Avrupa Bilim ve Teknoloji Dergisi
Sayı 27, S. 228-236, Kasım 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 27, pp. 228-236, November 2021

Research Article

Hibrit Açıklanabilir Yapay Zeka Tasarımı ve LIME Uygulaması

R.A. Selim Deliloğlu[1], Ayça Çakmak Pehlivanlı[2*]

[1] Mimar Sinan Güzel Sanatlar Üniv., Fen-Edebiyat Fakültesi, İstatistik Bölümü, İstanbul, Türkiye (ORCID: 0000-0003-1216-3918), selimdeliloglu@yahoo.com
[2*] Mimar Sinan Güzel Sanatlar Üniv., Fen-Edebiyat Fakültesi, İstatistik Bölümü, İstanbul, Türkiye, (ORCID: 0000-0001-9884-6538), ayca.pehlivanli@msgsu.edu.tr

(İlk Geliş Tarihi 28 Haziran 2021 ve Kabul Tarihi 21 Ağustos 2021)

(**DOI:** 10.31590/ejosat.959030)

ATIF/REFERENCE: Deliloğlu, R. A. S. & Pehlivanlı, A.Ç. (2021). Hibrit Açıklanabilir Yapay Zeka Tasarımı ve LIME Uygulaması. *Avrupa Bilim ve Teknoloji Dergisi*, (27), 228-236.

Öz

Günümüz teknolojisinin hızlı gelişimi ile yapay zeka günlük yaşantının bir çok alanında vazgeçilmez hale gelmiştir. Özellikle yanlış karar verme maliyetinin yüksek olduğu finans, sağlık, hukuk gibi alanlarda kullanılmaya başlamasına rağmen bu alanlardaki düzenlemeler nedeni ile kısıtlı düzeyde kullanılabilmektedir. Bu kısıtlamanın en temel nedeni de elde edilen yüksek performanslı sonuçların açıklanabilirliklerinin düşük olmasıdır. Bu çalışma kapsamında açıklanabilirliği yüksek hibrit bir tasarım önerilmiş, finans ve sağlık alanlarından elde edilmiş farklı veri setlerine uygulanmıştır. Şeffaf, hibrit ve açıklanabilirlik olmak üzere üç temel aşamada gerçekleştirilen hibrit yaklaşımın açıklanabilirlik aşamasında yerel olarak seçilen gözlemlerin tahmininde değişkenlerin etkisini belirlemek için LIME ölçütü kullanılmış ve sonuçlar yorumlanmıştır.

Anahtar Kelimeler: Açıklanabilir Yapay Zeka, LIME, Karar Ağaçları, Rasgele Orman.

Hybrid Explainable Artifical Intelligence Design and LIME Application

Abstract

With the rapid development of today's technology, artificial intelligence has become indispensable in many areas of daily life. Although it has started to be used in the areas such as finance, health and law, where the cost of making wrong decisions is high, it can be used at a limited level due to the regulations in these areas. The main reason for this limitation is the low explainability of the high-performance results obtained. Within the scope of this study, a hybrid design with high explainability was proposed and applied to different datasets obtained from the fields of finance and health. The LIME criterion was used to determine the effect of variables on the estimation of locally selected observations in the explainability phase of the hybrid approach, which was carried out in three basic stages as transparent, hybrid and explainability, and the results were interpreted.

Keywords: Explanaible Artificial Intelligence, LIME, Decision Trees, Random Forest.

[*] Sorumlu Yazar: ayca.pehlivanli@msgsu.edu.tr

1. Giriş

Günümüzde kurumların yapay zekayı, operasyonel çözümlerin yerine giderek daha fazla konumlandırdığı ve hızla iş süreçlerine entegre ettiği görülmektedir. Kurumlar yapay zekalarını ne kadar uzun ömürlü ve sürdürülebilir inşa ederlerse zaman içerisinde rekabette farklılaşma ve öne çıkma şansları o kadar artacaktır. Uzun ömürlü ve sürdürülebilir bir yapay zekanın ön koşulları araştırıldığında sadece sistemin performansının yüksek olması değil, bunun yanında şeffaf, adaletli, hesap verilebilir olmasının da kritik düzeyde önemli olduğu görülmektedir. Bu nedenle açıklanabilir yapay zekâ (explainable artificial intelligence) kavramı ortaya çıkmıştır. Açıklanabilir yapay zeka, etkilenen tüm paydaşların sonuçları yorumlayabilmesini sağlar. Regülasyonun ve piyasa şartlarının sık değiştiği ortamlarda yapay zekanın ömrünü ve sürdürülebilirliğini uzatacak bir diğer faktörün de dış şoka dirençli algoritmalar olduğu açıktır. Bu çalışmada dış şoklara dirençli, şeffaf, adaletli ve hesap verilebilir, açıklanabilir bir yapay zekanın kurulmasına olanak sağlayacak bir algoritma önerilmiştir. Özellikle yanlış karar verme maliyetinin yüksek olduğu; "şimdi al sonra öde" (fintech) şirketlerinin ve bankaların kredi risk hesaplamalarında, güvenlik, hukuk, tıp gibi alanlarındaki uygulamalarda bu çalışmada önerilen algoritmanın kullanılması, yapay zekadan elde edilen faydayı da sürekli kılacaktır. BM Dünya Çevre ve Kalkınma Komisyonu (WCED) sürdürülebilir kalkınmayı "gelecek nesillerin ihtiyaçlarını karşılama olanaklarını tehlikeye atmadan bugünün ihtiyaçlarını karşılayan kalkınma" şeklinde tanımlamıştır (Keeble, 1988). Bu tanım uzun ömürlü ve sürdürülebilir yapay zeka için "paydaşların gelecek dönemlerdeki ihtiyaçlarını karşılama olanaklarını tehlikeye atmadan bugünün ihtiyaçlarının karşılanması" şeklinde uyarlanabilir.

Yapay zekanın sonuçlarından etkilenen paydaşlar için katkının sürekli ve uzun ömürlü olmasını sağlayacak üç önemli özellik mevcuttur; güvenilir yapay zekâ, açıklanabilir yapay zekâ ve dış şoklara dayanıklı yapay zekâ. Bu üç özelliği içeren bir tasarım ve bu tasarımın tüm öğelerini içeren bir yaklaşım yardımıyla bir yapay zekâ geliştirmek ve uygulamak oldukça önemli bir çalışma alanı ortaya çıkarmıştır.

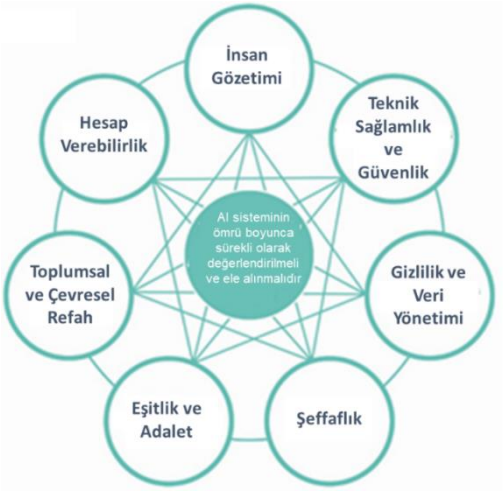
Güvenilir Yapay Zekâ

Avrupa komisyonu yapay zekâ uzmanları tarafından 2018 yılında yayınlanan çalışmada güvenilir yapay zekâ için gerekli etik kurallar yedi maddede özetlenmiştir (High-Level Expert Group on AI (AI HLEG), 2019). Şekil 1'de verilen her bir madde eşit düzeyde önemli olup birbirini destekler ve yapay zekanın yaşam döngüsü süresince de kullanılır (High-Level Expert Group on AI (AI HLEG), 2019).

Açıklanabilir Yapay Zeka

Son yıllarda giderek daha fazla gereksinim duyulan açıklanabilir yapay zekâ (Explainable AI ya da XAI) teknikleri, yapay zekâ sonuçlarının insanlar tarafından yorumlanmasını ve anlaşılmasını sağlayan teknikler ve yöntemler bütünü olarak nitelendirilebilir. Yapay zekanın hedefi ve uygulandığı alana göre açıklanabilir yapay zekâ gereksinimi de farklılık gösterir. Bir görselin kedi mi yoksa köpek mi olduğunun tahmin edildiği bir modeldeki açıklanabilirlik ihtiyacı düşükken, hastanın kanser olup olmadığını tahmin eden bir modeldeki açıklanabilirlik ihtiyacı oldukça yüksektir.

Açıklanabilir yapay zekâ ihtiyacı, özellikle hukuk, tıp ve finans gibi insan hayatını etkileyen alanlarda, kararları sadece kara kutu (blackbox) diyebileceğimiz açıklanamayan algoritmalara delege etmenin doğurduğu kritik operasyonel risklerden ortaya çıkmıştır (Malioutov et al., 2017)



Şekil 1. Güvenilir yapay zeka için etik kurallar.(High-Level Expert Group on AI (AI HLEG), 2019)

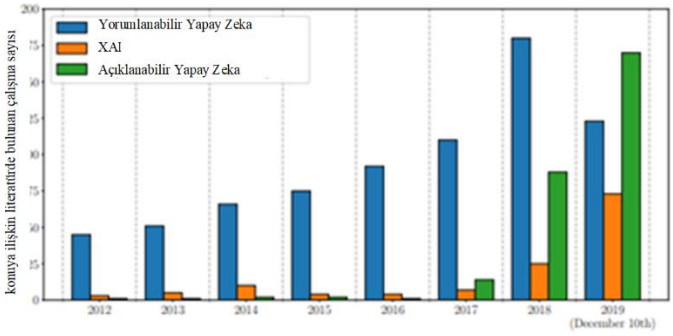
Bu risklerin gerçekleştiği bazı örnek durumlar Guidotti ve ark. tarafından 2018 yılında yayınlanan makalede özetlenmiştir (Guidotti et al., 2018);

- ABD'de serbest bırakılmadan önce hangi suçlunun daha tehlikeli olduğu tahmini uzun süredir yapılmaktadır. 1970'lere kadar ırk, milliyet ve renk üzerinden yapılan bu ayırımlar politik olarak kabul edilemez bulundu ve bu tutumdan vazgeçildi. Ancak suçluların tekrar suç işleme eğilimini tahmin eden modellerin ırk, milliyet ve renk üzerinden adaletsiz skorlar verdiği ve bu skorların kamu otoritelerince kullanıldığı görülüyor. Probublica'da yayınlanan bir makalede gerçekte riskin çok düşük olduğu durumlarda dahi siyahların beyazlara göre daha riskli olarak skorlanabildiği belirtilmiş ve bu skorlar baz alınarak haksız uygulamaların yapıldığına yönelik örnekler verilmiştir (Kirchner et al., 2016).
- Bloomberg'in 2016 senesinde yayınlanan bir raporuna göre Amazon'un ABD'nin büyük şehirlerindeki özel müşterilerine sunduğu aynı gün teslimat hizmetinin, siyah mahalleleri hariç tuttuğu görülmüştür. Amazon konuya ilişkin şöyle bir açıklama yapmıştır; "aynı gün nereye teslim edebileceğimizi belirleyen bir dizi faktör var. Bunlar, en yakın yerine getirme merkezine olan mesafeyi, bir bölgedeki yerel talebi, bir bölgedeki özel müşteri sayısını ve ulaştırmadaki ortaklarımızın pazar günleri dahi her gün saat 9:00'a kadar teslimat yapma kabiliyetlerini içerir." Dijital Business Insider sistesi bu yorumu, "özetle kararı veri veriyor" şeklinde değerlendirmiştir (Ingold & Spencer, 2016).
- 2002 senesinde "Consumer Federation of America"nın yayınladığı raporda, tüketicileri skorlayan üç firmanın (Experian, Trans Union ve Equifax) skorları karşılaştırılmış ve tüketicilerin %31'ine ait skorların 50 puandan fazla farklılaştığı tespit edilmiştir. Sonuçlar karşılaştırıldığında üç sistemin aynı tüketicileri birbirinden çok farklı ve yanlı skorladığı sonucu çıkmaktadır (Credit Score Accuracy and Implications for Consumers, 2002).

- Freitas'ın 2014 yılında yazdığı makalede de dost ve düşman tankları birbirinden ayıran bir modelin, geliştirme ve test aşamasında çok başarılı olduğu halde uygulamada çok kötü performans gösterdiği bir çalışmadan bahseder. Kötü performansının kök nedeni araştırıldığında düşman tankların fotoğrafının bulutlu günlerde, dost tanklarının fotoğraflarının ise güneşli günlerde çekildiği fark edilir.

Bütün bu örnekler, yapay zekanın bir sihir olmadığını, yapay zekâ projelerinin risklerin öngörüldüğü, sonuçların değerlendirildiği ve yorumlandığı bir süreç olduğunu göstermektedir (Freitas, 2014).

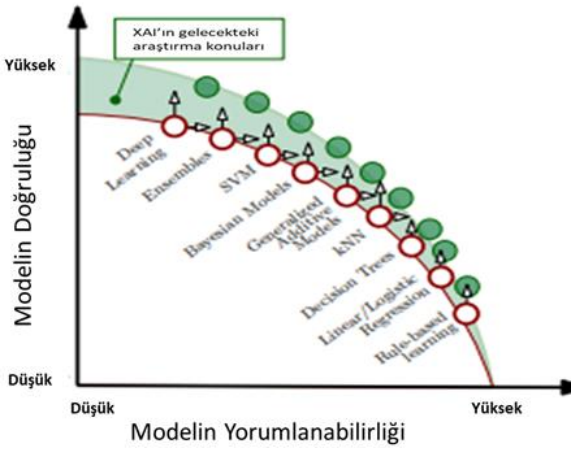
Arrieta ve ark. 2019 yılında yayınladıkları çalışmalarında Scopus R veri tabanındaki makalelerin başlık, özet ve anahtar kelimelerini analiz etmiş, açıklanabilir yapay zeka kavramlarının geçtiği yayın sayılarındaki gelişimi incelemişlerdir (Barredo Arrieta et al., 2020). Şekil 2'de verilen grafikte açıklanabilir yapay zekâ konusunda yayınlanan makalelerin son yıllarda hızla arttığı açıkça görülmektedir. Bu durumda yapay zekanın kullanımı yaygınlaştıkça açıklanabilir yapay zekaya olan ihtiyacın da giderek arttığı sonucu çıkartılabilir. Ayrıca yorumlanabilir yapay zekâ yayın sayısı 2019 senesinde azalırken açıklanabilir yapay zekâ kavramlarında kritik seviyede artış gözlenmektedir (Barredo Arrieta et al., 2020).



Şekil 2. Açıklanabilir yapay zeka kavramının geçtiği çalışmaların dağılımı (Barredo Arrieta et al., 2020)

Açıklanabilir yapay zeka performans açısından incelendiğinde genel olarak modellerin performansı arttıkça açıklanabilirliğinin düştüğü söylenebilir. Algoritmaların açıklanabilirliğiyle performansı arasındaki ilişkiyi özetleyen bir grafik Şekil 3'te verilmiştir. Kural bazlı sistemler genellikle performansı düşükken açıklanabilirliği en yüksek modellerdir. Derin öğrenme (deep learning) modelleri ise performansı yüksek olmasına karşın yorumlanması en zor modellerdir. Açıklanabilirlik problemi çözüldükçe performansı yüksek modellerin kullanım alanları da genişleyecektir. Bu nedenle açıklanabilir yapay zekânın gelecekteki araştırmalara giderek daha fazla konu olma potansiyeli çok yüksektir.

Yorumlanabilirlikle performans arasındaki ilişkiyi grafikte verildiği şekilde genellemenin doğru olmadığı yönünde görüşler de vardır. Rudin çalışmasında, bazı durumlarda karmaşık modellerinin performansının düşük olabileceğini, bu nedenle de ilk aşamada yorumlanabilir basit modellerin tercih edilmesi gerektiğini belirtmiştir. Basit bir model tercih edebilecekken yerine yorumlanması zor bir modelin tercih edilmesinin ilgili çalışma alanları bakımından ciddi zararlar verebileceğine de değinmiştir (Rudin, 2019).



Şekil 3. Modelin yorumlanabilirliği ile algoritmaların performansı arasındaki ilişki (Barredo Arrieta et al., 2020)

Açıklanabilir Yapay Zeka Teknikleri

Açıklanabilir Yapay Zeka (XAI) yeni bir alan olduğu için literatürde standart ve net bir terminolojisi de henüz tam anlamıyla oluşmamıştır. Guidotti ve ark.'nın 2018'de yayınladıkları makaledeki terminoloji baz alındığında, açıklanabilir yapay zeka tekniklerinin iki ana grupta toplandığı söylenebilir: tersine mühendislik (reverse engineering) ve şeffaf model tasarımı (transparent model design).

Tersine mühendislik de kendi içerisinde üç gruba ayrılır: modelin açıklaması (model explanation), çıktının açıklaması (outcome explanation) ve model denetlemesi (model inspection). Her bir teknik kendi içerisinde yapay zekayı açıklamaya yönelik farklı araç ve yaklaşımlar içerir (Guidotti et al., 2018).

- Taklit model, model açıklama grubundaki yaklaşımlardan olup, açıklanamayan kara kutu modelini taklit eden yorumlanabilir bir model bulmayı amaçlamaktadır. Sinir ağının kararlarını taklit eden bir karar ağacı örnek olarak verilebilir.
- Çıktı açıklama grubundaki yaklaşımlarda, tahminlerin kayıt düzeyinde adlandırılması hedeflenmektedir. 2016 yılında Ribeiro ve ark. tarafından önerilen aLIME yöntemi, tahmin değerlerini anlamak ve uygulamak için genellikle düşük çaba gerektiren eğer-ise (if-then) kurallarını tercih eder (Ribeiro et al., 2016a).
- Model denetleme grubundaki yaklaşımlarda ise, modelin tahminleri metin ya da görsel teknikler kullanılarak görüntülenir ve modelin bazı özelliklerinin bu şekilde anlaşılması/açıklanması hedeflenir.

Şeffaf model tasarımı ise doğrudan yorumlanabilen şeffaf bir model hedefler. Kural bazlı modeller ya da karar ağaçları modelleri yorumlanabilen şeffaf modellerdir (Guidotti et al., 2018).

Dış şoklara Dayanıklı Yapay Zekâ

Gambacorta ve ark. 2019 yılında yayımladıkları çalışmalarında, Çin'in önde gelen bir fintek şirketlerinin verilerini analiz ederek geleneksel modelleme yöntemleriyle makine öğrenmesi yöntemlerini karşılaştırmışlardır (Gambacorta et al., 2019). Karşılaştırma sonucunda, normal zamanlarda birbirine yakın performans gösteren iki modelin performansının dış şok sonrasında farklılaştığını gözlemlemişlerdir. Model performanslarındaki farklılığın kök nedeninin makine öğrenmesi algoritmasının doğrusal olmayan ilişkileri daha iyi yakalaması olduğu sonucuna varmışlardır.

Bu örnekten de görüldüğü üzere bazı durumlarda modelin sadece geliştirildiği dönemdeki koşulları açıklaması yanıltıcı olabilmektedir. Özellikle yanlış karar verme maliyetlerinin yüksek olduğu durumlarda regülasyonun değişmesi, piyasa koşullarının değişmesi gibi dış şoklar karşısında modelin nasıl davranacağı da kritik bir konudur. Çin'deki örneğe benzer şekilde eğer veri seti içerisinde modelin dış şok karşısındaki tepkisini gözlemleyecek bir veri yoksa, açıklanabilir yapay zekâ tekniklerinin de yardımıyla olası bir dış şok senaryosuna karşı modelin vereceği tepkinin öngörülmesi önerilir.

Bu çalışmada temel amaç, güvenilir, açıklanabilir ve dış şoklara dayanıklı yapay zekâ öğelerini içeren bir yaklaşım yardımıyla açıklanabilir bir yapay zeka tasarımı önermek ve çeşitli veri setleri üzerine uygulayarak kara kutu yöntemleri ile karşılaştırmaktır. Ayrıca, çok yeni ve hızla yaygınlaşan açıklanabilir yapay zeka kavram ve terminolojisine ilişkin kaynak oluşturarak ulusal literatüre katkıda bulunmak amaçlanmıştır.

2. Materyal ve Metot

Çalışma şeffaf model, hibrit model ve açıklanabilirlik olmak üzere üç temel aşamada gerçekleştirilmiştir. Şeffaf model aşamasında açıklanabilirliği yüksek yöntemler kullanılırken, hibrit model aşamasında performansı yüksek ancak açıklanabilirliği düşük kara kutu modelleri ve açıklanabilirlik aşamasında ise LIME kullanılmıştır. Bu çalışma kapsamında şeffaf model için karar ağaçları ve kara kutu model için ise rasgele orman algoritmaları tercih edilmiştir. Denetimli öğrenme yöntemlerinden olan karar ağaçları tümevarım yaklaşımı ile sınıflama yapan yorumlanabilirlik ve anlaşılabilirlik açısından oldukça yaygın kullanılan bir algoritmalardır. Genel olarak, bir çok karar ağacından gelen sonuçları birleştirerek sınıflama yapan rasgele orman algoritması ise ilk defa 2001 yılında Leo Breiman tarafından ortaya konmuştur (Breiman, 2001). Tek bir karar ağacına karşılık çok sayıda karar ağacına sahip olması nedeni ile karmaşık ve yorumlanabilirliği güç bir algoritmadır.

2.1. LIME (Local Interpretable Model-agnostic Explanations)

LIME, genel olarak herhangi bir makine öğrenme algoritmasına uygulanabilen ve model tahminlerini yorumlanabilir bir modele yaklaştırarak modelin açıklanabilmesi için önerilen görsel bir tekniktir (Ribeiro et al., 2016b). "Her karmaşık model yerel (local) ölçekte doğrusaldır" varsayımına dayanan bu yaklaşımın temel çalışma prensibi, veri setindeki gözlemlere benzer özellikte yeni gözlemler türetip, açıklayıcı değişkenlerin tahmin üzerindeki etkilerinin ve önem derecelerinin ortaya konması şeklinde özetlenebilir.

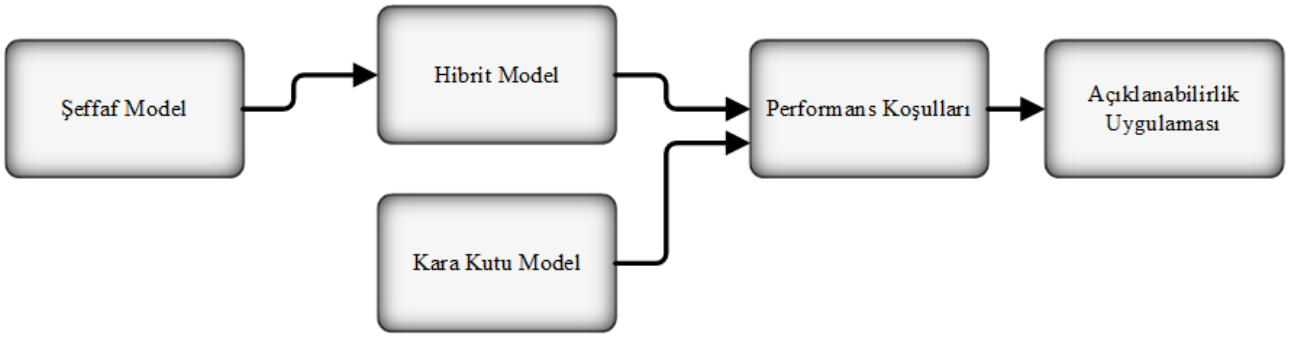
LIME algoritması genel olarak sayısal değişkenlerin kategorize edilmesi, veri setinin dağılımına benzer yeni gözlemlerin elde edilmesi ve elde edilen bu veri seti üzerinden açıklanabilir bir model geliştirerek değişkenlerin gözlem üzerindeki etkilerinin belirlenmesi aşamalarını takip eder (Garreau & von Luxburg, 2020).

LIME çıktısı, açıklayıcı değişlenlerin seçilen gözlemlerin tahminine katkısını özetler ve tahmin üzerinde en fazla etkiye sahip olan açıklayıcı değişkenlerin bulunabilmesine olanak sağlar.

2.2. Açıklanabilir Yapay Zeka Tasarımı (Hibrit Tasarım)

Bu çalışma kapsamında, çalışmanın temel amacında da belirtildiği üzere açıklanabilir bir yapay zeka tasarımı önerilmiştir. Bu tasarım genel olarak şefaf algoritmalar ile kara kutu algoritmalar arasında seçim yapmak zorunda kalan uygulayıcılara yarı şeffaf, performansı yüksek hibrit model öneren esnek bir çerçeve olarak düşünülebilir.

Önerilen hibrit modelin temel prensibi toplam bilginin mümkün olabildiği kadar açıklanabilir yapay zekâ teknikleri olarak belirtilen şeffaf algoritmalarla, geri kalan açıklanamayan bilginin ise performansı yüksek kara kutu algoritmalarla açıklanmasıdır.



Şekil 4. Açıklanabilir yapay zeka için önerilen hibrit tasarım akışı

Çalışmada önerilen açıklanabilirliği yüksek hibrit modelin elde edilmesi için Şekil 4'te verilen genel şemadaki adımların uygulaması Algoritma 1'de verildiği gibidir. (Algoritma mizanpaj nedeni ile bir sonraki sayfada verilmiştir.)

3. Uygulama

Bu çalışma kapsamında önerilen algoritma, herkese açık UCI veri deposundan elde edilen veri setleri kullanılarak sınanmıştır (Dua & Graff, 2019). Seçilen tüm veri setleri için, bu çalışma kapsamında önerilen hibrit algoritma adımları sırasıyla şeffaf model için karar ağacı, hibrit ve kara kutu aşamaları için ise rastgele orman (random forest) algoritması seçilerek uygulanmıştır. İlk üç adım sonucu elde edilen sonuçlar doğrultusunda Adım 4'te verilen üç koşulun geçerliliği her veri seti için kendi içinde kontrol edilmiştir.

Algoritma 1. Açıklanabilir yapay zeka için önerilen hibrit tasarım

Eğitim veri seti D_{mxd} : $\{x(i), y(i);\ i = 1, 2, 3, \dots m\}$ ve $x(i) = [x_1(i), x_2(i), \dots, x_d(i)]^T$

Geçerlilik veri seti G_{pxd} : $\{x(i), y(i);\ i = 1, 2, 3, \dots p\}$ ve $x(i) = [x_1(i), x_2(i), \dots, x_d(i)]^T$

$f(.)$: açıklanabilirliği yüksek algoritma (lojistik regresyon, kural tabanlı, karar ağacı,…)

$g(.)$: kara kutu algoritması (derin öğrenme, rasgele orman,…)

y_k : k. adımda eğitim sonucu elde edilen çıktı

y_k^G : k. adımda geçerlilik sonucu elde edilen çıktı

P_k : k. adımda eğitim veri seti ile elde edilen modelin geçerlilik veri seti üzerindeki performansı (başarı, F1 skoru, AUC,…)

Adım 1. *Şeffaf Model*

Eğitim veri seti üzerine açıklanabilirliği yüksek bir algoritma uygulanır ve model çıktısı (y_1) ve modelin geçerlilik veri seti ile performansı (P_1) elde edilir.

$y_1 = f(D_{mxd})$

$P_1 = f(G_{pxd})$

Adım 2. *Hibrit Model*

Şeffaf modelin uygulanmasıyla Adım 1'in sonunda oluşan şeffaf model çıktısı y_1 yeni bir açıklayıcı değişken olarak veri setine eklenir.

$D_{yeni} = D_{mx(d+1)}$: $\{x(i), y(i);\ i = 1, 2, 3, \dots m\}$ ve $x(i) = [x_1(i), x_2(i), \dots, x_d(i), y_1]^T$

Yeni veri seti üzerine performansı yüksek bir kara kutu algoritması uygulanır ve model çıktısı (y_2) ve modelin geçerlilik veri seti ile performansı (P_2) elde edilir.

$y_2 = g(D_{yeni})$

$G_{yeni} = G_{px(d+1)}$: $\{x(i), y(i);\ i = 1, 2, 3, \dots p\}$ ve $x(i) = [x_1(i), x_2(i), \dots, x_d(i), y_1^G]^T$

$P_2 = g(G_{yeni})$

Adım 3. *Kara kutu Model*

Eğitim veri seti üzerine Adım 2'de kullanılan $g(.)$ algoritması uygulanır ve model çıktısı (y_3) ve modelin geçerlilik veri seti ile performansı (P_3) elde edilir.

$y_3 = g(D_{mxd})$

$P_3 = g(G_{pxd})$

Adım 4. *Performans Koşulları*

İlk üç adım sonuçlarının aşağıdaki koşulları sağlayıp sağlamadığı kontrol edilir.

Koşul 1. $P_1 < P_2 < P_3$ eşitsizliğinin geçerli olma koşulu

Koşul 2. $P_2 \cong P_3$ olma koşulu

Koşul 3. Şeffaf model sonucu elde edilen ve hibrit model için girdi olarak eklenen y_1 hibrit modeli yani y_2'yi açıklamada en önemli değişken olmalıdır.

Bu koşulların en az birinin gerçekleşmemesi durumunda Adım 1'e tekrar geri dönülür.

Çalışmanın uygulaması Taiwan'da bir bankaya ait kredi kartı verisi üzerinden ayrıntılı olarak açıklanmış, diğer veriler tablo halinde özet olarak verilmiştir. Taiwan veri seti 23 açıklayıcı değişkenden ve bir hedef değişkenden oluşan 30000 gözleme sahiptir (Yeh & Lien, 2009). Veri setinde kredi kartı borcunu ödeyen müşteriler 1, ödemeyen müşteriler 0 olarak etiketlenmiş olup sınıf dağılımı sırasıyla %78 ve %22 şeklindedir. Veri setinde bulunan 23 değişkene ait ayrıntılar Tablo 1'de verilmiştir.

Tablo 1. Kredi kartı verisi değişken listesi

Değişken sıralaması	Değişken	Açıklama	Veri Tipi	Aldığı Değerler
X1	LIMIT_BAL	Müşterinin toplam kulllandığı kredi miktarı	Sayısal	Min: 10^4-Max:10^6
X2	SEX	Cinsiyet	Kategorik	Erkek: 1, Kadın: 2
X3	EDUCATION	Eğitim	Kategorik	İlk/Orta/Lise: 1, Üniversite: 2, Yüksek okul: 3, Diğer: 4
X4	MARRIAGE	Medeni Durum	Kategorik	Evli: 1, Bekar: 2, Diğer: 3
X5	AGE	Müşterinin Yaşı (Yıl)	Sayısal	
X6 - X11	PAY_0-PAY_6	Geçmiş dönem ödeme durumu {PAY_*i, i* = Eylül 2005(0), Ağustos 2005(2), … , Nisan 2005(6)}	Sayısal	Zamanında ödeme : -1 Gecikmeli ödeme (ay): 1,2, …, 8, 9
X12-X17	BILL_AMT1-BILL_AMT6	Fatura tutarı (Dolar) {BILL_AMT*i, i* = Eylül 2005(1), Ağustos 2005(2), …, Nisan 2005(6)}	Sayısal	Min : -339603 Max: 1664089
X18-X23	PAY_AMT_1-PAY_AMT_6	Geçmiş dönem ödeme miktarı (Dolar) {PAY_AMT_*i, i* = Eylül 2005(1), Ağustos 2005(2), …, Nisan 2005(6)}	Sayısal	Min :0 Max: 1684259

Veri seti içerisindeki 5000 gözlem geçerlilik için ayrılmış, kalan 25000 gözlemin de %60'ı eğitim (model geliştirme) ve %40'ı sınama amacıyla kullanılmıştır. Taiwan kredi kartı verisi için elde edilen model performansları Tablo 2'de F1 skoru, AUC ve GINI metrikleri ile verilmiştir. Algoritma 1'in performansa ilişkin son aşamasında verilen koşullar değerlendirildiğinde Koşul 1 ve Koşul 2'nin geçerlilik veri seti üzerinde sağlandığı görülmüştür.

Tablo 2. Model performanslarının karşılaştırılması

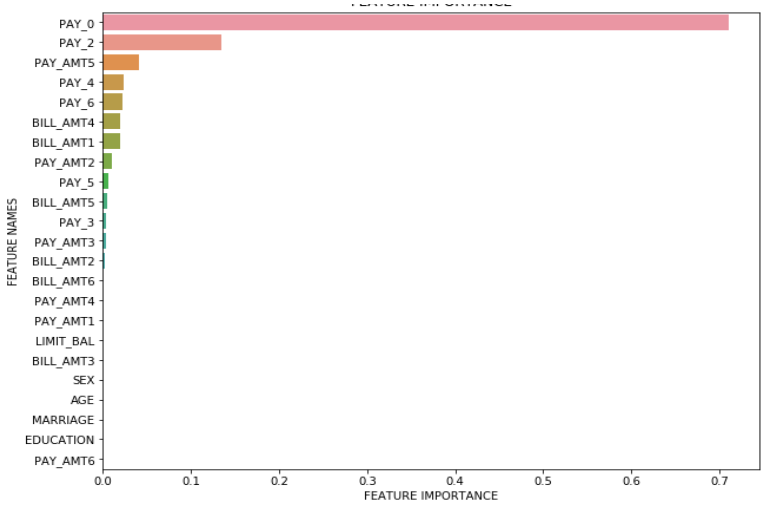
Model	F1 Skoru	AUC	GINI
Şeffaf Model (Karar Ağacı)	79.89%	74.4%	48.9%
Hibrit Model (Karar Ağacı ve Rastgele Orman)	80.2%	77.1%	54.2%
Kara Kutu (Rastgele Orman)	80.3%	77.4%	54.8%

√ **Koşul 1**. Şeffaf model performansı < Hibrit model performansı < Kara kutu model performansı

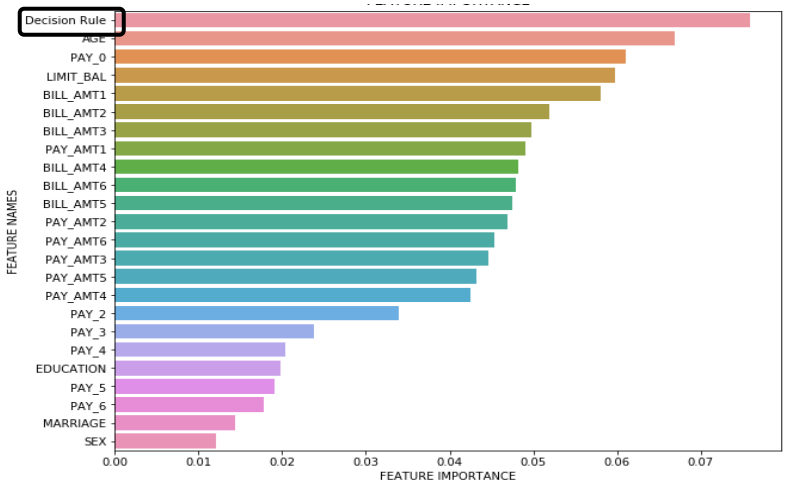
√ **Koşul 2**. Hibrit model performansı ≅ Kara kutu model performansı

Koşul 3'ün geçerliliğinin kontrolü ise değişken önem düzeyleri incelemesi ve LIME olmak üzere iki aşamada gerçekleştirilmiştir;

i) Şekil 5a'da verilen şeffaf model aşamasına ilişkin değişken önem sıralaması incelendiğinde ortaya çıkan modelin oldukça basit ve anlaşılır olduğu gözlemlenebilir. Bu model ile elde edilen ve Tablo 2'de verilen F1 Skoru %79.89 bulunmuştur. Adım 1 sonunda elde edilen bu modelin çıktısı Decision Rule değişkeni olarak adlandırılarak, Adım 2'de belirtildiği gibi hibrit modele yeni bir girdi olarak eklenmiş ve Şekil 5b'de verilen önem düzeylerine göre hibrit modeli açıklamada en önemli değişken olarak bulunmuştur.



Şekil 5a. Şeffaf modele ilişkin değişkenlerin önem düzeyi



Şekil 5b. Hibrit modele ilişkin değişkenlerin önem düzeyi

ii) Geçerlilik veri setinden rasgele 100 adet örneklem seçilmiş ve LIME algoritması uygulanarak her bir gözlemde hibrit model skorunu etkileyen ilk üç değişken araştırılmıştır. Tablo 3'te verildiği üzere Decision Rule değişkenini 100 gözlemin 94'ünde en önemli ilk üç değişken arasına girdiği gözlenmiştir.

√ **Koşul 3**. Decision Rule değişkeni, hem hibrit modelin kullandığı ve Tablo 2'de verilen rasgele orman algoritması sonuçlarına, hem de geçerlilik veri seti üzerinde uygulanan ve Tablo 3'te verilen LIME algoritması sonuçlarına dayanarak hibrit modeli açıklamada en önemli ilk üç değişkenden biri olarak bulunmuştur.

Tablo 3- Değişkenlerin mdoel skorunu etkileyen ilk üç önemli değişken arasına girme frekansı

	Değişken önem düzeyleri			
	1. Derece	2. Derece	3. Derece	Toplam
DECISION RULE	39	36	19	**94**
LIMIT_BAL	38	19	21	**78**
PAY_0	19	31	20	**70**
PAY_AMT2	3	6	15	**24**
PAY_AMT1		3	12	**15**
PAY_5	1	3	1	**5**
Diğer	0	2	12	14
Toplam	100	100	100	

Uygulama bölümünün başında da belirtildiği üzere, çalışma kapsamında Taiwan verisine ek olarak UCI veri deposundan elde edilen ve ayrıntıları Tablo 4'te verilen başka veri setleri de kullanılmıştır. Taiwan veri seti için uygulanan adımlar Tablo 4'de verilen veri setleri için de uygulanmış ve elde edilen hibrit modelin Adım 4'de verilen koşullara uygunluğu test edilmiştir.

Tablo 4-Hibrit modelin uygulandığı veri setleri

Veri Seti	Değişken Sayısı	Gözlem Sayısı	Bağımlı Değişken - Sınıf dağılımı
Breast Cancer Wisconsin(Wolberg & Mangasariant, 1990)	11	699	Tümör: İyi huylu (458) / Kötü huylu (241)
Statlog (Heart) (Dua & Graff, 2019)	13	270	Kalp hastalığı: Var (120) / Yok (150)
Credit Approval (Dua & Graff, 2019)	15	690	Kredi: Kabul (307) / Red (383)

Algoritma 1'de verilen Adım1, Adım2 ve Adım3 sırası ile Tablo 4'de verilen veri setlerine uygulanmıştır. İlk üç adım sonucunda elde edilen ve Tablo 5'te verilen sonuçlara göre kara kutu modeli tüm veri setlerinde en yüksek performansa, şeffaf model ise en düşük performansa sahiptir. Adım 4'te belirtilen koşullardan Koşul 1 ve Koşul 2 hibrit model performansının kara kutu model performansına yakın olması nedeniyle her bir veri seti ile elde edilen modellerde karşılanmıştır. Koşul 3 incelemesi Taiwan veri setine benzer şekilde her bir veri seti için gerçekleştirilmiş ve elde edilen tüm modellerde Decision Rule değikeninin en önemli değişken olması nedeniyle hibrit modellerde Koşul 3 de sağlanmıştır. Tablo 5, tüm sonuçları bütün olarak görebilmek adına Taiwan verisine ilişkin sonuçları da özet olarak içermektedir.

Tablo5- Modellerin Gini metriği ile karşılaştırması ve değişkenlerinönem seviyesi

	Gini			Şeffaf Modelin Önem Düzeyi	
	Şeffaf Model (Karar Ağacı)	Hibrit Model (Karar Ağacı ve RF)	Kara Kutu (RF)	Hibrit Model*	LIME**
Taiwan (kredi kart müşterileri)	48.9%	54.2%	54.8%	1	1
Breast Cancer Wisconsin	95.4%	98.3%	98.5%	1	2
Statlog (Heart)	68.9%	74.7%	73.9%	1	3
Credit Approval	89.7%	93.8%	94.8%	1	1

* Şeffaf model skor açıklayıcı değişkeninin hibrit model içerisindeki önem sırası

** Geçerlilik veri setinden seçilen100 adet örnekle üzerinden LIME algoritması ile elde edilen açıklayıcı değişkenlerin her bir gözlemdeki ilk üç değişken arasına girme sayılarına ilişkin önem sırası

4. Sonuç ve Tartışma

Makale kapsamında önerilen hibrit yaklaşım şeffaf model va kara kutu modelin birleşiminden oluşur. Yapay zeka büyük bir bina olarak düşünüldüğünde şeffaf model de hibrit modelin kolonları olarak kabul edilebilir. Şeffaf modeldeki değişken sayısının az tutulması, modele sezgisel olarak anlamlı ve dış etkenlere dayanıklı değişkenlerin seçilmesi kolonların kuvvetlenmesini, yapay zekanın da hem uzun ömürlü olmasını hem de dış etkenlerden en az düzeyde etkilenmesini sağlar. Hibrit modelin kara kutu model bileşeni ise performansın yüksek kılınmasını ve şeffaf modelin kullanmadığı değişkenlerin modelde kullanılmasını sağar.

Yapay zekaya dayalı sistemler giderek daha fazla günlük hayatımızın bir parçası olurken, zaman içerisinde bozulmaları ve onarılmaya gereksinim duymaları nedeni ile de hayatımızı olumsuz etkileme riskleri artmaktadır. Bu nedenle özellikle yanlış karar verme maliyetinin yüksek olduğu güvenlik, sağlık, finans ve hukuk gibi alanlarda bu riskin azaltılması daha da önem kazanmakdır. Sistemde bir bozulma olduğu zaman kök nedenin hızlı ve doğru şekilde bulunmasını kolaylaştıracağı için kara kutu model yerine makalede önerilen hibrit model ve/ya benzer yaklaşımların kullanılması sistemin sürdürülebilirliğini ve tutarlılığını koruyabileceği gibi elde edilen sonuçların açıklanabilir ve şeffaf olmasını da sağlayabilir.

Yapay zekanın sürdürülebilirliğini sağlayan bir diğer önemli özellik de sistemin kendi kendine öğrenebilmesi ve gerektiği durumlarda tıpkı insan zekasında olduğu gibi geri bildirimlere açık olmasıdır. Hibrit modelin kara kutu bileşeni kendi kendine öğrenmeyi sağlarken, şeffaf model bileşeni eğer-ise yapısıyla kurgulandığında geri bildirimleri de kural setine yansıtabilecektir.

Bu çalışmada önerilen modele ilişkin geliştirilmesi gereken noktalardan biri şefaf modelin performansının hibrit modelin açıklanabilirliğini ve performansını etkileme riskidir. Bu riskin Adım 4'te verilen koşulların sağlanmasıyla en aza indirgenmesi hedeflenmiştir. Ayrıca daha sonraki çalışmalarda LIME ölçütüne ek olarak kullanılabilecek diğer ölçütler de çeşitli sektörlerden elde edilen veriler üzerinde hem teorik hem de uygulamalı olarak

incelenecektir. Özellikle bankacılık ve sağlık alanlarındaki regülasyonlar nedeni ile performansı yüksek ancak açıklanabilirliği düşük kara kutu modellerinin kısıtlı kullanımı açıklanabilir yapay zeka yaklaşımları ile başka bir boyut kazanabilecektir.

Kaynakça

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5–32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Credit Score Accuracy and Implications for Consumers. (2002).

Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

Freitas, A. A. (2014). Comprehensible classification models. ACM SIGKDD Explorations Newsletter, 15(1), 1–10. https://doi.org/10.1145/2594473.2594475

Gambacorta, L., Huang, Y., Qiu, H., & Wang, J. (2019). How do machine learning and non-traditional data effect credit scoring? New evidence from Chinese fintech firm (Issue 834).

Garreau, D., & von Luxburg, U. (2020). Looking Deeper into Tabular LIME. http://arxiv.org/abs/2008.11092

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1–45. https://doi.org/10.1145/3236009

High-Level Expert Group on AI (AI HLEG). (2019). Ethics Guidelines for Trustworthy AI. In Ethics Guidelines for Trustworthy AI. Independent High-Level Expert Group on Artificial Intelligence Set up by the European Commission - Ethics Guidelines for Trustworthy AI. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

Ingold, D., & Spencer, S. (2016). Amazon Doesn't Consider the Race of Its Customers. Should It? https://www.bloomberg.com/graphics/2016-amazon-same-day/

Keeble, B. R. (1988). The Brundtland Report: "Our Common Future." In Medicine and War (Vol. 4, Issue 1, pp. 17–25). https://doi.org/10.1080/07488008808408783

Kirchner, L., Mattu, S., Larson, J., & Angwin, J. (2016). Machine Bias. Propublica, 1–26. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Malioutov, D. M., Varshney, K. R., Emad, A., & Dash, S. (2017). Learning Interpretable Classification Rules with Boolean Compressed Sensing. 95–121. https://doi.org/10.1007/978-3-319-54024-5_5

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). Nothing Else Matters: Model-Agnostic Explanations by Identifying Prediction Invariance. 30th Conference on Neural Information Processing Systems (NIPS2016). http://arxiv.org/abs/1611.05817

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). "Why Should I Trust You?" Explaining the Predictions of Any Classifier Marco. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016). Association for Computing Machinery, 13-17-Augu, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Wolberg, W. H., & Mangasariant, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology (linear programniing/pattern recognition/expert systems/cancer diagnosis). In Proc. Natl. Acad. Sci. USA (Vol. 87).

Yeh, I.-C., & Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020