YBS Ansiklopedi

www.YBSAnsiklopedi.com

Sayı ve Yıl Bilgisi Daha Sonra Eklenecektir

HermAI: Hermenötik Açıklanabilir Yapay Zeka

HermAI : Hermeneutic Explainable Artificial Intelligence

Sadi Evren SEKER*

OptiWisdom Inc. , evren@optiwisdom.com

Makine öğrenmesi modellerinin, özellikle de derin sinir ağlarının, karar alma süreçlerindeki içsel karmaşıklık, "kara kutu" olarak nitelendirilen bir şeffaflık sorununa yol açmaktadır. Mevcut Açıklanabilir Yapay Zeka (XAI) yaklaşımları, bu sorunu genellikle özellik atfı (feature attribution) yöntemleriyle ele alarak, bir karara hangi özelliklerin ne kadar etki ettiğini sayısal olarak ortaya koymaya odaklanır. Ancak bu analitik yaklaşım, özellikler arası bağlamı ve kararın ardındaki bütüncül mantığı sunmakta yetersiz kalmaktadır. Bu çalışma, felsefi hermenötik (yorum bilimi) geleneğinden esinlenerek, açıklanabilirliği statik bir rapordan dinamik bir diyalog sürecine dönüştüren yeni bir kavramsal çerçeve olan "Yorumlayıcı Döngü"yü (The Interpretive Loop) önermektedir. Çerçevemiz, bir modelin tekil, somut bir kararının ("parça") ve modelin genel davranışını yöneten soyut kuralların ("bütün") karşılıklı aydınlatılmasına dayanır. Bu amaçla, (1) gözlemlenen veri dağılımının kovaryans yapısını koruyan bağlamsal bir pertürbasyon mekanizması ve (2) hem tekil durumlar için eyleme geçirilebilir karşı-olgusal anlatılar üreten hem de modelin genel mantığını şeffaf vekil modellere damıtan çift modlu bir açıklama sistemi sunulmaktadır. Bu yaklaşımın, mevcut yöntemlerin sayısal skorlarının ötesinde, kullanıcı ile model arasında daha derin, bağlamsal ve güvenilir bir anlayış köprüsü kurma potansiyelini tartışıyoruz.

Anahtar Kelime veya Kelime Grupları: Açıklanabilir Yapay Zeka (XAI), Yorum Bilimi (Hermenötik), Karşı-Olgusal Açıklamalar, Model Şeffaflığı, İnsan-Merkezli Yapay Zeka.

* Yazar için dipnot metnini (varsa) buraya yerleştirin.

1 GİRİŞ

Yüksek performanslı makine öğrenmesi modelleri, tıp, finans ve hukuk gibi kritik alanlarda insan kapasitesini aşan başarılara imza atmaktadır. Ne var ki, bu modellerin artan karmaşıklığı, karar süreçlerinin insan sezgisi tarafından anlaşılamaz hale gelmesine neden olan bir "anlaşılabilirlik krizi" yaratmıştır [1]. Bu kara kutu sorunu, yalnızca akademik bir merak konusu değil, aynı zamanda etik, yasal ve pratik sonuçları olan temel bir güvensizlik kaynağıdır. Algoritmik kararlara yönelik hesap verebilirlik ve adalet talepleri, özellikle Avrupa Birliği'nin GDPR ve Yapay Zeka Yasası gibi düzenlemeleriyle somutlaşarak, açıklanabilirliği bir lüksten zorunluluğa dönüştürmüştür [2].

Bu ihtiyaca yanıt olarak geliştirilen LIME ve SHAP gibi öncü Açıklanabilir Yapay Zeka (XAI) araçları, bir modelin kararında hangi özelliklerin etkili olduğunu belirlemede önemli bir rol oynamıştır [3, 4]. Bu yöntemler, genellikle bir kararı, onu oluşturan özelliklerin katkılarının bir toplamı olarak sunan analitik ve indirgemeci bir yaklaşım benimser. Ancak bu "özellik atfı" paradigması, iki temel zorlukla karşı karşıyadır:

1. **Bağlamsal Kopukluk:** Özelliklerin birbirleriyle olan ilişkilerini göz ardı ederek yapılan analizler, gerçek dünyada var olması mümkün olmayan senaryolar üzerinden açıklamalar üretebilir ve bu da açıklamanın güvenilirliğini zedeler [5].
2. **Anlatısal Bütünlük Eksikliği:** Bir dizi sayısal skor, bir kararın ardındaki "hikayeyi" anlatmakta yetersiz kalır. Kullanıcılar, "Neden bu karar alındı?" sorusunun yanı sıra, "Farklı bir sonuç için ne değişmeliydi?" gibi daha eyleme dönük ve bütüncül sorulara yanıt arar.

Bu çalışma, mevcut yaklaşımların bu sınırlılıklarına, kökleri felsefeye dayanan bir alternatif sunmaktadır. Açıklanabilirliğin felsefi temelleri, modern hermenötiğin kurucusu kabul edilen Schleiermacher'in metin yorumlama teorilerine kadar uzanır [6]. Schleiermacher, bir metni anlamanın, hem metnin dilsel yapısını ("gramatik yorum") hem de yazarın zihinsel ve tarihsel bağlamını ("teknik yorum") anlamayı gerektiren çift yönlü bir süreç olduğunu savunur. Bu fikir, daha sonra Dilthey tarafından doğa bilimleri (erklären - açıklama) ile beşeri bilimler (verstehen - anlama) arasına bir ayrım koyarak geliştirilmiştir [7]. Ona göre, bir olguyu sadece neden-sonuç ilişkileriyle açıklamak, onun arkasındaki insani anlamı ve niyeti "anlamak" ile aynı şey değildir. Yapay zeka alanında bu ayrım, bir modelin tahminini sayısal olarak "açıklamak" ile o tahmini insan için "anlamlı" kılmak arasındaki farka tekabül eder.

Bu felsefi mirastan yola çıkarak, bu makale açıklanabilirliği tek yönlü bir raporlamadan, kullanıcı ile model arasında çift yönlü bir "anlama eylemine" dönüştürmeyi önermektedir. Bu diyalojik yaklaşım, özellikle yapılandırılmış argümantasyon sistemlerinde de karşılığını bulur. Örneğin, Şeker tarafından geliştirilen "Bilgisayarlı Argüman Delphi Tekniği", uzmanlar arasında tez ve antitezlere dayalı yapılandırılmış bir diyalog kurarak bir fikir birliğine veya daha derin bir anlayışa ulaşmayı hedefler [8]. Benzer şekilde, bizim önerdiğimiz Yorumlayıcı Döngü çerçevesi de, hem tekil kararların ("parça") hem de modelin genel mantığının ("bütün") birbirini aydınlattığı diyalojik bir yapı sunarak, kullanıcı ile model arasında bir anlayış ortaklığı kurmayı amaçlamaktadır.

2 LİTERATÜR TARAMASI: ANALİTİK PARADİGMADAN YORUMLAYICI DİYALOĞA AÇIKLANABİLİR YAPAY ZEKA

Bu bölümde, Açıklanabilir Yapay Zeka (XAI) alanının entelektüel kökenleri, mevcut teknik paradigmaları, bu paradigmaların doğurduğu felsefi tartışmalar ve alanın gelecekteki yörüngesi detaylı bir şekilde incelenecektir. Amacımız, XAI'ı yalnızca bir dizi algoritmik araç olarak değil, aynı zamanda insan anlama eyleminin doğasıyla ilgili daha derin soruları gündeme getiren disiplinlerarası bir araştırma alanı olarak konumlandırmaktır.

2.1 Açıklanabilirlik Paradoksu: Performans ve Şeffaflık İkilemi

Yapay zeka alanındaki güncel ilerlemeler, özellikle derin öğrenme mimarilerinin [9, 10] ve büyük dil modellerinin [11] yükselişiyle birlikte, bir "açıklanabilirlik paradoksu" ortaya çıkarmıştır. Bu paradoks, bir modelin tahmin doğruluğu (performance) ile içsel yorumlanabilirliği (interpretability) arasında gözlemlenen ters orantılı ilişkiyi ifade eder [12, 13]. Lineer regresyon veya karar ağaçları gibi basit modeller, doğaları gereği şeffaf olsalar da, karmaşık ve doğrusal olmayan örüntüleri yakalamada yetersiz kalırken; sinir ağları gibi karmaşık modeller, milyonlarca parametre arasındaki etkileşimler nedeniyle insan denetiminden uzaklaşan "kara kutulara" dönüşmektedir [14].

Bu ikilem, yalnızca teknik bir zorluk değil, aynı zamanda etik ve toplumsal bir açmazdır. Kredi başvurularının reddedilmesi [15], tıbbi teşhislerin konulması [16] veya adli risk değerlendirmeleri [17] gibi yüksek riskli alanlarda, hatalı veya yanlı bir kararın gerekçelendirilememesi, hesap verebilirlik ilkesini temelden sarsmaktadır [18, 19]. Bu bağlamda, Doshi-Velez ve Kim'in de vurguladığı gibi, yorumlanabilirlik, bir modelin diğer arzu edilen özelliklerini (örneğin, adalet, sağlamlık ve güvenilirlik) doğrulamak için bir ön koşul haline gelmiştir [20].

2.2 Post-Hoc Açıklanabilirlik Yöntemlerinin Sınıflandırılması ve Eleştirisi

Kara kutu modellerin yaygınlaşması, modelin kendisini değiştirmeden, eğitildikten sonra davranışını yorumlamayı amaçlayan "post-hoc" açıklama tekniklerinin gelişmesine yol açmıştır. Bu yöntemler, açıklamanın kapsamına (yerel/küresel) ve dayandığı tekniğe göre sınıflandırılabilir [21].

2.2.1Yerel Vekil Modeller ve Pertürbasyon Tabanlı Yöntemler

Post-hoc açıklanabilirliğin popülerleşmesindeki en önemli dönüm noktalarından biri, LIME (Local Interpretable Model-agnostic Explanations) olmuştur [4]. LIME'ın temel mantığı, karmaşık bir modelin herhangi bir tekil kararını, o kararın yerel komşuluğunda, lineer regresyon gibi basit ve yorumlanabilir bir "vekil model" (surrogate model) ile taklit etmektir. Bu yaklaşımın model-agnostik olması, onu son derece esnek kılmaktadır. Ancak, LIME'ın pertürbasyon stratejisinin (özelliklerin bağımsız olarak değiştirilmesi) gerçek veri dağılımını göz ardı etmesi, anlamsız veya gerçek dışı senaryolar üzerinden açıklamalar üretmesine neden olabilir [5, 22]. Ayrıca, pertürbasyon sürecinin stokastik doğası, aynı örnek için tekrar çalıştırıldığında farklı açıklamalar üretebilen "kararsızlık" sorununa yol açmaktadır [23, 24]. Bu kararsızlık, açıklamanın güvenilirliğini ciddi şekilde zedelemektedir.

2.2.2 Atıf Yöntemleri: SHAP ve Gradyan Tabanlı Yaklaşımlar

Açıklamaların teorik tutarlılığına yönelik artan talep, SHAP (SHapley Additive exPlanations) çerçevesinin geliştirilmesine yol açmıştır [3]. İşbirlikçi oyun teorisinden Shapley değerlerini ödünç alan SHAP, bir karara her bir özelliğin yaptığı marjinal katkıyı adil bir şekilde paylaştırır ve LIME'ın aksine, tutarlılık gibi arzu edilen teorik özelliklere sahiptir. SHAP, hem yerel hem de küresel açıklamalar için birleşik bir çerçeve sunmasıyla alanda bir standart haline gelmiştir.

Derin öğrenme modelleri için ise, özellikle görüntü ve metin işleme alanlarında, gradyan tabanlı yöntemler öne çıkmaktadır. Saliency Maps [25], bir karara ilişkin olarak her bir girdi pikselinin gradyanını görselleştirirken, Grad-CAM [26] ve türevleri [27, 28], modelin dikkatini haritalamak için daha gürültüsüz ve sınıfa özgü yöntemler sunar. Integrated Gradients [29] ve DeepLIFT [30] gibi daha gelişmiş atıf yöntemleri ise, gradyan doygunluğu gibi sorunları ele alarak daha güvenilir atıflar sağlamayı amaçlar. Ancak, bu yöntemlerin de temel çıktısı, bir karara neden olan "hikayeyi" anlatmaktan ziyade, etkili pikselleri veya kelimeleri vurgulayan bir "ısı haritası"dır.

2.2.3. Kural-Tabanlı ve Örnek-Tabanlı Açıklamalar

Sayısal skorların ötesinde, insan bilişine daha uygun açıklama türleri de araştırılmaktadır. Anchors [31], bir kararı "çapalayan" ve kapsamı belirli, yüksek kesinlikli IF-THEN kuralları üreterek, bir açıklamanın geçerlilik sınırlarını net bir şekilde ortaya koyar.

Örnek-tabanlı açıklamalar ise, bir kararı, eğitim setindeki prototipik veya etkili örneklere referansla açıklar [32, 33]. Ancak bu yaklaşımların en güçlüsü, "karşı-olgusal açıklamalar" (counterfactual explanations) olmuştur [6]. Bir karşı-olgusal, "Eğer girdi X yerine X' olsaydı, çıktı Y yerine Y' olurdu" şeklinde ifade edilir. Bu yaklaşım, kullanıcılara sadece bir kararın nedenini değil, aynı zamanda istenen bir sonuca ulaşmak için neyi değiştirmeleri gerektiğini de söyleyerek eyleme geçirilebilir bir yol haritası sunar [15, 34]. DiCE [35] gibi modern çerçeveler, tek bir karşı-olgusal yerine, kullanıcıya seçenek sunan çeşitli ve eyleme geçirilebilir karşı-olgusallar üretmeye odaklanmaktadır.

2.3 . Açıklamanın Felsefi Temelleri: Analitik ve Yorumlayıcı Gelenekler

XAI alanındaki teknik çeşitlilik, "açıklama" eyleminin doğasına ilişkin daha derin felsefi ayrımları da beraberinde getirir. Mevcut XAI paradigmaları, büyük ölçüde Anglo-Amerikan analitik felsefe geleneği içinde konumlandırılabilir. Bu gelenek, bir olguyu açıklamanın, onu daha temel, gözlemlenebilir veya ölçülebilir bileşenlere ayrıştırmak anlamına geldiğini varsayar.

Buna karşılık, Kıta Avrupası felsefesinden doğan hermenötik ve fenomenoloji gibi yorumlayıcı gelenekler, anlamanın mekanik bir ayrıştırma eylemi olmadığını, aksine bağlam, niyet ve yorumcunun kendi ön yargılarıyla şekillenen bütüncül bir süreç olduğunu savunur. Bu iki yaklaşım arasındaki temel farklar Tablo 2'de özetlenmiştir.

Tablo 2: Açıklamaya Yönelik Felsefi Yaklaşımların Karşılaştırılması

Boyut	Analitik Gelenek (örn. LIME/SHAP)	Yorumlayıcı Gelenek (örn. Hermenötik)
Açıklamanın Amacı	Bir kararın nedensel veya ilişkisel bileşenlerini nesnel olarak ortaya koymak. (Erklären - Açıklama)	Bir kararın, kullanıcının bağlamı içinde anlamlı hale gelmesini sağlamak. (Verstehen - Anlama)
Temel Birim	Öznitelik (Feature), Sayısal Skor	Anlatı (Narrative), Bağlam
Süreç	Analiz (Ayrıştırma)	Diyalog (Parça ve bütün arasında döngüsel ilişki)
Kullanıcı Rolü	Pasif Gözlemci	Aktif Katılımcı, Yorumcu
İdeal Açıklama	Nedensel olarak doğru ve nicel olarak kesin olan.	Kullanıcının zihinsel modeline uyan, eyleme geçirilebilir ve bütüncül olan.

Bu felsefi ayrım, XAI'nin geleceği için kritik bir yol ayrımına işaret etmektedir. Miller'ın sosyal bilimlerden derlediği bulgular, insanların doğal olarak aradığı ve sunduğu açıklamaların, basit özellik listelerinden ziyade, seçici, karşıtlıklara dayalı ve sosyal bir diyalog içinde şekillenen anlatılar olduğunu göstermektedir [36]. Bu bulgu, mevcut XAI araçlarının insan beklentileriyle ne kadar uyumsuz olabileceğini ortaya koymaktadır.

2.4 Diyalojik ve Etkileşimli XAI: Yeni Ufuklar

Literatürdeki bu boşluğu doldurmak üzere, açıklanabilirliği tek seferlik bir çıktıdan, kullanıcı ile sistem arasında devam eden bir diyalog sürecine dönüştürmeyi amaçlayan yeni bir araştırma dalgası ortaya çıkmaktadır. Bu "etkileşimli XAI" (iXAI) [37, 38] veya "diyalojik XAI" [39] yaklaşımları, kullanıcının "neden?" sorusunu takip eden "peki ya şöyle olsaydı?" veya "bana başka bir örnek göster" gibi ek sorular sormasına olanak tanır.

Bu alandaki çalışmalar, genellikle argümantasyon teorisinden [40, 41] ve diyalog modellerinden [42] beslenir. Amaç, kullanıcının zihinsel modelini ve bilgi seviyesini dikkate alarak açıklamaları kişiselleştiren sistemler yaratmaktır [43].

Örneğin, bir açıklamanın çok karmaşık olması durumunda, sistem daha basit bir analog sunabilir; bir açıklamanın tatmin edici olmaması durumunda ise, kullanıcı farklı bir perspektiften (örneğin, karşı-olgusal bir senaryo) ek bilgi talep edebilir. Bu diyalojik yaklaşım, aynı zamanda "güvenilir yapay zeka" (Trustworthy AI) kavramıyla da yakından ilişkilidir [44, 45]. Güven, statik bir özellik değildir; zaman içinde, tutarlı ve anlaşılır etkileşimlerle inşa edilir [46]. Kullanıcıların bir sisteme soru sorabildiği, onun mantığına meydan okuyabildiği ve tatmin edici yanıtlar alabildiği bir ortam, körü körüne bir kabullenmeden çok daha sağlam bir güven temeli oluşturur. Bu bağlamda, XAI'nin nihai hedefi, yalnızca şeffaflık sağlamak değil, aynı zamanda insan ve makine arasında etkili bir işbirliği ve ortak anlayış zemini yaratmaktır [47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59].

3 METODOLOJİ: YORUMLAYICI DÖNGÜYÜ OPERASYONEL HALE GETİRMEK

Önceki bölümde ortaya konan felsefi temelden hareketle, bu bölüm, hermenötik döngü kavramını somut, tekrarlanabilir ve pratik bir hesaplama çerçevesine nasıl dönüştürdüğümüzü detaylandırmaktadır. Amacımız, "anlama" eyleminin diyalojik doğasını, bir makine öğrenmesi modelinin davranışını yorumlamak için kullanılabilecek bir dizi algoritmik adıma ve yazılım mimarisine tercüme etmektir. Bu süreç, felsefi bir ilkenin (parça-bütün ilişkisi) operasyonel bir XAI metodolojisine dönüştürülmesini temsil eder. Önerdiğimiz Hermai çerçevesi, bu dönüşümü üç temel mimari ilke üzerine inşa eder: (1) Anlamlı bir diyalog için temel bir ön koşul olan bağlamsal gerçekçilik, (2) Hermenötik döngüyü doğrudan yansıtan çift modlu bir açıklama yapısı ve (3) Bu yapıyı araştırmacılar ve uygulayıcılar için erişilebilir kılan açık kaynaklı bir kütüphane.

3.1 Felsefi Temelden Teknik Mimarîye: Bağlam ve Diyalog

Hermenötik geleneğin temelinde, bir yorumun içinde bulunduğu "ufuk" veya bağlamdan ayrılamayacağı ilkesi yatar. Bir XAI çerçevesi için bu ilke, üretilen açıklamaların modelin öğrendiği veri dünyasının istatistiksel ve anlamsal gerçekliğine sadık kalması gerektiği anlamına gelir. Bu nedenle, metodolojimizin temeline, özellikler arasındaki ilişkileri göz ardı eden ve dolayısıyla "bağlam-körü" olan pertürbasyon tekniklerini reddeden bir yaklaşım yerleştirdik. Bunun yerine, veri dağılımının kovaryans yapısını onurlandıran **Bağlamsal Pertürbasyon Motoru**'nu mimarinin temel katmanı olarak tasarladık. Bu motor, anlamlı bir yorumlama diyaloğu için gerekli olan "gerçekçi" soruların sorulabilmesini sağlar.

Bu temel üzerine inşa edilen **Çift Modlu Yorumlama Sistemi**, hermenötik döngünün kendisini doğrudan modellemektedir. Bu mimari, kullanıcıya iki farklı, ancak birbirini tamamlayan "mercek" sunar:

1. **Yerel Anlatı Modu (LocalExplainer)**: Bu mod, döngünün "parça" analizine odaklanır. Kullanıcının, tekil ve somut bir model kararının "nedenlerini" ve "karşı-olgusal" alternatiflerini keşfetmesini sağlar. Bu, bir metindeki tek bir kelimenin anlamını derinlemesine incelemeye benzer.
2. **Genel Kural Modu (GeneralExplainer):** Bu mod, döngünün "bütün" sentezine odaklanır. Modelin genel davranışını yöneten temel kuralları ve stratejileri ortaya çıkararak, kullanıcının modelin "karakterini" veya genel mantığını anlamasına yardımcı olur. Bu da, tek tek kelimelerden yola çıkarak metnin ana fikrini kavramaya benzer.

Bu iki mod arasındaki geçiş, kullanıcının bir kararın bütün içindeki yerini anlamasını ve bütünün parçalar üzerindeki etkisini görmesini sağlayarak, mekanik bir açıklamadan daha derin bir "anlama" sürecini mümkün kılar.

3.2 Çerçevenin Temel Bileşenleri ve Algoritmik Uygulaması

Hermai çerçevesi, yukarıda bahsedilen felsefi ilkeleri somutlaştıran bir dizi algoritma ve modülden oluşur.

3.2.1Bağlamsal Pertürbasyon Motoru

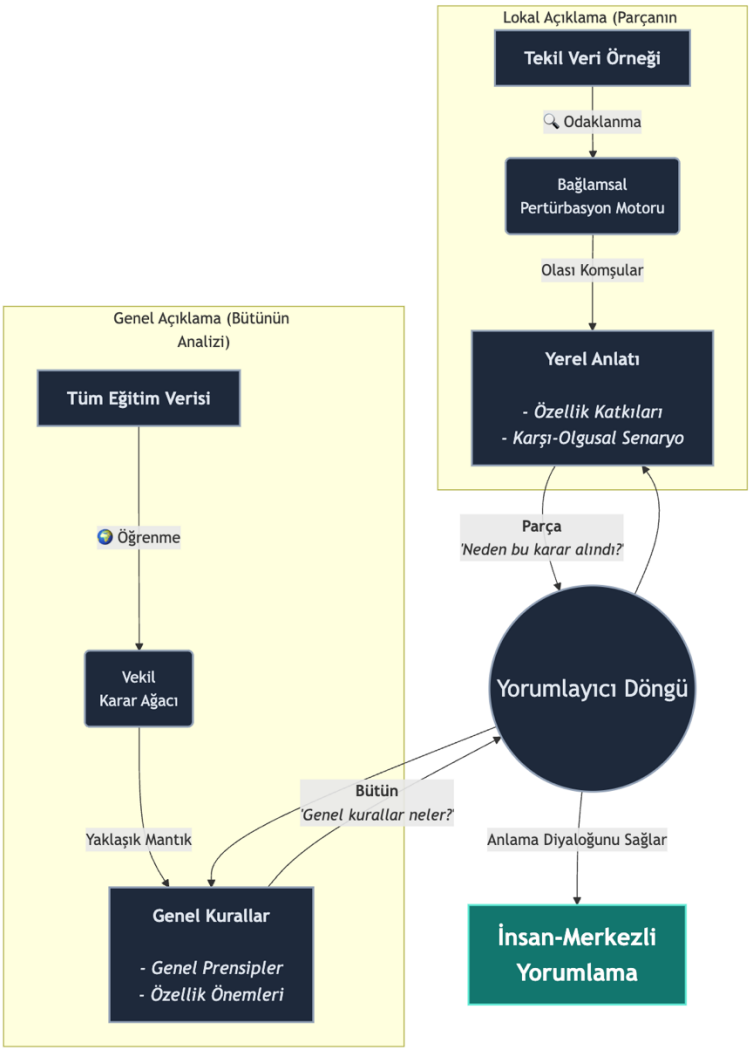
Bu motor, bir fit metodu aracılığıyla öncelikle eğitim verisinin istatistiksel dokusunu öğrenir. Sayısal özellikler için ortalama vektörünü ve özellikler arası lineer ilişkileri yakalayan kovaryans matrisini hesaplar. Kategorik özellikler için ise her bir kategorinin olasılık dağılımını çıkarır. Bir perturb işlemi istendiğinde, bu öğrenilmiş parametreleri kullanarak, hem orijinal örneğe Öklid mesafesi olarak yakın hem de veri manifolduna sadık olan çok değişkenli normal dağılımdan sentetik örnekler çeker. Bu, LIME'ın özelliklerin bağımsız olduğu varsayımına dayanan ve gerçek dışı veri noktaları yaratma riski taşıyan yaklaşımına karşı temel bir metodolojik üstünlük sağlar.

3.2.2. Çift Modlu Yorumlama Sistemi

- **Yerel Anlatı Modu (LocalExplainer)**: Bu modül, tek bir veri örneği (instance) için iki temel çıktı üretir:
    - Yerel Vekil Model: Bağlamsal pertürbasyon motoruyla üretilen yerel komşuluk üzerinde, orijinal örneğe yakınlığa göre ağırlıklandırılmış bir Lasso (L1-regularized) regresyon modeli eğitir. Lasso'nun seyreltme (sparsity) özelliği, en etkili özelliklerin seçilmesini sağlayarak gürültüsüz bir özellik önem skoru seti sunar.
    - Karşı-Olgusal Anlatı: Aynı yerel komşuluk içinde, modelin tahminini istenen zıt sonuca çeviren en yakın örneği verimli bir şekilde arar. Sonuç, "Bu karar X, Y, Z özelliklerinden ötürü alındı. Ancak, eğer Z özelliği Z' olsaydı, karar T olurdu." şeklinde, insan tarafından kolayca anlaşılabilen ve eyleme geçirilebilir bir anlatı olarak sunulur.
- **Genel Kural Modu (GeneralExplainer)**: Bu modül, kara kutu modelin genel mantığını "damıtmayı" amaçlar. Bunu, eğitim verisi (X_train) üzerindeki kara kutu model tahminlerini hedef değişken olarak kullanarak, sığ bir Karar Ağacı (DecisionTreeClassifier) gibi doğası gereği şeffaf olan bir küresel vekil model eğiterek yapar. Bu basit ağacın okunabilir IF-THEN kuralları, karmaşık modelin genellikle hangi özelliklere öncelik verdiğini, hangi etkileşimleri öğrendiğini ve genel karar stratejisinin ne olduğunu ortaya koyar.

3.3 Hermai Döngüsü: Yorumlayıcı Akışın Görselleştirilmesi ve Mantığı

Hermai çerçevesinin temelindeki diyalojik ve döngüsel anlama süreci, Şekil 1'de görselleştirilen "Yorumlayıcı Döngü" (veya Hermai Döngüsü) akış şeması ile somutlaştırılmıştır. Bu şema, yalnızca kütüphanenin teknik iş akışını değil, aynı zamanda dayandığı hermenötik felsefenin operasyonel bir modelini de temsil eder. Akış, birbirinden beslenen ve birbirini aydınlatan iki ana yoldan oluşur: tekil ve somut olan "parça" analizi ile genel ve soyut olan "bütün" analizi.



Figür 1: Hermai Çerçevesinin Yorumlayıcı Döngüsü.

Bu döngünün temel mantığı, anlama eyleminin tek yönlü bir bilgi alımından ziyade, farklı bilgi düzeyleri arasında gidip gelmeyi gerektiren dinamik bir süreç olduğu varsayımına dayanır. Akışın merkezinde yer alan Yorumlayıcı Döngü, bu dinamizmi temsil eder ve iki ana bilgi akışından beslenir:

- Parçadan Bütüne Giden Yol (Tümevarımsal Anlama): Kullanıcının anlama süreci genellikle somut bir problemle, yani tekil bir model kararıyla başlar. Akışın sol tarafında gösterildiği gibi, Lokal Açıklayıcı bu tekil veri örneğini alır. Bu noktada üretilen Yerel Anlatı, o karara özeldir; hangi özelliklerin etkili olduğunu ve sonucun nasıl değiştirilebileceğini söyler. Ancak bu "parça" bilgisi, bağlamdan kopuk olduğunda eksik kalır. Kullanıcı doğal olarak şu soruyu sorar: "Bu açıklama, modelin genel çalışma prensibiyle uyumlu mu, yoksa bir istisna mı?" Bu soru, kullanıcıyı döngünün diğer tarafına, yani "bütün"ü anlamaya iter.

- Bütünden Parçaya Giden Yol (Tümdengelimsel Anlama): Akışın sağ tarafında gösterilen Genel Açıklayıcı, modelin tüm eğitim verisi üzerindeki davranışını analiz ederek, onun genel stratejisini temsil eden Genel Kurallar'ı ortaya çıkarır. Bu "bütün" bilgisi, modelin önceliklerini ve genel mantığını soyut bir düzeyde sunar. Ancak bu soyut kurallar da, somut bir örnekle ilişkilendirilmedikçe havada kalabilir. Kullanıcı, "Bu genel kural, benim incelediğim spesifik durumda pratikte nasıl işledi?" sorusunu sorarak döngüyü tamamlar ve tekrar "parça" analizine döner.

Bu döngüsel akışın temel gerekçesi, her iki açıklama türünün de tek başına yetersiz olmasıdır. Yerel açıklama bağlamsağlarken, genel açıklama bu bağlamı doğrular ve genelleştirir. Bir kararın, modelin genel kurallarına uygun olduğunu görmek, o açıklamaya olan güveni artırır. Tersi durumda, bir yerel açıklamanın genel kurallarla çelişmesi, modelin tutarsız çalıştığına veya incelenen örneğin bir aykırı değer (outlier) olduğuna dair kritik bir ipucu sunar. Dolayısıyla, Hermai döngüsü, sadece iki farklı açıklama türü sunan bir araç değil, aynı zamanda bu iki tür arasında bir köprü kurarak, kullanıcının daha derin, eleştirel ve bütüncül bir anlayışa ulaşmasını sağlayan yapılandırılmış bir diyalog mekanizmasıdır. Bu sürecin nihai hedefi, şemanın en altında belirtildiği gibi, İnsan-Merkezli Yorumlama'ya ulaşmaktır.

3.4 Kütüphane Olarak Hermai: Pratik Uygulama ve Kullanım Senaryosu

Teorik çerçeveyi somut bir araca dönüştüren hermai Python kütüphanesi, PyPI (Python Package Index) üzerinden standart pip komutu ile kolayca kurulabilir:

```
pip install hermai
```

Kütüphanenin kullanımı, Yorumlayıcı Döngü'nün adımlarını takip edecek şekilde tasarlanmıştır. Aşağıdaki kod örneği, Titanic veri seti üzerinde eğitilmiş bir RandomForestClassifier modelini yorumlama senaryosunu göstermektedir.

ALGORİTMA 1: HermAI Python Kod Örneği

```
import seaborn as sns
from sklearn.ensemble import RandomForestClassifier
# Hermai kütüphanesinden gerekli sınıfları import edelim
from hermai.explainers import LocalExplainer, GeneralExplainer
from hermai.perturbations import TabularPerturbationGenerator
# ... (Veri yükleme ve model eğitme adımları varsayılıyor) ...
# black_box_model, X_train ve X_test hazır.
# --- Adım 1: Yorumlayıcı Döngüye Giriş ("Parça" Analizi) ---
# Açıklanacak tek bir yolcu seçelim.
instance_to_explain = X_test.iloc[0]
# Hermai'nin yerel açıklayıcısını hazırlayalım.
generator = TabularPerturbationGenerator(categorical_features=['pclass', 'sex', 'embarked'])
generator.fit(X_train)
local_explainer = LocalExplainer(black_box_model, generator)
# Bu tek yolcu için yerel bir anlatı üretelim.
local_explanation = local_explainer.explain(instance_to_explain)
# Üretilen anlatıyı yazdıralım.
print("--- YEREL AÇIKLAMA (PARÇA) ---")
print(local_explanation.narrative())
# Çıktı Örneği:
# Model predicted class 'Did not survive' with 88.00% confidence.
# Key Feature Contributions: 'sex=male' and 'pclass=3' decrease the probability.
# Counterfactual Analysis: To flip the prediction to 'Survived', one possibility
# is to change 'pclass' from 3 to 1 and 'fare' from 7.89 to 75.45.
# --- Adım 2: Bağlam Kazanma ("Bütün" Analizi) ---
# Şimdi, modelin genel kurallarını anlamaya çalışalım.
general_explainer = GeneralExplainer(black_box_model)
general_explanation = general_explainer.explain(X_train)
# Modelin genel mantığını temsil eden vekil karar ağacını görselleştirelim.
print("\n--- GENEL AÇIKLAMA (BÜTÜN) ---")
print("Modelin genel karar verme mantığı (yaklaşık):")
general_explanation.plot_surrogate_tree()
# Bu görsel, modelin genellikle önce 'sex', sonra 'pclass' ve 'age' gibi
# özelliklere baktığını gösteren bir kural ağacı çizer.
# --- Adım 3: Yorumlayıcı Döngünün Tamamlanması ---
# Artık, ilk başta analiz ettiğimiz tekil yolcunun ("parça") neden hayatta kalamadığını, modelin genel stratejisi
("bütün") ışığında daha iyi anlıyoruz. Modelin genel kuralı, 'erkek' ve '3. sınıf' yolcuların hayatta kalma olasılığını
düşük görmektir. Bizim incelediğimiz örnek de tam olarak bu profile uyduğu için, yerel açıklama genel kurallar
tarafından teyit edilmiş ve anlam kazanmış olur.
```

Bu kullanım senaryosu, Hermai'nin nasıl mekanik bir çıktı üretmenin ötesinde, kullanıcının bir kararın bağlamını anlaması ve modelin mantığıyla bir diyalog kurması için bir yapı sunduğunu göstermektedir.

REFERANSLAR

[1] Adadi, A. and Berrada, M. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access. 6, (2018), 52138–52160. DOI:https://doi.org/10.1109/ACCESS.2018.2870052.

[2] Selbst, A. D. and Powles, J. 2017. Meaningful information and the right to explanation. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT* '17). ACM, New York, NY, USA, 1–13. DOI:https://doi.org/10.1145/3155695.3156621.

[3] Lundberg, S. M. and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems, 30 (NIPS 2017). Curran Associates, Inc., 4765–4774.

[4] Ribeiro, M. T., Singh, S., and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135–1144. DOI:https://doi.org/10.1145/2939672.2939778.

[5] Molnar, C. 2020. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Lulu.com.

[6] Schleiermacher, F. D. E. 1998. Hermeneutics and Criticism and Other Writings. (A. Bowie, Ed. & Trans.). Cambridge University Press.

[7] Dilthey, W. 1989. Introduction to the Human Sciences. (R. A. Makkreel & F. Rodi, Eds.). Princeton University Press.

[8] Seker, S. E. 2015. Computerized Argument Delphi Technique. IEEE Access. 3, (2015), 368–380. DOI:https://doi.org/10.1109/ACCESS.2015.2424703.

[9] LeCun, Y., Bengio, Y., and Hinton, G. 2015. Deep learning. Nature. 521, 7553 (2015), 436–444.

[10] Krizhevsky, A., Sutskever, I., and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In Advances in Neural Information Processing Systems, 25 (NIPS 2012).

[11] Brown, T. B., et al. 2020. Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems, 33 (NeurIPS 2020).

[12] Arrieta, A. B., et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 58, (2020), 82–115.

[13] Guidotti, R., et al. 2018. A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys. 51, 5 (2018), Article 93.

[14] Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 1, 5 (2019), 206–215.

[15] Ustun, B., Spangher, A., and Liu, Y. 2019. Actionable recourse in linear classification. In Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT* '19).

[16] Holzinger, A., et al. 2019. Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 9, 4 (2019), e1312.

[17] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. 2016. Machine Bias. ProPublica.

[18] O'Neil, C. 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown.

[19] Elish, M. C. 2016. The Moral Crumple Zone: Cautionary Tales in Human-Robot Interaction. Engaging Science, Technology, and Society. 2, (2016), 40-60.

[20] Doshi-Velez, F. and Kim, B. 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

[21] Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics. 8, 8 (2019), 832.

[22] Hooker, S., et al. 2019. A Benchmark for Interpretability Methods in Deep Neural Networks. In Advances in Neural Information Processing Systems, 32 (NeurIPS 2019).

[23] Alvarez-Melis, D. and Jaakkola, T. S. 2018. On the Robustness of Interpretability Methods. ICML 2018 Workshop on Human Interpretability in Machine Learning (WHI 2018).

[24] Fel, T., Štrumbelj, E., and Benčina, D. 2022. How good is your explanation? Algorithmic stability measures for post hoc explanation methods. In IEEE Winter Conference on Applications of Computer Vision (WACV).

[25] Simonyan, K., Vedaldi, A., and Zisserman, A. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. ICLR 2014 Workshop.

[26] Selvaraju, R. R., et al. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[27] Chattopadhay, A., et al. 2018. Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks. In IEEE Winter Conference on Applications of Computer Vision (WACV).

[28] Omeiza, D., et al. 2019. Smooth Grad-CAM++: An Enhanced Inference Level Visualization for Deep Neural Networks. arXiv preprint arXiv:1908.01224.

[29] Sundararajan, M., Taly, A., and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In Proceedings of the 34th International Conference on Machine Learning (ICML 2017).

[30] Shrikumar, A., Greenside, P., and Kundaje, A. 2017. Learning Important Features Through Propagating Activation Differences. In Proceedings of the 34th International Conference on Machine Learning (ICML 2017).

[31] Ribeiro, M. T., Singh, S., and Guestrin, C. 2018. Anchors: High-Precision Model-Agnostic Explanations. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence.

[32] Kim, B., Wattenberg, M., and Gilmer, J. 2016. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In Proceedings of the 35th International Conference on Machine Learning (ICML 2018).

[33] Chen, C., et al. 2019. This Looks Like That: Deep Learning for Interpretable Image Recognition. In Advances in Neural Information Processing Systems, 32 (NeurIPS 2019).

[34] Karimi, A.-H., et al. 2020. A survey of algorithmic recourse: definitions, formulations, and challenges. arXiv preprint arXiv:2010.04050.

[35] Mothilal, R. K., Sharma, A., and Tan, C. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20).

[36] Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence. 267, (2019), 1-38.

[37] Langer, M., et al. 2021. What do we want from Explanations? A review of Human-Centered Explainable AI. arXiv preprint arXiv:2104.09705.

[38] Hohman, F., et al. 2019. Gamut: A Design Probe for Understanding Developer-Centric Explainable AI. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19).

[39] Walton, D. 2006. Fundamentals of critical argumentation. Cambridge University Press.

[40] Tolomei, G., Silvestri, F., and Jaimes, A. 2017. "My explanation is better than yours": a framework for structuring and evaluating explanations in XAI. In Proceedings of the First Workshop on Explainable AI (XAI).

[41] Cyras, K., et al. 2021. Argumentative XAI: A survey. arXiv preprint arXiv:2105.10932.

[42] Moore, J. D. 1995. Participating in Explanatory Dialogues: Interpreting and Responding to Questions in Context. MIT Press.

[43] Madumal, P., et al. 2020. A user-grounded evaluation of algorithmic recourse. arXiv preprint arXiv:2002.10313.

[44] High-Level Expert Group on AI. 2019. Ethics Guidelines for Trustworthy AI. European Commission.

[45] Jobin, A., Ienca, M., and Vayena, E. 2019. The global landscape of AI ethics guidelines. Nature Machine Intelligence. 1, 9 (2019), 389–399.

[46] Lee, J. D. and See, K. A. 2004. Trust in Automation: Designing for Appropriate Reliance. Human factors. 46, 1 (2004), 50–80.

[47] Cramer, H., et al. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. User Modeling and User-Adapted Interaction. 18, 5 (2008), 455–496.

[48] Kulesza, T., et al. 2015. Principles of explanatory debugging for intelligent systems. In Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15).

[49] Lipton, Z. C. 2018. The Mythos of Model Interpretability. Communications of the ACM. 61, 10 (2018), 36–43.

[50] Samek, W., et al. 2017. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. ITU Journal: ICT Discoveries. 1, 1 (2017).

[51] Murdoch, W. J., et al. 2019. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences. 116, 46 (2019), 22071–22080.

[52] Biran, O. and Cotton, C. 2017. Explanation and justification in machine learning: A survey. In IJCAI-17 Workshop on Explainable AI (XAI).

[53] Mittelstadt, B., Russell, C., and Wachter, S. 2019. Explaining explanations in AI. In Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT* '19).

[54] Gilpin, L. H., et al. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA).

[55] Vigano, E. and O'Sullivan, D. 2021. The Philosopher in the Loop: A Survey of the Role of Philosophy in Explainable AI. arXiv preprint arXiv:2111.00645.

[56] Sokol, K. and Flach, P. 2020. "One explanation does not fit all": A toolkit and taxonomy of AI explainability techniques. arXiv preprint arXiv:2009.03012.

[57] van der Waa, J., et al. 2021. Contrastive explanations for machine learning: A survey. ACM Computing Surveys. 54, 5 (2021), Article 105.

[58] Keane, M. T., and Smyth, B. 2020. Good Counterfactuals and Where to Find Them: A Case-Based Approach to Unlocking the Causal Secrets of Machine Learning. arXiv preprint arXiv:2012.00063.

[59] Hoffman, R. R., et al. 2018. Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.