellikle Destek Vektör Makineleri çekirdek işlevcilerinin işlem süreleri yaklaşık 2 ay sürmüştür. Fakat K-EYK ve Derin Öğrenme modeli işlem süreleri dakikalar içinde tamamlanmıştır. Bu da bazı algoritmaların kendilerine özgü karmaşık matematiksel işlem döngüleri nedeniyle büyük verilerde çok fazla zaman kaybı ve iş gücüne sebep olduğunu göstermektedir.

Verinin çok olması, verinin türü ve yapılacak işlemle ilişkili olmakla birlikte göreceli bir kavramdır. Büyük verilerin varlığı uygun model seçimine bağlı olarak performansı olumlu yönde etkileyecektir. Bu çalışmadaki verilerin de uygulanan modeller açısından yeterli seviyede olduğu değerlendirilmektedir. Ancak sadece iki yıla ait veri bulunmaktadır. Daha fazla yıla ait veri olduğunda yıllar arasında karşılaştırma yapılması da mümkün hale gelecektir. Bu durumun aynı zamanda, algoritmaların performans sonuçlarını da etkileyeceği tahmin edilmektedir.



KAYNAKLAR

- [1]. 3501 Sayılı Kanun. (1988, Ekim 10). Uyumu Sağlanmış (Armonize) Mal Tanımı ve Kodlama Sistemi Hakkında Uluslararası Sözleşmeye Katılmamızın Uygun Bulunduğuna Dair Kanun.
- [2]. Adak, M. (2010). Dış ticarete giriş. İstanbul Üniversitesi Açık ve Uzaktan Eğitim Fakültesi.
- [3]. Agrawal, R. (2014). K-Nearest neighbor for uncertain data. International Journal of Computer Applications (0975 – 8887), 13-16.
- [4]. Altıntaş, H. (2018). Gümrük vergilerinin yapısının, Türkiye gümrük ve avrupa gümrük mevzuatı çerçevesinde karşılaştırılması. Yayımlanmamış yüksek lisans tezi, Dumlupınar Üniversitesi Sosyal Bilimler Enstitüsü. Kütahya.
- [5]. Apostolidis-Afentoulis, V., & -Ina Lioufi, K. (2010). SVM classification with linear and RBF Kernels. 1-7.
- [6]. Asraf, H., Nooritawati, M. T., & Shah Rizam, M. S. (2012). A Comparative Study in Kernel-Based Support Vector Machine of Oil Palm Leaves Nutrient Disease. International Symposium on Robotics and Intelligent Sensors 2012 (IRIS 2012), 1353-1359.
- [7]. Balaban, M., & Kartal, E. (2015). Veri madenciliği ve makine öğrenmesi temel algoritmaları ve R dili ile uygulamaları. İstanbul: Çağlayan Kitabevi.
- [8]. Bhandari, A. (2020, June 16). AUC-ROC Curve in Machine Learning Clearly Explained. 08 Temmuz 2021 tarihinde <https://www.analyticsvidhya.com/:https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/> adresinden alınmıştır.
- [9]. Büyükgöral, A., & Türkoğlu, Y. (2015). Gümrük kanunu'na göre tatbik edilen para cezalarında zamanaşımı sorunu. Gümrük ve Ticaret Dergisi, 75-85.
- [10]. Castelli, V. (2001). Unsupervised learning. New York: IBM T.J. Watson Research Center.
- [11]. Chawla, N., Bowyer, K., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 321-357.
- [12]. Choi, Y. (2020, July 12). Customs fraud detection. 06 Temmuz 2021 tarihinde https://raw.githubusercontent.com/YSCHOI-github/Customs_Fraud_Detection/master/synthetic_data2.csv adresinden alınmıştır.
- [13]. Chomboon, K., Chujai, P., Teerarassamee, P., Keerdprasop, K., & Kerdprasop, N. (2015). An empirical study of distance metrics for k-nearest neighbor algorithm. Proceedings of the 3rd International Conference on Industrial Application Engineering, (s. 280-285). Thailand.
- [14]. Commission Implementing Regulation (EU) 2019/1811. (2019, October 23). Brussels: Official journal of the european union.
- [15]. Cover, T. M., & Hart, P. E. (January 1967). Nearest Neighbor Pattern Classification. IEEE Transactions On Information Theory, 21-27.
- [16]. Çoban, K. (2020). İthal Eşyasının Cif Kıymeti, İthal Eşyasının Gümrük Kıymeti, Gümrüklenmiş Değer ve Ceza Uygulamaları. Gümrük Müşaviri Dergisi.

- [17]. Daşkın, Y. Ç., Balık Şanlı, F., Özkan, T., & Kartal, H. (2018). Derin öğrenme yöntemleri ile yüz görüntülerinden yaş kestirimi. VII. Uzaktan Algılama-CBS Sempozyumu (UZAL-CBS 2018), 1-7.
- [18]. Demir, G. (2018, Nisan 20). www.galyademir.com. Makine Öğrenmesi Nedir (Machine Learning)? : 24 Ağustos 2021 tarihinde <https://galyademir.com/2018/04/20/makine-ogrenimi-nedir/> adresinden alınmıştır.
- [19]. Devenport, T. H., & Laurence, P. (2001). İş dünyasında bilgi yönetimi: Kuruluşlar ellerindeki bilgiyi nasıl yönetirler. (G. Günay, Çev.) İstanbul: Rota Yayınlar.
- [20]. Digiampietri, L. A., Roman, N. T., Meira, L. A., Filho, J. J., Ferreira, C. D., & Kondo, A. A. (2007). Uses of artificial intelligence in the brazilian customs fraud detection system. The Proceedings of the 9th Annual International Digital Government Research Conference, (s. 181-187). Brazil.
- [21]. Doğan, F., & Türkoğlu, İ. (2019). Derin öğrenme modelleri ve uygulama alanlarına ilişkin bir derleme. DÜMF Mühendislik Dergisi, 410-455.
- [22]. Domo. (2020, May 08). Data Never Sleeps 6.0. 01 Ağustos 2021 tarihinde www.domo.com: https://www.domo.com/assets/downloads/18_domo_data-never-sleeps-6+verticals.pdf adresinden alınmıştır.
- [23]. Döring, M. (2018, November 21). Data science blog. 26 Temmuz 2021 tarihinde <https://www.datascienceblog.net/>: <https://www.datascienceblog.net/post/machine-learning/specificity-vs-precision/> adresinden alınmıştır.
- [24]. Google Brain Team. (2-4 November 2016). Tensor Flow: A system for large-scale machine learning. 12th USENIX Symposium on Operating Systems Design and Implementation, 265-277.
- [25]. Gupta, N., Rawal, A., Narasimhan, D., & Shiwani, S. (2013). Accuracy, sensitivity and specificity measurement of various classification techniques on healthcare data. IOSR Journal of Computer Engineering, 70-73.
- [26]. Göldoğan, E. (2017). Çeşitli çekirdek fonksiyonları ile oluşturulan destek vektör makinesi modellerinin performanslarının incelenmesi: Bir klinik uygulama. Yayımlanmamış doktora tezi, İnönü Üniversitesi ve Mersin Üniversitesi Biyoistatistik ve Tıp Bilişimi Anabilim Dalı, Malatya.
- [27]. Gümrük Kanunu. (1999, Ekim 27). Gümrük Kanunu. 31-45.
- [28]. Gümrük yönetmeliği. (2009, Temmuz 10). Gümrük Yönetmeliği.
- [29]. Güneş, H., Bousmalis, K., Martinez, B., Nicolaou, M., Oikonomopoulos, A., Rudovic, O., & Tzimiropoulos, G. (2011). Machine Learning (Course 395). Computer Based Coursework Manual. London, U.K.
- [30]. Hanley, J. A., & McNeil, B. J. (1982). The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. 1. 29-36.
- [31]. Hilbert, M., & Lopez, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. New York: Science.
- [32]. Hinton, G. (2013). CSC2535 2013: Advanced Machine Learning Lecture 10 Recurrent neural networks. Toronto: University of Toronto.

- [33]. Hu, L. Y., Huang, M. W., Ke, S. W., & Tsai, C. F. (2016). The distance function effect on k nearest neighbor classification for medical datasets. Springer Plus, 1-9.
- [34]. IBM Cloud Education. (2020, July 15). Machine Learning. 23 Temmuz 2021 tarihinde [www.ibm.com: https://www.ibm.com/cloud/learn/machine-learning](https://www.ibm.com/cloud/learn/machine-learning) adresinden alınmıştır.
- [35]. IBM Cloud Education. (2020, August 19). Supervised Learning. 28 Temmuz 2021 tarihinde [www.ibm.com: https://www.ibm.com/cloud/learn/supervised-learning](https://www.ibm.com/cloud/learn/supervised-learning) adresinden alınmıştır.
- [36]. Ichi.pro. (2021, Mart 25). <https://www.ichi.pro/tr>. 15 Temmuz 2021 tarihinde <https://www.ichi.pro/tr/destek-vektor-makinesine-svm-giris-225431442261105> adresinden alınmıştır.
- [37]. Jose, I. (2018, November 08). KNN (K-Nearest Neighbors). 24 Temmuz 2021 tarihinde [towardsdatascience: https://towardsdatascience.com/knn-k-nearest-neighbors-1-a4707b24bd1d](https://towardsdatascience.com/knn-k-nearest-neighbors-1-a4707b24bd1d) adresinden alınmıştır.
- [38]. Khanna, S. K. (2019). Machine Learning v/s Deep learning. International Research Journal of Engineering and Technology (IRJET), 455-458.
- [39]. Kim, S., Kaist, T. D., Kaist, T.-D. M., Kaist, S. H., Kaist, S. P., Kaist, J. S., . . . Kaist, M. C. (2020). Take a Chance: Managing the Exploitation-Exploration Dilemma in Customs Fraud Detection via Online Active Learning. 1-10.
- [40]. Kim, S., Tsai, Y.-C., Singh, K., Choi, Y., Ibok, E., Li, C.-T., & Cha, M. (2020). DATE: Dual Attentive Tree-aware Embedding for Customs Fraud. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (s. 2-11). New York, NY, USA: Association for Computing Machinery.
- [41]. Kumar, A. (2020, July 28). SVM RBF Kernel Parameters With Code Examples. 20 Temmuz 2021 tarihinde [https://dzone.com: https://dzone.com/articles/using-jsonb-in-postgresql-how-to-effectively-store-1](https://dzone.com/articles/using-jsonb-in-postgresql-how-to-effectively-store-1) adresinden alınmıştır.
- [42]. LI, A., LI, Y.-x., & LI, X.-h. (2017). TensorFlow and Keras-based Convolutional Neural Network in Cat Image Recognition. Northeastern University, 1-5.
- [43]. Liew, A. (2007). Understanding data, information, knowledge and their inter-relationships. Journal of Knowledge Management Practice, 1-16.
- [44]. Lu, L. (2018). Optimal γ and C for e-Support vector regression with RBF kernels. 1-13.
- [45]. Lundell, J. F. (2019). Tuning hyperparameters in supervised learning models and applications of statistical. All Graduate Theses and Dissertations, 7594, 1-106.
- [46]. Majnik, M., & Bosnic, Z. (2013). ROC analysis of classifiers in machine learning: A survey, 531-558.
- [47]. Marshall, C., & Rossman, G. B. (1989). Designing qualitative research. USA: Sage Publications.
- [48]. MEB. (2011). Dış ticaret kavramları. Ankara: Milli Eğitim Bakanlığı.

- [49]. Nemire, B. (2014, March 11). Cuda-spotlight-gpu-accelerated-speech-recognition/. 20 Temmuz 2021 tarihinde <https://developer.nvidia.com/https://developer.nvidia.com/blog/cuda-spotlight-gpu-accelerated-speech-recognition/> adresinden alınmıştır.
- [50]. Özdamar, S. (2013). Dış ticarete yeni bir kavram:Hayali ithalat. Gümrük ve Ticaret Dergisi, 49-52.
- [51]. Price, D. (2020, July 17). Infographic: How much data is produced every day? 24 Temmuz 2021 tarihinde [www.cloudtweaks.com: https://cloudtweaks.com/2015/03/how-much-data-is-produced-every-day/](https://cloudtweaks.com/2015/03/how-much-data-is-produced-every-day/) adresinden alınmıştır.
- [52]. Richmond, B. (2006). Introduction to data analysis handbook. Washington: Migrant & Seasonal Head Start Technical Assistance Center.
- [53]. Roman, N. T., Ferreira, C. D., Meira, L., Rezende, R. C., Digiampietri, L. A., & Filho, J. J. (2009). Attribute-Value specification in customs fraud detection. The Proceedings of the 10th International Digital Government Research Conference, (s. 264-271). Puebla, Meksika.
- [54]. Sain, H., & Purnami, S. (2015). Combine sampling support vector machine for imbalanced data classification. The Third Information Systems International Conference, 59-66.
- [55]. Salomon, J., King, S., & Osborne, M. (2002). Framewise phone classification using support vector machines. 1-4.
- [56]. Samuel, A. (1959). Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development, 535-545.
- [57]. Sokolova, M., Japkowicz, N., & Szpakowicz, S. (January 2006). Beyond Accuracy, F-score and ROC: A family of discriminant measures for performance evaluation. Australasian Joint Conference on Artificial Intelligence (s. 1015-1021). Australia: Springer, Berlin, Heidelberg.
- [58]. Şahan, D. (2021, Şubat 24). [www.sahanndilara.medium.com](https://sahanndilara.medium.com/overfitting-a%C5%9F%C4%B1r%C4%B1-%C3%B6%C4%9Frenme-ve-underfitting-eksik-%C3%B6%C4%9Frenme-nedir-a7da43532312). 24 Ağustos 2021 tarihinde Denetimli (Supervised) ve Denetimsiz (Unsupervised) Öğrenme Nedir?: <https://sahanndilara.medium.com/overfitting-a%C5%9F%C4%B1r%C4%B1-%C3%B6%C4%9Frenme-ve-underfitting-eksik-%C3%B6%C4%9Frenme-nedir-a7da43532312> adresinden alınmıştır.
- [59]. Taylor, J. R. (1997). Error analysis. The study of uncertainties in physical measurement. U.S.A: Colorado University Science Books.
- [60]. TB. Yönerge. (2020, Şubat 26). Ticaret Bakanlığı Taşra Teşkilatı Çalışma Usul ve Esasları Hakkında Yönerge 52703313. Ankara.
- [61]. Ticaret Bakanlığı. (2021, Ağustos 12). Tarife arama motoru. 12 Ağustos 2021 tarihinde <https://uygulama.gtb.gov.tr/Tara>: <https://uygulama.gtb.gov.tr/Tara> adresinden alınmıştır.
- [62]. Uzun, F. (2016). Yeni gümrük kanunu taslağı neler getiriyor? Gümrük ve Ticaret Dergisi, 135-137.
- [63]. Uzun, F. (2019, Ekim 02). [www.nazaligundem.com](http://nazaligundem.com/tr/yayinlar/gumruk-kanununda-degisiklik-yapilmasina-dair-kanun-teklifinin-degerlendirilmesi?onlyTm=t&tm=6). 24 Ağustos 2021 tarihinde Gümrük Kanununda Değişiklik Yapılmasına Dair Kanun Teklifinin Değerlendirilmesi: <http://nazaligundem.com/tr/yayinlar/gumruk-kanununda-degisiklik-yapilmasina-dair-kanun-teklifinin-degerlendirilmesi?onlyTm=t&tm=6> adresinden alınmıştır.

- [64]. Vakili, M., Ghamsari, M., & Rezaei, M. (2019). Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification. 1-13.
- [65]. Vanhoeyveld, J., Martens, D., & Peeters, B. (2020). Customs Fraud Detection. Pattern Anal Applic 23, 1457-1477.
- [66]. Vasileska, L. (2012). Prevention and Detection of Customs Fraud.
- [67]. Walczak, S., & Cerpa, N. (2003). Artificial Neural Networks. S. Walczak, & N. Cerpa içinde, Encyclopedia of Physical Science and Technology (s. 631-645). Academic Press.
- [68]. Wang, J., Leung, K., Lee, K., & Wang, Z. (2012). Nonlinear Integrals with Polynomial Kernel and Its Applications. International Journal of Intelligent Systems, 48-68.
- [69]. Wang, Y., Li, Y., Song, Y., & Rong, X. (2020). The Influence of the Activation Function in a Convolution Neural Network Model of Facial Expression Recognition. Applied Sciences, 1-20.
- [70]. WCO. (2019, August 19). World Customs Organization. Temmuz 17, 2021 tarihinde <http://www.wcoomd.org/>: <http://www.wcoomd.org/-/media/wco/public/global/pdf/about-us/wco-members/list-of-members-with-membership-date.pdf?db=web> adresinden alınmıştır.
- [71]. Webber, R. J. (2013). The Evolution of Direct, Data and Digital Marketing.
- [72]. Wind, I. (2008). HS codes and the renewable energy sector. ICTSD Programme on Trade and Environment, 1-77.
- [73]. www.crsouza.com. (2010, March 17). Kernel Functions for machine learning applications. 13 Temmuz 2021 tarihinde [www.crsouza.com](http://crsouza.com): <http://crsouza.com/2010/03/17/kernel-functions-for-machine-learning-applications/#polynomial> adresinden alınmıştır.
- [74]. www.ichi.pro.tr. (2021, Haziran 21). 25 Ağustos 2021 tarihinde Çekirdeklerle Destek Vektör Makinesini Anlama: <https://ichi.pro/tr/cekirdeklerle-destek-vektor-makinesini-anlama-118325373705925> adresinden alınmıştır.
- [75]. www.keras.io. (2021, March 08). <https://keras.io>. 10 Temmuz 2021 tarihinde <https://keras.io>: <https://keras.io/about/> adresinden alınmıştır.
- [76]. www.muhasenews.com. (2019, Nisan 24). www.muhasenews.com. 19 Temmuz 2021 tarihinde Türk Gümrük Tarife Cetvelinde Kullanılan 12 Rakamlı Kodun İsmi Nedir ve Bu Sayılar Ne Anlama Gelmektedir?: <https://www.muhasenews.com/turk-gumruk-tarife-cetvelinde-kullanilan-12-rakamli-kodun-ismi-nedir-ve-bu-sayilar-ne-anlama-gelmektedir/> adresinden alınmıştır.
- [77]. www.nairametrics.com. (2011). Nigerian Import Duty. 29 Temmuz 2021 tarihinde <https://nairametrics.com>: <https://nairametrics.com/wp-content/uploads/2011/11/Nigerian-Import-Duty.pdf> adresinden alınmıştır.
- [78]. www.tarifemevzuati.com. (2021, Ağustos 16). Gümrük Tarife İstatistik Pozisyonu GTİP Nedir? 15 Temmuz 2021 tarihinde www.tarifemevzuati.com: <https://tarifemevzuati.com/tarife-hakkinda/gumruk-tarife-istatistik-pozisyonu-nedir/> adresinden alınmıştır.
- [79]. www.tradingeconomics.com. Nigeria GDP. www.tradingeconomics.com: 13 Temmuz 2021 tarihinde <https://tradingeconomics.com/nigeria/gdp> adresinden alınmıştır.

[80]. Xu, L., Skoularidou, M., Infante, A. C., & Veeramachaneni, K. (2019). Modeling Tabular data using Conditional GAN. NeurIPS, 2019.

[81]. Yılmaz, M. (2013). Küreselleşmenin oluşumuna zemin hazırladığı yeni ekonomik anlayış: Bilgi ekonomisi. Atatürk Üniversitesi İktisadi ve İdari Bilimler Fakültesi, 27(1), 241-255.

[82]. Yılmaz, Ö. (2018, Ağustos 13). www.lojiblog.com. 24 Ağustos 2021 tarihinde Gümrük İşlemlerinde Sıkça Karşılaşılan Sorunlar: <http://www.lojiblog.com/gumruk-islemlerinde-sikca-karsilasilan-cezalara-genel-bir-bakis/> adresinden alınmıştır.

