

Karar Ağaçlarının Aşırı Uyumla Sınavı: Karşılaştırmalı Bir Performans Analizi

Araştırma Makalesi/Research Article

 Gökhan KORKMAZ¹

¹Şırnak Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, İşletme Bölümü, Şırnak, Türkiye

gokhankorkmaz@istanbul.edu.tr

(Geliş/Received:09.12.2024; Kabul/Accepted:25.06.2025)

DOI: 10.17671/gazibtd.1594957

Özet— Genelleme kabiliyetine ket vuran aşırı uyum probleminden Yapay Sinir Ağları (YSA) yönteminde metriklerin optimizasyonu yoluyla, Destek Vektör Makinelerinde (DVM), Vapnik-Chervonenkis (VC) Katsayısı yardımıyla, Karar Ağaçlarında (KA) ise ön budama (Pre-Pruning) ve son budama (Post-Pruning) yöntemiyle kaçınılmaya çalışılmaktadır. Bu çalışmada Gini ve Entropi ve Logaritmik Kayıp Fonksiyonu (LKF) indeksli KA modelleri, budamasız, ön budama sonrası ve son budama sonrası olmak üzere üç ayrı evrede, gerçek hayattan alınan bir kredi veri kümesine ve yine bu veriden türetilmiş “dengelenmiş” veri kümesine, üçüncül olarak da bir tarım verisine uygulanmış ve sonuçları YSA, DVM ve Lojistik Regresyon (LR) sınıflandırıcılarının sonuçları ile mukayese edilmiştir. Uygulamada YSA ve KA metrikleri Bayes (Optuna) ile optimize edilmiştir. Verilerin hepsine SMOTE dengeleme yöntemi uygulanmıştır. KA, her üç evrede tüm indeksleriyle birlikte, her üç veri kümesinde diğer yöntemlerinkinden belirgin bir şekilde daha iyi sonuçlar vermiştir. Sonuçlar Friedman’la test edilmiş ve modeller arasındaki sonuç farkları her üç veri kümesinde de anlamlı çıkmıştır. Üç veri kümesinin ikisinde indeks sonuçları arasındaki fark istatistiksel olarak anlamlı değilken gözlem sayısı çok olan (11069) veri kümesinde ise anlamlı çıkmıştır. Çalışmanın sonuçlarına bakıldığında KA’nın kriterlerin tamamında (doğruluk, standart sapma, hassasiyet, geri çağırma, F1 ve AUC_ROC) verdiği sonuçların kalitesi göz önünde bulundurulduğunda güçlü ve pratik bir sınıflandırıcı olarak ön plana çıkmaktadır. Fakat ön budama ve son budama yöntemlerinin KA’nın genelleme kabiliyetine herhangi bir katkısının olmadığı sonucuna bu çalışma sonucunda varılmıştır.

Anahtar Kelimeler— karar ağaçları, yapay sinir ağları, destek vektör makineleri, sınıflandırma, budama

Testing Decision Trees with Overfitting: A Comparative Performance Analysis

Abstract— The overfitting problem, which hampers generalization capability, is addressed through metric optimization in Artificial Neural Networks (ANN), with the aid of the Vapnik-Chervonenkis (VC) coefficient in Support Vector Machines (SVM), and via pre-pruning and post-pruning methods in Decision Trees (DT). In this study, DT models indexed by Gini, Entropy, and Logarithmic Loss Function (LLF) were applied in three separate phases—before pruning, after pre-pruning, and after post-pruning—to a real-world credit dataset, a “balanced” version of this dataset, and, thirdly, an agricultural dataset. Their results were compared with those of the classifiers ANN, SVM, and Logistic Regression (LR). In the implementation, ANN and DT metrics were optimized using Bayesian (Optuna) optimization, and SMOTE balancing was applied to all datasets. Across all three datasets and indices, DT consistently delivered significantly better results than the other methods in all three pruning phases. The differences in model results were tested with the Friedman test and found to be statistically significant in all three datasets. While the differences in index results were not statistically significant in two of the datasets, they were significant in the dataset with a large number of observations (11069). Reviewing the study’s findings, it can be concluded that DT stands out as a strong and practical classifier, considering the quality of the results it produced across all criteria (accuracy, standard deviation, precision, recall, F1, and AUC_ROC). However, this study found that pre-pruning and post-pruning methods did not contribute to the generalization capability of DT.

Keywords— decision trees, artificial neural networks, support vector machines, classification, pruning

1. GİRİŞ (INTRODUCTION)

İş süreçlerinde ve hayatın genelinde elde edilen başarılı sonuçları, verilen doğru kararlara; başarısız sonuçları ise verilen yanlış kararlara bağlamak çok yanlış olmasa gerektir. Geleceği doğru bir formda inşa edebilmek için doğru kararlar alabilmek, hedefe ulaşması gereken bir lokomotif rayına oturtabilmek kadar önemli bir faaliyet olduğunu ayrıca belirtmek gerekir. Burada cevap aranması gereken önemli bir soru da: “tek bir doğru karar mı vardır?”

Doğru kararın/kararların sayısı bir yana, “doğru karara/kararlara giden yollar, onları almak zorunda olan insan sayısı kadar çoktur” da denilebilir. Peki doğruluğu bu kadar önemli olan bu kararlar nasıl alınır? Kararlar, iş süreçleri açısından, genellikle; geçmiş benzer vakalar var ise ve bu vakalar adına çözüme dönük etkin tecrübe birikimi var ise ve yine bu vakalara dair faydalı araştırma sonuçları varsa tüm bunların üzerine uzman kişinin mevcut duruma dönük yaklaşımı da dikkate alınarak bir karar kombinasyonu oluşturulur ve bu düğünden bir sonuç kararı çıkarılır. Çözülen vakaların sayısı ve çözüm önerilerine dair araştırma sayısı günümüz teknolojik ve bilişimsel imkanların sayesinde hızla artmaktadır. Bundan dolayı yeni alınan kararların daha isabetli olması beklenebilir. Yalnız karar vericiler için bu teknolojik imkanların getirdiği avantajların yanı sıra yine aynı yaygın teknolojik imkanlar doğrultusunda veri miktarında yaşanan muazzam artışın getirdiği “veri işleme” ihtiyacı, özellikle büyük veri yığınları için daha kompleks bir süreç dönüşerek önemli bir dezavantaj halini de alabilmektedir. İşte tam burada, büyük verileri modelleyerek verinin yükünden karar vericinin kurtulmasını sağlayacak, bir başka deyişle bu veri yığınının damıtılmış sonuç mesajını okuyabilecek karar destek sistemlerine, bir başka ifadeyle sınıflandırma ve tahmin araçlarına şiddetle ihtiyaç duyulmaktadır. Günümüzde bu çalışmalar genel anlamda “veri madenciliği” olarak adlandırılmaktadır. Burada karar vermenin/verebilmenin yanı sıra, YSA başta olmak üzere kara kutu (black box) olarak adlandırılan modellerin aksine, alınan bu kararın hesabının verilebilmesini de sağlayabilecek “şeffaflık” iş süreçleri açısından oldukça önem taşımaktadır. Alınan bu kararın nasıl, niçin ve hangi açıklayıcı değişkenlerin öncüllüğünde alındığının açıklanabilmesi, sağlık ve bankacılık başta olmak üzere birçok sektör için hayati önem taşımaktadır. Şeffaflık, üretilen kararın uygunluğunun ölçümü noktasında da önemli bir parametredir [1]. Bu arada literatürde, kara kutu olarak bilinen modellerin şeffaflığını artırmaya dönük çalışmaların sayısı hızla artmakta ve günden güne önemli bir araştırma sahasına dönüşmektedir.

Günümüzde SQL vb. bir veri tabanlarında depolanan ve konuya dönük modelleme çalışmaları için hammadde görevi gören büyük veri yığınlarının oluşturulması, sistematik bir biçimde depolanması, gerektiğinde erişiminin sağlanması “veri tabanı yönetim sistemleri” aracılığıyla sağlanmakta ve özel sektörde önemli bir iş koluna dönüşmektedir. Veri hacmi ve veri tabanı sayısındaki bu patlama, verilerin işlenmesi ve anlamlı (yararlı) çıktılara dönüştürülmesi süreçlerinin

otomatikleştirilmesini zorunlu kılmaktadır. Bu da veri madenciliğini akademik anlamda önemli bir araştırma sahasına, özel sektörde ise önemli bir uzmanlık alanına dönüştürmektedir.

Sınıflandırma ve tahmin, önemli veri sınıflarını tanımlayan modelleri çıkarmak veya gelecekteki veri eğilimlerini tahmin etmek için kullanılabilen başlıca iki veri analizi yöntemidir. Bu tür analizler, bize büyük veriler hakkında daha iyi bir anlayış sağlamaya yardımcı olabilir. Sınıflandırmada, birkaç sonuç kategorisinden birisine gözlemlerin ataması yapılmaktadır. Yani sürekli değerler arasından bir sayısal değeri değil; ayrık, sırasız yani sürekli sınıflardan birisini tahmin ederken, tahmin modelleri ise sürekli olan sayısal değerler arasından muhtemel en iyi sonucu seçen fonksiyonlardır [2], [3]. Sınıflandırma, analiz öncesinde kategorik sınıf isimleri hakkında varsayımlarda bulunmak, eğitim kümeleri ve sınıfların etiketleri temelinde bilgiyi tasnif etmek ve yeni elde edilebilir verileri sınıflandırmak için kullanılabilir [4].

Makine öğrenimi sahasında şimdiye dek birçok sınıflandırma ve tahmin yöntemi geliştirilmiştir. Bu yöntemler zaman içerisinde doğal seleksiyona uğrayarak belli başlı olan uygulamaların daha ön plana çıktığı söylenebilir. Bunlar YSA, DVM, K-En Yakın Komşu, Bayes, Genetik Algoritma, K-Ortalamalar, KA geleneksel anlamda yani istatistiki olarak ise Diskriminant Analizi (DA) ve LR gibi yöntemlerdir. Çoğu algoritma esasen bellekte yerleşiktir ve genellikle küçük bir veri boyutunu kapsamaktadır. Son veri madenciliği araştırmaları ise veri ambarlarında biriken yüksek hacimli verileri işleyebilen ölçeklenebilir sınıflandırma ve tahmin teknikleri geliştirerek bu tür çalışmaların boyutlarını daha da genişletmektedir. Sınıflandırmada, vakalar farklı gruplara yerleştirilir. Bu metodolojiyle geliştirilen modeller daha önce modellenmemiş olan yeni vakalar için sınıflandırma kurallarını oluşturur [2].

Veri madenciliği aslında yapay zekanın ve istatistiğin veri analizi için özelleştirilmiş olan kısmını teşkil ettiği söylenebilir. Yapay zekâ çalışmalarının altında, insanın yaşamını devam ettirmesindeki en önemli yetilerinden olan “zekâyı” modelleme yaklaşımı yatar. Yani, yapay zekâ çalışmalarının amacı, haddini aşan bir ifadeyle “işlevsel anlamda bir insan yaratmak” olduğu söylenebilir. Bu da insanı insan yapan zekâyı modellemekle eşdeğerdir.

Zekâ nedir ve ne işe yarar vb. konular, insana dair olgular olduğu için, genellikle bir fikir birliğine varılmış konular değildir. Fakat zekanın işlevselliği; insanların öncelikle çevresel değişimleri “algılama”, algı yoluyla elde edilen bu verilerin işlenmesi anlamına gelen “muhakeme” ve bu muhakeme sonucunda yapılması gerekenin belirlendiği bir “karar verme” süreci şeklinde açıklanabilir. Bundan sonraki aşamalar ise bir bakıma, zekaya değil insanın kas gücüne kalmaktadır. Dikkat edilirse, zekâ söz konusu olduğunda, üç önemli unsur ön plana çıkmaktadır. Bunlar; algılama, muhakeme ve karar vermedir. Konumuza tekrar dönecek olursak zekânın fonksiyonlarından ilki olan algılama aşamasını, veri tabanı yönetim sistemlerine,

muhakeme aşamasını ise veri madenciliği çalışmalarına ve karar verme sürecini de veri madenciliği aşamasında kullanılan makine öğrenmesi yöntemlerinin çıktılarını yorumlama aşamasına benzetebiliriz.

Karar vermeye dönük olan iş süreçlerin otomatikleştirilmesinden bahsetmiştik. Zekaya dönük olan bu süreçler insanda otomatiktir ve oldukça hızlıdır. Yapay zekanın hızı da Alan Turing'in ilk bilgisayarından bu yana yıllar içerisinde devinerek arttı. Bu arada yapay zekâ modelleri öğrendi ve öğrendikçe insana yakın ve hatta artık insan gruplarından daha iyi ve daha hızlı sonuçlar vermeye başladı. Bugün Google Translate gibi bir tercüme altyapısına ve gücüne insan olarak erişmek mümkün değildir. Çünkü bu yazılım internet üzerindeki tüm tercüme tecrübesini kullanarak çeviri yapmaktadır.

Yapay zekâ, tüm bu süreçleri modellemeler sayesinde insandan yani doğal zekadan öğrendi. Bu öğrenme ve modelleme sürecinin tersi de mümkündür. Veri madenciliği teknikleri, öğrenmeye dayalı teknikler olduğu için eğitim alanında da öğrenme sürecine ilişkin anlayışımızı geliştirmek için kullanılabilir ve öğrencilerin öğrenme sürecinde değişkenleri belirlemeye, çıkarmaya ve değerlendirmeye çalışırken bu tekniklerin yönteminden de ilham alınabilir [25]. Bu yaklaşım tüm makine öğrenimi yöntemleri için mümkün olabileceği gibi bu çalışmada ayrıntılı olarak ele alınan ve insanın karar verme sürecini en benzer şekilde modelleyen KA modelleri için de geçerlidir.

Özetle günümüz iş süreçlerinde doğruluk seviyesi yüksek kararlar verebilmek için, adeta bir tür petrol gibi yüzeye vuran veri yığınlarının içerisindeki bilgiye erişebilecek ve bu bilgilerden karar çıktısı oluşturabilecek etkin veri madenciliği süreçlerine, makine öğrenmesi yöntemlerine ve sınıflandırma ve tahmin algoritmalarına ihtiyaç duyulmaktadır. Bu çalışmada kredi riskini değerlendirme (credit scoring) problemi bir sınıflandırma problemine dönüştürülerek çözülmüştür. Söz konusu problem literatürde ön plana çıkan sınıflandırma yöntemlerinden olan YSA, DVM, KA ve LR uygulanmıştır. Literatürde bu yöntemlerin, bu problemin çözümüne dönük çok sayıda uygulaması bulunmaktadır. Bu çalışmada KA yönteminin Gini ve Entropi İndeksleri ve LKF uygulanmış ve performans sonuçları uygulanmış ve adı geçen diğer yöntemlerle sonuçları mukayese edilmiştir. Normalde LKF indeksten çok bir düzeltme faktörü olarak kullanılır. Bu çalışma da gini ve entropi indekslerinin yanı sıra düzeltme faktörlü üçüncü bir uygulama (indeks) olarak ele alınmış ve raporlarmıştır. LKF teamül açısından, Python programında da gini ve entropi indekslerinin bir üçüncüsü olarak da ele alınmaktadır ve yöntemin doğruluğuna katkıda bulunmaktadır.

Yapay zekanın bir alt kümesi olan makine öğrenmesi yöntemlerinin özellikle en çok kullanılanları, sağlam matematiksel arka plana sahip, verinin içerisinde yer alan ilişkileri modelleyebilecek, önemli, gelişmiş ve birçok sahadaki uygulama sonuçlarıyla kendini kanıtlamış yetenekli yöntemlerdir. Fakat her modelin olduğu gibi bu

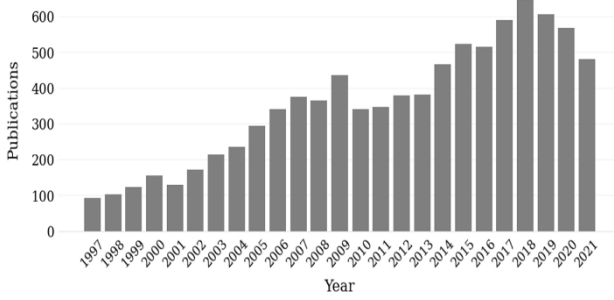
modellerin de dezavantajları bulunmaktadır. Bu dezavantajların en önemlisi ise, algoritmaların bu gücünden kaynaklanan “eğitim verisine aşırı uyum” (overfitting) problemidir. Bu yöntemlerin hepsi, kendi içerisinde farklı yaklaşımlarla ve uygulamalarla, aşırı uyumla mücadele etmekte ve kısmen başarılı olmaktadır. Fakat bu sorunun tam anlamıyla aşıldığı söylenemez. Dolayısıyla yapacağımız bu çalışma bir anlamda bu yöntemlerin aşırı uyumla olan mücadelesindeki başarılarını da mukayese etmek anlamına da geldiğini ifade etsek yanlış olmayacaktır.

Bu çalışmada öncelikle KA; yaklaşımıyla, işleyişi ve kullanım alanları, çeşitleri, avantajları ve dezavantajlarıyla ayrıntılı olarak ele alınmış sonrasında KA'nın gini ve entropi ve LKF indeksli algoritmaları, bir bankadan alınan kredi verilerine genel ve dengelenmiş olarak bir de tarım sektöründen üçüncü bir veri kümesine “budamasız”, “ön budamalı” ve “son budamalı” olarak uygulanmış ve tüm uygulama sonuçları YSA, DVM ve LR yöntemlerinin sonuçlarıyla mukayese edilmiştir.

Bu çalışmada gini, entropi ve LKF indekslerinin yanı sıra LKF yardımıyla veri kümesindeki gözlemlerin yani kredi müşterilerinin “riskli” ve “risksiz” şeklinde etiketlenen sonuç sınıflarına ataması yapılmış böylelikle; hem KA'nın kendi içinde kullanılan indeksler arasında hem de KA dışındaki makine öğrenmesi ve geleneksel istatistiki yöntemler arasında performans mukayesesi yapılarak indeksler ve modeller arasında, özellikle makine öğrenmesi yöntemlerinin tamamının başını ağrıtan aşırı öğrenme problemiyle mücadele konusunda, en iyi olanlarını tespit etmek amaçlanmıştır. Buradaki en iyiden kasıt, sadece sonuçların doğruluk oranları değil aynı zamanda çapraz doğrulamalar arasındaki standart sapmanın temsil ettiği model stabilizasyonu, modellerin zaman maliyeti, aşırı uyumla mücadeledeki başarısı ve uygulama kolaylığı da göz önünde bulundurulmuştur.

2. LİTERATÜR TARAMASI (LITERATURE REVIEW)

Şekil 1, bize KA'nın yıllar itibariye kullanımına ve yaygınlığına dair önemli fikirler vermektedir. Normalde bir yöntem tanındıktan kısa bir süre sonra zirve yapmakta ve zamanla, yöntem geliştirilmeye açık yani esnek bir yöntem değilse gelişen, diğer yöntemlerin gölgesinde kalarak kullanımı giderek azalmaktadır. Fakat Şekil 1'den de anlaşılacağı üzere KA, yıllar geçtikçe kullanımı giderek artan ve altmış yılı aşkın geçmişine rağmen kullanımı yaygın bir şekilde devam eden, eskimemiş, esnek (geliştirilmeye açık) bir sınıflandırma ve tahmin aracı olduğu söylenebilir [5].



Şekil 1. Karar ağaçlarının literatürdeki kullanım sıklığı
(Frequency of use of decision trees in literature)

Açıklama: Şekil 1, “TI=(tree*) AND TS=(classif* OR regress*OR decision)” arama terimiyle bulunan yayın sayısı, yayın yılına göre düzenlenmiştir. Bu sonuçlar yalnızca arama terimini ve mekâna göre bir filtrelemeyi (Bilgisayar Bilimi veya Mühendislik kategorileri) dikkate alır, bu nedenle uygulamalar da dahil edilmiştir. Grafik Web of Science veri tabanı aracılığıyla elde edilmiştir [5].

Biró ve Nédá (2020), gini ve indekslerinin matematiksel benzerliklerini ve bağlantılarını ele aldıklarını ve bu ikisini birbirine bağlayabilen yeni bir ölçü sunduklarını ve geliştirdikleri bu ölçünün sosyolojik ve ekonomik karmaşıklığın farklı yönlerini nicelleştirmede uygun şekilde hizmet edebileceği fikrini belirtmişlerdir. Mussard, Syete ve Terraza (2003), gini indeksinin entropiden çıkarılanlardan daha kullanışlı bir ölçüm olduğunu belirtmişlerdir [6].

Rahmatia, Avand, Yariyan, Tiefenbacher, Azareh ve Bui (2022), Gini, Entropi ve Oran Tabanlı Sınıflandırma Ağaçlarına (OTSA) (Ratio-Based Classification Trees – RBCT) dayalı yeni bir yaklaşımı denediklerini ve gini'nin en iyi sonuçları elde ettiğini, bunu entropinin takip ettiğini ve sonrasında OTSA'nın izlediğini tespit etmişler her üç tabanlı algoritmanın da inceledikleri problemde çok iyi olduğunu ayrıca eklemişlerdir [2].

Ryu ve Slottje (2023), entropi ölçülerine dayalı olarak iyi bilinen gini katsayısını türetmenin 10 yolunu göstermişler ve bundan sonraki çalışmalarda Entropi Tabanlı Gini (ETG) (Entropy-Based Gini - EBG) indeksinin kullanılabileceğini belirtmişlerdir. Huikun, Heping, Tao, Jun ve Deti (2016), Entropi Değer Yöntemi (Entropy Value Method – EVM) ve Dagum Gini katsayısına dayalı kırsal mekân fonksiyonunun zamansal-mekânsal örüntülerinin analiz etmişler ve iyi sonuçlar elde ettiklerini belirtmişlerdir [7].

Pal ve Mather (2003), KA'nın mimari oluşturma ve eğitim maliyetlerinin yüksek olduğu YSA uygulamasından avantajlı olduğunu ve tek değişkenli çalışmalar için sonuçların çok büyük veriler hariç belirgin derecede iyi olduğunu çok değişkenliye geçildiğinde ise bu belirginliğin azaldığını ve iş yüksek boyutlu verilere geldiğinde hem tek hem çok değişkenlide YSA ve maksimum olabilirlik sınıflandırıcıları kadar iyi performans göstermediğini tespit etmişlerdir [8].

Xu, Watanachaturaporn, Varshney ve Arora (2005), yumuşak sınıflandırma için Karar Ağacı Regresyon (KAR) (Decision Tree Regression - DTR) algoritmasını görüntü işlemede kullanmışlar ve sonuçların son derece ümit verici olduğunu belirtmişlerdir [9].

Zhang, Ma, Ock ve Qing (2022), istenmeyen çıktıları dikkate almak için Süper Verimlilik Boşluk Tabanlı Ölçüm (SVBTÖ) (Super Efficiency Slacks-Based Measure - SBM) modeli ve Dagum Gini katsayısı yönteminin bir kombinasyonunu kullanarak, 2005'ten 2018'e kadar Çin il panel verilerine dayalı endüstriyel yeşil teknoloji inovasyonunun verimliliğini ve bölgesel farklılıkları ölçtüklerini ve Dagum Gini katsayısı ayrıştırmasının, endüstriyel fikri mülkiyet haklarının (GTIE)'nin genel farkının 2005'ten 2018'e kadar dalgalanan bir düşüş eğilimi gösterdiğini belirtmiştir. Dagum Gini Katsayısı literatürde önemli yaygın bir katsayı olarak benimsenmiş ve bu katsayıyı kullanan çok sayıda çalışmaya rastlanmıştır [10].

Chakrabarty ve Cauwenberghs (2007) DVM (DVM) büyük marjlı sınıflandırmasını GiniDVM Maksimum Entropi Çok Sınıflı Olasılık Regresyonuna (MEÇSOR) (Quadratic Entropy Based Robust Multi-Class Probability Regression - QEBRMCPR) genişleterek etkili sonuçlar ürettiklerini belirtmişlerdir [11]. Wang, Peng ve Xi (2019), kurtosis, negatif entropi, gini indeksi ve Pürüzsüzlük İndeksi (Smoothness index) indeksleriyle makine arızalarının önceden tespiti ile alakalı veriler üzerinden gerçekleştirdikleri çalışmada, tüm bu indeksleri birleştirerek oluşturdukları indeksle daha iyi sonuç aldıklarını raporlamışlar ve bu endeksler arasındaki temel farkı, normalleştirilmiş kare zarfa sırasıyla farklı ağırlıkların uygulanması şeklinde yorumlamışlardır [27].

Jost (2006), ham çeşitlilik endekslerinin makul sonuçlar veremeyeceğini ancak endeksleri veya entropileri etkin tür sayılarına dönüştürerek genel örtüşme ölçülerinin, Jaccard İndeksi, Sørensen İndeksi, Horn Örtüşme İndeksi ve Morisita-Horn İndeksi olabileceğini belirtmiştir [12].

Daniya, Geetha ve Kumar (2020), kategorik hedefin, Entropi seçimini kullandığını, gini indeksinin bir bölme kuralı olduğunu Ve Regresyon Ve Karar Ağaçlarının (CART) performansını artırdığını belirtmişlerdir [13].

Fogel ve Feder (2018), LKF açısından toplu öğrenme sorununu ele almışlardır [14].

Vovk (2015), Olasılıksal tahmin çalışmalarında kullanılan standart üç kayıp fonksiyonu bulunmaktadır. Bunlar, LKF, Brier Kayıp Fonksiyonu (BKF) ve Küresel Kayıp Fonksiyonudur (KKF). Kayıp fonksiyonları arasında en seçici olanın LKF olduğunu fakat herhangi bir hesaplanabilir uygun kayıp fonksiyonu kullanılabileceğini ve veri dizisi hangi kayıp fonksiyonu altında oluşturulmuşsa o fonksiyonun uygun sonuç için kullanılması gerektiğini belirtmiştir [15].

Sonuç olarak literatürde gini ve entropi indeksleri başta olmak üzere çok farklı türde indeksler, çok fazla sayıda kullanılmıştır. Fakat bilinen temel formundaki gini ve entropi indekslerinin günümüzde de yaygın olarak kullanılmaya devam ettiği, bunun dışında Dagum İndeksi, Jaccard İndeksi, Sørensen İndeksi, Horn Örtüşme İndeksi ve Morisita-Horn İndeksi gibi daha özelleştirilmiş indekslerinde, literatürde tanındığı, bilindiği, kullanıldığı ve çalışma anlamında çok önemli sayılarda takipçilerinin bulunduğu şeklinde özetlenebilir.

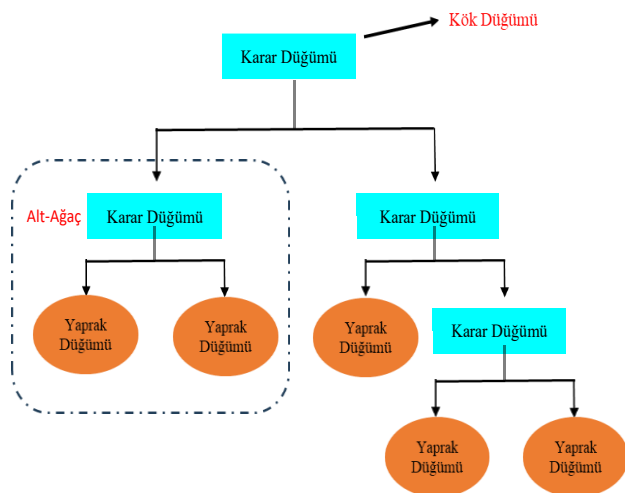
3. YÖNTEM (METHOD)

Bu çalışmada bir bankadan (gerçek hayattan) alınan kredi verilerine gini, entropi ve LKF özelinde KA uygulanmış ve sonuçları YSA'nın Çok Katmanlı Ağlar (ÇKA) (Multi Layer Perceptron – MLP) modelinin ve LR modelinin sonuçlarıyla mukayese edilmiştir.

3.1. Karar Ağaçları (Decision Trees)

3.1.1. Tanım (Definition)

Bir ağacın kökü, dalları ve nihayet yaprakları vardır. KA da bu şekilde şematize edilebilir. Kökte bir tane düğüm dallarda ve yapraklarda ise çok sayıda düğüm vardır. Düğümden kasıt bir niteliğin testidir. Test sonucu dal ve sınıfın üzerine etiket olarak yansır. Sonucu ise yapraktır. Kök düğüm tıpkı anayasa gibi ana kararı temsil eder ve bundan sonra verilecek tüm kararlar ana karar sonucu altında verilir. Kök düğüm Şekil 2'de görülebileceği, üzere ağaçtaki en üst düğümdür. Bir karar ağacında, her düğüm bir özelliği (niteliği), her bağlantı (dal) bir kararı (kural) ve her etiketin (yaprak) bir sonucu (kategorik veya sürekli değer) gösterdiği bir diyagram bir kararlar silsilesinin yol haritasıdır. Tüm fikir, tüm kararlar için böyle bir ağaç oluşturmak ve her yaprakta tek bir sonucu işlemektir [16].



Şekil 2. Karar Ağacının Yapısı (www.javatpoint.com)
(The Structure of Decision Trees)

KA, iç kısımdaki düğümlerin (kararların) sistematik sıralanışından oluşmaktadır. Bu düğümler mantıksal

testleri içerir. Uygulama her gözlem için düğümün kökünden başlayıp yapraklardan birinde son bulacak şekilde tasarlanır [3].

3.1.2. İşleyiş (Operation)

Denetimli öğrenmede iyi bir sınıflandırma ve regresyon modeli olan karar ağaçlarının işleyişini anlamak için öncelikle KA terminolojisini bilmek gerekir [17].

Kök Düğüm: KA kök düğüm ile başlar. İlk düğüm ilk karar, ilk nitelik testidir. Daha sonra iki veya daha fazla homojen kümeye bölünen tüm veri kümesini temsil eder.

Yaprak Düğüm: KA yaprak düğümle sona erer. Yani çıktı düğümdür. Bundan sonra budaklanma olmaz.

Bölme: Bölme, karar düğümünü/kök düğümünü verilen koşullara göre alt düğümlere bölme işlemidir.

Dal/Alt Ağaç: Ağacı bölerek oluşturulan bir ağaç.

Budama: Gereğinden fazla budaklanma aşırı karmaşıklığa bu da veri kümesini ezberlemeye varacağı için budama yapılır. Aşırı dallanma engellenir.

Üst/Alt Düğüm: Ağacın kök (ilk) düğüme üst düğüm, diğer düğümlere ise alt düğümler denir.

KA'nın işleyişi Şekil 2'de gösterilmektedir.

Kaynak: www.javatpoint.com

Adım-1: Ağacı, tüm veri setini içeren kök düğüm olan S ile başlatın.

Adım-2: Nitelik Seçim Ölçüsü'nü (NSÖ) kullanarak veri setindeki en iyi niteliği bulun.

Adım-3: S'yi, en iyi nitelikler için olası değerleri içeren alt kümelerine bölün.

Adım-4: En iyi niteliği içeren karar ağacı düğümünü oluşturun.

Adım-5: Adım -3'te oluşturulan veri setinin alt kümelerini kullanarak yinelemeli olarak yeni karar ağaçları oluşturun. Düğümleri daha fazla sınıflandıramayacağınız ve son düğümü yaprak düğüm olarak adlandıracağınız bir aşamaya ulaşana kadar bu işlemi sürdürün [18].

Karar ağaçlarının oluşturulmasındaki en önemli adım ağaçtaki dallanmanın hangi kritere veya kıstasa göre yapılacağı ya da hangi öz nitelik değerlerine göre ağaç yapısının oluşturulacağıdır. Literatürde bu problemin çözümü için geliştirilmiş çeşitli yaklaşımlar vardır. Bunlardan en önemlileri bilgi kazancı ve bilgi kazanç

oranı, gini indeksi, entropi indeksi gibi yöntemler olup bunun dışında birçok yöntem daha mevcuttur [19].

KA’nda nitelik seçimi ölçüsü, tuple’ların belirli bir düğümde nasıl bölüneceğini belirler ve bu nedenle bölme kuralları olarak da bilinir. Bölme nitelikleri sürekli değerli olabilir veya ikili ağaçlarla sınırlı olabilir. Sürekli değerli nitelikler için, bölme ölçütünün bir parçası olarak bir bölme noktası belirlenmelidir. İkili ağaçlar için ise bir bölme alt kümesi belirlenmelidir. Bölümleme için ağaç düğümü bölme kriteri ile etiketlenir, her bir kriter sonucu için dallar yetiştirilir ve tuple’lar buna göre bölümlendirilir [20].

Bilgi kazanımı şeklinde isimlendirilen faaliyet adından da anlayacağımız üzere, bir özellik tarafından sonuca sağlanan sınıflandırma katkısı şeklinde tanımlanabilir. Çalıştırma esnasında düğüm bölünür ağaç; bilgi kazanımı en yüksek olan özelliğe ait düğümü ilk önce açarak bilgi kazanımını mümkün olan en üst seviyeye çıkarır. Bilgi kazanımı entropi indeksi ile birlikte anılır. Matematiksel ifadesi şu şekildedir:

3.1.2.1. Entropi İndeksi (Entropy Index)

“Entropi, fizik, matematik, bilişim ve ilgili alanlarda dengeyi, dağılmayı, maksimum olasılık durumlarını ve bilginin optimum sıkıştırılmasını tanımlamak için kullanılır” [6]. “Entropiler kendi başlarına çeşitlilik değildir. Bunların etkin tür sayısına dönüştürülmesi, çeşitliliğin birleşik ve sezgisel bir yorumunun anahtarıdır” [12].

Entropi indeksi, veri setindeki keyfiliğin miktarını ölçer ve şu şekilde temsil edilir.

$$Entropi(s) = -P(evet) \log_2 P(evet) - P(hayır) \log_2 P(hayır)$$

ve s “örneklem sayısını”, P(evet) “evet alma şansını” ve P(hayır) “hayır alma şansını” ifade eder.

Bilgi Kazanımı:

$$Bilgi Kazanımı = Entropi(s) - [Ağırlıklı Ortalama * Entropi(Her Özellik)]$$

3.1.2.2. Gini İndeksi (Gini Index)

“Gini indeksi, bir toplumdaki sosyal ve ekonomik eşitsizlikler için yerleşik bir ölçüdür” [6].

Bu indeks bir karar ağacı oluşturulurken kullanılan “saflığı” ya da “safsızlığı” ölçmek içindir. KA karar alırken gini indeksi küçük olan nitelikler tercih edilir ve matematiksel formülü şu şekildedir [17]:

$$Gini Endeksi = 1 - f(x) = 1 - \sum_j p_j^2$$

“Ham çeşitlilik endeksleri çok çeşitli matematiksel davranışlar sergilediğinden, doğrudan genel bir çeşitlilik denklemine veya formülüne eklendiğinde hepsi makul sonuçlar veremez. Ham endeksleri gerçek çeşitliliklere dönüştürmek, çeşitliliği içeren anlamlı endeksten bağımsız genel denklemlerin, ölçümlerin veya formüllerin oluşturulmasını mümkün kılar” [12].

3.1.2.3. Logaritmik Kayıp Fonksiyonu (Logarithmic Loss Function)

Logaritmik kayıp (aynı zamanda çapraz entropi kaybı olarak da bilinir) metriği, sınıflandırma algoritmalarının performansını değerlendirmek için makine öğrenimi uygulamaları genelinde yaygın olarak kullanılır [21]. “Olasılıksal tahmin literatüründe kullanılan standart kayıp fonksiyonları log kayıp fonksiyonu, Brier kayıp fonksiyonu ve küresel kayıp fonksiyonudur; ancak, tahmin algoritmalarının karşılaştırılması için herhangi bir hesaplanabilir uygun kayıp fonksiyonu kullanılabilir” [15]. Bu kayıp fonksiyonlarından en çok kullanılanı yine LKF’dur [28].

$$Logloss = \frac{1}{N} \sum_{i=1}^N logloss_i$$

$$Logloss = -\frac{1}{N} \sum_{i=1}^N [y_i \ln p_i + (1 - y_i) \ln(1 - p_i)]$$

Bu çalışmada KA’nın gini ve entropi indeksleri ve LKF kullanılmıştır.

Gupta, Rawat, Jain, Arora ve Dhami (2017) yaptıkları çalışmada, veri madenciliğinde kullanılan çeşitli karar ağacı algoritmalarını analiz etmişler ve sonuç olarak her algoritmanın kendine özgü avantajları ve dezavantajları olduğunu bulmuşlardır. Burada önemli bir ölçü de şudur ki; algoritmaların verimliliği ayrıca algoritma tarafından karar ağacının harcanan zaman bilgisine de bağlıdır. Makale ayrıca bu öznelik seçim ölçülerinin hesaplanması için yöntemler de vermektedir. Halbuki YSA tarafında mevcut problem için kaç tane gizli katman, her katmanda ne kadar nöron kullanılırsa daha iyi sonuç elde edileceğine dair (araştırmacının YSA oluştururken bu ve buna benzer yedi parametreyi deneme yanılma yoluyla optimize etmesi gerekmektedir. Sonuç olarak, çalışmada karar ağacı tümevarımına yönelik bu algoritmaların duruma ve veri yapısına göre farklı türlerinin kullanılması gerektiğini belirtmişlerdir. Bu da tüm algoritma sonuçlarına bakılması gerekliliğini beraberinde getirmektedir [20].

3.1.3. Avantajları (Advantages)

KA yaklaşık altmış yıldır, tüm yapay zekâ modellerindeki çok yönlü ilerlemelere rağmen önemini ve etkinlik önceliğini kaybetmemiş, basitliği ve uygulanabilirliğiyle kullanımı giderek artan (bkz: Şekil 1) en popüler yöntemlerden birisidir. Model, mantıksal testlerin yerleştirildiği iç düğümlerin sistematik sıralanışından oluşmaktadır. Her gözlem için uygulama düğümünün kökünden başlar ve yapraklardan birinde son bulur. En önemlisi ise, YSA başta olmak üzere, diğer yöntemlerin aksine tüm süreç şeffaftır. Çok basit olduğu için güçlü bir yöntem değilmiş gibi algılanır. Düşük hesaplama maliyetleri, eksik değerlerle etkin bir biçimde başa çıkabilmesi, karışık verilerin kullanımına dönük işlenebilmesi ve son yıllarda ön plana çıkan şeffaflık noktasındaki karnesi yöntemin önemli avantajları arasındadır [3].

KA teknikleri, insan muhakemesine çok benzer. Gerçek hayatta da insanlar bir ana karar ve bu ana kararın ardından birçok alt karar oluşturarak uygulama modellerini oluştururlar. Bu yüzden anlaşılması çok kolaydır bu yüzden de kullanımı çok yaygındır [22]. KA’nda hem ayrık hem de sürekli değişkenler hedef değişkenler veya bağımsız değişkenler olarak kullanılabilir [23]. Bir başka deyişle KA ile hem sınıflandırma hem tahmin yapılabilir.

Karar ağaçlarını, hiyerarşik yapısı sayesinde, sınıflandırma kurallarına dönüştürmek oldukça kolaydır. Karar ağacının öğrenimi, bir öge hakkındaki gözlemleri ögenin hedef değeri hakkındaki sonuçlara eşleyen bir tahmin modeli olarak bir karar ağacı kullanma şeklindedir. Hedef değişkenin sonlu (kategorik) bir değer kümesi alabildiği ağaç modellerine sınıflandırma ağaçları denir, bu ağaç yapısında yapraklar sınıf etiketlerini, dallar ise bu sınıf etiketlerine yol açan özelliklerin birleşimlerini temsil eder. Karar ağacı, diğer sınıflandırma yöntemlerine kıyasla nispeten hızlı bir şekilde oluşturulabilir. SQL ifadeleri, veri tabanlarına etkili bir şekilde erişmek için kullanılabilen ağaçtan oluşturulabilir. KA sınıflandırıcıları, diğer sınıflandırma yöntemleriyle karşılaştırıldığında benzer veya daha iyi doğruluk elde eder [24].

Veri türü kısıtlaması yoktur. Hem sayısal hem de kategorik yapıya sahip değişkenleri esnek bir şekilde işleyebilir. Daha az veri temizliği gerekir. Karşılaştırmalı olarak daha az veri temizleme adımı gerektirir ve eksik değerlerden ve verilerden etkilenmez [17].

Şeffaflıklarının bir sonucu olarak yorumlanabilirlikleri nedeniyle, bir analist bir karar ağacını inceleyebilir ve sınıfları birbirinden ayıran önemli faktörleri belirleyebilir. Bu bağlamda, karar ağaçları keşifsel veri analizi için yararlı araçlardır ve giriş verileri, özellik alanı ve sınırlayıcı sınıflar arasındaki hiyerarşik ilişkileri incelemek için kullanılabilir. Özellikle, karar ağacı tahmin algoritmaları genellikle hızlıdır ve giriş verilerindeki gürültüye duyarsızdır ve bu nedenle uzaktan algılama arazi

eşleme problemlerinde bulunan büyük hacimli verileri sınıflandırmak için önemli bir faydaya sahiptir [25], [17].

KA algoritmaları kesinlikle parametrik değildir ve dolayısıyla girdi verilerinin dağılımıyla ilgili hiçbir varsayımında bulunmaz ve girdi özellikleri ile sınıf etiketleri arasındaki doğrusal olmayan ve gürültülü ilişkiler açısından esnek ve sağlam bir yapıya sahiptir [25]. Eksik verileri işlemenin yaygın, ancak yanlış bir yöntemi, eksik değerleri olan durumları hariç tutmaktır; bu hem verimsizdir hem de analizde önyargı oluşturma riski taşır. KA analizi, eksik verilerle iki şekilde başa çıkabilir: eksik değerleri diğer kategorilerle analiz edilebilecek ayrı bir kategori olarak sınıflandırabilir veya çok sayıda eksik değeri olan değişkeni, tahmin yapmak ve bu eksik olanları tahmin edilen değerle değiştirmek için hedef değişken olarak ayarlayan bir yerleşik karar ağacı modeli kullanılabilir [23].

Karar ağaçları optimalite noktasında başarısız olabilir. Optimal karar ağaçlarını optimal bir şekilde algoritma geliştirmede başarısızlığın, karar ağacı sorununun altında yatan doğayı doğru bir şekilde ele almadaki başarısızlığa atfedilebilir. Yani diğer birçok modelde olduğu gibi elde edilen başarısız sonuç araştırmacıların yanlış seçimlerinden de kaynaklanıyor olması muhtemeldir. Her modelin kendine özgü bir doğası, uyumlu olduğu bir veri yapısı, sınırları, yatkın olduğu parametreler ve başa çıkabildiği açıklayıcı değişken sayısı gibi faktörler bulunabilmektedir. Araştırmacıların bu sınırları daha geniş bağlamda ele alacak olursak, modellerin avantajlı ve dezavantajlı oldukları yönlerini dikkate alarak model seçimleri ve modellerini konfigüre etmeleri gerekmektedir [5].

KA’nın avantajları şu şekilde özetlenebilir:

- Basittir,
- Sınıflandırıcı oluşturması kolaydır,
- Gerçek hayattaki insani süreçleri modellediği için anlaşılması kolaydır,
- Sınıflandırma ve tahmin yapılabilir,
- Esnektir,
- Zaman maliyeti düşüktür,
- Şeffaftır, yorumlanabilir ve öne çıkan değişkenler belirlenebilir,
- Yaygındır,
- Varsayım testi gerektirmez, pratiktir. Bu da zaman maliyetini azaltır.
- Büyük veri yığınlarına uygulanabilir,
- Aşırı uyuma neden olan aşırı dallanma çeşitli algoritmalarla budanabilir,
- Sağlamdır,
- Verideki gürültüden etkilenmez/ az etkilenir. Her türlü veri (sayısal ya da kategorik) işlenebilir,
- Eksik verilerle başa çıkabilir,
- Parametrik değildir. Doğrusal olmayan örüntüleri de değerlendirebilir.

3.1.4. Dezavantajları (Disadvantages)

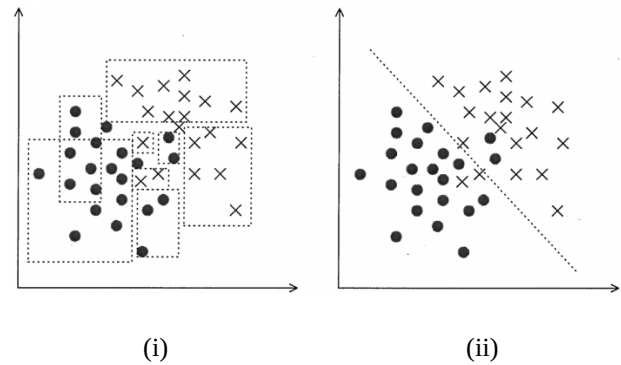
Makine öğrenimi yöntemlerinin geneli için çok önemli bir problem var ki, aşırı uyum diğer adıyla aşırı öğrenme (over fitting) olarak adlandırılır. Örnek vermek gerekirse bir YSA modeli oluştururken gizli katman sayısına araştırmacının deneme yanılma yoluyla optimal anlamda bir karar vermesi gerekir. Yalnız genel itibariyle basit içerikli problemlerde tek gizli katman yeterli olabilmektedir. Fakat biraz daha ileri eğrisel (nonlinear) diyebileceğimiz karmaşık içerikli veri yapılarında iki gizli katman yeterli olabilmektedir. Hemen her problemde üçe çıktığında modelin performansının düştüğü bilinmektedir. Elbette yedi ya da sekiz gizli katman kullanmak veya gizli katmanlardaki nöron sayısını yükseltmek mümkündür. Aynı durumu regresyon formülünün aşırı karmaşık yani bağımsız değişkenlerin üslerinin çok yükseltildiği bir senaryoda gözlemlemek mümkün olacaktır. Fakat bu şekilde yöntemin/yöntemlerin performansı daha da düşecektir. Bunun sebebi aşırı uyumdur. Model eğitim verisindeki örüntüleri öğrenir. Sonra verideki gürültüyü de öğrenir. Tabir yerindeyse veriyi ezberler. Bu da model bir başka veri setine uygulandığında eğitim verisine bağımlı hale gelmiş modelin performansının düşmesine sebep olur. Oysa ki, “iyi bir model, eğitim verisi üzerinde öğrenme derinliği ile yeni bir veriye uygulandığındaki genelleme yetisiyle arasında iyi bir denge kurabilmelidir. Aşırı uyum KA da dahil tüm yöntemlerin en önemli sorunudur ve hepsinin bu sorunu aşabilmek adına çok önemli varyasyonları bulunmaktadır. Vapnik ve Chervonenkis’in VC katsayısı buna en önemli örnektir ve teşbihte hata olmasın bir nevi Occam’ın usturasına giydirilmiş matematiksel bir kılıftır. Bunun tersi de geçerlidir. Karmaşık içerikli bir probleme tek gizli katmanlı bir YSA modeli uyguladığımızda optimize edilmiş bir sonuç alamayacağımız açıktır. Çünkü problem karmaşık, çözüm modeli ise basittir. Bu durumda da eksik öğrenme (under fitting) problemi yaşanır. Eksik uyum (eksik öğrenme) genel itibariyle yöntemler için çok sorun olmamakta fakat aşırı uyum uygulamacıların başını çok ağrıtabilmektedir [29]. Bir sinir ağının ağırlıklarını uydurmak veya en küçük karar ağacını bulmak, her ikisi de zor problemlerdir [26].

Yukarıdan aşağıya doğru tümevarım yolunu kullanan yöntemlerin önemli bir handikabı, eğer iyi bir genelleme kabiliyeti olsun isteniyorsa budama gerektirmesidir. Yani budama yoluyla aşırı uyumun engellenmesi gerekir. Budama yukarıdan aşağıya doğru ağaç büyütülürken ağaç karmaşıklığındaki cezaları göz önünde bulunduramamasından dolayı gereklidir. Çünkü güçlü bölünmeler daha zayıf bölünmelerin arkasında gizlenebilir. Burada karmaşıklık cezası çok yüksekse ve ilk daha zayıf bölünmenin seçilmesini engelliyorsa bu aşağı yönlü tümevarımla en iyi ağacın bulunamaması sonucunu doğuracaktır. Bu sorunun önüne geçmek için mümkün olan yaklaşım karmaşıklık cezasını kullanarak ağacı budamadan önce ağacı olabildiğince derine büyütmektir. Böylelikle zayıf bölünmelerin güçlü bölünmeleri perdelemesinin önüne geçilir. Fakat bu da eğitimin iki aşamada gerçekleştiği, bir dizi açgözlü kararlar büyüdüğü

ve ardından budama yapıldığı anlamına gelir. Karar ağacı yöntemleri, her bir bölmeyi izole bir şekilde oluşturmak için sezgisel yöntemleri yinelemeli olarak uygular; fakat bunun, veri kümesinin altta yatan özelliklerini iyi yakalayabildiğinin, bir başka deyişle modelin sezgiselliğinin iyi olduğunun bir garantisi yoktur [5].

Aşırı uyum, karar ağacı modelini etkileyen en önemli pratik sorunlardan biridir. Ancak, modelin parametre kısıtlamalarını budayarak ve ayarlayarak aşırı uyum sorunları kısmen azaltılabilir. KA, doğası gereği sürekli olan değişkenler için uygun değildir. Karar ağacı, değişkenleri çeşitli kategorilere ayırırken değerli bilgilerin bir kısmını kaybeder [17].

Şimdiye kadar sunulan algoritmalar, uygulanması kolay olduğu (çünkü bir seferde yalnızca bir özellik analiz edilir) ve türetilen karar ağacının anlaşılması nispeten kolay olduğu için çekici olan tek değişkenli bölümlere yöntemlerini kullanmaktadır. Tek değişkenli bölümlere yöntemlerinin yanı sıra, arama alanını eksen paralelinde bir seferde yalnızca bir özneliğe göre bölümlenmeyen, ancak bir öznelik kombinasyonuna dayalı eğik bölüm sınırları oluşturan bazı başarılı bölümlere yöntemleri de vardır. Eğik bölümlere, tek değişkenli yöntemlere uygulanabilir bir alternatif sunar. Tek değişkenli muadillerinin aksine, eğik bölümler, aşağıdakilerin kombinasyonları ile oluşturulur. Çok değişkenli yapıları nedeniyle, eğik yöntemler arama alanını bölümlendirmede çok daha fazla esneklik sunar; ancak bu esneklik aşırı uyuma neden olan daha yüksek karmaşıklık bedeliyle gelir (Bkz. Şekil 3)[24]. Özetlemek gerekirse, tek değişkenli bölümlere, basit, anlaşılır ve çoğu problem için kullanışlıyken çok değişkenli uygulamalara gibi geçildiğinde aşırı uyumdan kaçınmak tek değişkenliye oranla daha zor bir hal almaktadır.



Şekil 3. Sınıf etiketlerine karşılık gelen nitelik değerlerini ve şekli gösteren eksenler verildiğinde (i) eksen paralel ve (ii) eğik karar sınırları [24].

(Given axes representing the attribute values and shape corresponding to the class labels, (i) axis-parallel and (ii) oblique decision boundaries)

KA uygulanırken değişkenler arasında çok güçlü bir korelasyon mevcut ise bu korelasyon modelin istatistiklerini iyileştirir fakat bu arka planda veriden bilgi kazanımını artırmaz ve sonuçla nedensel olarak ilişkili

olmayan değişkenlerin seçilmesine yol açabilmesidir. Bu konuda ağaç oluşturulurken dikkatli olunmalıdır [23], [24].

KA'nın dezavantajlarını özetlemek gerekirse:

- Aşırı uyum problemi tamamen aşılmış değildir,
- Sürekli değişkenler için uygun değildir,
- Değişkenler arasındaki korelasyon güçlü ise dikkatli olunmalıdır.

4. UYGULAMA (APPLICATION)

4.1. Veri (Data)

Uygulama bir bankadan (gerçek hayattan) alınan kredi müşterilerinin verisine ve bir tarımsal veri kümesine uygulanmıştır. Analiz kredi verisinin “genel” haline ve verinin riskli müşteri sayısı ile risksiz müşteri sayısının eşitlendiği “dengelenmiş” veri kümesi olmak üzere iki kredi veri kümesine ayrı ayrı uygulanmış yani toplamda üç veri kümesi üzerinde çalışılmıştır.

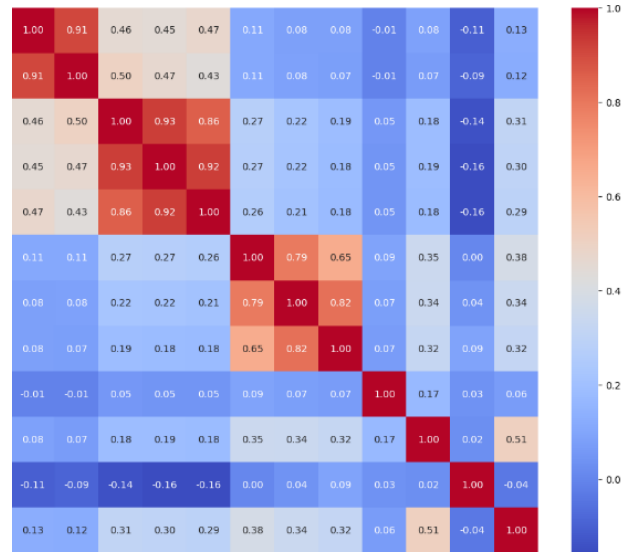
4.1.1. Veri (Data)

Tablo 1'e bakıldığında da görüleceği üzere, genel kredi veri kümesi toplam 11069 adet müşterinin kredi ödemelerini içermektedir. Bu müşterilerin yaklaşık 898 tanesi kredi borcunu ödemeyen yani temerrüde düşen veya zamanında ödemiş “riskli” müşteriler olup toplam müşteri kitlesinin yaklaşık %8,1'ine denk gelmektedir. Müşterilerin geriye kalan 10171 tanesi ise kredi borcunu zamanında ödeyen (temerrüde düşmeyen) “risksiz” müşteriler olup toplam müşterilerin yaklaşık %91,9'una tekabül etmektedir. İkincil olarak oluşturulan dengelenmiş veri kümesine baktığımızda 898 riskli ve yine eşit sayıda risksiz müşterilerden oluştuğu görülmektedir. Bu veri kümesi on bir (11) adet açıklayıcı (bağımsız) değişken içermektedir.

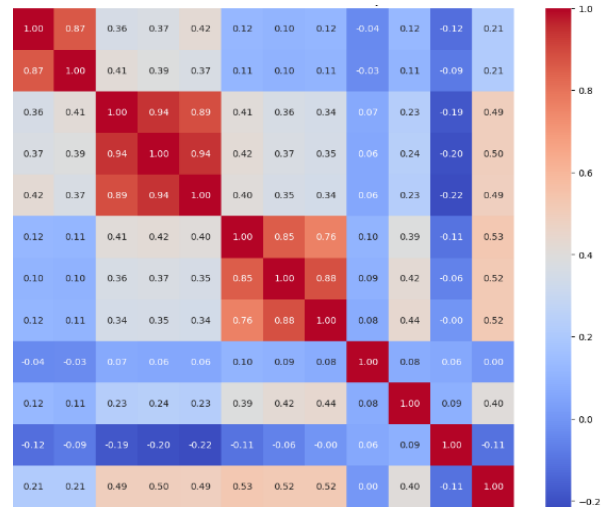
Tablo 1. Kredi Veri Kümeleri
(Credit Datasets)

Kredi Veri Kümeleri				
Müşteri Yapısı	Genel Veri Kümesi		Dengelenmiş Veri Kümesi	
	Sayısı	Oranı (%)	Sayısı	Oranı (%)
Riskli Müşteri	898	8,11	898	50,00
Risksiz Müşteri	10171	91,89	898	50,00
TOPLAM	11069	100,00	1796	100,00

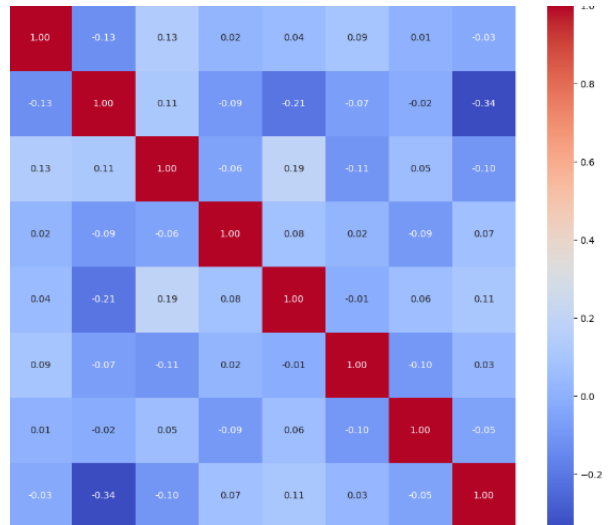
Şekil 1 ve Şekil 2'de sırasıyla genel ve dengelenmiş kredi verilerinin Şekil 3'de ise tarım verisine dair Kendal ısı (değişkenler arası ilişki) haritaları görülmektedir.



Şekil 1. Genel Kredi Verisi Isı Haritası
(General Credit Data Heatmap)



Şekil 2. Dengelenmiş Kredi Verisi Isı Haritası
(Balanced Credit Data Heatmap)



Şekil 3. Tarım Verisi Isı Haritası
(Agricultural Data Heatmap)

4.1.2. Tarım Veri Kümesi (Agricultural Dataset)

Tarım verisi farklı toprak değerleri için en uygun tarım ürününün belirlenmesine dönük hazırlanmış olan bir veri olup [www.kaggle.com](https://www.kaggle.com/datasets/madhuraatmarambhagat/crop-recommendation-dataset) (tam link: <https://www.kaggle.com/datasets/madhuraatmarambhagat/crop-recommendation-dataset>) sitesinden edinilmiştir. Tarım verisi 22 farklı tarım ürününün her biri için 100 gözlemden toplamda da 2200 gözlemden oluşmaktadır ve 7 adet bağımsız değişken içermektedir.

4.1.3. Analiz (Analysis)

KA yönteminin başarısını mukayese edebilmek için modern yapay zekâ yöntemlerinden YSA ve DVM ve geleneksel istatistiki yöntemlerden de LR ayrıca uygulanmıştır. Uygulamada min-max normalleştirilmesi yapılmış ve modellerin örneklem çekme kalitesinin iyileştirilmesi adına analizler on (10) kez çapraz doğrulanmıştır (cross validation: 10). Verinin %20'lik kısmı eğitim verisi olarak ayrılmıştır. Uygulamaların tamamında SMOTE veri dengeleme işlemi yapılmıştır. YSA ve KA yöntemlerinin metrikleri Bayes (Optuna) yapılarak optimize edilmiştir. Bu optimizasyon çalışmalarının sonuçlarına yani YSA ve KA'nın optimal metriklerine ekteki tablolardan ulaşılabilir. Hangi yöntemin ya da indeksin daha iyi sonuçlar verdiğini kontrol etmek için sonuçlar doğruluk (accuracy), standart sapma (accuracy std), hassasiyet (precision), geri çağırma (recall), F1 ve son olarak ROC eğrisi altında kalan alan (area under the curve – AUC_ROC) olmak üzere toplam 6 kritere göre kıyaslanmıştır. Sonuçlardaki farklılığın istatistiksel anlamlılığı Friedman testiyle gerçekleştirilmiştir. Ayrıca her üç veri kümesi için ısı haritası hazırlanmış ve değişkenler arasındaki ilişkiler ortaya çıkarılmıştır. Uygulama Python programının, Jupiter IDE'sinin 3.6.3 sürümünde gerçekleştirilmiştir.

4.1.4. Analiz Sonuçları (The Results Of The Application)

Tablo 2'de genel kredi verisine ait Tablo 3'de dengelenmiş kredi verisinin (riskli ve risksiz müşteri sayıları eşitlenerek dengelenmiş bir veri olup ayrıca SMOTE uygulaması da tüm verilere uygulanmıştır) uygulama sonuçları ve Tablo 4'de de tarım verisinin uygulama sonuçları görülmektedir.

Tablo 2. Genel Kredi Veri Kümesinin Analiz Sonuçları
(The Results Of General Credit Dataset)

Genel Kredi Verisi Uygulama Sonuçları							
Budama	Yöntemler		Doğruluk	Std. Sapma	Hassasiyet	Geri Çağırma	AUC_ROC
-	YSA		94.04	0.31	85.47	68.87	75.50
	DVM		93.67	0.26	85.42	65.49	76.06
	LR		93.85	0.35	85.61	66.79	73.82
Yok	KA	Gini	95.68	0.49	89.59	79.57	83.12
	KA	Entropi	95.67	0.40	88.61	79.68	83.63
	KA	LKF	95.69	0.44	89.06	79.69	83.18
Ön	KA	Gini	93.78	0.51	79.40	76.64	80.25
	KA	Entropi	94.18	0.43	81.78	77.32	81.26
	KA	Log-loss	93.93	0.46	79.06	76.79	80.94
Son	KA	Gini	87.12	0.75	63.89	71.92	71.96
	KA	Entropi	87.40	1.02	64.06	72.02	73.26
	KA	Log-loss	87.39	0.96	64.43	71.96	72.95

Not: Tüm sonuç tablolarında en iyi sonuçlar koyu **puntolu**, ikinci en iyi (indeksten bağımsız) sonuçlar *italik*, KA dışındaki en iyi sonuçlar ise altı çizili olarak verilmiştir.

Tüm sonuçlara bakıldığında sadece genel kredi verilerinin sonuçlarında (Tablo 2) DVM yönteminin en düşük standart sapmayı, LR yönteminin en yüksek hassasiyet değerini ve tarım verisinde en iyi AUC_ROC değerini YSA ve DVM yöntemlerinin verdiği bunun dışındaki tüm sonuçlarda, tüm kriterlerde KA yönteminin budamalı, ön budamalı ya da budamasız formları ve indeksleriyle en iyi sonuçları verdiği gözlemlenmektedir.

Tablo 3. Dengelenmiş Kredi Veri Kümesinin Analiz Sonuçları
(The Results Of Balanced Credit Dataset)

Dengelenmiş Kredi Verisi Uyg. Sonuçları							
Budama	Yöntemler		Doğruluk	Std. Sapma	Hassasiyet	Geri Çağırma	AUC_ROC
-	YSA		67.76	2.50	76.75	67.76	75.80
	DVM		61.53	3.48	62.14	61.53	72.32
	LR		61.53	3.33	62.15	61.53	72.96
Yok	KA	Gini	79.17	2.53	84.04	79.17	84.90
	KA	Entropi	79.01	2.70	83.73	79.00	84.66
	KA	LKF	79.01	2.70	83.73	79.00	84.66
Ön	KA	Gini	78.78	2.93	83.88	78.83	84.40
	KA	Entropi	78.51	2.71	83.34	78.50	84.09
	KA	LKF	77.50	2.35	82.07	77.50	83.17
Son	KA	Gini	72.38	2.42	75.97	72.38	73.59
	KA	Entropi	73.66	2.85	78.31	73.66	79.10
	KA	LKF	73.66	2.85	78.31	73.66	79.10

Budamasız modelin genel kredi verisinde (Tablo 2) ve dengelenmiş kredi verisinde (Tablo 3) en iyi sonuçları ve ön budamalı modelin ise tarım verisinde (Tablo 4) en iyi sonuçları verdiği görülmektedir. Tablo 2 ve 3'de en iyi ikinci sonuçları ön budamalı yöntemin Tablo 4'de ise budamasız yöntemin verdiği görülmektedir. Sadece son budamalı model Tablo 4'de en iyi ikinci AUC_ROC değerini vermiştir.

Tablo 4. Tarımsal Veri Kümesinin Analiz Sonuçları
(The Results Of Agricultural Dataset)

TARIM VERİSİ UYGULAMA SONUÇLARI							
Budama	Yöntemler	Doğruluk	Std. Sapma	Hassasiyet	Geri Çağırma	F1	AUC_ROC
-	YSA	98.45	0.65	98.61	98.45	98.45	99.99
	DVM	98.27	0.64	98.47	98.27	98.26	99.99
	LR	94.32	1.08	95.12	94.32	94.22	99.80
Yok	KA Gini	98.95	0.54	99.08	98.95	98.95	99.45
	KA Entropi	98.73	0.88	98.81	98.73	98.72	99.33
	KA LKF	98.73	0.88	98.81	98.73	98.72	99.33
	KA Gini	99.00	0.49	99.11	99.00	98.99	99.63
Ön	KA Entropi	98.41	1.14	98.54	98.41	98.40	99.47
	KA LKF	98.59	0.85	98.70	98.59	98.59	99.50
	KA Gini	98.05	1.20	98.30	98.05	98.04	99.66
Son	KA Entropi	98.05	1.20	98.30	98.05	98.04	99.66
	KA LKF	97.05	1.15	97.45	97.05	96.99	99.71

Son budamalı KA, yalnızca Tablo 4’de en iyi ikinci AUC_ROC değerini vermiştir. Yapılan uygulamalarda ccp_alpha değeri düşükken son budamalı yöntem, budamasız yöntemle birebir aynı sonuçları vermektedir. Sonrasında ccp_alpha değerleri 0,1 üzerine çıkarıldığında değerleri genel ve dengeli kredi veri kümesinde budamasız ve ön budamalı modellere göre bariz düşmüş fakat tarım verisinde böyle bir düşüşün meydana gelmediği görülmüştür.

Tablo 5. Friedman Testi Sonuçları
(The Results Of Friedman Test)

Friedman Testi Sonuçları			
Veri	Test İstatistiği	p değeri	Yorumu
Genel Kredi	42.9315	0.0000	Anlamlı Farklılık Var
Dengeli Kredi	47.5157	0.0000	Anlamlı Farklılık Var
Tarım	27.9939	0.0000	Anlamlı Farklılık Var

KA yönteminin bu bariz diyebileceğimiz ölçüde daha iyi sonuçlar vermiş olması istatistiksel olarak anlamlılığı Friedman testiyle sorgulandığında (Tablo 5) evet her üç veri setinde modeller arasında anlamlı farklılığın olduğu 0.05’in altındaki p değerleri ile gözlemlenmiştir.

Tablo 6. KA’nın İndeks Sonuçları Arasındaki Friedman Testi Sonuçları
(The Results Of Friedman Test)

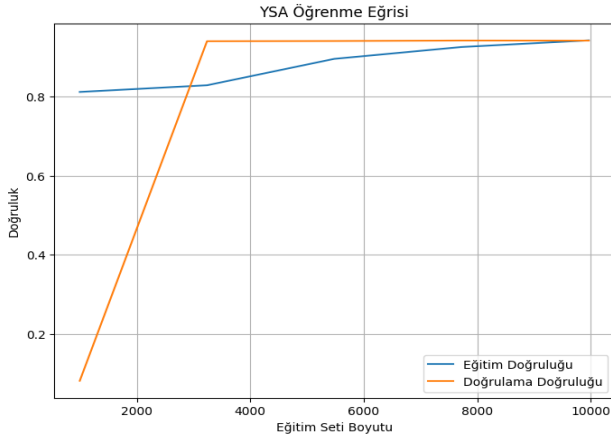
KA’nın İndeks Sonuçlarının Friedman Testi Sonuçları		
Veri	p değeri	Yorumu
Genel Kredi	0.0002	Anlamlı Farklılık Var
Dengeli Kredi	0.1316	Anlamlı Farklılık Yok
Tarım	0.1409	Anlamlı Farklılık Yok

Bu çalışma KA yönteminin diğer yöntemlerle kıyasına dayanmakta ve KA sonuçlarına etki edebilen üç farklı indeksle gerçekleştirilmiştir. Peki bu indeks farklılıkları KA sonuçlarına ne kadar etki etmektedir sorusuna cevap arandığında, Tablo 2, Tablo 3 ve Tablo 4’e tekrar bakılacak olursa, bu sonuç farklılıklarının yöntemlerden birini, diğerinin önüne ya da arkasına genel itibariyle taşımadığı görülecektir. Yine de indeks sonuçları arasında farklılıklar bulunmaktadır. Tablo 5’de Friedman testi sonuçlarına baktığımızda, genel kredi verisi -ki en hacimli (11069

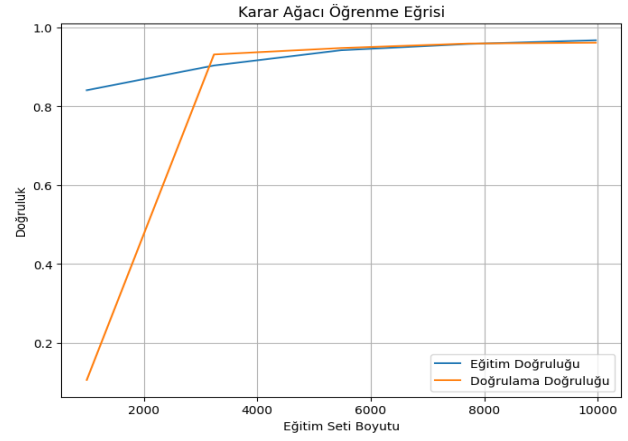
gözlem) veri olup- uygulama sonuçlarında bu farklılıklar anlamlı çıkmış diğer iki veri kümesinde ise anlamlı çıkmamıştır.

Karar ağaçlarının aşırı uyumla mücadele yeteneğine gelecek olursak öğrenme eğrileri (learning curves) tam kanıt olmamakla birlikte önemli fikirler sağlayabilir. Şekil 4’de genel veri kümesi için ve budamasız KA için çizdirilmiş bir öğrenme eğrisi görülmektedir. Eğitim doğruluğu (mavi) ve doğrulama doğruluğu (turuncu) çizgilerinin birlikte hareket etmesi aşırı uyum olmaması olarak kabul edilir. Nitekim Şekil 4’de bu özellikle 4000 gözlemden sonra oldukça iyi bir oranda sağlanmaktadır.

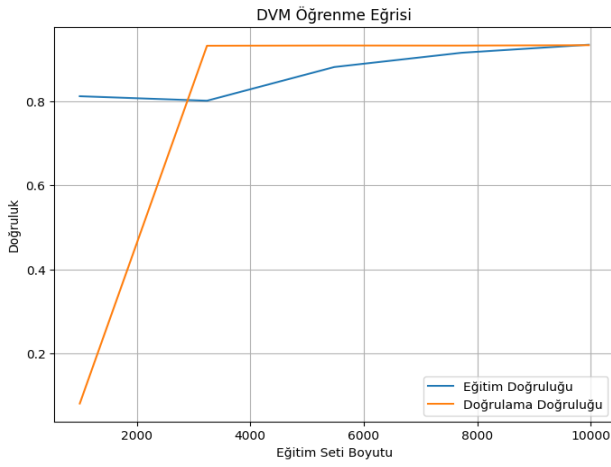
Bu iki çizginin ayrı ayrı hareket etmesi ise aşırı uyumla eşdeğer kabul edilmektedir. Nitekim Şekil 5’e bakıldığında budamasız KA için yaklaşık olarak 550. veriye kadar 1 gibi çok yüksek doğruluk oranı ile aşırı uyumu direk olarak ele vermektedir. Fakat yaklaşık 1300 veriden sonra aşırı uyum probleminin kalmadığı görülmektedir.



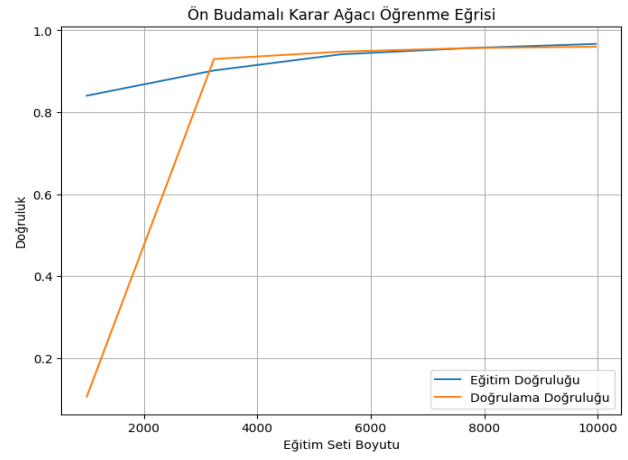
Şekil 4. Genel Kredi Verisi Öğrenme Eğrisi (YSA)
(General Credit Data Learning Curve, ANN)



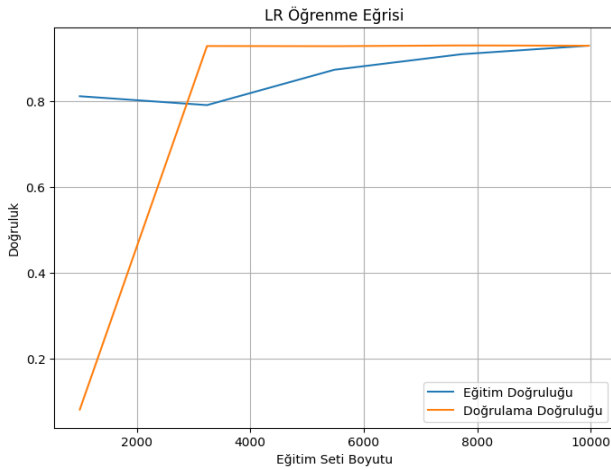
Şekil 4. Genel Kredi Verisi Öğr. Eğrisi (Budamasız, Gini)
(General Credit Data Learning Curve, Unpruning, Gini)



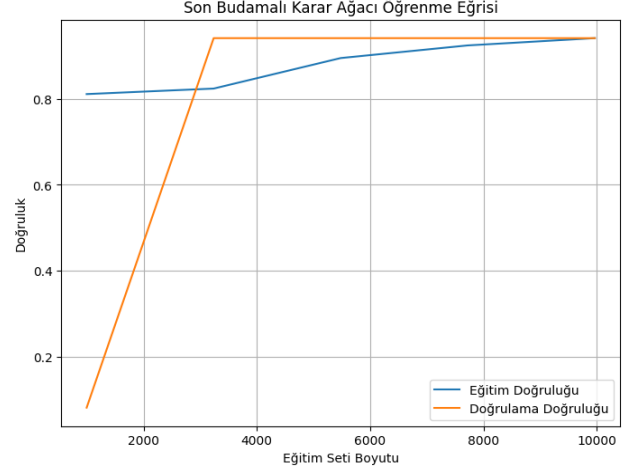
Şekil 5. Genel Kredi Verisi Öğrenme Eğrisi (DVM)
(General Credit Data Learning Curve, DVM)



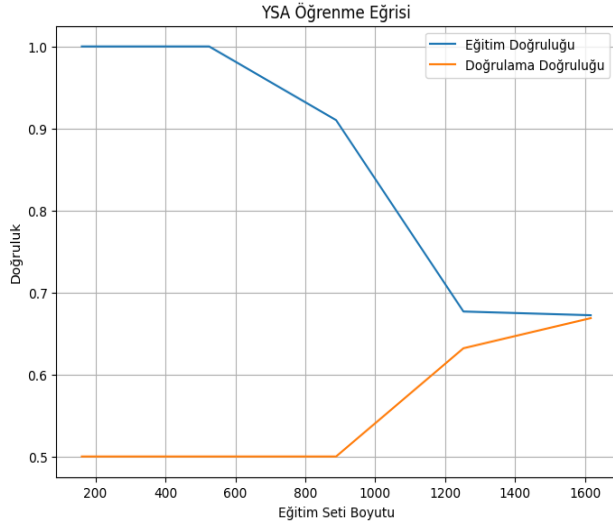
Şekil 5. Genel Kredi Verisi Öğr. Eğrisi (Ön Bud., Gini)
(General Credit Data Learning Curve, Unpruning, Gini)



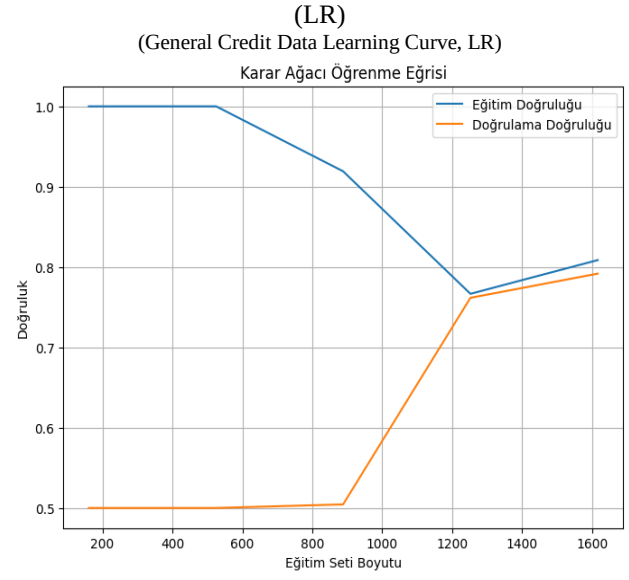
Şekil 6. Genel Kredi Verisi Öğrenme Eğrisi (LR)
(General Credit Data Learning Curve, DVM)



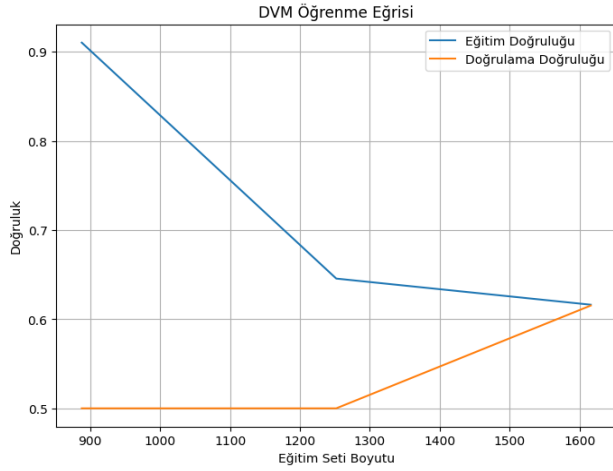
Şekil 6. Genel Kredi Verisi Öğr. Eğrisi (Son Bud., Gini)
(General Credit Data Learning Curve, Post-Pruning, Gini)



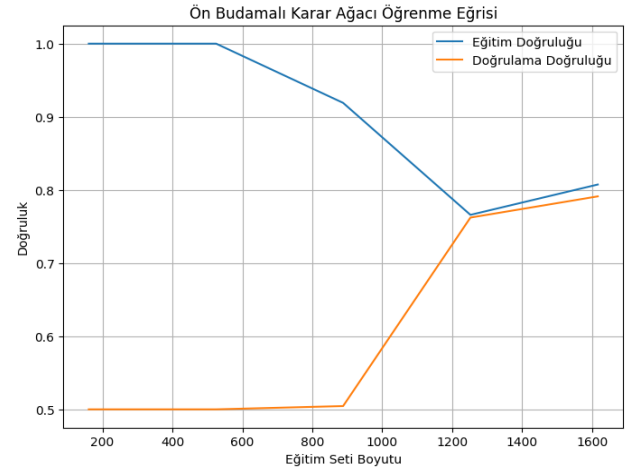
Şekil 4. Dengelenmiş Kredi Verisi Öğrenme Eğrisi
(YSA)
(General Credit Data Learning Curve, ANN)



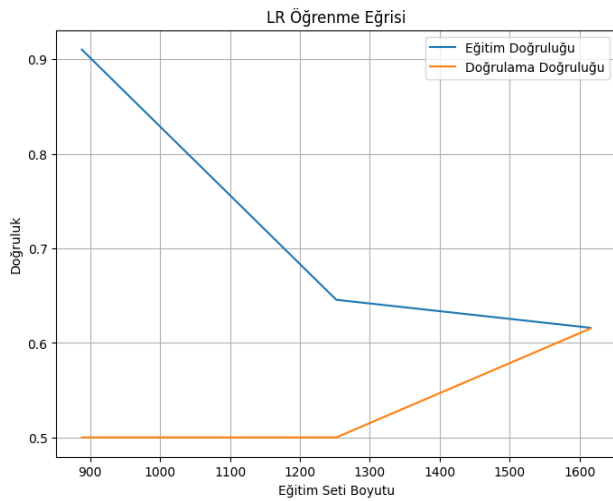
Şekil 7. Dengelenmiş Kredi Verisi Öğrenme Eğrisi
(Budamasız, Gini)
(Balanced Credit Data Learning Curve, Unpruning, Gini)



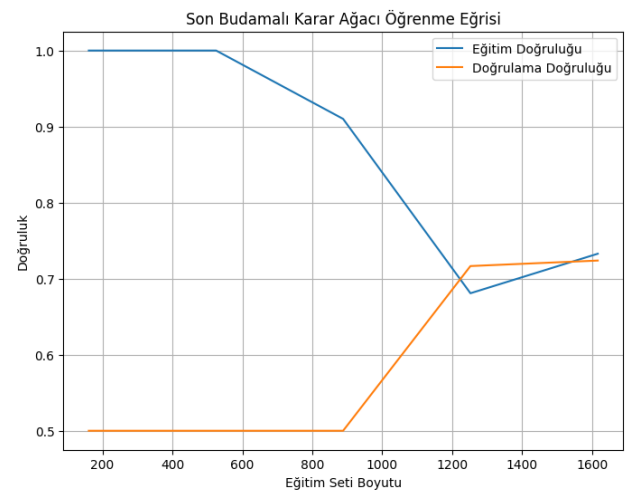
Şekil 5. Dengelenmiş Kredi Verisi Öğrenme Eğrisi
(DVM)
(General Credit Data Learning Curve, DVM)



Şekil 8. Dengelenmiş Kredi Verisi Öğrenme Eğrisi
(Ön Budamalı, Gini)
(Balanced Credit Data Learning Curve, Pre-pruning, Gini)



Şekil 6. Genel Kredi Verisi Öğrenme Eğrisi



Şekil 9. Dengelenmiş Kredi Verisi Öğrenme Eğrisi

(Son Budamalı, Gini)

(Balanced Credit Data Learning Curve, Post-pruning, Gini)

Tarım verisi de dengelenmiş kredi verisiyle yaklaşık aynı gözlem sayısında (dengelenmiş kredi: 1796, tarım: 2200) olduğu için ve KA'nın diğer iki indeksi (entropi ve LKF) sonuçları benzerlik taşıdığı için grafikleri buraya alınmamıştır.

Şekillerin genel yorumunu yapacak olursak her iki veri kümesinde de tüm yöntemler yaklaşık olarak aynı miktarda veri elde ettikten sonra bir genelleme kabiliyeti elde edebilmektedirler YSA'da optuna ile optimize edilmiş yani bir nevi budanmış haliyle Şekil 4 ve 7'de görülmektedir. Fakat Şekil 7,8 ve 9'a bakıldığında budama neticesinde KA'nın artan bir performansına rastlanmadığı bu çalışmada kullanılan veri kümelerinden elde edilen sonuçlara dayanarak söylenilebilir. Çünkü hem genel kredi verisinde hem dengelenmiş kredi verisinde ön ya da son budama sonrası artan bir grafiksel (genelleme) performansına bu çalışmada rastlanmamaktadır. KA yöntemi de az veri söz konusu olduğunda aşırı uyum tuzağına düşmekte ve fakat gözlem sayısı arttıkça aşırı uyumdan uzaklaşan ve genelleme kabiliyeti artan bir yöntem olarak gözükmemektedir.

5. SONUÇ VE ÖNERİLER (CONCLUSION AND RECOMMENDATIONS)

Yapılan analizde, modellerin uygulandığı her üç (genel kredi, dengelenmiş kredi ve tarım) veri kümesinde, KA'nın her üç indekste (entropi, gini, LKF) ve her üç evrede de (budamasız, ön budama sonrası ve son budama sonrası) en iyi değerleri veren yöntemler KA'nın çeşitli uygulamaları ve indeksleri olmuştur. Yapılan Friedman testinde modellerin sonuçları arasındaki farklar anlamlı (Bkz. Tablo 5) çıkmıştır. Sadece indeksler arasındaki sonuç farklılıkları ise gözlem sayısı çok olan veri kümesinde anlamlı çıkarken gözlem sayıları az olan veri kümelerinde anlamsız çıkmıştır (Tablo 6).

Modellerin aşırı uyumla baş edebilme yani genelleme yetileri ise öğrenme eğrileri yardımıyla karşılaştırılmıştır. Öğrenme eğrilerinden anlaşıldığı kadarıyla ağaç tabanlı olan ve olmayan yöntemlerin de hepsi belli ve hemen hemen aynı sayıda gözlem edindikten sonra aşırı uyum alanından çıkıp genelleme kabiliyeti geliştirebilmektedirler. Bu çalışmanın sonuçlarına göre KA yönteminde kullanılan ön budama ve son budama yöntemlerinin yöntemin genelleme kabiliyetini artırıcı bir etkisi gözlemlenmemiştir (Bkz. Şekil 4, Şekil 5, Şekil 6, Şekil 7, Şekil 8, Şekil 9). Fakat bu olumsuz diyebileceğimiz sonuç, KA'nın tüm metriklerde en iyi sonuçları veren yöntem olduğu bulgusunu da gölgelememelidir.

DVM yönteminin düşük veriyle iyi sonuçlar bir yöntem olarak ön plana çıkan bir yöntem olduğu görüşü bu çalışmada gözlemlenmemiştir. Ama genel veri kümesinde en düşük standart sapmayı veren yöntem olmuştur. Bu da yöntemin stabilitesine işaret etmektedir.

Bir sınıflandırıcının performans düşüklüğünün birçok nedeni olabilir. Eksik veri, gürültülü veri, verinin toplanması ya da işlenmesi ile alakalı süreç hataları, parametre hataları, verinin modele uygunluğu, aşırı uyum, eksik uyum, verinin az olması, verinin az açıklayıcı değişkenin çok olduğu boyut laneti (curse of dimensionality) ve uyumsuz mimari ilk akla gelen sorunlardır. Dolayısı ile bir sınıflandırıcı başarısız olduğunda bunun sebebinin net olarak belirlemek zordur. Fakat bildiğimiz bir şey var ki aşırı uyum güçlü modeller başta olmak üzere tüm modeller için önemli bir handicap oluşturmaktadır. Sınıflandırıcının başarısızlık sebebinin ne kadarının aşırı uyum kaynaklı olduğunu belirlemek de zordur. Fakat bu problemle başa çıkmak için üzerinde önemli teoriler geliştirilen KA, DVM ve YSA gibi yöntemlerin bu problemi salt anlamda çözümler bile iyi mücadele ettikleri literatürden anlaşılanlar arasındadır. Fakat bu çalışmada literatür de yer alan ve KA'nın içinde bulunduğu alternatiflerden diğer bir yöntemin en iyi sonucu verdiğini diğer birtakım çalışmalarda KA'nın ümit verici olduğu belirten çalışmalarla yarı uyumlu sonuçlar elde edilmiştir. Buradaki yarı uyumdan kasıt, KA'nı aşırı uyumdan sakındırmak ve bu şekilde genelleme yeteneğini artırmak için geliştirilen ön budama ve son budama tekniklerinin modelin genelleme yetisine olumlu katkısının olmaması bu sonucun yanı sıra ise KA'nın birçok metrikte en iyi sonuçları vermiştir. Literatürle uyumlu olan diğer bir nokta ise indeks konusudur ve çalışmada en yaygın üç indeks kullanılmış diğer ağaç tabanlı olmayan yöntemlerden daha iyi sonuçlar elde edilmiştir. İndeks sonuçları arasında büyük verideki fark anlamlı küçük verilerdeki fark ise anlamsız çıkmıştır (Tablo 6).

Bu çalışma ekseninde aşağıdaki sorular cevaplamak olursa:

- KA, ağaç tabanlı olmayan diğer yöntemlere kıyasla etkin sonuçları istatistiksel olarak anlamlı bir yöntem midir?
_Evet
- KA'nda kullanılan ön budama ve son budama teknikleri yöntemin aşırı uyumuna engel olarak genelleme kabiliyetini artırmakta mıdır?
_Hayır
- KA'nda en çok kullanılan gini, entropi ve LKF indekslerinin modelin sonuçlarına katkısı anlamlı mıdır?
_Veri büyükse evet

Yöntemler nasıl evrilirse evrilsin, hala veri kümelerinin tamamında, salt anlamda, her şart ve koşulda en iyi sonucu veren bir yöntem söz konusu olmadığı literatürde araştırmalarda hala pek çok yöntemin sonuçlarının birlikte değerlendiriliyor olmasından anlamak mümkün görünmektedir. Bu yüzden güncel araştırmalarda da bir sınıflandırma probleminde tüm sınıflandırıcıların sonuçlarına bakılması faydalı olacaktır.

Sonuç olarak gerek kredi riski probleminde gerekse diğer sınıflandırma ve regresyon problemlerinde KA'nın da sonuçlarını kontrol etmeden girilecek uygulamalarda sistemin doğruluk seviyesinden taviz verilebilir. KA'nın

basit ama güçlü ve etkili yapısı bu araştırma sonuçlarında da kendisini göstermiştir. Bu çalışmada kullanılan indekslerin yanı sıra Dagum, Jaccard, Sørensen İndeksi, Horn Örtüşme ve Morisita-Horn İndeksi gibi daha özelleştirilmiş indekslerinde, literatürde tanındığı, bilindiği, kullanıldığı ve çalışma anlamında çok önemli sayılarda takipçilerinin bulunduğu literatürden anlaşılmaktadır. Aslında burada temel indekslerin yanı sıra bu yeni indekslerin de bir literatür çalışması istediği ve tüm indekslerin muhtelif veri kümeleri üzerinde budamasız, ön budamalı ve son budamalı olarak denenmesi, indeks bilgisi ve aşırı uyumdan kaçınma bakımından önemlidir ve bundan sonraki çalışmanın/çalışmaların konusu olabilir.

KAYNAKLAR (REFERENCES)

- [1] S. Mussard, F. Seyte, M. Terraza, "Decomposition of Gini and the generalized entropy inequality measures", *Economics Bulletin*, 4(7), 1–6, 2003.
- [2] O. Rahmatia, M. Avand, P. Yariyan, J. P. Tiefenbacher, A. Azareh, D. T. Bui, "Assessment of Gini-, entropy- and ratio-based classification trees for groundwater potential modelling and prediction", *Geocarto International*, 37(12), 3397–3415, 2022.
- [3] V. G. Costa, C. E. Pedreira, "Recent advances in decision trees: an updated survey". *Artificial Intelligence Review*, 56, 4765–4800, 2023.
- [4] L. Zhang, X. Ma, Y. S. Ock, L. Qing, "Research on Regional Differences and Influencing Factors of Chinese Industrial Green Technology Innovation Efficiency Based on Dagum Gini Coefficient Decomposition", *Land*, 11, 1–20, 2022.
- [5] Bertsimas, J. Dunn, "Optimal classification trees", *Mach Learn*, 106, 1039–1082, 2017.
- [6] S. T. Biró, Z. Nédá, "Gintropy: Gini Index Based Generalization of Entropy", *Entropy*, 22, 1–13, 2020.
- [7] K. Ryu, D. J. Slottje, "Ten Ways to Specify a Gini Coefficient Using Entropy", *Annals of Financial Economics*, 18(1), 1–19, 2023.
- [8] M. Pal, Ve P.M. Mather, "An assessment of the effectiveness of decision tree methods for land cover classification", *Remote Sensing of Environment*, 86(4), 554–565, 2003.
- [9] M. Xu, P. Watanachaturaporn, P. K. Varshney, M. K. Arora, "Decision Tree Regression For Soft Classification Of Remote Sensing Data", *Remote Sensing Of Environment*, 97(3), 322–336, 2005.
- [10] L. Zhang, X. Ma, Y. S. Ock, L. Qing, "Research on Regional Differences and Influencing Factors of Chinese Industrial Green Technology Innovation Efficiency Based on Dagum Gini Coefficient Decomposition", *Land*, 11, 1–20, 2022.
- [11] S. Chakrabarty, G. Cauwenberghs, "Gini Support Vector Machine: Quadratic Entropy Based Robust Multi-Class Probability Regression", *Journal of Machine Learning Research*, 8, 813–839, 2007.
- [12] L. Jost, "Entropy and Diversity". *Opinion*, 3, 363–375, 2006.
- [13] T. Daniya, M. Geetha, K. Suresh, "Classification And Regression Trees With Gini Index", *Advances in Mathematics: Scientific Journal*, 9(10), 8237–8247, 2020.
- [14] Y. Fogel, M. Feder, "Universal Batch Learning with Log-Loss". *2018 IEEE International Symposium on Information Theory (ISIT)*, 6, 2018.
- [15] Vovk, V. "The Fundamental Nature of the Log Loss Function", *Fields of Logic and Computation II*, Kitap Bölümü, Kahire, Mısır, 307–318, 2015.
- [16] H. Patel, P. Prajapati, "Study and Analysis of Decision Tree Based Classification Algorithms", *International Journal of Computer Sciences and Engineering (JCSE)*, 6(10), 74–78, 2018.
- [17] M. Bansal, A. Goyal, A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning", *Decision Analytics Journal*, 3, 1–21, 2022.
- [18] www.javatpoint.com, <https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm>, 27.11.2024.
- [19] T. Kavzoğlu, ve İ. Çölkesen, "Karar Ağaçları İle Uydu Görüntülerinin Sınıflandırılması: Kocaeli Örneği", *Harita Teknolojileri Elektronik Dergisi*, 2(1), 36–45, 2010.
- [20] Gupta, A. Rawat, A. Jain, A. Arora, N. Dhami, "Analysis of Various Decision Tree Algorithms for Classification in Data Mining", *International Journal of Computer Applications*, 163(8), 15–19, 2017.
- [21] Aggarwal, S. Kasiviswanathan, Z. Xu, O. Feyisetan, N. Teissier, "Label Inference Attacks from Log-loss Scores", **38. Uluslararası Makine Öğrenmesi Konferansı Bildirileri, PMLR**, 139, 120–129, 2021.
- [22] S.B. Kotsiantis, "Decision trees: a recent overview". *Artificial Intelligence Review*, 39, 261–283, 2011.
- [23] Y.Y. Song, Y. Lu, "Decision Tree Methods: Applications For Classification And Prediction", *Shanghai Arch Psychiatry*, 27(2), 130–135, 2015.
- [24] H. Sharma, S. Kumar, "A Survey on Decision Tree Algorithms of Classification in Data Mining", *International Journal of Science and Research (IJSR)*, 5(4), 2094–2097, 2016.
- [25] M. A. Friedl, ve C. E. Brodleyf, "Decision Tree Classification of Land Cover from Remotely Sensed Data", *Remote Sensing of Environment*, 61(3), 399–409, 1997.
- [26] T. Dietterich, *ACM Computing Surveys*, 27(3), 326–327, 1995.
- [27] D. Wang, Z. Peng, L. Xi, "The sum of weighted normalized square envelope: A unified framework for kurtosis, negative entropy, Gini index and smoothness index for machine health monitoring", *Mechanical Systems and Signal Processing*, 140, 1–10, 2020.
- [28] Dembla, [www.towards datascience.com, https://towardsdatascience.com/intuition-behind-log-loss-score-4e0c9979680a](https://towardsdatascience.com/intuition-behind-log-loss-score-4e0c9979680a), 26.11.2024.
- [29] Corfield, B. Schölkopf, V. Vapnik, "Popper, Falsification And The VC Dimension", *Max Planck Institute for Biological Cybernetics, Teknik Rapor*, 145, 1–4, 2005.

Ek. 1. Tablo 7. YSA Yönteminde Kullanılan Gizli Katman ve Nöron Sayıları
(Number of Hidden Layers and Neurons Used in ANN Method)

YSA Yönteminde Uygulanan Gizli Katman Ve Nöron Sayıları							
(5,)	(5, 5)	(5, 10)	(10,)	(10, 5)	(10, 10)	(10, 20)	(20,)
(20, 5)	(20, 10)	(20, 20)	(30,)	(30, 5)	(30, 10)	(30, 20)	(10, 30)
(20, 30)	(30, 30)	(40,)	(40, 10)	(40, 20)	(40, 30)	(40, 40)	(50,)
(50, 10)	(50, 20)	(50, 30)	(50, 40)	(50, 50)	(50, 100)	(50, 75)	(75, 50)
(100,)	(100, 50)	(100, 100)	(5, 5, 5)	(10, 10, 10)			

Ek. 2. Tablo 8. YSA Metriklerinin Optimizasyon Sonuçları
(Preliminary Pruning Results of Decision Trees)

YSA Metriklerinin Optimizasyon Sonuçları					
Veri	Gizli Katman Sayısı	1. Gizli Katman Nöron Sayısı	2. Gizli Katman Nöron Sayısı	Alfa (Alpha) Katsayısı	Başlangıç Öğrenme Oranı (learning_rate_init)
Genel Kredi	2	30	5	0.011091539	0.000338884
Dengeli Kredi	2	40	40	0.000129042	0.000584349
Tarım	2	30	30	4.81E-05	0.008164213

Ek. 3. Tablo 9. Ön Budamalı KA Metriklerinin Optimizasyon Sonuçları
(Optimization Results of Pre-Pruned KA Metrics)

Ön Budamalı KA Metriklerinin Optimizasyon Sonuçları					
Veri	İndeks	Maksimum Derinlik (max_depth)	Bölünme İçin Minimum Örnek Sayısı (min_samples_split)	Yaprak Düzümü İçin Minimum Örnek Sayısı (min_samples_leaf)	Optimizasyon Aralığı
Genel Kredi	Gini	20	12	3	1-20
	Entropi	18	2	6	1-20
	LKF	20	10	6	1-20
Dengeli Kredi	Gini	18	3	1	1-20
	Entropi	15	3	1	1-20
	LKF	18	10	1	1-20
Tarım	Gini	13	7	3	1-20
	Entropi	9	4	1	1-20
	LKF	20	10	6	1-20

Ek. 4. Tablo . Son Budamalı KA Metriklerinin Optimizasyon Sonuçları
(Optimization Results of Post-Pruned KA Metrics)

Son Budamalı KA Metriklerinin Optimizasyon Sonuçları		
Veri	İndeks	Alfa Değerleri (ccp_alpha)
Genel Kredi	Gini	0.02414583
	Entropi	1.88E-02
	LKF	0.025378674
Dengeli Kredi	Gini	0.010383085
	Entropi	1.11E-02
	LKF	0.011209106
Tarım	Gini	0.010055386
	Entropi	1.27E-02
	LKF	0.012740215