Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 29, Sayı 1, 72-83, 2025

Süleyman Demirel University
Journal of Natural and Applied Sciences
Volume 29, Issue 1, 72-83, 2025

DOI: 10.19113/sdufenbed.1596624

Veri Madenciliğinde Karar Ağaçları İndükleyicilerinin Likert Ölçekli Verilerde Sınıflandırma Performansı

Özer ÖZDEMİR*[1], Ferdi KARAKÜTÜK[2], Aslı KAYA KARAKÜTÜK[3]

[1]Eskişehir Teknik Üniversitesi, Fen Fakültesi, İstatistik Bölümü, 26555, Eskişehir, Türkiye

[2]Türkiye İstatistik Kurumu, Zonguldak Bölge Müdürlüğü, 67000, Zonguldak, Türkiye

[3]Eskişehir Teknik Üniversitesi, Rektörlük, 26555, Eskişehir, Türkiye

(Alınış / Received: 05.12.2024, Kabul / Accepted: 02.03.2025, Online Yayınlanma / Published Online: 25.04.2025)

Anahtar Kelimeler
Veri madenciliği,
Karar ağaçları,
Likert ölçekli veri,
Türkiye aile yapısı araştırması,

Öz: Madencilik terimi ile benzer anlam taşıyan veri madenciliği, sorunların çözülmesine, eğilimlerin tahmin edilmesine, risklerin azaltılmasına ve yeni fırsatlar bulunmasına yardımcı olmak için muazzam miktarda bilgi ve veri setini analiz etme, yararlı zekayı keşfetme sürecidir. Bu çalışmada veri madenciliğinin muazzam yeteneklerinden faydalanarak Likert ölçekli veri tiplerinde bilgi keşfi yapılması amaçlanmıştır. Farklı veri madenciliği tekniklerinin Likert ölçekli veri türleri üzerinde sınıflandırma başarısını karşılaştırmak üzere veri seti olarak Türkiye Aile Yapısı Araştırması (TAYA) seçilmiştir. Yapısı gereği dengesiz olan veri seti üzerinde ilk olarak dengesizlik giderilmeden sınıflandırma yapılmış ardından sınıflar arası dengesizlik giderilmiş ve sınıflama analizine etkisi gözlemlenmiştir. Sınıflar arası dengesizliği giderebilmek amacıyla yeniden örnekleme ve veri tamamlama yöntemi ile toplam örnek hacmi değiştirilerek üç farklı veri seti oluşturulmuştur. Oluşturulan veri setlerinde sınıflandırma başarısı en yüksek olan algoritmanın CART algoritması olduğu görülmüştür. Dengesizlik giderilmeden yapılan sınıflandırmada ise RepTree algoritmasının daha başarılı sonuçlar ürettiği görülmüştür.

Classification Performance of Decision Tree Inducers in Data Mining on Likert Scale Data

Keywords
Data mining,
Decision trees,
Likert scale data,
Family structure survey in Turkey,

Abstract: Data mining, which has a similar meaning to the term mining, is the process of analyzing enormous amounts of information and data sets and discovering useful intelligence to help solve problems, predict trends, mitigate risks and find new opportunities. This study aims to utilize the enormous capabilities of data mining for knowledge discovery in Likert-scale data types. To compare the classification success of different data mining techniques on Likert-scale data types, the Turkish Family Structure Survey (TFSS) was selected as the data set. On the data set, which is imbalanced due to its structure, firstly, classification was performed without removing the imbalance, then the imbalance between classes was removed and its effect on the classification analysis was observed. In order to eliminate the imbalance between classes, three different data sets were created by changing the total sample volume with resampling and data completion method. It was observed that the algorithm with the highest classification success in the created data sets was the CART algorithm. RepTree algorithm was found to produce more successful results in the classification without removing the imbalance.

1. Giriş

Dijitalleşme çağı, sayısallaştırılmış bilgilerin işlenmesini, saklanmasını, dağıtılmasını ve iletilmesini kolaylaştırmaktadır [1]. İlgili teknolojilerdeki önemli ilerlemeler ve bunların hayatın farklı alanlarında sürekli genişleyen kullanımları ile, farklı özelliklerde çok büyük miktarda veri toplanmaya ve veri

*Sorumlu yazar: ozerozdemir@eskisehir.edu.tr

tabanlarında saklanmaya devam edilmektedir. Bu tür verilerin depolanma hızı da paralel olarak artış göstermektedir. Devasa veri hacimlerinden bilgi keşfi ciddi iş ve zaman yükü doğurmaktadır. Veri madenciliği kavramı, devasa veri hacmine gömülü bilgi patlamasını anlamlandırma girişimi olarak hayatımıza girmiştir. Veri madenciliği kavramına ilişkin literatürde birçok tanım bulunmaktadır. Bazı araştırmacılar veri madenciliğini şu şekilde tanımlamaktadır: Gundecha ve Tan'a göre, büyük ölçekli verilerde yararlı veya eyleme geçirilebilir bilgileri keşfetme sürecidir [2, 3]. Zaki ve Meira'ya [4] göre veri madenciliği, büyük ölçekli verilerden açıklayıcı, anlaşılır ve tahmine dayalı modellerin yanı sıra anlayışlı, ilginç ve yeni kalıpları keşfetme sürecidir [4]. Veri madenciliğinin diğer bir tanımı Özer ve Sprinkhuizen-Kuyper [5] ve Garcia v.d. [6] göre, büyük miktarda veriden ilginç (önemsiz olmayan, üstü örtülü, önceden bilinmeyen ve potansiyel olarak yararlı) kalıpların veya bilginin çıkarılmasıdır. Sınıflandırma önemli bir veri madenciliği problemidir. Bir sınıflandırma probleminde, her biri bir dizi özniteliğe sahip bir dizi örnekten oluşan eğitim seti adı verilen bir girdi veri kümesi bulunmaktadır. Nitelikler, nitelik değerleri sıralandığında süreklidir veya nitelik değerleri sıralanmadığında kategoriktir. Kategorik niteliklerden biri, sınıf etiketi veya sınıflandırma niteliği olarak adlandırılır. Amaç, modelin eğitim veri kümesinden değil yeni verileri sınıflandırmak için kullanılabileceği şekilde, diğer niteliklere dayalı olarak bir sınıf etiketi modeli oluşturmak için eğitim veri kümesini kullanmaktır [7].

Veri madenciliği teknikleri sınıflandırma problemlerinde sıklıkla kullanılmaktadır. Özellikle farklı veri tiplerinin sınıflandırılmasına ilişkin literatürde çeşitli çalışmalar bulunmaktadır. Likert ölçekli veri türlerinde de veri madenciliği teknikleri çalıştırılmıştır. Bu amaçla yapılan bazı araştırma örneklerini sıralamak gerekirse, Brefelean'ın çalışması [8], bir fakültedeki farklı uzmanlık öğrencileri (yüksek lisans, doktora vb.) üzerinde yapılan anketlerden toplanan verilerle, karar ağaçları aracılığıyla üniversite sonrası çalışmalarla eğitimlerine devam etme tercihlerini farklılaştırmak ve tahmin etmek amacıyla bir J48 algoritması analiz aracının uygulanmasını temsil etmektedir. Mardikyan ve Batur (2011) [9] yaptıkları çalışmada, iki farklı veri madenciliği tekniği olan adımsal regresyon ve karar ağaçlarını kullanarak öğretim elemanlarının öğretim performansının değerlendirilmesiyle ilişkili faktörleri araştırmışlardır. Veriler Boğaziçi Üniversitesi Yönetim Bilişim Sistemleri bölümü öğrencilerinin anket değerlendirmelerinden anonim olarak toplanmıştır. Sonuçlar, değerlendirme formundaki öğretim elemanı ile ilgili soruları özetleyen bir faktörün, öğretim elemanının çalışma durumunun, dersin iş yükünün, öğrencilerin devam durumunun ve formu dolduran öğrencilerin yüzdesinin öğretim elemanının öğretim performansının önemli boyutları olduğunu göstermektedir. Kuzey (2012), [10] karar destek sistemleri, karar ağaçları ve destek vektör makineleri (DVM) yöntemlerini 2011 yılında bilgi işlem çalışanları üzerinde yapılan anket yolu ile elde edilmiş veriler üzerinde çalıştırmıştır. Elde edilen ilgili veri kümesi kullanılarak DVM ve Karar ağaçları modelleri uygulanarak, modellerin performansları değişik araçlar ile kıyaslanmıştır. Analiz sonuçlarına göre, DVM ile C&R Tree karar ağacı modelli en yüksek doğruluk oranına sahip olduğu görülmüştür. Şehribanoğlu ve Saygın (2016), [11] yapmış oldukları yüksek lisans tezinde, veri madenciliği süreci, yöntemleri ve sınıflandırma yöntemleri altında yer alan karar ağaçları algoritmalarını TÜİK tarafından yürütülen Yaşam Memnuniyeti Araştırması B grubu mikro verileri kullanarak incelemişlerdir. CHAID algoritmasının daha fazla açıklayıcı değişken ile karar ağacı oluşturmasından dolayı CHAID algoritması daha araştırmacı yapıya sahip olduğu gözlemlenmiştir. Beernaert (2021) [12] yapmış olduğu çalışmasında, anket verilerinin analizini genişletmeyi sağlayan bazı teknikler denemiştir. Özellikle, daha sağlam olan ve anket verilerinin analizine olanak tanıyan bazı veri madenciliği ve makine öğrenimi tekniklerini doğrusal olmayanlar için önermiştir. Çalışmada kullanılan veriler, 1010 Slovakyalı'nın müzik ve film tercihlerinden fobilere, hayata bakış açılarından harcama alışkanlıklarına kadar hayatın birçok farklı yönüyle ilgili 150 soruyu yanıtladığı Genç İnsanlar Anketi'dir. Herhangi bir boyut indirgemesi olmaksızın Extreme Gradient Boosting algoritmasının en iyi performansı gösterdiği ve onu Rastgele Orman algoritmasının izlediği ortaya çıkmıştır. Destek Vektör Makineleri kullanılması durumunda PCA hariç, boyut azaltma yöntemlerinin burada tahmin performansı üzerinde hiçbir faydası olmadığı kanıtlanmıştır. Bezek-Gürü ve arkadaşları (2023), [13] yaptıkları çalışmada, Rastgele Orman ve MARS yöntemlerini kullanarak öğrencilerin PISA 2018 okuma becerilerini etkileyen faktörleri belirlemeyi ve tahmin yeteneklerini karşılaştırmayı amaçlamışlardır. Deney sonuçları MARS yönteminin Rastgele Orman yönteminden daha iyi performans gösterdiğini ortaya koymuştur. Çalışmanın eğitim bilimleri araştırmalarında veri madenciliğinin kullanımı için rol model teşkil edeceği vurgulanmıştır. Ogedengbe ve arkadaşları (2024), [14] yaptıkları çalışmada, Ortak Soru Öznitelik Budama (CQAP) adı verilen ve standart Apriori algoritmasını, öznitelik ayrıştırmasına gerek kalmadan 5'li Likert ölçeği veri kümelerini işleyecek şekilde genişleterek geliştiren ve böylece katılımcıların görüşlerinin sıralı doğasını koruyan yeni bir madencilik tekniği önermişlerdir. Yapılan deneyler sonucunda, değerlendirilen örnek veri kümeleri üzerinde önerilen CQAP tekniğinin son teknoloji Apriori algoritmasına karşı üstün performans potansiyelini gösterdiğini belirtmişlerdir.

Bu çalışmanın amacı, veri madenciliği tekniklerinin Likert ölçekli veriler üzerinde uygulanmasına ilişkin çalışmaların literatürde yetersiz olması nedeniyle, bu bağlamda değerli bilgiler kazandırmaktır. Çalışmada, veri madenciliği algoritmalarından karar ağaçları yapısı ele alınarak, bu yapıların Üçlü Likert (Ordinal)

ölçeğine sahip veri setleri üzerindeki sınıflandırma performansını ölçmek, sınıflar arası dengesizliğin olduğu veri setlerinde sınıflandırma performanslarını karşılaştırmak ve dengesizliğin giderildiği durumda sınıflandırma performansının nasıl değiştiğini ortaya koymak amaçlanmıştır. Bu minvalde, çalışmada, veri seti olarak Türkiye İstatistik Kurumu Başkanlığı tarafından yürütülen Türkiye Aile Yapısı Araştırması (TAYA) seçilmiştir. TAYA veri seti yapısı gereği dengesiz bir veri setidir ve anketin amacı, Türkiye'deki ailelerin yapısını, bireylerin aile ortamındaki yaşam biçimlerini ve bireylerin aile hayatına ilişkin değer yargılarını tespit etmektir. Dolayısıyla bu çalışma, aynı zamanda sınıflandırma analizi sonucunda "Türk aile yapısının güncel durumunu ve demografik çerçevesini" sunmaktadır. Ayrıca, 2006 yılında TÜİK tarafından 18 ve daha yukarı yaştaki bireylerle ilk kez gerçekleştirilen Aile Yapısı Araştırması, 2011 ve 2016 yıllarında tekrarlanmış ve anket verileri betimsel ve çeşitli ileri istatistiksel teknikler (çoklu regresyon, lojistik regresyon vb.) aracılığı ile anlamlandırılmaya çalışılmıştır [15]. Ancak, yapılan tarama sürecinde ilgili araştırma üzerinde veri madenciliği teknikleri ile çalışılmadığı görülmüştür. Bu husus, çalışmamızın özgünlüğünü oluşturmaktadır.

2. Materyal ve Metot

Bu bölümde; veri madenciliği teknolojisi ve öneminden, uygulanan karar ağacı algoritmalarından kısaca bahsedilecektir. Güçlü veri analiz araçlarına duyulan ihtiyaçla birleşen veri bolluğu, tarih açısından zengin ancak bilgi yönünden fakir bir durum olarak görülmektedir. Büyük ve sayısız veri tabanlarında toplanan ve depolanan, hızla büyüyen, muazzam miktardaki veri, güçlü araçlar olmadan insan kavrayışını çok aşmıştır. Buna bağlı olarak, büyük veri tabanlarında toplanan veriler, nadiren yeniden ziyaret edilen veri arşivleri olan veri yığınları haline gelmiştir. Sonuç olarak, önemli kararlar genellikle veri tabanlarında depolanan bilgi açısından zengin verilere değil, karar vericinin sezgisine dayalı olarak alınır, çünkü karar vericinin çok büyük miktardaki verilerin içine gömülü değerli bilgileri çıkaracak araçları yoktur. Bununla birlikte, bilgileri bilgi tabanına manuel girmek ne yazık ki, önyargılara ve hatalara eğilimlidir ve son derece zaman alıcı ve maliyetlidir. Veri analizi yapan veri madenciliği araçları, iş stratejilerine, bilgi tabanlarına ve bilimsel araştırmalara büyük ölçüde katkıda bulunan önemli veri modellerini ortaya çıkarmaktadır. Veri madenciliği, çeşitli sorgular oluşturma ve muhtemelen veri tabanlarında depolanan büyük miktarlardaki verilerden genellikle daha önce bilinmeyen faydalı bilgiler, modeller ve eğilimleri çıkarma işlemidir.

2.1. Karar ağaçları ve indükleyicileri

Basit bir ifade ile karar ağacı, denetimli sınıflandırma problemlerinde yaygın olarak kullanılan önceden tanımlanmış bir hedef değişkene sahip bir öğrenme algoritması türü olarak tanımlanabilir. Sınıflandırma problemlerine mükemmel bir şekilde uydukları için karar ağaçları güçlü yön olarak kabul edilmektedir [17]. Birçok uygulamada, örnekleri doğru bir şekilde sınıflandırmak için yalnızca oluşturulan sınıflandırma modelini kullanmak yetmez, aynı zamanda modelin incelenmesi de istenebilir. Bu durum tahminlerin açıklanması, değiştirilmesi veya mevcut bazı arka plan bilgileriyle birleştirilmesi ile gerçekleşmektedir. Modelin hem yüksek sınıflandırma doğruluğunun hem de insanlar tarafından okunabilirliğinin gerekli olduğu bu tür uygulamalarda, çoğu veri madencisinin tercih edeceği bariz yöntem karar ağaçları olmaktadır. Karar ağacı algoritmaları uzun yıllardır incelenmektedir ve özellikle çok sayıda iyileştirme ve varyasyonun önerildiği veri madenciliği algoritmalarına sahiptir. Bu nedenle, aynı model temsilini ve algoritma işlem şemalarını paylaşan, ancak birkaç ayrıntıda farklılık gösterebilen bir algoritma ailesinden bahsedilebilir. Bu çeşitlilik için alan, genellikle karar ağacı modelleri oluşturmak için gerçekleştirilen, karar ağacı büyütme ve budamadan (pruning) oluşan iki aşamalı işlemle artırılmaktadır.

Karar ağacı, bir sınıflandırma modelini temsil eden hiyerarşik bir yapıdır. İç ağaç düğümleri, etki alanını bölgelere ayırmak için uygulanan bölmelere karşılık gelir ve uç düğümler, yeterince küçük veya yeterince tek biçimli olduğuna inanılan bölgelere sınıf etiketlerini atamaktadır [16]. Karar ağaçları, her testte sayısal bir özelliğin bir eşik değerle karşılaştırıldığı, bir dizi temel testi verimli ve tutarlı bir şekilde birleştiren ardışık bir modeldir [18]. Kavramsal kuralların oluşturulması, sinir ağındaki düğümler arasında yer alan bağlantılara ait ağırlıkların oluşturulmasından çok daha kolaydır [19]. Ağırlıklı olarak gruplama amaçları için karar ağaçları kullanılır. Ayrıca karar ağaçları, veri madenciliğinde sıklıkla kullanılan bir sınıflandırma modelidir [20]. Sürecin en kritik hususu, karar ağacı yapısını oluştururken hangi algoritmadan yararlanılacağının belirlenmesidir. Literatürde çeşitli karar ağacı algoritması bulunmaktadır.

2.1.1. C4.5 algoritması (J48)

C4.5, en iyi bilinen ve en yaygın kullanılan karar ağacı algoritmalarından biridir [21]. Doğruluk seviyesi, işlenecek veri hacminden bağımsız olarak yeterince yüksektir. Karar ağaçlarını ve diğer öğrenme algoritmalarını karşılaştıran en son çalışmalardan biri, C4.5'in çok iyi bir hata oranı ve hız kombinasyonuna sahip olduğunu göstermektedir [22], [23]. Algoritma, tamamlanmamış bir eğitim veri setini işleme ve boyutunu küçültmek ve karar yolunu optimize etmek için sonuçtaki karar ağacını budama yeteneğine sahiptir [24]. C4.5 ayrıca sürekli niteliklerle başa çıkma yeteneğine de sahiptir. Binarizasyon sürecini kullanarak sürekli öznitelikleri yönetmektedir. Bu nitelikler, verileri iki aralığa ayıran eşik değerleri kullanan ayrıklarla değiştirilir [25]. C4.5 algoritması ID3 algoritmasının varisidir ancak bu algoritma

çalışmasında kayıp verileri hesaba dahil etmeyerek ID3 algoritmasından farklılaşmaktadır.

2.1.2. CART (C&RT) algoritması

Sınıflandırma ve Regresyon Ağacı (kısaca CART), yaygın kullanımda olan çok popüler bir ağaç tabanlı yöntemdir. CART her zaman bir ikili ağaç oluşturur, yani terminal olmayan her düğümün iki alt düğümü vardır. Bu, birden çok alt düğüme izin verebilen genel ağaç tabanlı yöntemlerin tersidir. Ağaç tabanlı yöntemin ve özellikle CART'ın büyük bir çekiciliği, karar sürecinin biz insanların nasıl karar verdiğimize çok benzemesidir. Bu nedenle, ağaç tarzı karar sürecinden çıkan sonuçları anlamak ve kabul etmek kolaydır. Bu sezgisel açıklayıcı güç, ağaç yönteminin asla ortadan kalkmayacağının en büyük nedenlerinden biridir. CART algoritmasının bir başka çekici yönü de lojistik regresyon veya destek vektör makinesi gibi çeşitli girdi verisi türlerine izin vermesidir. Böylece girdi verileri, fiyat veya alan gibi sayısal değişkenleri ve ev tipi veya konumu gibi kategorik değişkenleri karıştırabilir. Bu esneklik, CART algoritmasını çeşitli uygulamalarda tercih edilen bir araç haline getirmektedir. CART'ın belirli bir ağaç oluşturma yöntemi vardır. İlk olarak, her zaman girdinin tek bir özelliği (değişkeni) boyunca bölünmektedir. Giriş vektörüx 'in $x = (x_1, \dots, x_d)$ biçimindedir ve burada her x_i sayısal, kategorik (nominal) veya ayrık sıralı değişken olabilmektedir. Her düğümde, CART her zaman bu değişkenlerden birini seçmekte ve düğümü yalnızca o değişkenin değerlerine göre bölmektedir. Başka bir deyişle, CART, lojistik regresyonun yaptığı gibi değişkenleri birleştirmez. CART'ın bölme yaptığı başka bir özel yol da her zaman iki alt düğüm oluşturmasıdır- ikili bölme- ki bu da bir ikili ağaçla sonuçlanmaktadır.

2.1.3. CHAID (Ki-Kare Otomatik Algılama Dedektörü) Algoritması

Adından da anlaşılacağı gibi, CHAID bir Ki-kare bölme kriterine dayanmaktadır. Ki-kare otomatik etkileşim detektörü (CHAID) algoritması, eldeki her türlü problemi çözmeye uyarlanabildiği için en çok kullanılan denetimli öğrenme yöntemlerinden biri olarak kabul edilmektedir. Daha spesifik olarak, Ki-karenin p-değerini kullanmaktadır. Kass tarafından tanıtılan CHAID algoritması, AID ve THAID'in bir evrimi olarak, günümüzde bu daha önceki istatistiksel denetimli ağaç yetiştirme teknikleri arasında en popüler olanıdır [26]. CHAID algoritmasının temel fikri, örnekleri verilen hedef değişkene ve seçilen özellik indeksine (tahmin değişkeni gibi) göre en uygun şekilde bölmek ve ki-kare testinin önemine göre otomatik olarak yargılamak için olasılık tablosunu gruplandırmaktır. Kriter, bölme tarafından oluşturulan g düğümleri arasındaki ortalama değerlerdeki fark için F istatistiğinin p değeridir ve denklem (1) ile hesaplanmaktadır:

$$F = \frac{BSS/(g-1)}{WSS/(n-g)} \sim F_{(g-1),(n-g)} \qquad (1)$$

Denklem 1'de WSS notasyonu, varyans analizinde genellikle kalıntı veya kareler toplamı olarak bilinmektedir ve denklem (2) ile hesaplanmaktadır

$$WSS = \sum_{j=1}^{g} \sum_{i=1}^{n_j} (y_{ij} - \underline{y_j})^2 \qquad (2)$$

Denklem 2'de $\underline{y_j}$ düğümündeki y_{ij}'lerin ortalama değeridir. BSS notasyonu ise gruplar arası kareler toplamı olarak adlandırılır.

2.1.4. NbTree algoritması

NbTree algoritması, karar ağacı sınıflandırıcıları ile Naive Bayes sınıflandırıcıları arasında bir melezdir. Öğrenilen bilgiyi, yinelemeli olarak oluşturulmuş bir ağaç biçiminde temsil etmektedir. Bununla birlikte, yaprak düğümler, tek bir sınıfı tahmin eden düğümlerden ziyade Naive Bayes kategorileştiricileridir [27]. NbTree'nin temel fikri, yüksek boyutlu noktalar için indeks anahtarı olarak Öklid norm değerini kullanmaktır. Saf Bayes'in öznitelik bağımsızlığı varsayımı tüm eğitim verilerinde her zaman ihlal edilse de yerel eğitim verileri içindeki bağımlılıkların tüm eğitim verilerindekinden daha zayıf olması beklenebilir. Böylece, NbTree [28], karar ağacı sınıflandırıcılarının ve Naive Bayes sınıflandırıcılarının avantajlarını bütünleştiren, oluşturulmuş karar ağacının her bir yaprak düğümü üzerinde saf bir Bayes sınıflandırıcısı oluşturmaktadır. Basitçe söylemek gerekirse, öncelikle, eğitim verilerinin her bir bölümünün bir ağaç yaprak düğümü tarafından temsil edildiği eğitim verilerini bölümlere ayırmak için karar ağacını kullanır ve ardından her bölüm üzerinde saf bir Bayes sınıflandırıcısı oluşturmaktadır. Karar ağaçlarının oluşturulmasında temel bir konu, ağacın terminal olmayan her bir düğümündeki öznitelik seçim ölçüsüdür. Yani, karar ağaçlarının oluşturulmasında uç olmayan her bir düğümün ve bir bölünmenin faydasının ölçülmesi gerekmektedir.

2.1.5. RepTree algoritması

Azaltılmış hata budama (REP) ile bir karar ağacının birleştirilmesiyle oluşturulan Azaltılmış Hata Budama Ağacı ("REPT") temelde hızlı karar ağacı öğrenimidir ve bilgi kazanımına veya varyansı azaltmaya dayalı bir karar ağacı oluşturmaktadır. Bu yaklaşımda DT, modellemek için eğitim veri setini kullanmaktadır; karar ağacının performansı yüksek olduğunda, REP ağaç yapısı karmaşıklığını en aza indirmektedir. Budama yöntemi, geriye dönük aşırı uyum problemlerini açıklamaktadır.

2.1.6. RandomTree algoritması

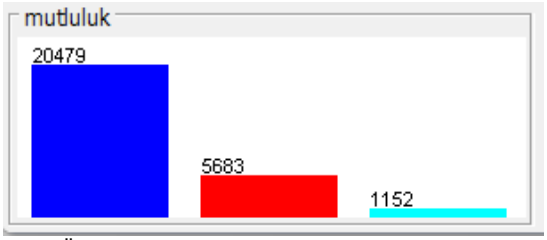
Rastgele Ağaçlar, esasen makine öğrenimindeki iki algoritmanın birleşimidir; tek model ağaçlar ve

rastgele orman. Model ağaçları, her bir yaprağın, bu yaprak tarafından tanımlanan yerel altuzay için optimize edilmiş doğrusal bir modele sahip olduğu karar ağaçlarıdır. Rastgele ağaç denetimli bir sınıflandırıcıdır; birçok bireysel öğrenici üreten bir topluluk öğrenme algoritmasıdır. Rastgele bir ormanda, her düğüm, o düğümde rastgele seçilen tahmin edicilerin alt kümesinin en iyisi kullanılarak bölünmektedir. Rastgele ağaçlar Leo Breiman ve Adele Cutler tarafından üretilmiştir [29]. Algoritma hem sınıflandırma hem de regresyon problemleriyle başa çıkabilmektedir. Rastgele ağaçlar, orman adı verilen ağaç tahmincilerinin bir koleksiyonudur.

3. Bulgular

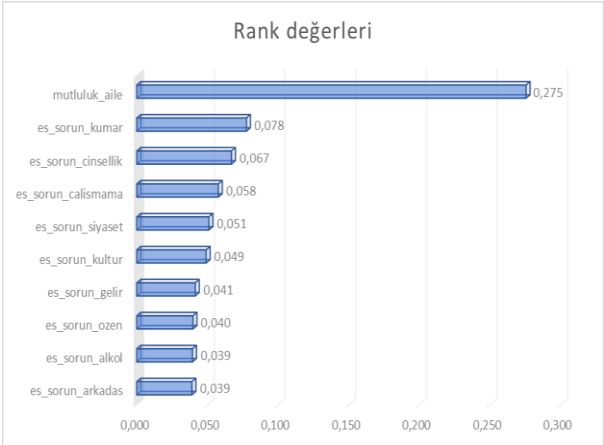
3.1. Üçlü likert ölçeğe sahip değişkenler için sınıflandırma sonuçları

Çalışmanın amacı, üçlü Likert ölçeğe sahip veri seti üzerinde veri madenciliği algoritmalarının sınıflandırma performanslarını karşılaştırmaktır. Bu kapsamda ilk olarak, TAYA araştırmasında kullanılan öznitelikte ölçekler herhangi bir okula gitmeyen, ilköğretim, lise ve yükseköğretim mezunu olarak tanımlanmış ve bu tanımlamaya uygun veri birleştirilmesi gerçekleştirilmiştir. Beşli Likert ölçek düzeyiyle veri alınan kişisel mutluluk, aile içi mutluluk ve kişisel gelirden memnuniyet düzeylerinde üçlü Likert ölçek kullanılmış ve veri birleştirme işlemi gerçekleştirilmiştir. TAYA'da eşlerin kendi aralarında yaşadıkları sorunların sıklıklarının düzeyini ölçmek amacıyla kullanılan beşli Likert ölçeği üç düzeyli mutluluk Likert ölçeğine dönüştürülmüş ve bu tanımlamaya uygun veri birleştirme işlemi gerçekleştirilmiştir.



Şekil 1. Üçlü Likert ölçeğine sahip veri setinde mutluluk düzeyine ait sınıf frekansları

Veri seti WEKA 3.8.5 paket programında NumericToNominal filtresi kullanılarak sınıfı temsil eden özniteliğin kategorik olarak temsili sağlanmıştır. Ardından AttributeSelection filtresinde Information Gain (Bilgi Kazancı) algoritması ve Ranker (Sıralama) metodu kullanılarak her bir algoritma için en değerli 10 öznitelik seçilmiştir. Seçilen öznitelikler Şekil 2' te sunulmuştur.



Şekil 2. Mutluluk sınıflandırması için bilgi kazancı metodu ile seçilen 10 adet öznitelik (üçlü Likert ölçekli veri setinde)

Şekil 2 incelendiğinde "aile içinde mutluluk" değişkeninin sınıflandırma başarısında en etkili öznitelik olduğu görülmüştür. Öznitelik seçme adımından sonra yukarıda bahsi geçen tüm veri madenciliği algoritmaları çalıştırılmış ve bazı temel parametrelere göre genel sınıflandırma sonuçları Tablo 1 ve sınıf bazlı sonuçları Tablo 2 ile gösterilmiştir.

Tablo 1. Algoritmaların sınıflandırma sonuçları (genel)

	Kappa	TP	FP	Kesin.	F ölçüt	ROC	Doğru.
J48	0,550	0,843	0,334	0,829	0,828	0,784	0,842
CHAID	0,552	0,843	0,330	0,829	0,833	0,801	0,842
CART	0,552	0,843	0,332	0,829	0,829	0,779	0,841
Nb TREE	0,531	0,831	0,324	0,815	0,819	0,796	0,831
Rand. TREE	0,541	0,839	0,337	0,824	0,824	0,784	0,839
Rep TREE	0,552	0,843	0,333	0,830	0,829	0,798	0,843

Tablo 1 incelendiğinde RepTREE algoritmasının (%84.3) ile en başarılı sonuç verdiği gözlemlenmiştir. F-ölçütünde kapsamında CHAID, CART, RepTREE algoritmalarının sınıflandırmada daha başarılı olduğu gözlemlenmiştir. Kappa değeri ele alındığında bütün algoritmalar için değerin 0.4-0.6 aralığında olduğu görülmüştür. Bundan dolayı Kappa istatistiği sonuçları gözlenen uyumun tesadüfen gerçekleşmediğini göstermiştir.

Tablo 2. Algoritmaların sınıflandırma sonuçlarına ilişkin bilgiler (üçlü Likert ölçek, sınıf bazlı)

	TP Oran	FP Oran	Kesin.	F-Ölçüt	ROC Alanı	Sınıf
J48 (C4.5)	0,95	0,43	0,869	0,908	0,783	Mutlu
	0,593	0,057	0,733	0,66	0,770	Orta
	0,165	0,005	0,592	0,258	0,750	Mutsuz
CHAID	0,949	0,424	0,87	0,908	0,804	Mutlu
	0,594	0,058	0,73	0,655	0,789	Orta

	0,187	0,006	0,586	0,283	0,762	Mutsuz
CART	0,949	0,427	0,869	0,908	0,783	Mutlu
	0,593	0,057	0,733	0,656	0,771	Orta
	0,18	0,005	0,597	0,276	0,754	Mutsuz
Nb Tree	0,932	0,412	0,872	0,901	0,80	Mutlu
	0,602	0,072	0,687	0,642	0,788	Orta
	0,168	0,009	0,441	0,244	0,775	Mutsuz
Rep Tree	0,95	0,428	0,869	0,908	0,802	Mutlu
	0,595	0,057	0,733	0,657	0,788	Orta
	0,173	0,005	0,61	0,269	0,772	Mutsuz
Rand. Tree	0,948	0,433	0,868	0,906	0,791	Mutlu
	0,585	0,059	0,723	0,646	0,771	Orta
	0,164	0,006	0,546	0,252	0,719	Mutsuz

Tablo 2' de bütün algoritmalar da en yüksek F ölçütü değerine sahip sınıfın mutlu sınıfı olduğu görülmüştür.

Tablo 3. RepTREE ile elde edilen karışıklık matrisi

Mutlu	Orta	Mutsuz	Sınıflar
19450	951	78	**Mutlu**
2253	3381	49	**Orta**
674	279	199	**Mutsuz**

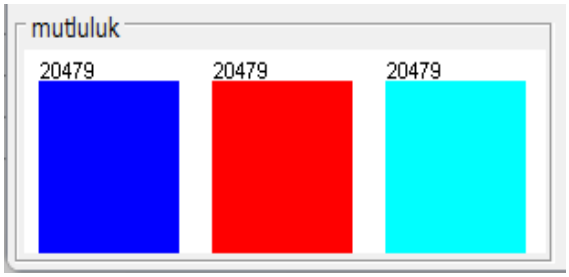
Tablo 3 incelendiğinde "mutlu" sınıfının en çok "orta" sınıfı ile yanlış sınıflandırıldığı, "orta" ve "mutsuz" sınıflarının en çok "mutlu" sınıfı ile karıştığı görülmektedir. Karışıklık matrisi incelendiğinde veri setinde "mutlu" sınıfı lehine dengesizliğin olduğu ve bu dengesizliğin algoritmaların sınıflandırma başarısını olumsuz yönde etkilediği görülmüştür.

3.2. Dengeli hale dönüştürülen veri setleri için sınıflandırma sonuçları

Üçlü Likert ölçekli veri seti üzerinde yapılan uygulama sonuçlarında sınıflar arasında dengesizliğin bulunduğu ve bu dengesizliğin algoritmaların sınıfları doğru sınıflandırmalarında olumsuz yönde etkilediği görülmüştür. Bu bölümde sınıflar arası dengesizliğin ortadan kaldırılması amacıyla önce en yüksek örnek sayısına sahip "mutlu" sınıfı örnek sayısı baz alınarak diğer sınıflarda veri tamamlama işlemi yapılmıştır. Ardından "orta" sınıfı örnek sayısı baz alınarak "mutlu" sınıfı üzerinde yeniden örnekleme (Resample) ve "mutsuz" sınıfı üzerinde veri tamamlama işlemi yapılmıştır. Son olarak "mutsuz" sınıfı örnek sayısı baz alınarak diğer sınıflar üzerinde yeniden örnekleme (Resample) işlemi yapılarak veri setleri dengeli hale dönüştürülmüştür.

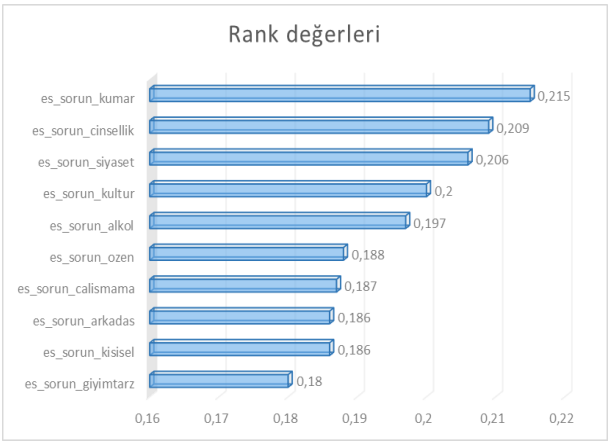
3.2.1. En yüksek örnek sayısına sahip mutlu sınıfı örnek sayısı baz alınarak yeniden örneklenen veri setinde mutluluk düzeyi sınıflandırma sonuçları

Veri setinde sınıf dengesizliğinin giderilmesi için "orta" ve "mutsuz" sınıflarına ait örnek sayıları artırılarak mutlu sınıfının örnek sayısına eşitlenmiştir. Şekil 3 incelendiğinde "mutlu" sınıfı örnek sayısı baz alınarak 5683 örnek sayısına sahip "orta" sınıfında 14796 adet ve 1152 örnek sayısına sahip "mutsuz" sınıfında 19327 adetlik veri tamamlama işlemi uygulanmıştır.



Şekil 3. En çok örneğe sahip sınıf örnek sayısı baz alınarak elde edilen veri setine ait mutluluk düzeyi sınıf frekansları

Şekil 3'te görüldüğü üzere sınıflar toplam örnek sayısı 61437 olmuştur. Oluşturulan veri seti WEKA 3.8.5 paket programında NumericToNominal filtresi kullanılarak sınıfı temsil eden özniteliğin kategorik olarak temsili sağlanmıştır. Ardından AttributeSelection filtresinde Information Gain (Bilgi Kazancı) algoritması ve Ranker (Sıralama) metodu kullanılarak her bir algoritma için en değerli 10 öznitelik seçilmiştir. Seçilen öznitelikler Şekil 4 ile sunulmuştur.



Şekil 4. En yüksek sınıf örneğine eşitlenmiş veri setinde mutluluk sınıflandırması için bilgi kazancı metodu ile seçilen 10 adet öznitelik

Şekil 4 incelendiğinde "eşler arası kumar sorunu" değişkeninin sınıflandırma başarısında en etkili öznitelik olduğu görülmüştür. Veri seti ile ilgili veri madenciliği algoritmalarının bazı temel parametrelere göre genel sınıflandırma sonuçları Tablo 4 ve sınıf bazlı sonuçları Tablo 5 ile gösterilmiştir.

Tablo 4. Algoritmaların sınıflandırma sonuçlarına ilişkin bilgiler (mutlu sınıfı örnek sayısı baz alınan, genel)

	Kappa	TP	FP	Kesin.	F ölçüt	ROC	Doğru.
J48	0,435	0,624	0,188	0,584	0,584	0,77	0,624

CHAID	0,459	0,639	0,18	0,591	0,553	0,783	0,639
CART	0,460	0,64	0,18	0,594	0,536	0,781	0,640
Nb TREE	0,443	0,629	0,189	0,585	0,582	0,782	0,629
Rand. TREE	0,411	0,608	0,196	0,582	0,588	0,682	0,607
Rep TREE	0,437	0,625	0,188	0,587	0,59	0,776	0,625

Tablo 4 incelendiğinde CART algoritmasının en yüksek doğrulukla sınıflandırma başarısı elde ettiği görülmüştür. Sınıf dengesizliği giderilmek üzere uygulanan işlem sonrası elde edilen sonuçlar, üçlü Likert ölçekli dengesiz veri seti sonuçlarıyla karşılaştırıldığında Kappa istatistiği, Kesinlik, F ölçütü, ROC alanı ve Doğruluk değerlerinde düşüş gözlemlenmiştir.

Tablo 5. Algoritmaların sınıflandırma sonuçlarına ilişkin bilgiler (mutlu sınıfı örnek sayısı baz alınan, sınıf bazlı)

	TP Oran	FP Oran	Kesin.	F -Ölçüt	ROC Alanı	Sınıf
J48	0,978	0,155	0,759	0,855	0,924	Mutlu
	0,22	0,143	0,435	0,292	0,618	Orta
	0,673	0,266	0,558	0,61	0,769	Mutsuz
CHAID	0,977	0,156	0,759	0,854	0,926	Mutlu
	0,073	0,044	0,455	0,126	0,635	Orta
	0,867	0,342	0,559	0,68	0,786	Mutsuz
CART	0,974	0,156	0,757	0,852	0,921	Mutlu
	0,036	0,02	0,469	0,066	0,636	Orta
	0,911	0,363	0,557	0,69	0,787	Mutsuz
Nb Tree	0,969	0,147	0,767	0,857	0,927	Mutlu
	0,18	0,12	0,429	0,253	0,634	Orta
	0,737	0,29	0,559	0,636	0,788	Mutsuz
Rep Tree	0,976	0,155	0,759	0,854	0,927	Mutlu
	0,237	0,15	0,441	0,308	0,621	Orta
	0,661	0,258	0,562	0,607	0,779	Mutsuz
Rand. Tree	0,981	0,161	0,753	0,852	0,922	Mutlu
	0,371	0,241	0,435	0,401	0,522	Orta
	0,47	0,186	0,558	0,511	0,603	Mutsuz

Tablo 5'te bütün algoritmalar da en yüksek F ölçütü değerine sahip sınıfın mutlu sınıfı olduğu görülmüştür.

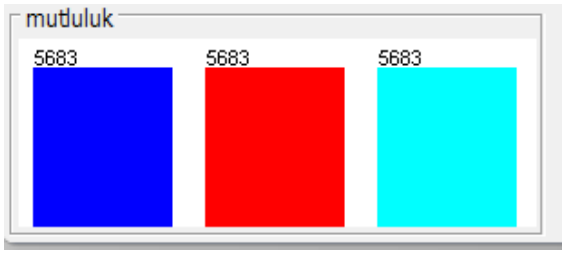
Tablo 6. CART algoritması ile elde edilen karışıklık matrisi

Mutlu	Orta	Mutsuz	Sınıflar
19952	199	328	**Mutlu**
5213	730	14536	**Orta**
1195	626	18658	**Mutsuz**

Tablo 6 incelendiğinde "mutlu" ve "mutsuz" sınıflarında yüksek oranda doğru sınıflandırma yapıldığı görülürken "orta" sınıfının en çok "mutsuz" sınıfı ile karıştığı görülmektedir.

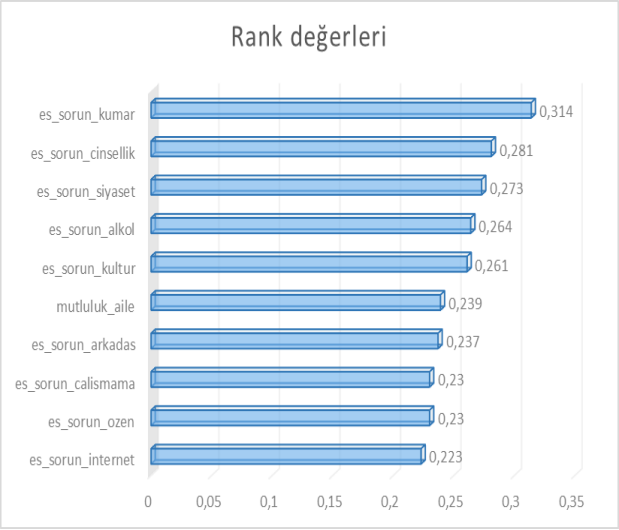
3.2.2. Orta mutluluk düzeyi sınıf örnek sayısı baz alınarak yeniden örneklenen veri setinde mutluluk düzeyi sınıflandırma sonuçları

Veri setinde "orta" sınıf örnek sayısı baz alındığında, "mutlu" sınıfı örnek sayısı yeniden örnekleme yöntemi yardımıyla "mutsuz" sınıfı örnek sayısı ise veri tamamlama yöntemi yardımıyla "orta" sınıfı örnek sayısına eşitlenmiştir. Şekil 5 incelendiğinde "orta" sınıfı örnek sayısı baz alınarak, 20479 örnek sayısına sahip "mutlu" sınıfı içerisinden rastgele yeniden örnekleme yöntemi ile 5683 adet alınmış ve 1152 örnek sayısına sahip "mutsuz" sınıfında 4531 adetlik veri tamamlama işlemi uygulanmıştır. Şekil 5'te görüldüğü üzere sınıfların toplam örnek sayısı 17049 olmuştur.



Şekil 5. Mutluluk düzeylerinden Orta sınıfı örnek sayısı baz alınarak yeniden örneklenen veri setine ait mutluluk düzeyi frekansları

Sınıf dengesizliğinin ortadan kaldırılması amacıyla oluşturulan veri setlerinden ikincisi olan "orta" sınıfı örnek sayısı baz alınarak oluşturulan veri seti WEKA 3.8.5 paket programında NumericToNominal filtresi kullanılarak sınıfı temsil eden özniteliğin kategorik olarak temsili sağlanmıştır. Ardından AttributeSelection filtresinde Information Gain (Bilgi Kazancı) algoritması ve Ranker (Sıralama) metodu kullanılarak her bir algoritma için en değerli 10 öznitelik seçilmiştir. Seçilen öznitelikler Şekil 6 ile sunulmuştur.



Şekil 6. Orta sınıf örneğine eşitlenmiş veri setinde mutluluk

sınıflandırması için bilgi kazancı metodu ile seçilen 10 adet öznitelik

Şekil 6 incelendiğinde "eşler arası kumar sorunu" değişkeninin sınıflandırma başarısında en etkili öznitelik olduğu görülmüştür. Veri seti ile ilgili veri madenciliği algoritmalarının bazı temel parametrelere göre genel sınıflandırma sonuçları Tablo 7 ve sınıf bazlı sonuçları Tablo 8 ile gösterilmiştir.

Tablo 7. Algoritmaların sınıflandırma sonuçlarına ilişkin bilgiler (orta sınıfı örnek sayısı baz alınan, genel)

	Kappa	TP	FP	Kesin.	F ölçüt	ROC	Doğru.
J48	0,677	0,785	0,108	0,798	0,78	0,873	0,784
CHAID	0,669	0,78	0,11	0,794	0,775	0,878	0,779
CART	0,678	0,785	0,107	0,8	0,781	0,876	0,785
Nb TREE	0,645	0,764	0,118	0,784	0,761	0,876	0,764
Rand. TREE	0,666	0,778	0,111	0,792	0,774	0,864	0,778
Rep TREE	0,674	0,783	0,109	0,797	0,779	0,877	0,783

Tablo 7 incelendiğinde CART algoritmasının bu dengeli veri setine ait sınıflandırma sonuçlarında da en yüksek doğrulukla sınıflandırma başarısı elde ettiği görülmüştür. Sınıf dengesizliği giderilmek üzere uygulanan işlem sonrası elde edilen sonuçlar, üçlü Likert ölçekli dengesiz veri seti sonuçlarıyla karşılaştırıldığında Kappa istatistiği ve ROC alanı değerlerinde artış, Kesinlik, F ölçütü ve Doğruluk değerlerinde düşüş gözlemlenmiştir.

Tablo 8. Algoritmaların sınıflandırma sonuçlarına ilişkin bilgiler (orta sınıfı örnek sayısı baz alınan, sınıf bazlı)

	TP Oran	FP Oran	Kesin.	F-Ölçüt.	ROC Alanı	Sınıf
J48 (C4.5)	0,897	0,2	0,691	0,781	0,866	Mutlu
	0,589	0,063	0,825	0,687	0,829	Orta
	0,867	0,06	0,878	0,873	0,924	Mutsuz
CHAID	0,9	0,206	0,687	0,779	0,874	Mutlu
	0,579	0,063	0,821	0,679	0,833	Orta
	0,86	0,062	0,874	0,867	0,927	Mutsuz
CART	0,894	0,201	0,69	0,779	0,868	Mutlu
	0,593	0,062	0,828	0,691	0,835	Orta
	0,869	0,06	0,879	0,874	0,926	Mutsuz
Nb Tree	0,902	0,236	0,656	0,76	0,874	Mutlu
	0,572	0,073	0,797	0,666	0,827	Orta
	0,816	0,045	0,9	0,856	0,926	Mutsuz
Rep Tree	0,896	0,204	0,687	0,778	0,87	Mutlu
	0,589	0,063	0,823	0,686	0,833	Orta
	0,863	0,059	0,88	0,871	0,927	Mutsuz
Rand. Tree	0,9	0,209	0,683	0,777	0,867	Mutlu
	0,594	0,075	0,799	0,681	0,826	Orta
	0,839	0,05	0,893	0,865	0,900	Mutsuz

Tablo 8'de daha önce yapılan diğer uygulama sonuçlarından farklı olarak bütün algoritmalar da en yüksek F ölçütü değerine sahip sınıfın "mutsuz" sınıfı olduğu görülmüştür.

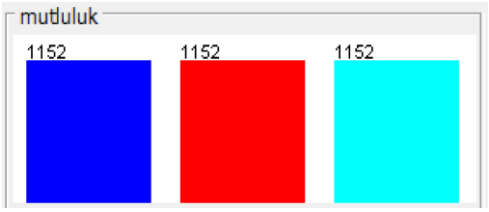
Tablo 9. CART algoritması ile elde edilen karışıklık matrisi

Mutlu	Orta	Mutsuz	Sınıflar
5079	343	261	**Mutlu**
1894	3369	420	**Orta**
386	359	4938	**Mutsuz**

Tablo 9 incelendiğinde "mutlu" ve "mutsuz" sınıflarında yüksek oranda doğru sınıflandırma yapıldığı görülürken "orta" sınıfının en çok "mutlu" sınıfı ile karıştığı görülmektedir.

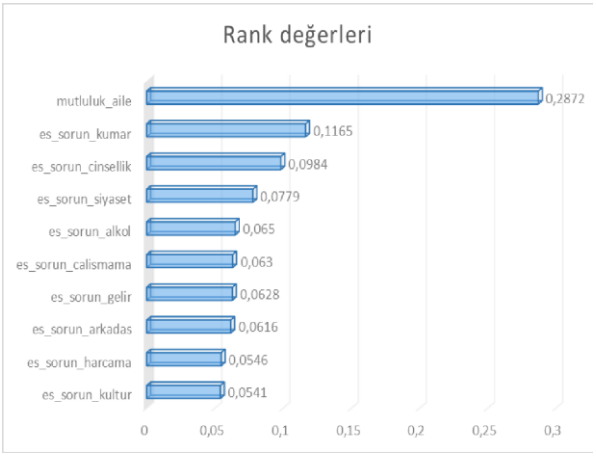
3.2.3. En düşük örnek sayısına sahip mutsuz sınıfı örnek sayısı baz alınarak yeniden örneklenen veri setinde mutluluk düzeyi sınıflandırma sonuçları

Veri setinde "mutlu" ve "orta" sınıfı örnek sayıları rastgele yeniden örnekleme yöntemiyle "mutsuz" sınıfının örnek sayısına eşitlenmiştir. Şekil 5.9 incelendiğinde "mutsuz" sınıfı örnek sayısı baz alınarak, 20479 örnek sayısına sahip "mutlu" sınıfı içerisinden 1152 adet örnek alınmış ve 5683 örnek sayısına sahip "orta" sınıfı içerisinden 1152 adet örnek alınarak veri seti oluşturulmuştur. Şekil 7'de görüldüğü üzere sınıfların toplam örnek sayısı 3456 olmuştur.



Şekil 7. Mutluluk düzeylerinden Mutsuz sınıfı örnek sayısı baz alınarak yeniden örneklenen veri setine ait mutluluk düzeyi frekansları

Veri seti üzerinde "mutsuz" sınıfı örnek sayısı baz alınarak oluşturulan dengeli veri seti WEKA 3.8.5 paket programında NumericToNominal filtresi kullanılarak sınıfı temsil eden özniteliğin kategorik olarak temsili sağlanmıştır. Ardından AttributeSelection filtresinde Information Gain (Bilgi Kazancı) algoritması ve Ranker (Sıralama) metodu kullanılarak her bir algoritma için en değerli 10 öznitelik seçilmiştir. Seçilen öznitelikler Şekil 8 ile sunulmuştur.



Şekil 8. Mutsuz sınıf örneğine eşitlenmiş veri setinde mutluluk sınıflandırması için bilgi kazancı metodu ile seçilen 10 adet öznitelik

Şekil 8 incelendiğinde "aile içi mutluluk" değişkeninin sınıflandırma başarısında en etkili öznitelik olduğu görülmüştür. Veri seti ile ilgili veri madenciliği algoritmalarının bazı temel parametrelere göre genel sınıflandırma sonuçları Tablo 10 ve sınıf bazlı sonuçları Tablo 11 ile gösterilmiştir.

Tablo 10. Algoritmaların sınıflandırma sonuçlarına ilişkin bilgiler (mutsuz sınıfı örnek sayısı baz alınan, genel)

	Kappa	TP	FP	Kesin.	F ölçüt	ROC	Doğru.
J48	0,480	0,654	0,173	0,704	0,642	0,762	0,654
CHAID	0,479	0,652	0,174	0,688	0,643	0,779	0,652
CART	0,485	0,657	0,172	0,693	0,648	0,772	0,657
Nb TREE	0,468	0,646	0,177	0,663	0,64	0,78	0,646
Rand. TREE	0,466	0,644	0,178	0,677	0,634	0,755	0,644
Rep TREE	0,482	0,655	0,173	0,69	0,646	0,775	0,654

Tablo 10 incelendiğinde CART algoritmasının bu dengeli veri setine ait sınıflandırma sonuçlarında da en yüksek doğrulukla sınıflandırma başarısı elde ettiği görülmüştür. Sınıf dengesizliği giderilmek üzere uygulanan işlem sonrası elde edilen sonuçlar, üçlü Likert ölçekli dengesiz veri seti sonuçlarıyla karşılaştırıldığında Kappa istatistiği, ROC alanı, Kesinlik, F ölçütü ve Doğruluk değerlerinde düşüş gözlemlenmiştir.

Tablo 11. Algoritmaların sınıflandırma sonuçlarına ilişkin bilgiler (mutsuz sınıfı örnek sayısı baz alınan, sınıf bazlı)

	TP Oran	FP Oran	Kesin	F-Ölçüt	ROC Alanı	Sınıf
J48	0,931	0,345	0,574	0,71	0,79	Mutlu
	0,574	0,138	0,675	0,62	0,75	Orta
	0,457	0,037	0,861	0,597	0,745	Mutsuz
CHAID	0,903	0,322	0,584	0,709	0,811	Mutlu
	TP Oran	**FP Oran**	**Kesin**	**F-Ölçüt**	**ROC Alanı**	**Sınıf**
	0,575	0,146	0,663	0,616	0,766	Orta
	0,48	0,054	0,817	0,604	0,761	Mutsuz
CART	0,911	0,329	0,581	0,709	0,802	Mutlu
	0,564	0,129	0,686	0,619	0,761	Orta
	0,494	0,057	0,813	0,614	0,755	Mutsuz
Nb Tree	0,843	0,286	0,596	0,698	0,809	Mutlu
	0,559	0,159	0,638	0,596	0,761	Orta
	0,535	0,087	0,755	0,626	0,77	Mutsuz
Rep Tree	0,905	0,325	0,582	0,709	0,806	Mutlu
	0,564	0,137	0,674	0,614	0,761	Orta
	0,494	0,056	0,814	0,615	0,757	Mutsuz
Rand. Tree	0,905	0,329	0,579	0,706	0,8	Mutlu
	0,551	0,143	0,658	0,6	0,738	Orta
	0,477	0,061	0,796	0,596	0,728	Mutsuz

Tablo 11'de bütün algoritmalar da en yüksek F ölçütü değerine sahip sınıfın mutlu sınıfı olduğu görülmüştür.

Tablo 12.CART algoritması ile elde edilen karışıklık matrisi

Mutlu	Orta	Mutsuz	Sınıflar
1050	47	55	**Mutlu**
426	650	76	**Orta**
332	251	569	**Mutsuz**

Tablo 12 incelendiğinde "mutlu" sınıfının yüksek oranda doğru sınıflandırıldığı, "orta" ve "mutsuz" sınıfının en çok "mutlu" sınıfı ile karıştığı görülmektedir. Üçlü Likert ölçekli dengesiz veri seti ile bu veri setinin yeniden örnekleme ve veri tamamlama yöntemleri kullanılarak dengeli hale getirilmiş iki farklı versiyonunda, CART algoritmasına ait sınıf bazlı sonuçlar Tablo 13 ile sunulmaktadır. Tablo 13 incelendiğinde, Likert ölçekli veri setinde bulunan sınıflara ait TP metrik oranları ve F-ölçütü değerleri karşılaştırıldığında "mutlu" sınıfının en yüksek başarı ile "mutsuz" sınıfının ise en düşük başarı ile sınıflandırılan kategori olduğu görülmektedir. Likert ölçekli dengesiz veri setine ait sonuçlar ile sınıflar arasındaki dengesizliğin giderilmesi amacıyla oluşturulan veri setlerine ait sonuçlar karşılaştırıldığında, "Orta" sınıfının örnek sayısının baz alındığı veri setinde, CART algoritmasının "mutsuz" sınıfını en yüksek F-ölçütü ve ROC alanı değeriyle en başarılı şekilde sınıflandırdığı görülmüştür.

Tablo 13. Üçlü Likert ölçekli veri setlerinde CART algoritmasına ait sınıf bazlı sonuçları

Veri Seti	TP Oran	FP Oran	Kesinlik	F-Ölçüt	ROC Alanı	Sınıf
Üçlü Likert Ölçekli Dengesiz Veri Seti	0,949	0,427	0,869	**0,908**	**0,783**	**Mutlu**
	0,593	0,057	0,733	0,656	0,771	**Orta**
	0,18	0,005	0,597	0,276	0,754	**Mutsuz**
	0,974	0,156	0,757	**0,852**	**0,921**	**Mutlu**

"Mutlu" Sınıfı Örnek Sayısı Baz Alınan Dengeli Veri Seti	0,036	0,02	0,469	0,066	0,636	**Orta**
	0,911	0,363	0,557	0,691	0,787	**Mutsuz**
"Orta" Sınıfı Örnek Sayısı Baz Alınan Dengeli Veri Seti	0,894	0,201	0,69	0,779	0,868	**Mutlu**
	0,593	0,062	0,828	0,691	0,835	**Orta**
	0,869	0,06	0,879	**0,874**	**0,926**	**Mutsuz**
"Mutsuz" Sınıfı Örnek Sayısı Baz Alınan Dengeli Veri Seti	0,911	0,329	0,581	**0,709**	**0,802**	**Mutlu**
	0,564	0,129	0,686	0,619	0,761	**Orta**
	0,494	0,057	0,813	0,614	0,755	**Mutsuz**

4. Tartışma ve Sonuç

Bu çalışmanın amacı, veri madenciliği algoritmalarının üçlü Likert (Ordinal) ölçeğe sahip veri seti üzerindeki sınıflandırma performansını ölçmek, sınıflar arası dengesizliğin olduğu veri seti üzerinde performanslarını karşılaştırmak ve bu dengesizliğin giderildiği durumda algoritmaların sınıflandırma performansının nasıl değiştiğini incelemektir. Bu amaç doğrultusunda, Türkiye İstatistik Kurumu Başkanlığı tarafından yürütülen Türkiye Aile Yapısı Araştırması (TAYA) verileri kullanılmıştır. İki aşamada gerçekleşen deneylerden ilkinde, genel memnuniyeti ifade eden öznitelikler arasından en değerli 10 öznitelik seçilmiştir.

İkinci aşamada ise mutluluk sınıflandırması değerlendirilmiş; yeniden örnekleme ve veri tamamlama yöntemleri ile sınıflandırma başarı yüzdesinin nasıl değiştiği incelenmiştir. "Genel mutluluk durumu düşünüldüğünde hangisi ailenizi en iyi ifade eder?" özniteliği dengesiz veri setinde ve "mutsuz" sınıfı örnek sayısı baz alındığında yeniden örnekleme ile oluşturulan dengeli veri setinde en yüksek rank değerine sahip olmuştur. Benzer olarak "mutlu" sınıfı ve "orta" sınıfı örnek sayısı baz alınarak oluşturulan dengeli veri setleri için en yüksek rank değeri "Eşinizle kumar alışkanlığı sebebiyle ne sıklıkla sorun yaşadınız?" özniteliğine aittir. Öznitelik seçiminin ardından, veri madenciliği teknikleri kullanılarak sınıflandırma analizi gerçekleştirilmiştir. Dengesiz veri setinde mutluluk düzeyi, özniteliği çeşitli algoritmalar ile sınıflandırılmış ve sınıflar arası dengesizliğin ve kategori sayısının az olduğu veri seti türünde RepTREE algoritmasının daha başarılı sınıflandırma yaptığı gözlemlenmiştir. Sınıflandırma başarısının yanı sıra Kappa istatistiği ve dengesiz veri setleri için F-ölçütü parametreleri de karşılaştırılmıştır. Kappa istatistiği, tüm algoritmalar için sınıflar arasında orta düzeyde uyum olduğunu göstermiştir. F-ölçütü açısından ise CART, CHAID ve RepTREE algoritmalarının diğer algoritmalara göre daha doğru sınıflandırma yaptığı belirlenmiştir. "Mutlu" sınıfı örnek sayısı baz alınarak diğer sınıflar üzerinde veri tamamlama yöntemiyle dengeli hale dönüştürülen veri seti için mutluluk düzeyi özniteliği aynı algoritmalar ile sınıflandırılmış ve dengeli veri seti türünde CART algoritmasının daha başarılı sınıflandırma yaptığı gözlemlenmiştir. "Orta" sınıfı örnek sayısı baz alınarak diğer sınıflar üzerinde veri tamamlama ve yeniden örnekleme yöntemleriyle dengeli hale dönüştürülen veri seti için mutluluk düzeyi özniteliği aynı algoritmalar ile sınıflandırılmış ve CART algoritmasının daha başarılı sınıflandırma yaptığı gözlemlenmiştir. "Mutsuz" sınıfı örnek sayısı baz alınarak diğer sınıflar üzerinde yeniden örnekleme yöntemiyle dengeli hale dönüştürülen veri seti için mutluluk düzeyi özniteliği aynı algoritmalar ile sınıflandırılmış CART algoritmasının daha başarılı sınıflandırma yaptığı gözlemlenmiştir. Yapılan deneylere göre çalışmada ulaşılan temel bulgular aşağıda özetlenmiştir;

- CART algoritması, üçlü Likert ölçekli dengesiz veri seti hariç tüm durumlarda en başarılı sınıflandırma sonuçları vermiştir. Anket verilerinin sınıflandırılması probleminde CART algoritmasının kullanılmasının daha uygun olacağı düşünülmektedir.
- Likert ölçekli veri setlerinde sınıflar arası dengesizlik giderildiğinde, FP metriğinin dengesiz veri setlerine göre daha düşük olduğu dolayısıyla dengeli veri setlerinde I. Tip hata yapma oranı daha düşük olduğu gözlemlenmiştir.
- Sınıflar arası dengesizlik söz konusu olduğunda kesinlik parametresi kategori sayısı az olan veri setinde daha anlamlı olduğu görülmüştür. Dengesizliği giderilen veri setlerinde sınıf örnek sayısı ortalama örnek hacmine yakın olan veri setinde daha yüksek kesinlik değeri elde edilmiştir.

Ayrıca, gerçekleştirilen deneyler sayesinde Türk aile yapısının demografik çerçevesi ve evli bireylerin mutluluk düzeylerine katkıda bulunan faktörler incelenmiştir. Elde edilen rank değerlerine göre, evli çiftlerin kişisel mutluluk düzeylerini etkileyen faktörlerin başında aile içi mutluluk düzeyi ve kumar alışkanlığından dolayı sorun yaşanması olduğu görülmüştür. Ailesi mutlu olan bireylerin çoğunlukla kişisel olarak mutlu olduğu gözlemlenmiştir. Ailesi mutsuz olan bireyler de ise eşler arasında kültürel farklılıklar, siyasi görüş farklılıkları, cinsel uyumsuzluk ve alkol alışkanlığı durumları bireyin kişisel mutluluğunu olumsuz yönde etkileyen faktörler olarak belirlenmiştir. Bu çalışma, veri madenciliği teknikleri kullanılarak dengeli ve dengesiz Likert ölçekli veri setlerinin sınıflandırılmasının mümkün olduğunu göstermektedir. Türkiye Aile Yapısı Araştırması verilerinin, dengesiz veri setleri için geliştirilen diğer algoritmalarla incelenmesi hususu başka çalışmalara konu olarak önerilmektedir.

Ayrıca edinilen bilgilere dayalı olarak anket sorularında yapılabilecek iyileştirmeler veya eklenebilecek yeni sorular ile sınıflandırma başarısının artırılmasına olanak sağlayacağı düşünülmektedir.

Teşekkür

Veri setinin çalışmada kullanılmasına olanak sağladığı için Türkiye İstatistik Kurumu Başkanlığı'na teşekkür ederiz.

Etik Beyanı/Declaration of Ethical Code

Çalışmanızı sunacağınız dile uygun olarak aşağıdaki ifadelerin makalede yer alması gerekmektedir.

Bu çalışmada, "Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesi" kapsamında uyulması gerekli tüm kurallara uyulduğunu, bahsi geçen yönergenin "Bilimsel Araştırma ve Yayın Etiğine Aykırı Eylemler" başlığı altında belirtilen eylemlerden hiçbirinin gerçekleştirilmediğini taahhüt ederiz.

Kaynakça

[1] Salzberg, S. L., Searls, D. B., Kasif, S. 1998. Computational Methods in Molecular Biology. Amsterdam: Elsevier Sciences B.V.,368.

[2] Gundecha, P., Liu, H. 2012. Mining Social Media: A Brief Introduction, In Informs TutORials in Operations Research, 1-17.

[3] Tan, P. N., Steinbach, M., Kumar, V. 2006. Introduction to Data Mining. Pearson: 1st Edition, Wesley, Boston, 719.

[4] Zaki, M. J., Meira, Jr. W. 2014. Data Mining and Analysis: Fundamental Concepts and Algorithms. 1st Edition, Cambridge: Cambridge University Press, 660.

[5] Ozer, P., Sprinkhuizen-Kuyper, I. G. 2008. Data algorithms for classification. Radboud University Nijmegen, Artificial Intelligence, BSc Thesis, Netherlands.

[6] García, E., Romero, C., Ventura, S., Calders, T. 2007. Drawbacks and solutions of applying association rule mining in learning management systems. CEUR Workshop Proceedings, 305, 13-22.

[7] Srivastava, A., Han, E. H., Kumar, V., Singh, V. 1999. Parallel Formulations of Decision-Tree Classification Algorithms. Data Mining and Knowledge Discovery, 3, 237–261.

[8] Bresfelean, V. 2007. Analysis and Predictions on Students' Behavior Using Decision Trees in Weka Environment, 29th International Conference on Information Technology Interfaces, Cavtat, Croatia, 51- 56.

[9] Mardikyan, S., Badur, B. 2011. Analyzing Teaching Performance of Instructors Using Data Mining Techniques. Informatics in Education, 10 (2), 245–257.

[10] Kuzey, C. 2012. Veri madenciliğinde destek vektör makinaları ve karar ağaçları yöntemlerini kullanarak bilgi çalışanlarının kurum performansı üzerine etkisinin ölçülmesi ve bir uygulama. İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, Doktora Tezi, İstanbul.

[11] Şehribanoğlu, S., Diler, S. 2016. Veri madenciliği süreçleri ve karar ağaçlari algoritmalari ile bir uygulama. Van Yüzüncü Yıl Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, Van.

[12] Beernaert, B. 2021. Using machine learning techniques for analyzing survey data. Ghent University, Faculty of Science, Master Thesis, Belgium.

[13] Bezek-Gürü, Ö., Şevgin, H., Kayri, M. 2023. Reviewing the Factors Affecting PISA Reading Skills by Using Random Forest and MARS Methods. International Journal of Contemporary Educational Research, 10(1), 181-196.

[14] Ogedengbe M. T., Junaidu S. B., Kana A. F. D., 2024. Association Rule Mining on Likert's Scale Data using a Novel Attributes Pruning Technique. International Journal of Scientific Research in Computer Science and Engineering, 12(4), 54-65.

[15] Aile, Çalışma ve Sosyal Hizmetler Bakanlığı. 2018. Türkiye Aile Yapısı İleri İstatistik Analizi, Ankara.

[16] Silahtaroğlu, G. 2013. Veri Madenciliği Kavram ve Algoritmaları. Papatya Yayınevi, 333.

[17] Sullivan, W. 2017. Machine Learning for Beginners Guide Algorithms: Supervised & Unsupervised Learning. Decision Tree & Random Forest Introduction. CreateSpace Independent Publishing Platform.

[18] Damanik, I. S., Windarto, A. P., Wanto, A., Andani, S. R., Saputra, W. 2018. Decision Tree Optimization in C4.5 Algorithm Using Genetic Algorithm. The International Conference on Computer Science and Applied Mathematics, October 10-12, Indonesia.

[19] Gupta, G. 2014. A Self-Explanatory Review of Decision Tree Classifiers. IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014), May 09-11, India, 1–7.

[20] Gavankar, S. S., Sawarkar, S. D. 2017. Eager Decision Tree. 2nd International Conference for Convergence in Technology (I2CT), 07-09 April, Mumbai, 837–840.

[21] Lu, Z., Wu, X., Bongard, J. C. 2015. Active Learning through Adaptive Heterogeneous Ensembling.

IEEE Transactions on Knowledge and Data Engineering, 27(2), 368–81.

[22] Lim, T. S., Loh, W. Y., Shih, Y. S. 2000. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. Machine Learning, 40(3), 203–28.

[23] Hssina, B., Merbouha, A., Ezzikouri, H., Erritali, M. 2014. A Comparative Study of Decision Tree ID3 and C4.5. International Journal of Advanced Computer Science and Applications, 4(2),13-19.

[24] García Laencina, P. J., Abreu, P. H., Abreu, M. H., Afonso, N. 2015. Missing Data Imputation on the 5-Year Survival Prediction of Breast Cancer Patients with Unknown Discrete Values. Computers in Biology and Medicine, 59, 125–33, Doi: 10.1016/j.compbiomed.2015.02.006.

[25] Behera, H. S., Mohapatra, D. P. 2015. Computational Intelligence in Data Mining. Volume 1: Proceedings of the International Conference on CIDM, 5-6 December, 410s.

[26] Kass, G. V. 1980. An Exploratory Technique for Investigating Large Quantities of Categorical Data. Applied Statistics, 29(2), 119–127.

[27] Kohavi, R. 1996. Scaling up the Accuracy of Naive Bayes Classifiers: A Decision-Tree Hybrid. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, AAAI Press, 202–207.

[28] Nor, W. M. H., Salleh, M., Omar, A. H. 2013. A Comparative Study of Reduced Error Pruning Method in Decision Tree Algorithms. IEEE International Conference on Control System, Computing and Engineering (ICCSCE), 392–397.

[29] Pfahringer, B. 2010. Random Model Trees: An Effective and Scalable Regression Method. University of Waikato, Department of Computer Science, Working Paper, New Zealand.