Araştırma Makalesi
Zeki Sistemler Teori ve Uygulamaları Dergisi 6(1) (2023) 1-11
DOI: 10.38016/jista.1140499

Makine Öğrenmesinde Kategorik Veri Kodlama Tekniğinin Kullanımına Alternatif Bir Çözüm Yöntemi

Ender Şahinaslan[1], Mustafa Günerkan [2], Önder Şahinaslan [3*]

[1] Trakya Üniversitesi, Bilgisayar Mühendisliği Bölümü, Edirne, Türkiye

[2] Maltepe Üniversitesi, Bilgisayar Mühendisliği Bölümü, İstanbul, Türkiye

[3] Maltepe Üniversitesi, Bilişim Bölüm Başkanlığı, İstanbul, Türkiye

dr.endsa@gmail.com, mgunerkan@gmail.com, ondersahinaslan@maltepe.edu.tr

Öz

Makine öğrenimi, derin öğrenme algoritmaları kullanarak insan zekâsını taklit eden bir teknolojidir. Öğrenme algoritmaları yalnızca sayısal veri kümeleri üzerinde çalışır. Kategorik veri kümeleri nitel veya nicel verilerden oluşur. Nitel veri setlerinin öğrenme algoritmalarında kullanılabilmesi için veri setinin sayısallaştırılması gerekmektedir. Sayısallaştırma için etiket kodlama, sıralı kodlama, toplam kodlama, ikili kodlama ve sıcak kodlama gibi birçok kodlama tekniği vardır ancak bu kodlama teknikleri performans, maliyet ve kullanım açısından bazı güçlükler ve yetersizlikleri barındırmaktadır. Diğer taraftan bir kodlama tekniği ile elde edilen eğitim çıktısının orijinalinin bilinmesine ihtiyaç duyulabilmektedir. Bu çalışma, kategorik verilerin sayısallaştırılmasında kodlama tekniklerinin kullanılmasından kaynaklanan yetersizliklere çözüm olabilecek, daha özgün ve daha iyi performansa sahip bir altyapı oluşturma arayışının bir sonucu olarak ortaya çıkmıştır. Geliştirilen yöntem uluslararası bir lojistik firmada 7 farklı kategoride toplam 46 kategorik özellik ve 80.154.139 adet veri üzerinden uygulanmıştır. Testlerin sonucuna göre veri setleri bazında %23.07 ile %300.13 arasında toplamda %153.62 performans kazancı elde edilmiştir. Bu sonuçlar, geliştirilen yöntemin daha başarılı ve uygulanabilir olduğunu göstermektedir. Çalışma, yüksek performans kazancı ve özgün yapısı ile benzer alanlarda kolaylıkla kullanılabilecek bir yapıya sahiptir. Makine öğrenmesinde kodlama tekniklerinin kullanımına alternatif bir çözüm sunmuştur.

Anahtar kelimeler: Kodlama, Makine Öğrenimi, Sistem Geliştirme, Teknoloji ve Yenilik, Veri Yönetimi.

An Alternative Solution Method to Using Categorical Data Encoding Technique in Machine Learning

Abstract

Machine learning is a technology that mimics human intelligence using deep learning algorithms. Learning algorithms only work on numerical datasets. Categorical datasets consist of qualitative or quantitative data. In order for qualitative data sets to be used in learning algorithms, the data set must be digitized. There are many coding techniques for digitization, such as label coding, sequential coding, total coding, binary coding and hot coding, but these coding techniques have some difficulties and inadequacies in terms of performance, cost and use. On the other hand, it may be necessary to know the original of the training output obtained with a coding technique. This study has emerged as a result of the search for a more original and better performing infrastructure that can be a solution to the inadequacies arising from the use of coding techniques in the digitization of categorical data. The developed method was applied on a total of 46 categorical features and 80.154.139 pieces of data in 7 different categories in an international logistics company. According to the results of the tests, a total of 153.62% performance gain was obtained between 23.07% and 300.13% on the basis of data sets. The study has a structure that can be used easily in similar areas with its high performance gain and original structure. It offered an alternative solution to the use of coding techniques in machine learning.

Keywords: Encoding, Machine Learning, Systems Development, Technology and Innovation, Data Management.

* Sorumlu yazar.
E-posta adresi: ondersahinaslan@maltepe.edu.tr

Alındı : 4 Temmuz 2022
Revizyon : 14 Ekim 2022
Kabul : 7 Kasım 2022

1. Giriş (Introduction)

Yapay zekâ ve makine öğrenmesi, akıllı sistemlerin inşasında ve geliştirilmesinde önemli rollere sahiptir. Makine öğrenimi, derin öğrenme algoritmaları kullanılarak insan zekâsını taklit eden bir bilgisayar teknolojisidir. Bu yeni teknoloji bilgisayar, bilişim, istatistik, bilgi ve kontrol teorisi, psikoloji, felsefe gibi birçok disiplin ve fikirlerden yararlanır (Mitchell, 1997). Makine öğrenimi, belirli görevleri gerçekleştirmek üzere sistemlerin verilerden öğrenebileceği, örüntüleri tanımlayabileceği ve minimum insan müdahalesi ile kararlar alabileceği fikrine dayanan bir yapay zekâ dalıdır (SAS, 2022). Makine öğrenimi yinelemeli verileri analizi sonucunda analitik bir model oluşturur. Kullanılan algoritmalar güvenilir sonuçlara ulaşmada önceki hesaplamaları kullanır. Öğrenme algoritmaları girdi olarak sayısal verilerden oluşan öznitelik matrisine ihtiyaç duyar. Kategorik verilerin kullanılabilir öznitelik matrisine dönüştürülebilmesi için kodlama yöntemleri kullanılır (Cerda vd., 2018). Bu kodlama tekniklerinin kullanımı bir çözüm sağlasa da bu defa kategorik veriler için hangi sayısal değerin atandığının bilinememesi durumuyla karşılaşılabilmektedir. Bu durum veri önişleme evresinde anlaşılması ve çözülmesi gereken problemlerden biridir. Veri ön işleme süreci, verilerle ilgili sorunları gidermek ve veri analiz öncesine hazırlık yapmak için gerçekleştirilir (Famili vd., 1997). Kullanılacak algoritmaların daha doğru ve verimli sonuçlar üretmesine yardımcı olmak için veriler, asıl veri analizinden önce bir dizi ön işleme tabi tutulur (MarketResearch, 2022). Veri setleri güvenilir, tekrar edilebilir ve sayısal değerlerden oluşmalıdır. Verilerin sayısallaştırılmasında etiket kodlama, sıcak kodlama, sıralı kodlama, ikili kodlama, Helmert kodlama ve Hash kodlama gibi bilinen kodlama teknikleri vardır. Bu tekniklerden *Etiket Kodlama* tekniğinde her etikete alfabetik sıraya göre benzersiz bir tam sayı atanır (Sethi, 2022). Atanan değerler 0 ile başlar ve kategorik veri tür sayısından bir eksik olacak şekilde verilir (Scikit-Learn, 2022). Özniteliklerin sıralı olmadığı durumlarda yaygın kullanılan *One-Hot Kodlama* tekniğinde: her kategori için ikili sütun oluşturulur ve kategoriler her bir özellikteki benzersiz değerlere göre türetilir. Bu kategoriler manuel olarak da belirlenebilir (ScikitLearn-OneHotEncoder, 2022). *Ordinal Kodlama* tekniğinde: mevcut kategorilerin sayısının bilindiği durumda her kategoriye bir tam sayı atanarak nitelikler sıralı tam sayılara dönüştürülür (ScikitLearn-OrdinalEncoder, 2022). *Binary Kodlama* tekniğinde: kategorik veriler için her kategoriye sayısal bir değer atanır, ardından ikili gösterime dönüştürülür ve bu ikili biçimde temsil edilir (Seger, 2018). *Helmert Kodlama* tekniğinde: kategorik bir değişkenin her seviyesi, sonraki seviyelerin ortalaması ile karşılaştırılır (Potdar vd., 2017). *Hash Kodlama* tekniğinde: farklı boyuttaki girdiler için sabit değerler üretilir ve bu değerler öznitelik olarak kullanılır (Turcanik ve Javurek, 2016).

Bu çalışma Türkiye dış ticaretinin yaklaşık %8'ini gerçekleştiren uluslararası lojistik bir firmanın 2010-2021 yılları arasındaki beyanname veri setleri üzerinde gerçekleştirilmiştir. Nitel özelliğe sahip kategorik verilerin varlığı, bu verilerin sayısallaştırma ihtiyacını doğurmuştur. Sayısallaştırma sürecinde kodlanması kolay ve performanslı olan etiket kodlama tekniği kullanılmıştır. Etiket kodlama tekniği kullanımı sonucunda elde edilen eğitim sonuçlarında kategorik verilere hangi sayısal değerlerin atandığı tam olarak elde edilememiştir. Bu tekniğin kullanımından kaynaklanan yetersizlik ve düşük performans problemi nedeniyle alternatif yöntem, yapı ve çözüme ihtiyaç duyulmuştur. Bu ihtiyacın giderilmesine yönelik araştırmalar sonucunda kategorik verilerin sayısallaştırmasında bilinen tekniklere alternatif olabilecek bir yöntem geliştirilmiştir. Her iki yöntem yaklaşık 80 milyon veri üzerinden aynı laboratuvar koşullarında süre performans başarım testlerine tabii tutulmuştur. Bu çalışmada önerilen yöntemin ciddi performans kazancına yardımcı olduğu görülmüştür.

Çalışmanın literatür bölümünde konuyla ilgili güncel çalışmalara, materyal ve yöntemler bölümünde çalışma ortamı, yöntem, veri seçimi, veri seti üzerinden yürütülen analiz, kodlama, uygulama ve test çalışmalarına ve kıyaslama yapabilmek için en çok bilinen ve kullanılan etiket kodlama çalışmasına yer verilmiştir. Çalışma sonucunda elde edilen sonuçlar bulgu bölümünde, sonuçların değerlendirilmesi ve çıkarımlar tartışma ve sonuç bölümlerinde sunulmuştur.

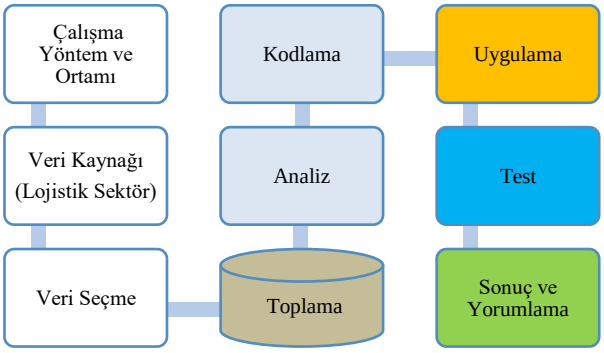
2. Literatür (Literature)

Çalışma konusuyla ilgili yapılan literatür taraması sonucunda günümüzde oldukça popüler ve farklı farklı alanlarda uygulama imkanı bulan makine öğrenmesine dayalı yöntemlerin uygulandığı çalışmalara yeterli sayıda ulaşılmasına rağmen kodlama teknikleri üzerine yürütülen çalışmalara sınırlı sayıda ulaşılabilmiştir. (Jackson & Agrawal, 2019), çalışmasında *Etiket kodlama*, *One-Hot kodlama ve Binary kodlama* tekniklerinin performansını ele almışlardır. Çalışma sonucunda en iyi performansı *Etiket kodlama* en düşük performansı ise tekniğinde *One Hot kodlama* tekniğinde bulmuşlardır. (Shen & Shafiq, 2019), kategorik verilerin sayısallaştırılmasında *Etiket ve One-Hot kodlama* tekniklerini kullanmışlardır. Kodlama süresi bakımından Label kodlama tekniğini daha avantajlı bulmuşlardır. (Jiang vd., 2020), üretim verileri üzerinden verim tahmini gerçekleştirmede kullandıkları kodlama teknikleri kıyaslamışlardır. (Yu vd., 2020), kredi sınıflandırmasında kayıp veri problemini çözmek için *One-Hot kodlama* tekniğine dayalı bir veri ön işleme yöntemi önermişlerdir. (Chandradeva vd., 2019), finansal işlem dolandırıcılığı tespitinde denetimli makine öğrenme algoritmaları kullanarak sahte işlem olaylarını %83 oranında tespit edebilmişlerdir. Kategorik verilerin kodlanmasında, Etiket kodlama tekniğinin makine öğrenme algoritması tarafından

yorumlanma dezavantajı nedeniyle *One-Hot kodlama* tekniği kullanmışlardır. (Nerlikar vd., 2020), bilgisayar tabanlı saldırıları tespit ve analiz etmeyi amaçlamışlardır. Verimlilik için veri önişleme aşamasında ilgisiz özellikleri kaldırmışlardır. *One-Hot kodlama tekniği* kullandıkları çalışmalarında %96,5 doğruluk oranı elde etmişlerdir (Chen vd, 2020), bulut ortamlarındaki ağlara izinsiz girişlerin tespitine odaklanmışlar, kodlama tekniği olarak *Etiket kodlama ve One-Hot kodlama* teknikleri kullanmışlardır. Oluşturdukları model ile %99,71 doğruluk değerine ulaşmışlardır. (Li, 2018), geçmiş kira verilerini kullanarak konut piyasası için objektif bir ölçüm sağlayamaya yarayan aylık konut kirasını doğru tahmin etmeye çalışmıştır. Çalışmasında verileri modele uygun hale getirmek için *Etiket kodlama ve One-Hot kodlama* tekniğini kullanmıştır. Kategori sayısının çok fazla olmadığı durumlarda *One-Hot kodlama* tekniğini önermiştir. (Chakrabarty, 2019) American Airlines iç hat uçuş bilgilerinim analizinde veri madenciliği ve makine öğrenimi yaklaşımları üzerinden uçuşun varış gecikmesini tahminine çalışmışlardır. Veri önişleme sürecinde *Etiket kodlama ve One-Hot kodlama* tekniklerini kullanmışlardır. (Li vd., 2020), hasta sağlık verileri üzerinden kanser klinik son nokta tahmini yapmada derin öğrenmeye dayalı bir yaklaşım önerdikleri çalışmada *Etiket kodlama ve One-Hot kodlama* tekniklerini kullanmıştır. (Sharma vd., 2020), ağ güvenliğine yönelik saldırıların tespitinde sistemin normal davranışından herhangi bir sapma olup olmadığını anlamaya yönelik çalışmada makine öğrenme algoritmalarından yararlanmışlardır. Çalışma veri ön işleme süreçlerinde *Label kodlama ve One-Hot kodlama* tekniklerini kullanmışlardır. Çalışmada kullandıkları algoritmalarda %98 ile %100 arasında değişen doğruluk oranları elde etmişlerdir. (Al-Shehari ve Alsowail, 2021), kurum içi siber saldırıların doğuracağı zararların etkisini tespit etmede karar ağacı ve en yakın komşuluk gibi bilinen öğrenme tabanlı algoritmaları kullanmışlardır. Veri ön işleme süreçlerinde Etiket kodlama ve One-Hot kodlama tekniklerinden yararlanmışlardır. (Günerkan vd., 2022), gümrük beyanname oluşturma sürecinde öğrenmeye dayalı algoritmaların performansını ölçtükleri çalışmalarında kategorik verilerin çokluğu nedeniyle sayısallaştırma sürecinde Etiket kodlama tekniğini kullanmıştır. (Reilly vd, 2022), 10 yıllık bir dönemi kapsayan, ev içi yangın yaralanmalarının kategorik bir veri setini kullanmışlardır. Algoritmanın doğru sonuçlar üretmesinde Label kodlama, One-Hot kodlama ve Ordinal kodlama gibi kodlama tekniklerinden çalışmaya uygun olanın seçiminin önemine vurgu yapmışlardır. (Yılmaz Yalçıner ve Gelen Mert, 2021), bir konaklama işletmesi için doluluk oran tahmininde yapay sinir ağlarını kullanmışlardır. (Şahinaslan vd., 2022), Naive Bayes sınıflandırma algoritması aracılığıyla Youtube sosyal medya uygulamasında yer alan 15.082 veri setini üzerinden %65,56 oranında doğru sonuca ulaşmışlardır. (Kıran vd., 2022), kullandıkları veri özniteliklerini dikkate alarak çalışmalarında derin öğrenme yöntemini tercih etmişlerdir. (Karasulu vd., 2022), insan kulağı görüntülerinden cinsiyeti belirlemek için derin öğrenme tabanlı melez bir yaklaşım geliştirmişler. Sayısal verilerin Bu çalışmalarda kullanılan kodlama tekniklerine ilişkin özel bir bilgiye rastlanmamıştır.

3. Materyal ve Yöntem (Material and Method)

Makine öğrenme algoritmalarının çalıştırılabilmesi için veri setlerindeki özniteliklerin sayısal değerlere sahip olması gerekir. Nitel özelliklere sahip kategorik verilerin sayısallaştırılması gerekir. Sayısallaştırma için birçok kodlama tekniği mevcuttur ancak bu kodlama teknikleri performans, maliyet ve kullanımından kaynaklı bazı zorluk ve yetersizlikleri barındırmaktadır. Bu sorunun çözümüne katkı sunacak yeni bir yol-yöntem arayışına ihtiyaç duyulmuştur. Uluslararası bir lojistik firmada karşılaşılan benzer bir sorunun çözümünde alternatif bir yöntem geliştirilmiştir. Elde edilen başarılı sonucun literatüre kazandırılması amaçlanmıştır. Yürütülen çalışmalara ait ana süreç aşamaları Şekil 1'de gösterilmektedir.



Şekil 1. Çalışma süreç aşamaları (Study process stages)

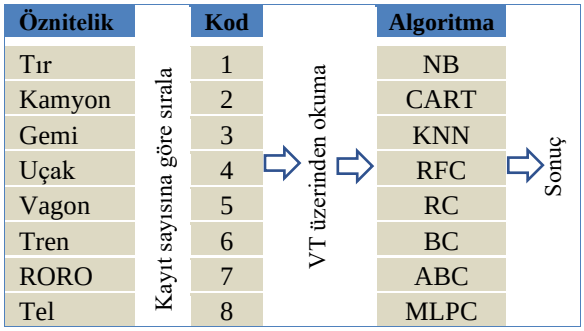
Yürütülen çalışmanın ana süreçleri çalışma ortamın hazırlanması, veri kaynağının belirlenmesi, seçilen verilerin toplanması, öz niteliklerin belirlemesi, veri analizi, kodlama, uygulama, test ve sonuçların değerlendirme aşamalarından oluşmaktadır.

3.1. Çalışma ortamı (Study environment)

Bu çalışma, Windows 10 Pro 64 bit işletim sistemi, işlemci Intel(R) Xeon(R) Gold 6154 CPU @ 3.00 GHz 2.99 GHz işlemcili, x64 tabanlı işlemci mimarisine ve 11.7 GB yüklü belleğe(RAM) sahip olan bir bilgisayar üzerinde gerçekleştirilmiştir. Çalışılan veri setleri ve eşleştirme tabloları için MSSQL Server Management Studio v17.9 VT yönetim sistemi kullanılmıştır. Veri öznitelik analiz ve tablo kayıt işlemlerinde T-SQL kullanılmıştır. Uygulama geliştirme, kodlama ve test işlemlerinde Spyder IDE 5.1.5 editörü ve Python 3.9.7 64 bit programlama dili kullanılmıştır. Performans testleri VT'nın az kullanıldığı gün ve saatte gerçekleştirilmiştir.

3.2. Yöntem (Method)

Bu çalışmada mevcut bir takım kodlama tekniklerinin kullanımına alternatif olabilecek bir yöntem geliştirilerek buna ait uygulama ve testler gerçekleştirilmiştir. Geliştirilen yönteme ait işlem aşamaları Şekil 2'de gösterilmektedir.

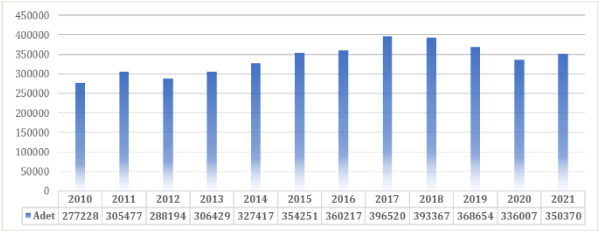


Şekil 2. İşlem adımları (Process steps)

Çalışmada kullanılacak verilerin ön inceleme ve analizi sonucunda kategorik veri öznitelikleri belirlenir. Veri setinin öznitelikleri belirlenirken delphi metodu kullanılmıştır. Uzmanlardan yararlanılarak bir uzlaşma veya karara varılmıştır. Burada da gümrük mevzuatı bilgisi olan kişilere bilgi alanı tahmininde kullanılacak özniteliklerin tespiti için sorular sorularak modeller belirlenmiştir. Modellerin ilk eğitim sonuçları değerlendirilerek özniteliklerin üzerinden geçilerek değişiklikler yapılmıştır. Bu süreç kabul edilebilir başarım oranlarına ulaşıncaya kadar tekrarlanarak modellere son hali verilmiştir. Bu öznitelik kayıtları, veri tabanı(VT) üzerinde oluşturulan tablo üzerinde kaydedilir. Her bir özniteliğe ait veri kayıt sayıları hesaplanır. Elde edilen kayıt sayıları büyükten küçüğe sıralanır. Sıralamaya uygun şekilde en fazla veri hacmine sahip olan veri kategorisine 1, ardından gelene 2, daha sonra 3 şeklinde ardışık olarak artırım yapılır. Belirlenen tam sayı değeri ilgili tablo üzerinde açılmış olan 'kod' veri alanına işlenir. Üretilen bu sayısal kodlar makine öğrenme uygulamalarında kullanılır. Geliştirilen bu yöntemde öğrenme algoritmaları kullanımından önce her hangi bir kodlama tekniği kullanımına ihtiyaç duyulmaz. Geliştirilen yöntemin bilinen bir kodlama yöntemi ile kıyaslaması yapılarak sonuçların değerlendirilmesi gerçekleştirdi.

3.3. Veri kaynağı (Data source)

Çalışmada uluslararası bir lojistik firmanın 2010 ve 2021 yılları arasındaki beyanname üzerinde yer alan 7 ayrı veri setinde toplam 80154139 adet kayıt kullanılmıştır. Çalışmada kullanılan veri setlerinden 'taşıma şekli' veri seti türünü detay olarak incelediğimizde bu veri türünde toplam 4064131 adet veri tespit edilmiştir. Bu verilerin 2010-2021 yılları arasındaki dağılım grafiği Şekil 3'de gösterilmektedir.



Şekil 3. Taşıma şekli verilerinin yıllara göre dağılımı (Distribution of mode of transport data by years)

Sunulan bu verilere setinde en az veri 277.288 adetle 2010 yılında gerçekleşirken en fazla veri 396.520 adetle 2017 yılında gerçekleşmiştir. Yıllık ortalaması yaklaşık 340.000 adettir. Çalışmada kullanılan diğer veri seti tür ve adetleri veri seçme bölümünde açıklanmaktadır.

3.4. Veri seçimi (Data selection)

Bu çalışma lojistik veri kaynağı beyanname kayıtları üzerinde yer alan kategorik verilerden seçilen 'teslim şekli', 'taşıma şekli', 'sevk ülke', 'döviz türü', 'birim', 'uluslararası anlaşma', 'ödeme şekli' öznitelikleri üzerinde gerçekleştirilmiştir. Seçilen bu yedi farklı veri setine ait kayıt sayıları Tablo 1'de gösterilmektedir.

Tablo 1. Veri kümesi kayıt sayıları (Dataset record counts)

Veri Seti	Kayıt Sayısı
Teslim Şekli	4034742
Taşıma Şekli	4064131
Sevk Ülke	4031130
Döviz Türü	4065936
Birim	22015778
Uluslararası Anlaşma	22000635
Ödeme Şekli	19941787
Toplam	80154139

3.5. Veri Toplama (Data collection)

Sürecin bu aşamasında beyanname kayıtlarından seçimi yapılan kategorik veriler MS SQL VT üzerinde oluşturulan eşleştirme tablosu üzerinde toplanmıştır. Öznitelikler bilgileri gruplanarak sayıları geçici bir tabloya alınmıştır. Geçici tablo üzerindeki veriler en çok olan bilgiye göre tersten sıralanarak eşleştirme tablosuna eklenmiştir. Kategorik veri taşıma türü özniteliğine ait verilerin eşleştirme tablosunda toplanmasına ilişkin T-SQL sözdizimi örneği Şekil 4'de gösterilmiştir.

```
select bordervehiclemode, sayi=COUNT(*)
into #gec
from MLGumrukDosyaTasimaSekli group by bordervehiclemode

insert into gnlparm1 (modul, grup, kod, aciklama, updtime, upduser)
select 'ML','VEHMODE',bordervehiclemode,
ROW_NUMBER() OVER(ORDER BY sayi desc),GETDATE(),3676
from #gec order by sayi desc

drop table #gec
```

Şekil 4. Veri toplama sözdizimi örneği (Data collection syntax example)

3.6. Analiz (Analysis)

Çalışma kapsamında ele alınan her bir kategorik veri tür ve kayıt sayıları bakımından analiz edilerek öz nitelik sayılarına ulaşmaya çalışıldı. Kategorik temelde toplanan verilerin analiz ve değerlendirmesi sonucunda toplam 53 adet öznitelik belirlendi. Bu verilere ait belirlenen öznitelik sayıları Tablo 2'de gösterilmektedir

Tablo 2. Veri seti öznitelik sayıları (Dataset attribute counts)

Veri Seti	Öznitelik Adet
Teslim Şekli	9
Taşıma Türü	8
Sevk Ülke	10
Döviz Türü	6
Birim	7
Uluslararası Anlaşma	6
Ödeme Şekli	7

3.7. Kodlama (Coding)

Bu aşamada kullanılan yöntem; her bir kategorik veri üzerinde yer alan öznitelikler bazında kayıt sayıları tespit edilip, büyükten küçüğe sıralandıktan sonra kayıt sayısı en yüksek olana 1, ardından gelene 2, daha sonraki sayıya 3 olacak şekilde ardışık tamsayı kodlaması yapılarak eşleştirme tablosuna işlenir. Taşıma türü verilerine ait veri seti kod ve kayıt sayıları Tablo-3'de verilmektedir.

Tablo 3. Kodlama örneği (Coding example)

Taşıma Türü	Kod	Kayıt Sayısı
TIR	1	1471167
KAMYON	2	1057390
GEMI	3	833475
UCAK	4	617526
VAGON	5	79251
TREN	6	4240
RORO	7	1071
TEL	8	11

Taşıma türüne göre kayıt sayıları büyükten küçüğe sıralanmıştır. Sıralama sonucunda göre 1.471.167 adet veri ile en fazla kayıt sayısına sahip 'TIR' taşıma türü kod değerine (1) tamsayı değeri ile kodlanmıştır. KAMYON 1.057.390 adet veri ile ikinci büyük veriye sahip olup buna karşılık kod değeri (2), 833.475 adet veri ile üçüncü sırada GEMİ(3), UCAK(4), VAGON(5), TREN(6), RORO(7), elektrik hatları üzerinden gerçekleştirilen elektrik ithalat-ihracatı anlamına gelen TEL özniteliğine 8 ataması yapılmıştır. Benzer şekilde her bir kategorik veri üzerinden ayrı bir çalışma yapılarak ilgili kategorik verilerin özniteliklerine kod ataması yapılarak VT eşleştirme tablosuna kaydedilmiştir. Böylece her bir veri setinde yer alan öznitelikler farklı bir tamsayı ile kodlanmıştır.

3.8.Uygulama (Practice)

Uygulamada beyanname kategorik verileri ile eşleştirme tablosunda yer alan kodlama verileri anahtar alanlar üzerinden eşleştirildi. Böylece başka her hangi bir kodlama tekniğine ihtiyaç duyulmadan uygulamaların çalıştırılması mümkün oldu. Sayısallaştırılmış olan bu veriler VT eşleştirme tablosundan okunarak makine öğrenme algoritmaları üzerinde çalıştırıldı. Eğitim sonuçları ikili(binary) dosyada saklandı ve tahminleme parametresi olarak kullanıldı. Ele alınan kategorik verilerden biri olan taşıma şekli veri setindeki özniteliklere ait sayısal(kod) değerlerinin VT üzerinden okunmasına dair Python programında yazılan SQL sorgusuna ait sözdizimi örneği Şekil 5'de gösterilmektedir.

```
import pandas as pd
import pypyodbc
from sklearn.model_selection import train_test_split

connection=pypyodbc.connect('DRIVER=SQL SERVER;SERVER=SERVER;DATABASE=DBName;UID=UId;PWD=Password')
c = connection.cursor()
c.execute('''
          SELECT M.firmnumber
                , customsoffice =P1.EncodedCode
                , regime = P2.EncodedCode
                , declaration1 = P3.EncodedCode
                , declaration2 = P4.EncodedCode
                , entryoffice = P5.EncodedCode
                , internalvehiclemode = P6.EncodedCode
                , bordervehiclemode = P7.EncodedCode
                , bordervehiclecode
          FROM MLGumrukDosyaTasimaSekli M
          LEFT OUTER JOIN gnlparm1 P1 ON (P1.Modul = 'ML' AND P1.Type = 'CUSTOFFICE' AND P1.Code = M.customsoffice)
          LEFT OUTER JOIN gnlparm1 P2 ON (P2.Modul = 'ML' AND P2.Type = 'REGIME' AND P2.Code = M.regime)
          LEFT OUTER JOIN gnlparm1 P3 ON (P3.Modul = 'ML' AND P3.Type = 'DECL1' AND P3.Code = M.declaration1)
          LEFT OUTER JOIN gnlparm1 P4 ON (P4.Modul = 'ML' AND P4.Type = 'DECL2' AND P4.Code = M.declaration2)
          LEFT OUTER JOIN gnlparm1 P5 ON (P5.Modul = 'ML' AND P5.Type = 'CUSTOFFICE' AND P5.Code = M.entryoffice)
          LEFT OUTER JOIN gnlparm1 P6 ON (P6.Modul = 'ML' AND P6.Type = 'VEHMODE' AND P6.Code = M.internalvehiclemode)
          LEFT OUTER JOIN gnlparm1 P7 ON (P7.Modul = 'ML' AND P7.Type = 'VEHMODE' AND P7.Code = M.bordervehiclemode)
          ''')
```

Şekil 5. Veritabanı okuma sözdizimi (Database read syntax)

Çalışmada ele alınan kategorik verisi için benzer SQL söz dizim cümleleri yazılarak uygulanmıştır. Böylece öğrenme algoritmalarının kullanabileceği sayısal değerlere ulaşılmıştır. Kurulan bu altyapı üzerinden veri seti eğitim çalışmaları yürütülmüştür. Gaussian Naive Bayes(NB), karar ağaçları (CART), en yakın komşuluk(KNN), rastgele orman(RFC), ridge(RC), torbalama(BC), artırma(ABC) ve MLPC sınıflandırma algoritmaları kullanılarak taşıma şekli veri setinin eğitimine ilişkin Python sözdizimi örneği Şekil 6'da gösterilmektedir.

```
models = []
models.append(('NB', GaussianNB()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('RFC', RandomForestClassifier()))
models.append(('RC', RidgeClassifier()))
models.append(('BC', BaggingClassifier()))
models.append(('ABC', AdaBoostClassifier()))
models.append(('MLPC', MLPClassifier()))
test_size = 0.25
seed = 7
results = []
names = []
for name, model in models:
    X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=test_size, random_state=seed)
    model.fit(X_train, Y_train)
    result = model.score(X_test, Y_test)

    results.append(result)
    names.append(name)
    msg = "%s: %f" % (name, result)
    print(msg)
```

Şekil 6. Veri seti eğitimi sözdizimi (Dataset training syntax)

Eğitim çalışmalarında 'test_size' parametresi %25 olarak seçilmiştir. Seçilen test size oranı, farklı test size oranlarına göre en iyi sonucu vermiştir. Rastgele sayı üreteci değerinin verildiği 'seed' parametresi 7 seçilmiş ve yine yapılan denemelerde en iyi sonuç bu değer ile alınmıştır. Önerilen modelde herhangi bir kodlama tekniği kullanımına ihtiyaç duyulmamıştır. Eğitim

verilerinin ikili dosya olarak saklamasında Pickle kütüphanesinden yararlanılmıştır. İkili dosya saklama sözdizimi örneği Şekil 7'de gösterilmektedir.

```
dataframe = pd.DataFrame(c.fetchall(), columns = ['firmnumber', 'customsoffice', 'regime',
                                                  'declaration1', 'declaration2', 'entrysoffice',
                                                  'internalvehiclemode', 'bordervehiclemode', 'bordervehiclecode'])
array=dataframe.values
X = array[:,0:8]
Y = array[:,8]

model=BaggingClassifier()
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.25, random_state=7)
model.fit(X_train, Y_train)
filename='TasimaSekli_Model.pickle'
dump(model, open(filename, 'wb'))
```

Şekil 7. Dosya saklanma sözdizimi (File storage syntax)

Makine öğrenme algoritmaları tahmininde sayısal değerler kullanır. Çalışmada VT eşleştirme tablosundaki kategorik verilerin sayısal kod karşılıkları kullanılarak tahminleme gerçekleştirildi. Taşıma şekli kategorik verisinin torbalama sınıflandırma model tahminine ilişkin Python sözdizimi örneği Şekil 8'de gösterilmiştir

```
tasimasekli_model=BaggingClassifier()
filename='TasimaSekli_Model.pickle'
tasimasekli_model = load(open(filename, 'rb'));
predictedbordervehiclecode = tasimasekli_model.predict([[dfTT.get('firmnumber')[0],
                                                          dfTT.get('customsoffice')[0],
                                                          dfTT.get('regime')[0],
                                                          dfTT.get('declaration1')[0],
                                                          dfTT.get('declaration2')[0],
                                                          dfTT.get('entryoffice')[0],
                                                          dfTT.get('internalvehiclemode')[0],
                                                          dfTT.get('bordervehiclsmode')[0]]])[0]
if predictedbordervehiclecode != dfTT.get('bordervehiclecode')[0] :
    insertsql = '''INSERT INTO MLIslemSonuc (islemid, islemversiyon, islem, sirket, yil, departman,
                                             dosyano, girilendeger, tahmindeger, instime, insuser)
                   VALUES (?, ?, 5, ?, ?, ?, ?, ?, ?, GETDATE(), ?)
                '''
    c.execute(insertsql, [islemid, islemver, int(dfI.get('sirket')[0]),
                          int(dfI.get('yil')[0]), dfI.get('departman')[0],
                          int(dfI.get('dosyano')[0]), dfTT.get('bordervehiclecode')[0],
                          predictedbordervehiclecode, user])
    c.commit()
```

Şekil 8. Model tahmini sözdizimi (Model prediction syntax)

3.9.Test (Test)

Bu çalışmada önerilen yöntem ile etiket kodlama tekniği ayrı ayrı performans başarım testlerine tabii tutulmuştur. Bu testler aynı veri setleri ve çalışma koşullarında gerçekleştirilmiştir

3.10. Etiket kodlama tekniği (Label encoding)

Performans testlerinin kıyaslanması amacıyla kullanımı kolay ve bilinen etiket kodlama tekniği tercih edilmiştir. Bu teknik kategorik verileri alfabetik olarak sıralar ve ilk sıradaki bilgiye 0 değeri atayarak devam eder (Li, 2018). Çalışmada kullanılan taşıma türü veri seti etiket kodlama tekniğine uygun şekilde kodlandığında oluşan kod değer ve kayıt sayıları Tablo 4'te gösterilmektedir.

Tablo 4. Etiket kodlama örneği (Label encoding example)

Taşıma Türü	Kod	Kayıt Sayısı
GEMI	0	833475
KAMYON	1	1057390
RORO	2	1071
TEL	3	11
TIR	4	1471167
TREN	5	4240
UCAK	6	617526
VAGON	7	79251

Kullanılan veri setlerinden taşıma şekli veri setinde yer alan kategorik verilerin sayısallaştırılmasında kullanılan etiket kodlama tekniğine dair Python programı üzerinde taşıma şekli veri setinin kodlamasına ait sözdizimi örneği Şekil 9'da gösterilmektedir.

```
connection=pypyodbc.connect('DRIVER=SQL SERVER;SERVER=SERVER;DATABASE=DBName;UID=UId;PWD=Password')
c = connection.cursor()
c.execute('''
          SELECT firmnumber, customsoffice, regime, declaration1, declaration2,
                 entryoffice, internalvehiclemode, bordervehiclemode, bordervehiclecode
          FROM MLGumrukDosyaTasimaSekli
          ''')
dataframe = pd.DataFrame(c.fetchall(), columns = ['firmnumber', 'customsoffice', 'regime', 'declaration1'
                                                  'declaration2', 'entryoffice', 'internalvehiclemode',
                                                  'bordervehiclemode', 'bordervehiclecode'])
le = LabelEncoder()
dataframe.customsoffice=le.fit_transform(dataframe.customsoffice)
dataframe.regime=le.fit_transform(dataframe.regime)
dataframe.declaration1=le.fit_transform(dataframe.declaration1)
dataframe.declaration2=le.fit_transform(dataframe.declaration2)
dataframe.entryoffice=le.fit_transform(dataframe.entryoffice)
dataframe.internalvehiclemode=le.fit_transform(dataframe.internalvehiclemode)
dataframe.bordervehiclemode=le.fit_transform(dataframe.bordervehiclemode)
```

Şekil 9. Etiket kodlama sözdizimi (Label encoding syntax)

Önerilen model altyapısında sadece VT okuma süre maliyeti varken kodlama yöntemi kullanılması durumunda ek olarak kodlama süre maliyeti de söz konusudur.

4. Bulgular (Results)

Çalışmada, lojistik bir firmanın 2010-2021 yılları arasında üretilen gümrük beyannameleri üzerinde belirlenen 46 kategorik özniteliğe sahip yedi veri setini oluşturan 80154139 adet veri ile çalışılmıştır. Çalışma, sayısal olmayan verilerin öğrenme algoritmalarında kullanılıp, eğitilebilmesi için sayısal veriye dönüştürmede bilinen kodlama tekniklerinden etiket kodlama yöntemi ile bu tekniğin süre maliyetini iyileştirmek için geliştirilen ve bu çalışmada sunulan yöntemin başarım değerlerini görebilmek için test ve kıyaslama çalışması yürütülmüştür. Yapılan kıyaslamalardan birisi veri seti öznitelikleri bakımından olup, buna ilişkin sonuçlar Tablo 5'de gösterilmektedir. Veri setlerindeki kategorik öznitelik sayısı (n=46), sayısal öznitelik sayısı(n=7) bulunmuştur.

Tablo 5. Veri seti öznitelik sayıları (Dataset attribute counts)

Veri Seti	Sayısal Öznitelik(n)	Kategorik Öznitelik(n)
Teslim Şekli	1	8
Taşıma Türü	1	7
Sevk Ülke	1	9
Döviz Türü	1	5
Birim	1	6
Uluslararası Anlaşma	1	5
Ödeme Şekli	1	6
Toplam	7	46

Bu çalışmada önerilen yöntemle bilinen kodlama tekniklerinden etiket kodlama tekniği aynı çalışma platformları aynı veri setlerinin Tablo 5'de belirtilen öznitelikler üzerinden ayrı ayrı performans testine tabii tutulmuştur. Gerçekleştirilen testler sonucunda elde edilen bulguları iki ayrı yöntem üzerinden ele alındığında şu şekilde bulunmuştur.

4.1.Önerilen yöntem sonuçları (Suggested method results)

Bu çalışmada kodlama teknik kullanımına alternatif olarak sunulan yöntemin uygulanmasında her hangi bir kodlama tekniği kullanılmadığından kodlama süresine yönelik bir süre performans ölçümü söz konusu olmadığından bu yönde bir başarım testi gerçekleştirilmemiştir. VT okuma başlangıç ve bitiş sürelerinin tespitine yönelik çalışılan testlerden elde edilen bulgular Tablo 6' da sunulmaktadır. VT okuma sürelerine ait bulgular saat, dakika, saniye cinsindendir.

Tablo 6. Önerilen yöntem VT okuma süreleri (Suggested method DB reading times)

Veri Seti	Başlama zamanı	Bitiş zamanı
Teslim Şekli	13:37:32.671657	13:38:21.362509
Taşıma Türü	13:46:08.992893	13:46:54.534876
Sevk Ülke	13:52:09.026631	13:53:03.291126
Döviz Türü	13:56:59.626856	13:57:36.069051
Birim	14:01:59.033092	14:06:10.337286
Uluslararası A.	14:48:46.971833	14:52:05.053243
Ödeme Şekli	15:02:31.388658	15:05:47.188537

4.2. Etiket kodlama sonuçları (Label coding results)

Kategorik veriler etiket kodlama tekniği kullanılarak sayısallaştırılmıştır. Etiket kodlama yöntemin uygulanmasında süre performansını etkileyen iki temel unsur vardır. Birincisi seçilen kodlama tekniğinin kullanımından kaynaklı süre, ikincisi VT okuma süresidir. Bu yöntemde uygulama süre performans değeri ölçümlenirken bu iki işlemde geçen sürelerin toplamı dikkate alınır. Çalışılan veriler üzerinden etiket kodlama işlemine başlama ve bitiş süreleri saat, dakika, saniye cinsinden Tablo 7'de gösterilmektedir.

Tablo 7. Etiket kodlama süreleri (Label encoding times)

Veri Seti	Başlama zamanı	Bitiş zamanı
Teslim Şekli	13:43:53.993317	13:44:11.116580
Taşıma Türü	13:48:46.550163	13:49:00.272876
Sevk Ülke	13:55:27.136143	13:55:48.097437
Döviz Türü	13:59:45.200095	13:59:53.715373
Birim	14:28:26.834826	14:34:57.854834
Uluslararası A.	14:58:43.443704	15:00:03.465403
Ödeme Şekli	15:13:09.899093	15:14:17.992492

Etiket kodlama yöntemi uygulanılan bir makine öğrenme uygulamasında çalışma süresini etkileyen diğer unsur VT okuma süresidir. Bu çalışmada kullanılan veri seti öznitelikleri üzerinde etiket kodlama tekniği kullanılarak gerçekleştirilen testler sonucunda elde edilen VT okuma sürelerine ait bulgular Tablo 8'de verilmektedir. Tabloda yer alan VT okuma başlama-bitiş süreleri saat, dakika, saniye cinsindendir.

Tablo 8. Etiket kodlama tekniği VT okuma süreleri (Label encoding technique DB reading times)

Veri Seti	Başlama zamanı	Bitiş zamanı
Teslim Şekli	13:43:04.062545	13:43:53.992317
Taşıma Türü	13:48:01.683201	13:48:46.550163
Sevk Ülke	13:54:28.217643	13:55:27.136143
Döviz Türü	13:59:08.864947	13:59:45.200095
Birim	14:18:12.281741	14:28:26.792822
Uluslararası A.	14:53:07.702644	14:58:43.380701
Ödeme Şekli	15:07:04.156517	15:13:09.855097

5. Tartışma (Discussion)

Makine öğrenme uygulamalarında VT eşleştirme tablosuna işlenmiş olan ilgili kategorik veri özniteliklerin sayısal değerleri kullanılmıştır. Bu çalışmada kodlama tekniklerinin kullanımına alternatif olarak önerdiğimiz yöntemde herhangi bir kodlama tekniğine ihtiyaç duymadan kullanılan 46 özniteliğe sahip yedi veri setini oluşturan 80.154.139 adet veri üzerinden eğitim ve tahmin çalışmaları başarılı bir şekilde yürütülmüştür. Sayısallaştırma sürecinde en çok kullanılan kategorik veri için en düşük sayısal değer kullanımında daha yüksek oranda başarılı sonuç elde edildiği deneyimlenmiştir.

Önerdiğimiz yeni yaklaşımın başarım sonucunu görmek için aynı ortam ve veriler üzerinde etiket kodlama tekniği kullanılarak testler gerçekleştirilmiştir. Karşılaştırmada etiket kodlama tekniğinin seçilmesinde bu tekniğin bilinen, kullanımı kolay ve daha önce deneyimlediğimiz bir teknik olması etkendi. Karşılaştırma amacıyla yapılan test sonuçları VT okuma, kodlama ve birlikte değerlendirilmiştir.

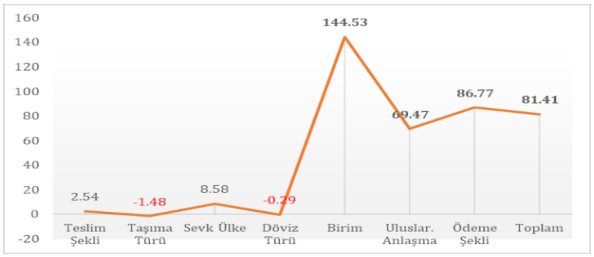
5.1. VT okuma süre kazançları (DB read time gains)

VT okuma süre kazançları bakımından elde edilen bulgular çalışmamızda önerdiğimiz yöntem ile bilinen etiket kodlama tekniği bakımından kıyaslanma sonuçları Tablo 9'da gösterilmektedir. VT okuma süre değerlerinin milisaniye cinsinden verilmiştir. Kazanç yüzdesi ise önerilen yöntemin bilinen etiket kodlama yöntemi karşısında elde ettiği performans kazanç yüzdesidir.

Tablo 9. VT okuma performans süreleri (DB reading times)

Veri Seti	Önerilen yöntem	Etiket kodlama	Kazanç (%)
Teslim Şekli	48691	49930	2,54
Taşıma Türü	45542	44867	-1,48
Sevk Ülke	54264	58919	8,58
Döviz Türü	36442	36335	-0,29
Birim	251304	614511	144,53
Uluslararası A.	198081	335678	69,47
Ödeme Şekli	195800	365699	86,77
Toplam	830124	1505939	81,41

VT okuma performans kazanç oranları incelendiğinde önerilen yöntemin taşıma türü(%1,48) ve döviz türünde(%0,29) oranında performans oranı düşük oranda da olsa negatif bir sonuç elde edilmiştir. Ancak bu verilere nazaran oldukça daha yüksek veriye sahip birim(%144,53), ödeme şekli(%86,77) ve uluslararası anlaşma(%69,47) veri setlerinde ise yüksek kazanç değerleri elde edilmiştir. Çalışma veri setinin tümü (n=80.154.139) üzerinden değerlendirmeye tabii tutulduğunda %81,41 gibi yüksek bir oranda başarım elde edilmiştir. Elde edilen VT okuma kazanç yüzdelerine ilişkin grafik Şekil 10'da gösterilmektedir.



Şekil 10.VT okuma kazanç yüzde grafiği (DB reading gain percentage graph)

VT okuma üzerine gerçekleştirilen zaman performans kazanç yüzdeleri incelendiğinde 'taşıma türü' ve 'döviz türü'ne ait veri setlerinde önerdiğimiz yöntemin diğer yönteme göre önemsiz sayılabilecek bir oranda da olsa başarısız olduğu görülmektedir. Ancak diğer veri setlerinde özellikle kayıt sayısı fazla olan veri setlerinde ciddi performans artışı sağladığı görülmüştür.

Önerdiğimiz yöntemde VT okuma sürecinde özniteliklerin sayısal değerlerini almak için ek olarak eşleştirme tablosuna bağlantı kurma maliyetiyle karşılaşılacak olsa da bunun uygulama çalışma süresi üzerinde önemli bir etkisinin olmayacağı yönündeki görüş kabul edilmiştir.

5.2. Kodlama süre maliyetleri (Encoding time costs)

Makine öğrenme uygulamasında bilinen bir kodlama tekniğinin kullanımında; verinin kodlanma işlemi uygulama esnasında gerçekleştirildiğinden bu kodlama işleminden kaynaklı bir süre maliyeti söz konusudur. Veri setleri öznitelikleri üzerinden etiket kodlama tekniği kullanarak gerçekleştirdiğimiz test sonucunda kodlamada geçen sürelerin milisaniye biriminde gösterimi Tablo 10'da gösterilmektedir.

Tablo 10. Kodlama başarım süreleri (Encoding success times)

Veri Seti	Önerilen yöntem	Etiket kodlama (ms)
Teslim Şekli	N/A	17123
Taşıma Türü	N/A	13723
Sevk Ülke	N/A	20961
Döviz Türü	N/A	8515
Birim	N/A	391020
Uluslararası A.	N/A	80022
Ödeme Şekli	N/A	68093

Uygulamada karşılaşılan bu kodlama süre maliyetlerinin aksine önerdiğimiz yöntemin uygulanması durumunda veri sayısallaştırma işlemi makine uygulamalarının çalışma esnasında değil uygulamalar çalıştırılmadan önce gerçekleştirilmekte ve VT eşleştirme tablosu üzerine kaydedilmektedir. Bu yüzden önerdiğimiz yöntemde makine öğrenme uygulamalarının çalıştırılması sırasında her hangi bir kodlama yapılmadığından süre maliyeti söz konusu değildir.

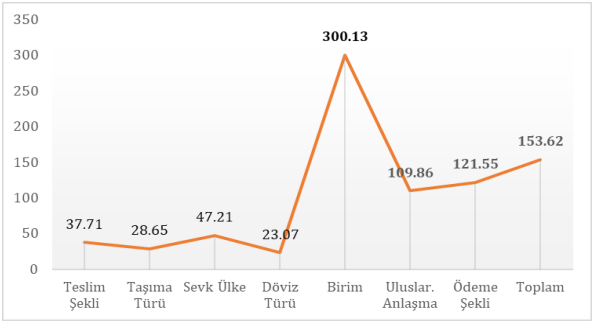
5.3. Uygulama toplam süre kazançları (Application total time earnings)

Bir makine öğrenme uygulaması bakımından yapılacak değerlendirmede toplam süre kazanç değerlendirmesinde bulunulur. Önerilen yöntemde etkili olan sadece VT okuma performans sonuçları dikkate alınırken her hangi bir kodlama yöntemi kullanılması durumunda VT okuma süresi yanında kodlama da geçen süre toplamın birlikte değerlendirilmesi gerekmektedir. Elde edilen uygulama süre performans kazanım sonuçları milisaniye biriminden Tablo 11'da gösterilmektedir.

Tablo 11. VT okuma performans süreleri (DB read performance times)

Veri Seti	Önerilen yöntem	Etiket kodlama	Kazanç (%)
Teslim Şekli	48691	67053	37,71
Taşıma Türü	45542	58590	28,65
Sevk Ülke	54264	79880	47,21
Döviz Türü	36442	44850	23,07
Birim	251304	1005531	300,13
Uluslararası A.	198081	415700	109,86
Ödeme Şekli	195800	433792	121,55
Toplam	830124	2105396	153,62

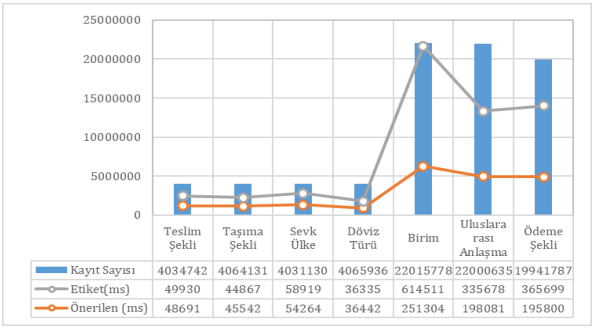
Elde edilen uygulama süre performans değerleri incelendiğinde veri seti bazında en düşük %23,07 oranla döviz türü veri kategorisinde, en yüksek oran ise birim öznitelik veri türünde %300,13 bulunmuştur. Uygulama bütününe etki eden toplam süre bakımında ele alındığında toplam süre başarım oranı %153,62 bulunmuştur. Uygulamanın bütünü üzerinden sağlanan süre performans gücü ve edinilen zaman kazancının ana nedeni uygulanan yöntemlerin farklılığından kaynaklanmaktadır. Önerdiğimiz ve başarılı bir şekilde uyguladığımız bu yeni yöntemde uygulama esnasında herhangi bir kodlama tekniği kullanılmamaktadır. Başarım testlerini yapıldığı etiket kodlama tekniğinden bağımsız uygulama esnasında kullanılacak her bir kodlama tekniğinin bir süre maliyeti olacaktır. Önerilen yöntemde ise veri sayısallaştırma işlemi uygulama öncesi VT üzerinde gerçekleştirilmektedir. Uygulama üzerinde çalışılan testlerden elde edilen süre değerleri üzerinden elde edilen süre kazanç başarım yüzde grafiği Şekil 11'de gösterilmektedir.



Şekil 11. Uygulama zamanı kazanç yüzde grafiği (Application time gain percent graph)

5.4. Süre performansına etki unsurlar (Factors affecting time performance)

Her iki yöntemde veri seti öznitelikleri üzerinden elde edilen çalışma süreleri veri öznitelik ve hacmi bakımından incelenmiştir. Bu amaçla kategorik veri sayısallaştırmasında, önerilen yöntem ile etiket kodlama yönteminin bu veri setleri üzerindeki çalışma sürelerinin milisaniye biriminden elde edilen bulgular üzerinden oluşturulan tablo ve grafik Şekil 12'de gösterilmiştir.



Şekil 12. Veri seti adet, yöntem ve çalışma süreleri (Data set quantity, method and run times)

Çalışılan testlerden elde edilen bulgular değerlendirildiğinde kategorik veri özniteliklerindeki kayıt sayılarındaki artışın çalışma süresi üzerinde etkili olduğu, kullanılan yöntem ve teknik açısından ise bir birlerine benzer paralellikte artış ve azalış gösterdiği gözlemlenmektedir. Ancak veri seti hacmi arttıkça önerdiğimiz yöntemin daha az sürede daha başarılı performans gösterdiği tespit edilmiştir. Birim ve uluslararası anlaşma kategorilerinde gözlenen benzer veri adetlerine sahip iki veri setindeki belirgin olan süre ayrışmasının nedeni irdelendiğinde ilgili veri setlerindeki öznitelik sayısı ve veri niteliğinin etkili olduğu saptanmıştır. Diğer bir taraftan birim veri seti öznitelik sayısı bakımından en az kayıt sayısı ve performans artışı (%300,13) bakımından en yüksektir. Yine kayıt sayısı fazla olan birim, uluslararası anlaşma ve ödeme şekli veri setlerindeki performans artışı dikkat çekici derecede yüksek bulunmuştur. Uygulama süre performans artışının öznitelik sayısı ile ters, kayıt sayısıyla doğru orantılı bir ilişkiye sahip olduğu görülmüştür.

Çalışma bulgularımızı diğer araştırma sonuçlarıyla değerlendirmek istediğimizde makine öğrenmesine dayalı güncel pek çok çalışma olmasına rağmen önerdiğimiz yöntem ve bunun sağladığı süre performans kazancına dair her hangi bir çalışmaya rastlayamadık. Diğer taraftan makine öğrenme algoritmalarına dayalı uygulamaların başarım ve performansında veri seti büyüklüğünün etkisinin yer aldığı (Tekin & Tunalı, 2019) tarafından yapılan çalışmada, veri seti hacminin sınıflandırma başarımı üzerinde etkili olduğu, veri setinin küçük olmasının başarımı düşürdüğü vurgulanmıştır. (Calp ve Akcayol, 2020), risk tahminine yönelik yapay sinir ağları ile geliştirdikleri internet tabanlı uygulamanın başarım test çalışma sonucuna göre farklı alanlarda daha fazla veri sayısı kullanımının daha yüksek doğruluk düzeyinde performanslı sonuçlara ulaşmışlardır. Bu çalışmalar birlikte değerlendirdiğinde makine öğrenme algoritmalarının kullanıldığı uygulamalarda veri hacminin sınıflandırma başarısı üzerinde olumlu bir etkiye sahip olduğuna ulaşabiliriz. (Takçı, 2018), kalp krizinin önceden tahminine yönelik yaptığı çalışmada makine öğrenme algoritmalarının uygulanmasında öznitelik seçiminin işlem süresi üzerinde etkili olduğunu vurgulamıştır. (Bilgin ve Oğuz, 2021) gerçekleştirdikleri çalışmada veri sınıflandırma ve kümelenmesinden kaynaklı bellek tüketimini azaltmaya yönelik yeni bir yöntem önermişlerdir. (Ma ve Zhang, 2020), Londra seyahat veri setini kullanarak seyahat tercih seçimini tahmin etmeye yönelik çalışmada kategorik özelliklerin daha iyi modellenme ve anlaşılmasında kodlamanın önemine vurgu yaparak Label kodlama ve One-Hot gibi kodlama tekniklerinin başarısının sınırlı kalabileceğine dikkat çekmişlerdir. Tüm bu çalışmalar veri hacmi, veri seti özniteliği, kodlama tekniği ve sınıflandırma yöntemi, çalışma ortam ve altyapısının uygulama performans ve yetkinliği üzerinde etkili olduğunu göstermektedir. Çalışmamızda sunduğumuz alternatif çalışma yönteminde her hangi bir kodlama tekniğinin kullanılmaması, veri seti özniteliği seçimi ve VT alt yapısı sayesinde diğer çalışmalarda belirtilen kazançlara ek kazanç elde edilmiştir.

Özetle makine öğrenimi uygulamalarında kullanılan kodlama tekniklerine alternatif olarak çalışılan ve başarılı bulunan bu çalışma yöntem ve kurulan altyapı sayesinde veriler daha okuma aşamasındayken sayısal olarak alınmaktadır. Çalışmamızda sunmuş olduğumuz yöntem veri okuma süresini daha efektif hale getirmiştir. Yine önerdiğimiz bu yöntemde herhangi bir kodlama tekniğinin kullanılmamış olması kodlama tekniği kullanan yaklaşımlara göre ciddi süre performans kazancı sağlamıştır. Çalışma, yüksek performans kazancı ve özgün yapısı ile benzer veri seti ve çalışma alanlarında kolay bir şekilde uygulanabilecek benzersiz bir yapıya sahiptir. Makine öğrenmesinde kodlama tekniklerinin kullanımına alternatif ve etkili bir çözüm geliştirilerek literatüre kazandırılmıştır.

6. Sonuç (Conclusions)

Bir makine öğrenme uygulamasında sayısal olmayan verilerin öğrenme algoritmalarında kullanılabilmesi için sayısal veriye dönüşmesi gerekmektedir. Bu amaçla bir takım kodlama yöntemleri kullanılmaktadır. Bu bilinen yöntemlerin kullanımının bazı kolaylıklar yanında uygulama üzerinde belli bir süre maliyeti vardır. Günümüzde neredeyse her alanda zamanla adeta yarışıldığı bir dönemde en küçük bir zaman kaybı ciddi derecede olumsuz bir etkiye sahip probleme dönüşebilmektedir. Bu çalışma, uluslararası faaliyet gösteren, Türkiye iç piyasanın yaklaşık %8 işlem hacmine sahip bir lojistik bir firmada yürütülen makine öğrenmesi uygulama çalışmalarında rastlanan süre probleminin iyileştirilmesine yönelik yapılan bir araştırma sonucunda üretilmiştir. 2010-2021 yılları arasında üretilen gümrük beyannameleri üzerinde belirlenen 46 özniteliğe sahip yedi veri setini oluşturan 80.154.139 adet veri ile çalışılmıştır. Sunduğumuz bu alternatif çalışma önerisi ile bilinen kodlama tekniklerinden etiket kodlama tekniği aynı çalışma ortam ve veriler üzerinden ayrı ayrı süre performans testlerine tabii tutulmuştur. Sonuç olarak, makine öğrenme uygulamalarında kullanılan kodlama yöntemine alternatif bir çalışma modeli çalışılmıştır. Geliştirilen bu yeni yaklaşımla bilinen etiket kodlama tekniği uygulama süre performans düzeyleri karşılaştırmalı olarak incelenmiştir. Bu süre başarım test sonuçlarına göre alternatif olarak önerdiğimiz yöntemin bilinen kodlama tekniğinin kullanımına karşı kategorik veri bazında en az %23,07 en fazla %300,13 oranında başarı göstermiştir. Uygulama bütünü üzerinden yapılan değerlendirmelerde daha büyük veri setinde daha yüksek başarım oranlarına ulaşılmıştır. Uygulama bütünü üzerinden yapılan değerlendirmelerde daha büyük veri setinde daha yüksek başarım oranlarına (%153,62) ulaşılmıştır.

Bu çalışma, sahada karşılaşılan gerçek bir problemin çözümüne yönelik araştırma sonucunda geliştirilen, uygulanabilir yapısı, elde edilen kazanç ve başarılı sonuçların varlığı, elde edilen deneyim ve tecrübelerin aktarılması, makine öğrenim çalışmalarında kodlama tekniği kullanımına alternatif bir yöntem önerisi getirmesi ve bu alanın gelişimine katkı sunacak nitelikte olması bakımından kıymetlidir. Çalışma, elde edilen yüksek orandaki süre performans kazancı yanında kolay sayılabilen kullanımıyla benzer saha çalışmalarında uygulanabilir nitelikte olup kodlama tekniklerinin kullanımına alternatif bir çözüm önerisi sunmaktadır.

Kaynaklar (References)

Al-Shehari T., Alsowail R. A., 2021. An insider data leakage detection using one-hot encoding, synthetic minority oversampling and machine learning techniques. Entropy, 23(10), 1258, doi:10.3390/e23101258

Bilgin, T., Oğuz, M., 2021. A new approach to minimize memory requirements of frequent subgraph mining algorithms. Politeknik Dergisi, 24(1), 237-246

Calp, M., Akcayol, M., 2020. Design and Implementation of Web Based Risk Management System Based on Artificial Neural Networks for Software Projects: WEBRISKIT. Pamukkale Univ Muh Bilim Derg., 26(5), 993-1014

Chakrabarty, N., 2019. A data mining approach to flight arrival delay prediction for american airlines. 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON). doi:10.1109/iemeconx.2019.8876970

Cerda, P., Varoquaux, G., Kégl, B., 2018. Similarity encoding for learning with dirty categorical variables. Machine Learning, 107(8-10), 1477–1494. doi:10.1007/s10994-018-5724-2

Chandradeva, L. S., Jayasooriya, I., Aponso, A. C., 2019. Fraud Detection Solution for Monetary Transactions with Autoencoders. National Information Technology Conference(NITC). doi:10.1109/nitc48475.2019.9114519

Chen, L., Xian, M., Liu, J., & Wang, H., 2020. Intrusion detection system in cloud computing environment. International Conference on Computer Communication and Network Security (CCNS). doi:10.1109/ccns50731.2020.00037

Famili, A., Shen, W.-M., Weber, R., Simoudis, E., 1997. Data preprocessing and ıntelligent data analysis. Intelligent Data Analysis, 1(1), 3–23. doi:10.3233/ida-1997-1102

Günerkan M., Şahinaslan E., Şahinaslan Ö., 2022. Gümrük beyannamesi sürecinde öğrenmeye dayalı algoritmaların etkinliğinin incelenmesi. Acta Infologica, doi: 10.26650/acin.1057060

Jackson, E., & Agrawal, R., 2019. Performance evaluation of different feature encoding schemes on cybersecurity logs. IEEE, 1-9. doi:10.1109/southeastcon42311.2019.9020560

Jiang, D., Lin, W., Raghavan, N., 2020. A novel framework for semiconductor manufacturing final test yield classification using machine learning techniques. IEEE 197885–197895. doi:10.1109/access.2020.3034680

Karasulu, B., Yücalar, F., Borandag, E., 2022. İnsan kulağı görüntüleri kullanarak cinsiyet tanıma için derin öğrenme tabanlı melez bir yaklaşım. Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, 37 (3) , 1579-1594 . doi: 10.17341/gazimmfd.945188

Kıran, E. , Karasulu, B. & Borandag, E. (2022). Gemi Çeşitlerinin Derin Öğrenme Tabanlı Sınıflandırılmasında Farklı Ölçeklerdeki Görüntülerin Kullanımı . Journal of Intelligent Systems: Theory and Applications , 5 (2) , 161-167 . DOI: 10.38016/jista.1118740

Li, J., 2018. Monthly housing rent forecast based on lightgbm (light gradient boosting) model. International Journal of Intelligent Information and Management Science, 7(6). http://www.hknccp.org/Public/upload/goods/2019/09-03/5d6e145f40393.pdf

Li, Y., Zhu, Z., Wu, H., Ding, S., & Zhao, Y., 2020. CCAE: Cross-field categorical attributes embedding for cancer clinical endpoint prediction. Artificial Intelligence in Medicine, 107, doi:10.1016/j.artmed.2020.101915

MarketResearch., 2022. Types of data & measurement scales: nominal, ordinal, ınterval, and ratio. "https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio", 13.05.2022

Ma, Y., Zhang, Z. 2020. Travel mode choice prediction using deep neural networks with entity embeddings. IEEE, 8, 64959-64970, doi: 10.1109/access.2020.2985542.

Mitchell, T. M., 1997. Machine learning. New York: McGraw-Hill

Nerlikar, P., Pandey, S., Sharma, S., Bagade, S., 2020. Analysis of intrusion detection using machine learning techniques. International Journal of Computer Networks and Communications Security, 8(10), 84-93

Potdar, K., Pardawala, T.S., Pai, C.D., 2017. A comparative study of categorical variable encoding techniques for neural network classifiers. International journal of computer applications, 175(4), 7-9. doi:10.1207/s15328031us0301_3

Reilly, D., Taylor, M., Fergus, P., Chalmers, C., Thompson, S., 2022. The categorical data conundrum: Heuristics for classification problems - A case study on domestic fire injuries. IEEE Access, 10, 70113-70125.

Sharma, N., Bhandari, H.V., Yadav, N.S., Shroff, H.V.J., 2020. Optimization of IDS using filter-based feature selection and machine learning algorithms". Int. J. Innov. Technol. Explor. Eng, 10(2), 96-102.

SAS., 2022. Makine Öğrenimi Nedir ve Neden Önemlidir, "https://www.sas.com/tr_tr/insights/analytics/machine-learning.html ", 15.06.2022

Scikit-Learn., 2022. sklearn.preprocessing.LabelEncoder. scikit-learn:https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html, 13.05.2022

ScikitLearn-OneHotEncoder., 2022. One Hot Encoder "https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html#sklearn.preprocessing.OneHotEncoder", 13.05.2022

ScikitLearn-OrdinalEncoder., 2022. Ordinal Encoder. "https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OrdinalEncoder.html#sklearn.preprocessing.OrdinalEncoder", 13.05.2022

Seger, C., 2018. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing. "https://www.diva-portal.org/smash/get/diva2:1259073/Fulltext01.pdf"

Sethi, A., 2022. Categorical encoding | one hot encoding vs label encoding. "https://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn", 13.05.2022

Shen, J., Shafiq, M. O., 2019. Learning mobile application usage - A deep learning approach. 18th IEEE International Conference On Machine Learning And Applications (ICMLA). doi:10.1109/icmla.2019.00054

Şahinaslan, Ö., Dalyan, H., Şahinaslan, E., 2022. Naive bayes sınıflandırıcısı kullanılarak youtube verileri üzerinden çok dilli duygu analizi. Bilişim Teknolojileri Dergisi, 15(2), 221-229. doi: 10.17671/gazibtd.999960

Takçı, H., 2018. Improvement of heart attack prediction by the feature selection methods, Turkish Journal of Electrical Engineering and Computer Science, 26 (1), 1-10

Tekin, M., Tunalı, V., 2019. Prioritization of software development demands with text mining techniques. Pamukkale Univ Muh Bilim Derg., 25(5), 615-620

Turcanik, M., Javurek, M., 2016. Hash function generation by neural network. 1-5. 10.1109/NTSP.2016.7747793

Yılmaz Yalçıner, A., Gelen Mert, M.B., 2021. Estimating the occupancy rate of an accommodation business using artificial neural networks . Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü Dergisi , (47) , 209-218 . doi: 10.30794/pausbed.828902

Yu, L., Zhou, R., Chen, R., Lai, K. K., 2020. Missing data preprocessing in credit classification: one-hot encoding or imputation? Emerging Markets Finance and Trade, 1–11. doi:10.1080/1540496x.2020.1825935