OKU Fen Bilimleri Enstitüsü Dergisi
5(2): 753-766, 2022

OKU Journal of The Institute of Science and Technology, 5(2): 753-766, 2022

Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi

Osmaniye Korkut Ata University Journal of The Institute of Science and Technology

Bireylerin Gelir Dağılım Seviyelerinin Makine Öğrenmesi Teknikleri ile Belirlenmesi

Sait Taner CANİBEY[1]*, Onur SEVLİ[2]

[1] Burdur Mehmet Akif Ersoy Üniversitesi, Mühendislik Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Burdur, Türkiye
[2] Burdur Mehmet Akif Ersoy Üniversitesi, Mühendislik Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Burdur, Türkiye

[1]https://orcid.org/0000-0002-5781-6372
[2]https://orcid.org/0000-0002-8933-8395
*Sorumlu yazar: stcanibey@gmail.com

Araştırma Makalesi

Makale Tarihçesi:
Geliş tarihi: 22.12.2021
Kabul tarihi: 20.01.2022
Online Yayınlanma: 18.07.2022

Anahtar Kelimeler:
Gelir seviyesi tahminlemesi
Makine öğrenmesi
Ekonomik yaşam

ÖZ

Toplumların ekonomik durumlarını değerlendirme üzerine yapılan çalışmalar mevcut durumu tespit etmek, yaşam koşullarını iyileştirmek ve geleceğe dönük stratejiler geliştirmek açısından son derece önemlidir. Bu konuda farklı disiplinler çeşitli araştırma, analiz ve tahminleme çalışmaları gerçekleştirmektedir. Bu çalışmada farklı ülkelerden toplanmış olan sosyoekonomik verilerine bağlı olarak bireylerin ekonomik seviyeleri üzerine makine öğrenmesi temelli tahminlemeler gerçekleştirilmiştir. Kullanılan veri setinde bireylerin yaş, çalışma durumu, eğitim seviyesi, medeni durumu, mesleği, ırkı, cinsiyeti, haftalık çalışma süresi ve gelir seviyesini gösterir sınıf yer almaktadır. KNN, DVM, Rastgele Orman ve Naive Bayes algoritmaları kullanılarak elde edilen ölçümler farklı metrikler açısından değerlendirilmiş ve karşılaştırılmıştır. En iyi doğruluk değeri %97.36 olarak Naive Bayes algoritması ile elde edilmiştir. Bu çalışma sosyoekonomik tahminlemeler konusunda çalışacak olan araştırmacılar için makine öğrenmesi temelli başarılı bir model örneği sunmaktadır.

Determining the Income Distribution Levels of Individuals with Machine Learning Techniques

Research Article

Article History:
Received: 22.12.2021
Accepted: 20.01.2022
Published online: 18.07.2022

Keywords:
Income level
Machine learning
Economic life

ABSTRACT

Studies on the evaluation of the economic situation of societies are extremely important in terms of determining the current situation, improving living conditions and developing strategies for the future. Different disciplines carry out various research, analysis and forecasting studies on this subject. In this study, machine learning-based predictions were made regarding the economic levels of individuals based on the socioeconomic data collected from different countries. The data set used includes the age, employment status, education level, marital status, occupation, race, gender, weekly working time, and income level of the individuals. On this data set, estimations were made on the income level of individuals by using different machine learning algorithms. The measurements obtained using KNN, SVM, Random Forest and Naive Bayes algorithms were evaluated and compared in terms of different metrics. The best accuracy value was obtained with the Naive Bayes algorithm as 97.36%. This study provides an example of a successful model based on machine learning for researchers who will work on socioeconomic forecasting.

To Cite: Canibey ST., Sevli O. Bireylerin Gelir Dağılım Seviyelerinin Makine Öğrenmesi Teknikleri ile Belirlenmesi. Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi 2022; 5(2): 753-766.

Giriş

Küreselleşen dünya düzeninde her geçen gün ekonomik şartlar gerek ülkeler bazında, gerekse bireyler bazında zorlaşmaktadır. Zorlaşan bu şartlar gelir seviyelerine göre toplumsal sınıfların aralarındaki uçurumları artırmaktadır. Küresel olarak 1990'lı yılların ekonomileri göz önünde alındığında gelir seviyelerine göre toplumu oluşturan düşük, orta ve yüksek sınıflar daha dengeli iken, 2010 yılı ve sonrasında orta seviye giderek daralmış ve düşük gelir seviyesi ile yüksek gelir seviyesi arasındaki makas giderek artmıştır. Özellikle son yıllarda yaşanan pandemi nedeniyle ekonomik koşulların iyi yönetilememesi sonucu orta seviye neredeyse yok olmuş, dağılımdaki eşitsizlik daha da büyümüştür.

Ülkelerin ekonomik durumlarını değerlendirmek sürdürülebilirliği sağlamak, yaşam koşullarını iyileştirmek, geleceğe dönük stratejiler oluşturmak açısından son derece önemlidir. Buna yönelik olarak ekonomi temelinde farklı disiplinlerin de iş birliği ile çok sayıda çalışma yürütülmektedir. Bilgisayar bilimleri içerisinde gelişen yapay zekâ teknikleri farklı alanlarda olduğu gibi ekonomi konusunda gerçekleştirilen çalışmalara da farklı bir boyut kazandırmaktadır. Bir yapay zekâ disiplini olan makine öğrenmesinin, tahminleme çalışmalarında kullanımı giderek artmaktadır.

Makine öğrenmesi, bilgisayarların belirli bir amaç doğrultusunda toplanan veri kümeleri üzerinde gerekli mantıksal, matematiksel, istatistiki işlemleri gerçekleştirerek, veriler arasındaki örüntüleri ortaya çıkarıp karşılaşılan yeni durumlar için başarılı tahminler yapabilmelerini sağlamak için kullanılan tekniklerdir (Öztemel, 2012).

Bireylerin ve ülkelerin sosyoekonomik durumları belirlenirken çeşitli yollarla doğrudan bireylerden elde edilen bilgiler kullanılır. Bu veriler geleneksel istatistiki yöntemlerle analiz edilebileceği gibi, bilgisayar destekli tahminleme sistemleri ile ele alınarak, hem mevcut durum değerlendirmesi yapılıp, hem de geleceğe dönük çıkarımlar yapılabilmektedir.

Gelir seviyesi tespiti yapan firmalar farklı çıkarımlar yaparken, bankalar kredi notunun hesaplanmasında farklı veri setleri üzerinden tahminlemeler yapmaktadır. Bu tahminlemeler üzerinden toplumsal ve bireysel durumları inceleyerek geleceğe dönük stratejiler oluşturmaktadır.

Literatürde yapılan çalışmaların çoğu toplanan veriler ya da işlem kayıtları üzerinden bireylerin ya da toplumun mevcut durumunu betimlemeye yönelik çalışmalardır. Bu çalışmalarda toplumun gelir seviyesi, banka hesap durumları (Ardic ve ark., 2011), harcama kalemleri, öğrenim durumunun ekonomiye etkisi (Doğrul, 2009) gelir dağılımı eşitsizliğinin toplumsal refaha ve çalışmaya başlama yaşındaki düşüşe etkisi (Karaman ve Özçalık, 2007) gibi durumlara dair çıkarımlarda bulunulmuştur.

Literatürdeki mevcut çalışmalar incelendiğinde bu alanda yapay zekâ destekli tahminleme çalışmalarının henüz yeteri kadar çok olmadığı görülmektedir. Bu çalışma ile benzer veri setini kullanan bir çalışmada Meng (1994), bireylerin gelir durumlarının 50 bin dolar üzerinde olup olmama durumuna göre karar ağaçlarına dayalı bir tahminleme çalışması gerçekleştirmişlerdir (Meng,1994). Gerçekleştirilen sınıflandırmada hata oranını %37,21 olarak bulmuşlardır.

Küreselleşen dünya düzeninde gelir seviyesi eşitsizliği özellikle ABD gibi kalabalık nüfusa sahip ülkelerde ekonomik kriz endişeleri yaratmaktadır. Ülkeler bireylerin gelir seviyesi eşitsizliklerini

azaltmak için farklı tedbirler almaya çalışmaktadırlar. Yoksulluğu ortadan kaldırmak, gelir seviyesi eşitsizliği için alınması gereken en önemli tedbirlerden biridir. Bu sorundan hareketle Chakrabarty ve Biswas (2018), makine öğrenmesi tekniklerinden yararlanarak bireylerin gelir seviyesi üzerinde bir tahminleme çalışması gerçekleştirmişlerdir. Gradient Boosting Classifier kullanarak gerçekleştirdikleri sınıflandırmalarda %88.16 doğruluk değerine ulaşmışlardır (Chakrabarty ve Biswas, 2018).

Bu çalışmada nüfus ve gelir durumu bilgilerinden oluşan bir veri seti üzerinden bireylerin gelir durumlarının tespitine yönelik bir sınıflandırma ve tahminleme çalışması gerçekleştirilmiştir. Dört farklı makine öğrenmesi algoritması ile gerçekleştirilen çalışmanın bulguları farklı metrikler ile raporlanmıştır. Bu çalışma ile ekonomik durum ve gelir seviyesi tahminleme konusunda çalışmalar yapan araştırmacılara bilgisayar destekli modeller ve makine öğrenmesi teknikleri ile gerçekleştirilebilecek çalışmalar hakkında yol gösterici bir örneğin oluşturulması amaçlanmıştır.

Materyal ve Metot

Veri Seti

Çalışmada kullanılan veri seti Backer tarafından oluşturulmuş olup, ülkelere göre nüfus ve gelir durumu bilgilerini içermektedir. Veri seti University of California Irvine (UCI) veri deposunda kamuya açık şekilde paylaşılmaktadır (UCI, 2021). Toplam 15 öznitelikten oluşan veri setinde 32561 örnek bulunmaktadır. Özniteliklerden 14 adedi bireyin çeşitli özelliklerini gösterir girdiler olup, hedef öznitelik ise bireyin kazan durumudur. Kazancın 50 bin doları geçip geçmeme durumuna göre iki farklı sınıf mevcuttur.

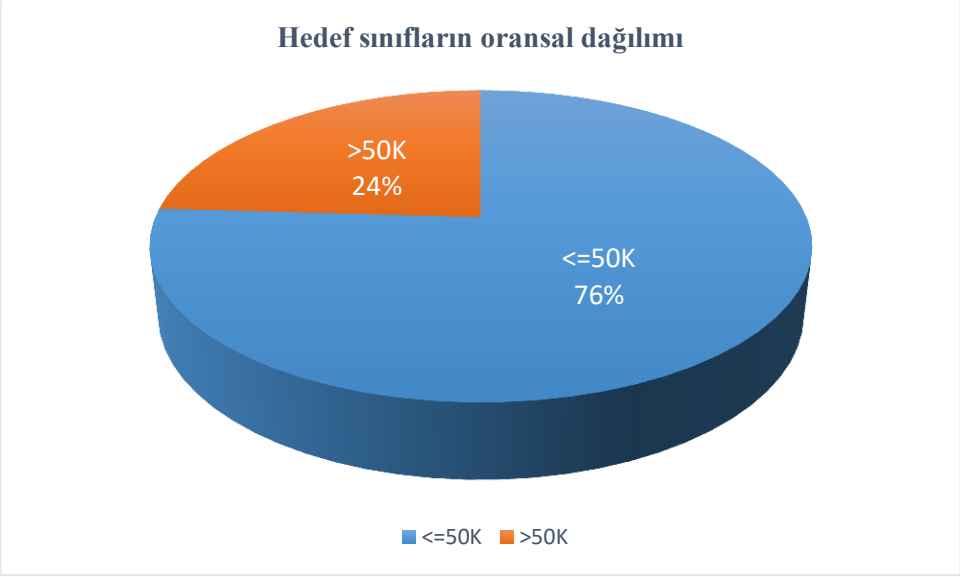
Veri setinde yer alan öznitelikler ve açıklamaları Tablo 1'de verilmiştir.

Tablo 1. Veri seti içeriklerin detayları

Öznitelik	Açıklama	Türü
Yaş(Age)	Bireyin yaşı.	Sayısal-sürekli
Çalışma durumu (Workclass)	Bireyin çalışma durumu. "Self-emp-not-inc, Federal-gov, Self-emp-inc, State-gov, Without-pay, Local-gov ve Never-worked" şeklinde çalışma sınıflarını içerir.	Kategorik
Son ağırlık (Fnlwgt)	Sayım bürosu tarafından atanan ağırlık değeridir. Farklı demografik yapıya sahip ırkları ayırmak için belirlenmiştir.	Sayısal-sürekli
Eğitim durumu (Education)	Eğitim durumu. "Some-college, Bachelors, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9^{th}, 7^{th}-8^{th}, 12^{th}, 1^{st}-4^{th}, 10^{th}, Doctorate, Masters ve 5^{th}-6^{th}, Preschool" şeklinde değerler alır.	Kategorik
Eğitim gördüğü süre (Education-num)	Eğitim süresi.	Sayısal-sürekli
Medeni durum (Marital-status)	Birerylerin medeni durumunu. "Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-civ-spouse ve Married-AF-spouse" değerlerini içerir.	Kategorik

Meslek(Occupation)	Meslek bilgisi. “Craft-repair, Other-service, Tech-support, Sales, Exec-managerial, Prof-specialty, Machine-op-inspct, Adm-clerical, Farming-fishing, Priv-house-serv, Handlers-cleaners, Protective-serv, Transport-moving ve Armed-Forces.” değerlerini içerir.	Kategorik
İlişiki durumu(Relationship)	İlişki durumu. “Husband, Wife, Not-in-family, Other-relative, Own-child ve Unmarried.” değerlerini içerir.	Kategorik
Irk(Race)	Irk bilgisi. “White, Amer-Indian-Eskimo, Other, Asian-Pac-Islander ve Black.” değerlerini içerir.	Kategorik
Cinsiyet(Sex)	Cinsiyet bilgisi. “Female, Male” değerlerini içerir.	Kategorik
Sermaye kazancı(Capital-gain)	Sermaye kazancı.	Sayısal-sürekli
Sermaye kaybı(Capital-loss)	Sermaye kaybı.	Sayısal-sürekli
Haftalık çalışma saati(Hours-per-week)	Haftalık çalışma süresi.	Sayısal-sürekli
Ülke(Native-country)	Bireylerin ülkeleri. “Trinadad&Tobago, Cambodia, England, Puerto-Rico, Canada, United-States, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Mexico, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Yugoslavia, El-Salvador, Peru, Thailand, Hong ve Holand-Netherlands” değerlerini içerir.	Kategorik
Maaş(Salary)	Yıllık gelir seviyesinin 50 bin doların altında veya üstünde olma durumu. (>50K, <=50K). Tahminlenecek hedef değişken budur.	Kategorik

Tahminlenecek hedef öznitelik Salary alanıdır. Bu alanın 50 bin dolardan büyük veya küçük olma durumuna bağlı olarak sahip olduğu iki farklı sınıfın oransal dağılımı Şekil 1’de verilmiştir.



Şekil 1. Salary özniteliği sınıf dağılımı

Veri setinde yer alan sayısal özelliklerin istatistiki karakteristikleri ise Tablo 2'de verilmiştir.

Tablo 2. Verilerin istatistiki karakteristiği

	Yaş	**Eğitim süresi**	**Sermaye kazancı**	**Sermaye kaybı**	**Haftalık çalışma saati**
veri sayısı	32561	32561	32561	32561	32561
ortalama	38.58	10.08	1077648.84	8730383	40,44
standart sapma	13.64	2.57	7385292.09	402960219	12,35
en küçük	17	1	0	0	1
%25	28	9	0	0	40
%50	37	10	0	0	40
%75	48	12	0	0	45
en büyük değer	90	16	99999	4356	99

Kullanılan Sınıflandırma Yöntemleri

Bu çalışmada bireylerin gelir durumlarını tespit etmek için K-En Yakın Komşuluk (KNN), Destek Vektör Makinesi (DVM), Rastgele Orman (RO) ve Naive Bayes (NB) olmak üzere 4 farklı sınıflandırma algoritması kullanılmıştır. Sınıflandırma işlemlerinde 5 kat çapraz doğrulama uygulanarak elde edilen sonuçların ortalamaları raporlanmıştır.

KNN, sınıflandırma problemlerinin çözümünde yaygın olarak kullanılan, uzaklığa dayalı bir makine öğrenmesi algoritmasıdır (Aksoy, 2021). Öğrenme işlemi ve test işlemleri olarak ayılan veri seti içinden öğrenme gerçekleştirilir. Öğrenme işlemi kümenin elemanları arasındaki k adet veriyi ve bunlar arasındaki uzaklıkların değerlerini hesaplayarak, en yakın olan sınıfa yeni üyenin dahil edilmesini sağlar (Pala ve ark., 2019).

Çalışmada kullanılan diğer bir algoritma olan DVM, doğrusal olarak ayrıştırılabilen iki farklı sınıfın üyelerinin destek vektörleri olarak bilinen karar düzlemi ile sınıf sınırını tanımlayan, örnekler arasındaki maksimum uzaklığın tespit edilmesi ilkesi üzerine kurulu bir sınıflandırma algoritmasıdır (Akben ve ark., 2010).

Kullanılan diğer algoritma olan RO, 2001 yılında Leo Breiman tarafından ortaya çıkarılmış bir modeldir. RO, veri seti ve öznitelikleri çok sayıda parçaya ayrılarak birden fazla ağaç üzerinde işlenerek çözüme ulaşır (Breiman, 2001).

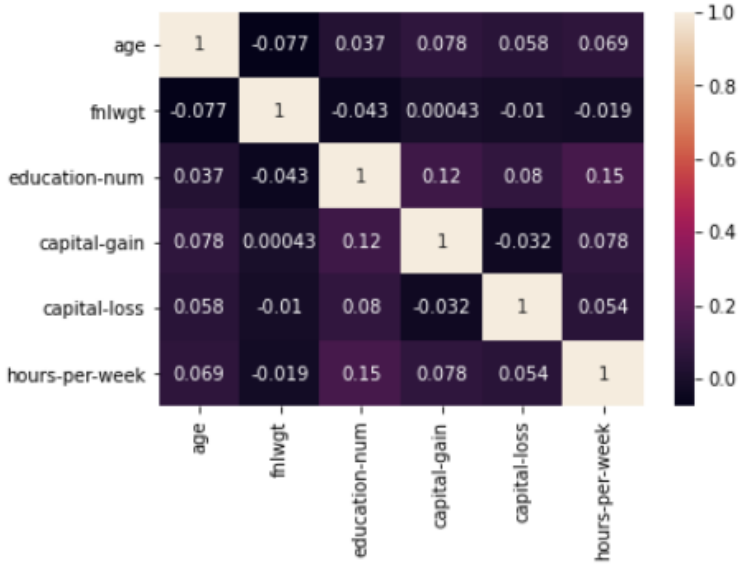
Çalışmada kullanılan diğer bir yöntem olan NB algoritması 19.yüzyılın başlarında Thomas Bayes tarafından ortaya atılan bir koşullu olasılık formülü olan Bayes teoremine dayanır (Eş. 1). NB, Bayes teoremi üzerine kurulu, makine öğrenmesinde ve veri madenciliğinde sık kullanılan yöntemlerden biridir (Zhang, 2005)

Bayes Teoremi Formülü → $P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)}$ (1)

Veri Analizi ve Araştırma Bulguları

Kullanılan veri seti 15 öznitelik ve 32561 satırdan oluşmaktadır. Hedef değişken Salary özniteliğine göre geliri 50K üzeri olan 7841 kayıt, 50K ve altında ise 24720 kayıt bulunmaktadır. Yıllık gelir durumu 50K'nın üzerindeki bireylerin yaş karakteristiği incelendiğinde yaş ortalamasının 44 olduğu, gelir durumu 50K ve altında olan bireylerin yaş ortalamasının ise 36 olduğu görülmektedir. Veri setindeki en küçük yaş 19 en yüksek ise 90'dır.

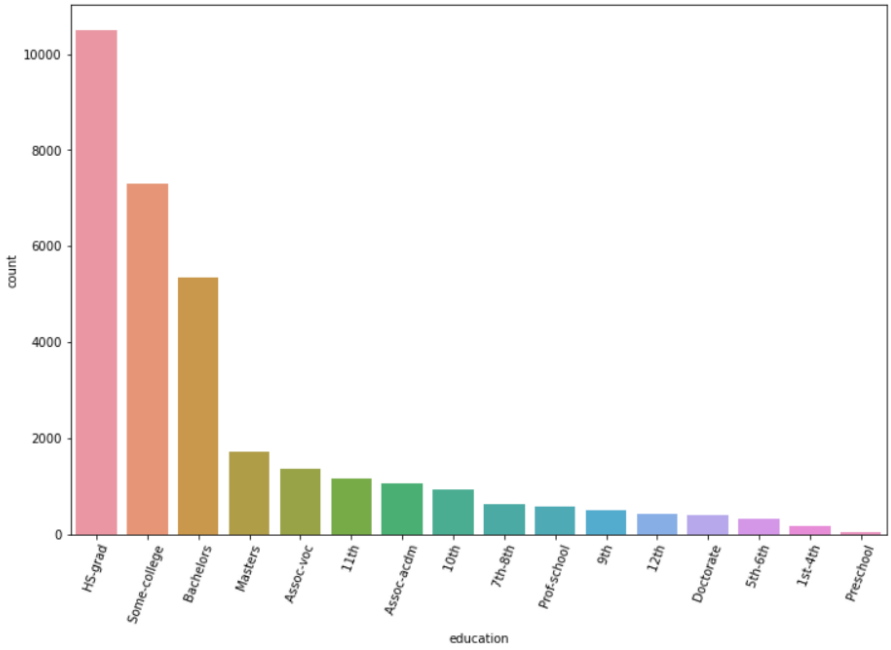
Veri setindeki sayısal alanların korelasyonunu gösterir matris Şekil 2'de verilmiştir.



Şekil 2. Korelasyon matrisi

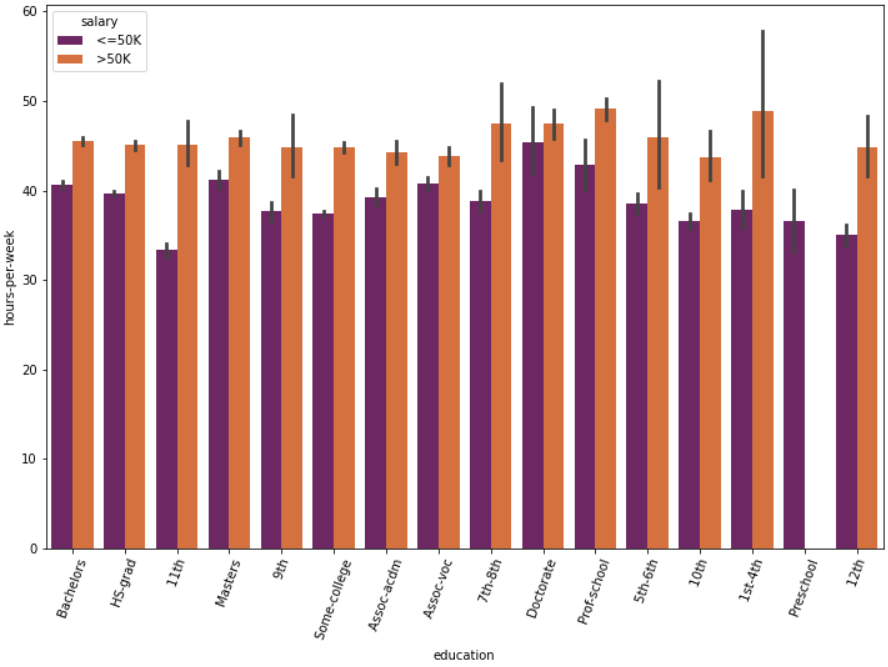
Veri setindeki hedef değişken salary özniteliği üzerinde en çok etkisi olan hours-per-week özniteliğidir. Veri setindeki education-num özniteliği ise hedef değişkeni en çok etkileyen ikincil öznitelik konumundadır.

Veri setindeki bireylerin eğitim durumlarına göre dağılımları incelendiğinde HS-grad 10501 adet, Some-college 7291 adet, Bachelors 5355 adet, Masters 1723 adet, Assoc-voc 1382 adet, 11^{th} 1175 adet, Assoc-acdm 1067 adet, 10^{th} 933 adet, 7^{th}-8^{th} 646 adet, Prof-school 576 adet, 9^{th} 514 adet, 12^{th} 433 adet, Doctorate 413 adet, 5^{th}-6^{th} 333 adet, 1^{st}-4^{th} 168 adet, Preschool 51 adet olarak dağılım gösterdiği gözlenmiştir (Şekil 3).



Şekil 3. Eğitim seviyelerine göre birey sayısı dağılımı

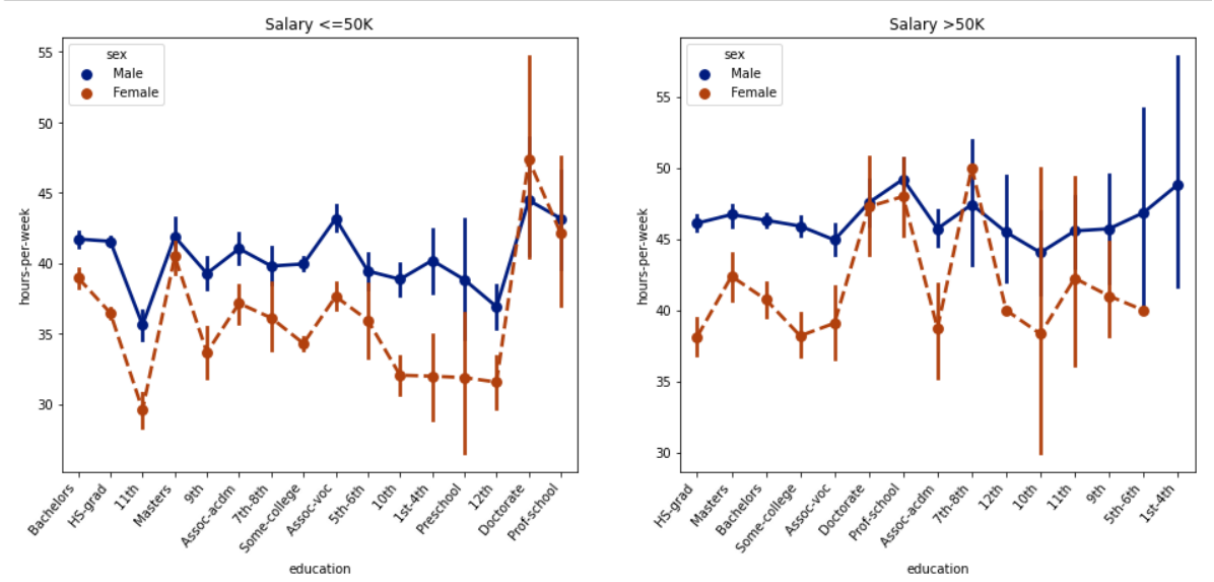
Eğitim durumları ve haftalık çalışma sürelerine göre bireylerin yıllık gelirleri incelendiğinde Geliri 50 bin dolar üzeri olup; haftada en az çalışan kesim 10^{th} olarak etiketli eğitim sınıfına sahip bireyler iken en çok çalışan kesim ise eğitim durumu Prof-school olanlardır. Geliri 50bin dolar ve altında olup haftada en az çalışan kesim eğitim sınıfı 11^{th} olanlar, en çok çalışan kesim ise eğitim durumu Doctorate olanlardır (Şekil 4).



Şekil 4. Eğitim durumu ve gelir seviyelerine göre bireylerin haftalık çalışma süreleri

Gelir seviyesine göre farklı gruplardaki bireylerin eğitim durumla ve cinsiyetlerine göre sınıflandırıldığındaki dağılımları incelendiğinde, erkeklerin her iki gelir grubunda da genel olarak daha

çok çalıştığı gözlenmiştir. Özellikle Doctorate grubunda çalışan kadınların haftalık çalışma süreleri daha uzundur (Şekil 5).



Şekil 5. Gelir seviyeleri ve cinsiyetlerine göre farklı grupların eğitim türü ve haftalık çalışma süresi karşılaştırması

Veri setindeki kategorik olan "education", "marital-status", "workclass", "occupation", "race", "sex", "relationship", "native-country" öznitelikleri analiz öncesinde kodlanmıştır. Ardından 4 farklı makine öğrenmesi algoritması kullanılarak sınıflandırmalar gerçekleştirilmiştir. Sınıflandırma sonuçları farklı metrikler açısından raporlanmıştır. Sınıflandırma sonuçlarının değerlendirilmesinde kullanılan ölçütler şu şekildedir:

Karmaşıklık matrisi, veri seti üzerindeki doğruluğu bilinen değerlerin test verileri üzerindeki sonuçlarıyla uygulanan modelin performans sonuçlarını değerlendirmek için kullanılan bir ölçüttür (Tablo 3). Karmaşıklık matrisi doğru pozitif, yanlış pozitif, yanlış negatif ve doğru negatif değerlerinden oluşur. Doğru pozitifler, gerçek değeri 1 ve modelin tahmin ettiği değeri 1 olan verilerdir. Yanlış pozitifler, gerçek değeri 0 ve modelin tahmin ettiği değeri 1 olan verilerdir. Yanlış negatifler, gerçek değeri 1 ve modelin tahmin ettiği değeri 0 olan verilerdir. Doğru negatifler, gerçek değeri 0 ve modelin tahmin ettiği değeri 0 olan verilerdir.

Tablo 3. Karmaşıklık matrisi

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

Karmaşıklık matrisinden çıkan sonuçlara göre doğruluk, duyarlılık, belirleyicilik ve keskinlik oranları hesaplanır. Kesinlik ve duyarlılık değerlerinin harmonik ortalamasından F1 puanı elde edilir.

Doğruluk, sınıflandırıcının ne sıklıkla doğru tahmin ettiğinin bir ölçüsüdür. Duyarlılık, sınıflandırıcının ne kadar gerçek pozitif değeri doğru tahmin ettiğinin bir ölçüsüdür. Belirleyicilik, sınıflandırıcının ne kadar gerçek negatif değeri olduğunu tahmin ettiğinin bir ölçüsüdür. Keskinlik, tüm sınıflardan, doğru olarak ne kadar tahmin edildiğinin bir ölçüsüdür. F1 puanı, sınıflandırıcının ne kadar iyi performans gösterdiğinin bir ölçüsüdür. F1 puanı kesinlik ve duyarlılık arasındaki dengeyi ifade eder. Doğruluk, duyarlılık, belirleyicilik, kesinlik metriklerine ait formüller Eş.2, Eş.3, Eş.4 ve Eş.5'te verilmiştir.

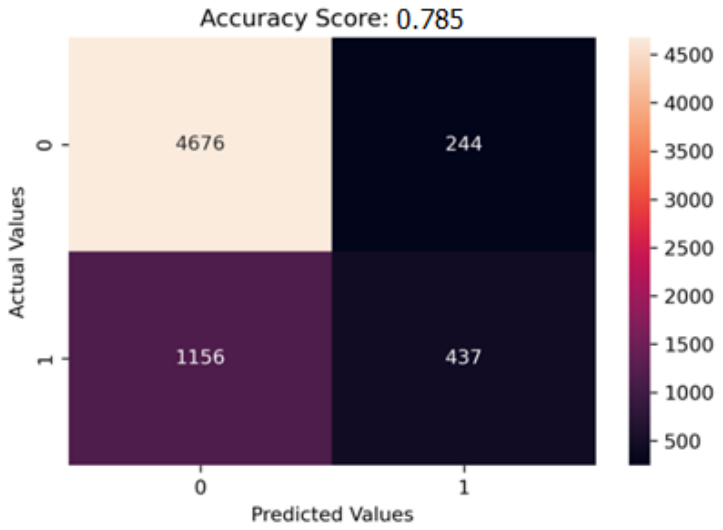
$$\text{Doğruluk} = \frac{DP+DN}{DP+DN+YP+YN} \quad (2) \qquad \text{Duyarlılık} = \frac{DP}{DP+YN} \quad (3)$$

$$\text{Belirleyicilik} = \frac{DN}{DN+YP} \quad (4) \qquad \text{Kesinlik} = \frac{DP}{DP+YP} \quad (5)$$

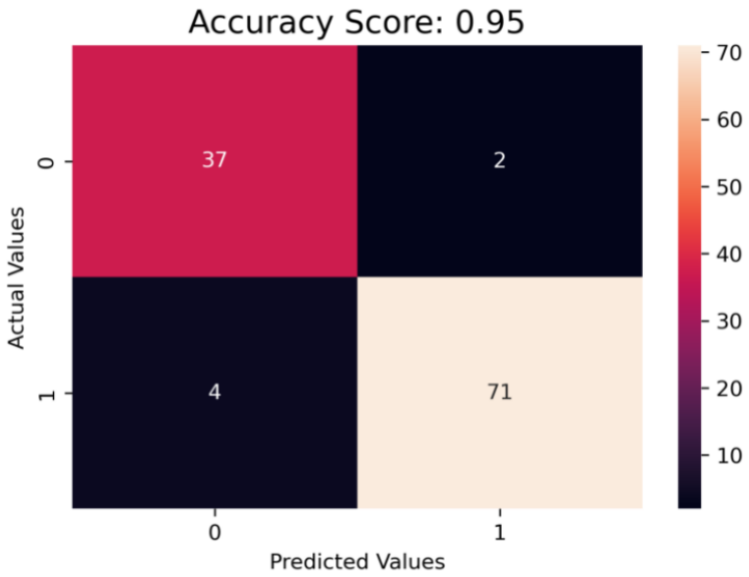
Duyarlık ve kesinliğin harmonik ortalaması olan F1 puanı ise Eş.6'daki formüle göre hesaplanır.

$$F1 = 2 * \frac{\text{Duyarlılık} * \text{Kesinlik}}{\text{Duyarlılık}+\text{Kesinlik}} \quad (6)$$

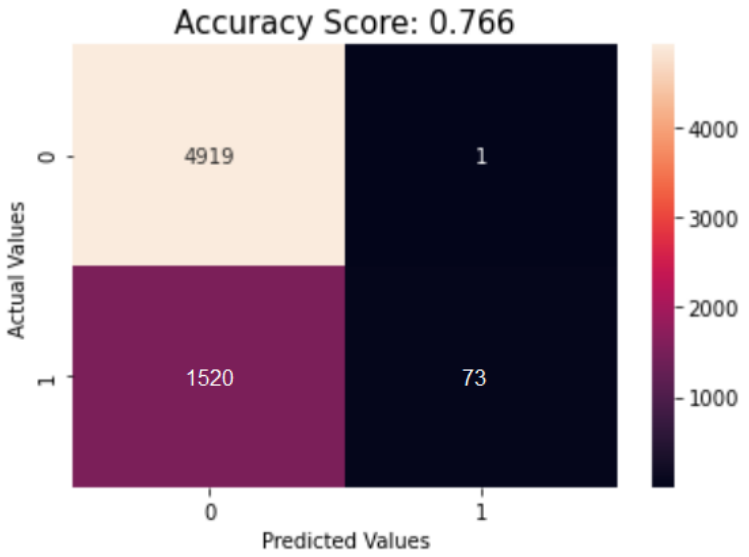
Veri seti üzerinde KNN, DVM, RO ve NB algoritmaları ile yapılan sınıflandırmalar sonucunda elde edilen karmaşıklık matrisleri sırayla Şekil 6, Şekil 7, Şekil 8 ve Şekil 9'da verilmiştir. Sınıflandırma işlemlerinde tüm algoritmaların parametreleri varsayılan değerleri ile kullanılmıştır.



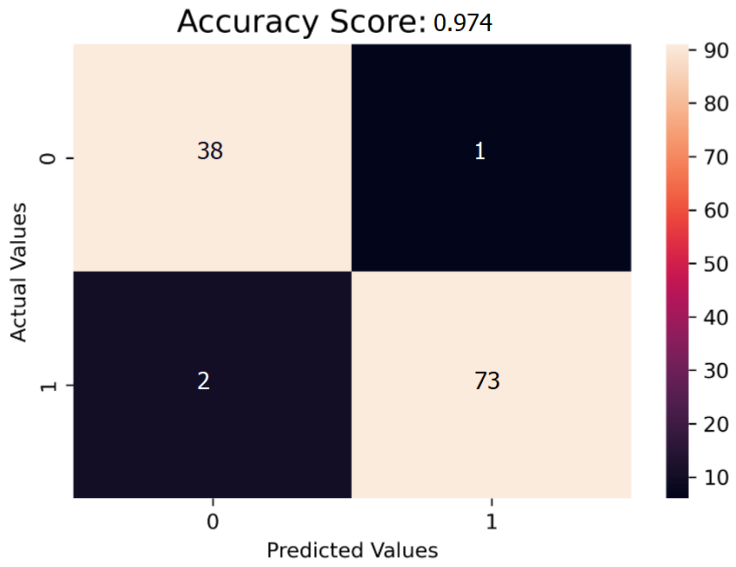
Şekil 6. KNN karmaşıklık matrisi



Şekil 7. DVM karmaşıklık matrisi



Şekil 8. Rastgele Orman karmaşıklık matrisi



Şekil 9. Naive Bayes karmaşıklık matrisi

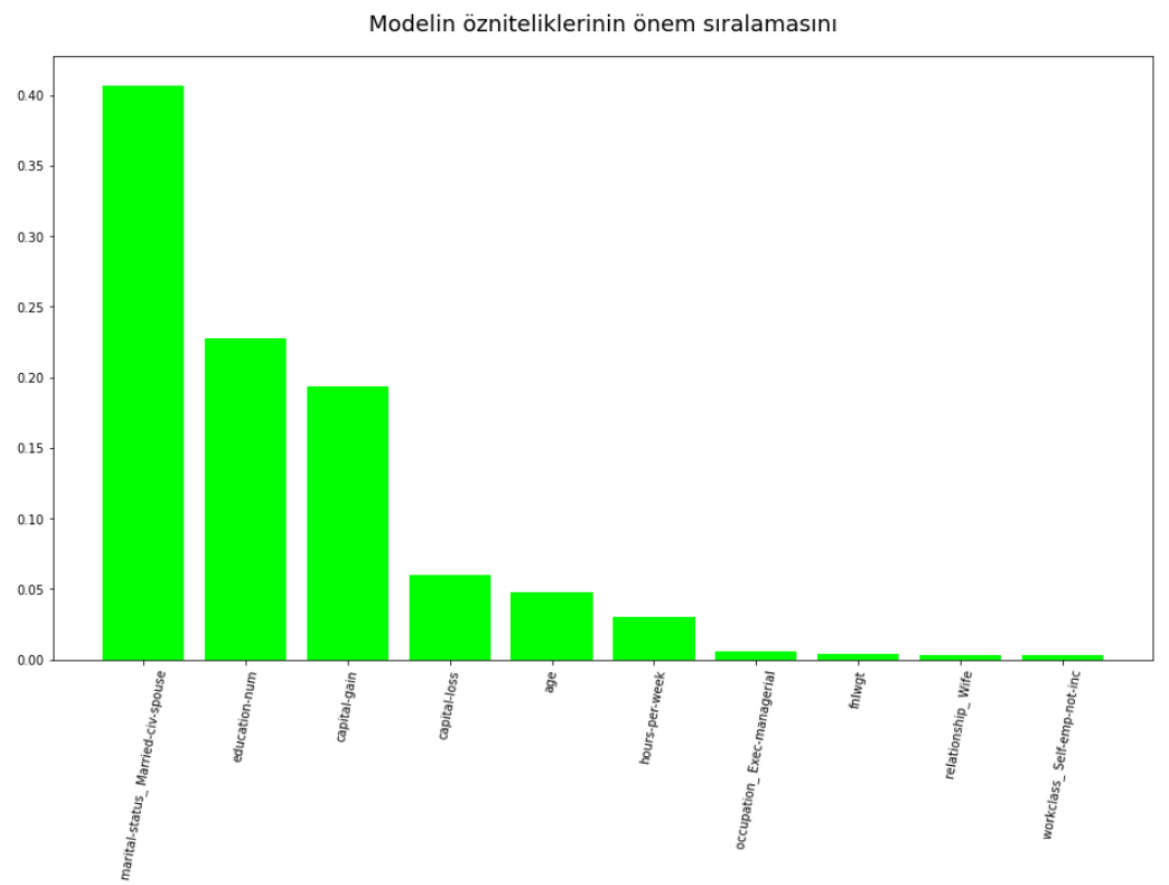
Karmaşıklık matrisleri üzerinden elde edilen hesaplanan diğer metriklerin değerleri Tablo 4'te verilmiştir.

Tablo 4. Kullanılan algoritmaların performans ölçümleri

Kullanılan Model	DP	YP	YN	DN	DOĞRULUK	DUYARLILIK	BELİRLEYİCİLİK	KESİNLİK	F1
KNN	4676	244	1156	437	0,7850	0,8017	0,6417	0,9504	0,87
DVM	37	2	4	71	0,9473	0,9024	0,9726	0,9487	0,92
Rastgele Orman (RO)	4919	1	1520	73	0,7664	0,7639	0,9864	0,9997	0,87
Naive Bayes (NB)	38	1	2	73	0,9736	0,95	0,9864	0,9743	0,96

Doğruluk değerleri incelendiğinde en başarılı modelin %97,36 oranıyla NB olduğu görülmektedir. İkinci başarılı model ise %94,73 doğrulukla DVM olmuştur. Üçüncü sırada %78,50 ile KNN yer alırken, doğruluk açısından en düşük başarıyı %76,64 ile RO sergilemiştir. Aynı veri seti üzerinde Chakrabarty ve Biswas (2018) tarafından gerçekleştirilen çalışmada elde edilen en yüksek doğruluk değeri %88,16 iken, bu çalışmada kullanılan NB ve DVM modelleri bu değerin üzerinde bir doğruluk sağlamışlardır. Doğru ve yanlış sınıfları ayırt etmedeki başarıyı gösterir diğer metriklerin dengesini ifade eden F1 başarı incelendiğinde NB algoritması %96 oranında başarı göstererek diğer algoritmalardan daha iyi sonuç vermiştir. Naive Bayes algoritmasından sonra en iyi sonucu %92 ile DVM göstermiştir. KNN ve RO algoritmalarında ise %87 başarı elde edilmiştir. Tüm metrikler açısından değerlendirildiğinde en başarılı modelin NB olduğu görülmektedir. Genel başarı sıralaması ise şu şekilde yapılabilir: NB>DVM>KNN>RO.

Sınıflandırma işlemlerinin ardından veri setindeki girdi özniteliklerin hedef değişkeni tahminleme üzerindeki etkileri incelenmiştir. Özellik önemlerini gösterir grafik Şekil 10'da verilmiştir.



Şekil 10. Tahminlemeye etki eden öznitelik önem sıralaması

Girdi özniteliklerin hedef değişkeni tahminlemedeki önemini gösterir sonuçlar incelendiğinde önem düzeyi en yüksek öznitelik Marital-status (evlilik durumu), en düşük olanın ise workclass (çalışma sınıfı) olduğu gözlenmiştir.

Sonuç

Bu çalışmada ülkelerin gelir durumu hakkında tahminlemeler yaparak, geliştirilecek stratejilere yardımcı olmak amacıyla makine öğrenmesi temelli bir çalışma gerçekleştirilmiştir. Bireylerin demografik bilgileri ve gelir bilgilerini içeren veri seti üzerinde KNN, DVM, RO ve NB algoritmaları ile sınıflandırmalar gerçekleştirilmiştir. Modellerin başarıları farklı metrikler açısından değerlendirilmiştir. 5 kat çapraz doğrulama uygulanarak gerçekleştirilen sınıflandırmalar sonucu en başarılı modelin %97,36 doğrulukla NB olduğu görülmüştür.

Sınıflandırmaya etki eden özelliklerin önem dereceleri incelendiğinde, hedef değişken olan gelir seviyesi düzeyi üzerinde en yüksek etkiye sahip olan özniteliğin bireylerin evlilik durumu olduğu görülmüştür.

Veri seti üzerinde yapılan incelemelerden çıkan bir diğer sonuç ise haftalık çalışma süresinin yıllık gelire doğrudan etki etmediğidir. Veri seti filtreleme sonuçları incelendiğinde, gelir düzeyi 50 bin doların altında ve haftada en çok çalışan kesimin doktora düzeyindeki eğitime sahip kadınlardan oluştuğu gözlemlenmiştir.

Bu çalışma gelir durumu tespiti konusunda çalışacak olan araştırmacılar için makine öğrenmesi temelli bir model örneği sunmaktadır. Gelecek çalışmalarda farklı veri setleri üzerinde, farklı algoritmalar kullanarak tahminleme çalışmalarının yapılması hedeflenmektedir.

Çıkar Çatışması Beyanı

Makale yazarları herhangi bir çıkar çatışması olmadığını beyan eder.

Araştırmacıların Katkı Oranı Beyan Özeti

Yazarlar makaleye benzer oranda katkı sağlamış olduğunu beyan eder.

Kaynakça

Akben S., Subasi A., Kıymık M. EEG işaretleri ile migren tanısında yapay sinir ağları ve destek vektör makineleri sınıflandırma yöntemlerinin karşılaştırılması. IEEE 18. Sinyal İşleme ve İletişim Uygulamaları Kurultayı Bildiri Kitabı, 2010.

Aksoy B. Estimation of energy produced in hydroelectric power plant ındustrial automation using deep learning and hybrid machine learning techniques. Electric Power Components and Systems 2021; 49(3): 213-232.

Ardic OP., Heimann M., Mylenko N. Access to financial services and the financial inclusion agenda around the world: a cross-country analysis with a new data set. World Bank Policy Research Working Paper 2011; 5537.

Breiman L. Random forests. Machine learning 2001; 45(1): 5-32.

Chakrabarty N., Biswas S. A statistical approach to adult census income level prediction. In 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018; 207-212.

Doğrul N. Gelir Seviyeleri farklı illerde eğitimin ekonomik büyümeye etkisi. Dumlupınar Üniversitesi Sosyal Bilimler Dergisi 2009; 23.

Karaman B., Özçalık M. Türkiye'de gelir dağılımı eşitsizliğinin bir sonucu: çocuk işgücü. Yönetim ve Ekonomi: Celal Bayar Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi 2007; 14(1): 25-41.

Meng H. Census income data set. 1994 classification using decision tree.

Öztemel E. Yapay sinir ağlari. PapatyaYayincilik, Istanbul, 2003.

Pala MA., Çimen ME., Boyraz ÖF., Yildiz MZ., Boz AF. Meme kanserinin teşhis edilmesinde karar ağacı ve knn algoritmalarının karşılaştırmalı başarım analizi. Academic Perspective Procedia 2019; 2(3): 544-552.

UCI. https://archive.ics.uci.edu/ml/datasets/adult. Erişim tarihi: 13.11.2021

Zhang H. Exploring conditions for the optimality of naive Bayes. International Journal of Pattern Recognition and Artificial Intelligence 2005; 19(02): 183-198.