Arda YAKAKAYI (19253519)
Bilgisayar Mühendisliği
01.12.2021

AÇIKLANABİLİR YAPAY ZEKA

(EXPLAINABLE ARTIFICIAL INTELLIGENCE)

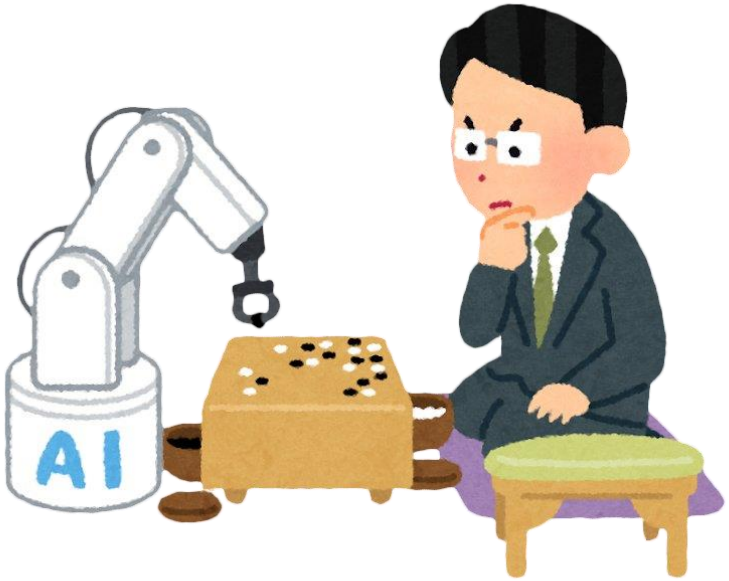
İÇİNDEKİLER

İÇİNDEKİLER

Sunum Videosu için tıklayınız.

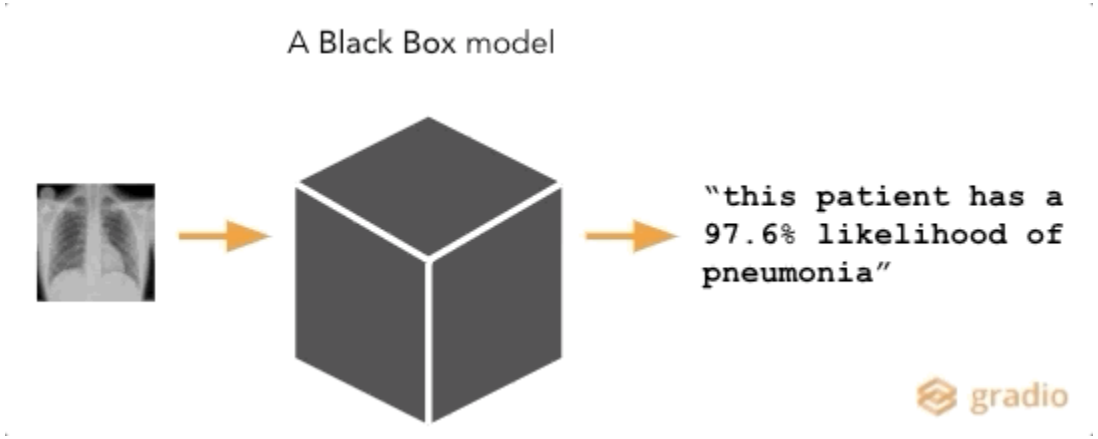
GIRIŞ

- Günümüzde Yapay zeka bir çok sektörün merkezinde yer alıyor.
- Yapay zeka günümüzde anlama, sorgulama, öğrenme ve adapte edilebilirlik gibi yönleriyle geliştiriciler için büyük bir öneme sahiptir.
- Pek çok sektörde neredeyse hiç insan müdahalesinin kalmadığı durumlar bulunmaktadır.
- Ancak Yapay Zeka zamanla insan hayatını önemli ölçüde etkiler duruma geldi ve bu durum aldığı kararların açıklanması ihtiyacını doğurdu.



GIRIŞ

- Derin Sinir Ağları (DNN'ler) gibi opak karar sistemlerinin ortaya çıkması verimli öğrenme algoritmalarının ve bunların devasa parametrik alanlarının birleşimi olan opak karar sistemlerinin tercih edilmesine yol açmıştır.
- Derin Sinir Ağları yüzlerce katmana ve milyonlarca parametreye sahip olması nedeniyle **Kara Kutu Modeli** olarak kabul görmüştür.
- Kara kutunun tersi olan **şeffaflık**, bir modelin çalıştığı mekanizmayı doğrudan anlama arayışıdır.



YORUMLANABİLİRLİK VE MAKİNE ÖĞRENİMİ

- Yorumlanabilirliğin makine öğrenimi modellerinde tercih edilmesi şu 3 nedene dayandırılabilir;

  - Yorumlanabilirlik, karar vermede tarafsızlığın sağlanmasına yardımcı olur, yani eğitim veri setindeki önyargıyı tespit etmek ve sonuç olarak düzeltmeye yardımcı olur.

  - Yorumlanabilirlik, tahmini değiştirebilecek olası çelişkili karışıklıkları vurgulayarak sağlamlığın sağlanmasını kolaylaştırır.

  - Yorumlanabilirlik, yalnızca anlamlı değişkenlerin çıktıyı çıkarsadığı, yani model akıl yürütmede altta yatan doğru bir nedenselliğin var olduğunu garanti eden bir sigorta işlevi görebilir.

O HALDE AÇIKLANABİLİR YAPAY ZEKA;

1) Yüksek düzeyde bir öğrenme performansını korurken daha açıklanabilir modeller üreten bir makine öğrenimi teknikleri paketi oluşturmayı önerir (örneğin, tahmin doğruluğu)

2) İnsanların yeni nesil yapay zekaya sahip ortakları anlamasını, uygun şekilde güvenmesini ve etkin bir şekilde yönetmesini sağlar.

TERMINOLOJI'YE GIRIŞ

TERMINOLOJI'YE GIRIŞ

Yorumlanabilirlik (Interpretability): Bir modelin, gözlemciler için anlamlı olduğu seviyeyi gösteren, bahsi geçen modelin pasif bir özelliğidir.

Açıklanabilirlik (explainability): bir modelin, işlevlerini netleştirmek veya detaylandırmak amacıyla gerçekleştirdiği herhangi bir eylem veya prosedürü ifade eden, bahsi geçen modelin aktif bir özelliğidir.

Anlaşılabilirlik (Understandability - Intellibility): Gözlemcilerin, bir modelin işlevini, o modelin iç yapısını, sahip olduğu algoritmik araçları açıklamaya gerek kalmadan anlayabilmesidir.

Anlaşılabilirlik (Comprehensibility): Bir öğrenme algoritmasının, öğrenilen bilgiyi gözlemci tarafından anlaşılabilir bir şekilde temsil etme becerisi.

Şeffaflık (Transparency): Bir modelin kendi başına anlaşılabilmesi.

AÇIKLANABILIR YAPAY ZEKA'YA GIRIŞ

AÇIKLANABILIR YAPAY ZEKA NEDIR?

"XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners"

D. Gunning

AÇIKLANABILIR YAPAY ZEKA NEDIR?

- Belirli bir hedef kitle göz önüne alındığında, açıklanabilirlik, bir modelin işleyişini net veya anlaşılması kolay hale getirmek için sunduğu ayrıntılar ve nedenlerdir.

- Modelin karmaşıklığını azaltmak veya çıktılarını basitleştirmek için herhangi bir yol, bir XAI yaklaşımı olarak düşünülmelidir.

“Açıklanabilir bir Yapay Zeka, işleyişini net veya anlaşılması kolay hale getirmek için ayrıntılar veya nedenler üreten bir Zekadır.”

NEDEN AÇIKLANABILIR YAPAY ZEKA?

- Açıklanabilirlik, yapay zekanın pratik uygulaması açısından günümüzde karşılaştığı ana engellerden biridir.
- Araştırma topluluğu ile iş sektörleri arasındaki boşluk.
- Bilgiye ulaşmada sağladığı kolaylık.

AÇIKLANABILIRLIK NE IÇIN? KIMLER IÇIN?

- Açıklanabilir Yapay Zeka zamanla çok fazla sayıda kitleye hitap etmeye başladı.
- Tüm bu farklı hedef kitleler açıklanabilirliğin gerekliliği açısından bir fikir oluşmasına katkı sağlar.

XAI Goal	Main Target Audience
Trustworthiness	Domain experts, users of the model affected by decisions
Causality	Domain experts, managers and executive board members, regulatory entities/agencies
Transferability	Domain experts, data scientists
Informativeness	All
Confidence	Domain experts, developers, managers, regulatory entities/agencies
Fairness	Users affected by model decisions, regulatory entities/agencies
Accessibility	Product owners, managers, users affected by model decisions
Interactivity	Domain experts, users affected by model decisions
Privacy awareness	Users affected by model decisions, regulatory entities/agencies

- *Tablo (x.x) Yukarıdaki tabloda açıklanabilirliğe dair tanımlamalar ve hedefledikleri kitleler görülmektedir.*

AÇIKLANABILIRLIK NE IÇIN? KIMLER IÇIN?

Güvenilirlik (Trustworthiness): Bir modelin belirli bir problemle karşılaştığında amaçlandığı gibi hareket edip etmeyeceğinin güveni olarak düşünülebilir.

“Her açıklanabilir modelin güvenilir olması beklenir. Ancak her güvenilir model açıklanamaz”

AÇIKLANABILIRLIK NE IÇIN? KIMLER IÇIN?

Nedensellik (Causality): Veri değişkenleri arasında bir neden sonuç ilişkisi bulmak. Bir ML modeli yalnızca öğrendiği veriler arasındaki korelasyonları keşfeder ve bu nedenle bir neden-sonuç ilişkisini ortaya çıkarmak için yeterli olmayabilir.

Bununla birlikte, nedensellik **korelasyon*** içerir, bu nedenle açıklanabilir bir ML modeli, nedensellik çıkarım teknikleri tarafından sağlanan sonuçları doğrulayabilir veya mevcut veriler içindeki olası nedensel ilişkilerin ilk sezgisini sağlayabilir.

Korelasyon: *olasılık kuramı ve istatistikte iki değişken arasındaki doğrusal ilişkinin yönünü ve gücünü belirtir.*

AÇIKLANABILIRLIK NE IÇIN? KIMLER IÇIN?

Aktarılabilirlik (Transferability): Öğrenme Modellerinin aktarılabilir olması önemlidir.

Bilgilendiricilik(Informativeness): Açıklanabilir ML modelleri, ele alınan problem hakkında bilgi vermelidir.

Güven (Confidence): Açıklanabilir bir model, çalışma rejiminin güvenilirliği hakkında bilgi içermelidir.

Adalet (Fairness): Açıklanabilir bir ML modeli, eldeki modelin adil veya etik bir analizine izin vererek, bir sonucu etkileyen ilişkilerin net bir görselleştirilmesini önerir.

AÇIKLANABILIRLIK NE IÇIN? KIMLER IÇIN?

Erişilebilirlik (Accessibility): Açıklanabilir modeller ile, ilk bakışta anlaşılmaz görünen algoritmalarla uğraşmak zorunda kaldıklarında uzman olmayan kullanıcıların yükünü hafifletmek amaçlanır.

Etkileşim (Interactivity): Açıklanabilir Yapay Zeka ve buna değinen modellerin hedeflerinden biri de ince ayar yapma ve modellerle etkileşim kurma yetenekleridir.

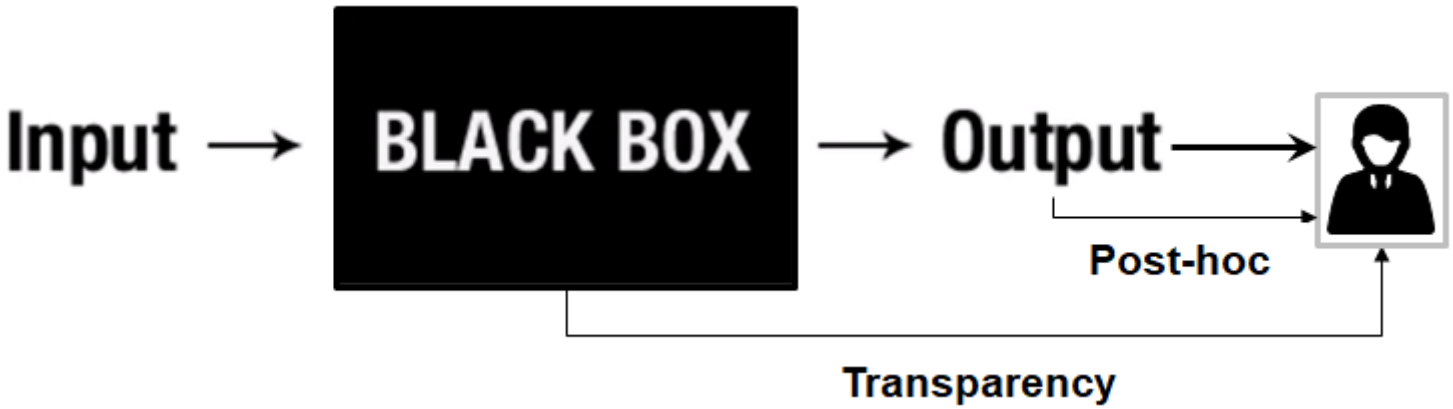
Gizlilik Bilinci (Privacy Awareness): Makine öğrenimi modellerinde açıklanabilirliğin sağladığı avantajlardan biri de mahremiyeti değerlendirme yeteneğidir.

MODELLERİN SINIFLANDIRILMASI

MODELLERIN SINIFLANDIRILMASI

- Modeller, farklı yazar ve uzmanlar tarafından pek çok açıdan sınıflandırılabilmektedir.

- Açıklanabilirliklerine göre en yaygın sınıflandırmada Modeller 2 grupta incelenir;
  - Şeffaf Modeller
  - POST-HOC açıklanabilir modeller



ŞEFFAFLIK

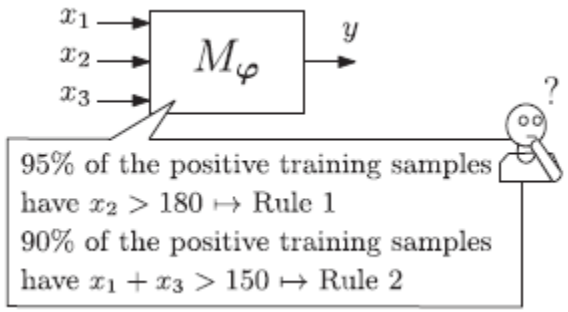
- Şeffaflık, 3 seviyede incelenmektedir;
  - Algoritmik şeffaflık
  - Ayrıştırılabilirlik
  - Simule Edilebilirlik

- Şeffaf modeller kendi başlarına bir dereceye kadar yorumlanabilirlik sağlar.

- Bu sınıfların her biri kendi öncüllerini içerir. Örneğin simüle edilebilir bir model aynı zamanda ayrıştırılabilir ve algoritmik olarak şeffaf bir modeldir.

ALGORITMIK ŞEFFAFLIK:

- Kullanıcının, girdi verilerinden herhangi bir çıktı üretmek için model tarafından takip edilen süreci anlama yeteneği ile ilgilenir.
- Lineer bir regresyon şeffaf kabul edilir, çünkü kapsamı anlaşılabilir ve mantıklıdır, bu da kullanıcının karşılaşabileceği her durumda modelin nasıl davranacağını anlamasını sağlar.
- Algoritmik olarak şeffaf modeller için en gerekli şart, modelin matematiksel analiz ve yöntemlerle çözümlenebilir olması gerektiğidir.

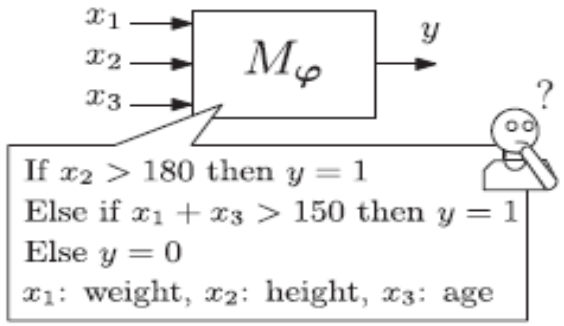


Regresyon ya da regresyon analizi, *ilgilenilen iki veya daha fazla değişken arasındaki ilişkiyi analiz etmenize ve anlamamıza yardımcı olan istatistiksel bir yöntemdir.*

Stokastik *kelimesi , rastgele bir olasılıkla bağlantılı bir sistem veya süreç anlamına gelir. Dolayısıyla, Stokastik Gradyan İnişinde, her bir yineleme için tüm veri seti yerine rastgele birkaç örnek seçilir.*

AYRIŞTIRILABILIR ŞEFFAFLIK:

- Bir modelin her bir parçasını (girdi, parametre ve hesaplama) açıklama yeteneği anlamına gelir.
- Bu özellik, bir modelin davranışını anlama, yorumlama veya açıklama yeteneğini güçlendirebilir.
- Her model bu özelliği yerine getiremez.
- Ayrıştırılabilirlik, her girdinin kolayca yorumlanabilir olmasını gerektirir.
- Algoritmik olarak şeffaf bir modelin ayrıştırılabilir olması için eklenen kısıtlama, modelin her parçasının ek araçlara ihtiyaç duymadan bir insan tarafından anlaşılabilir olması gerektiğidir.

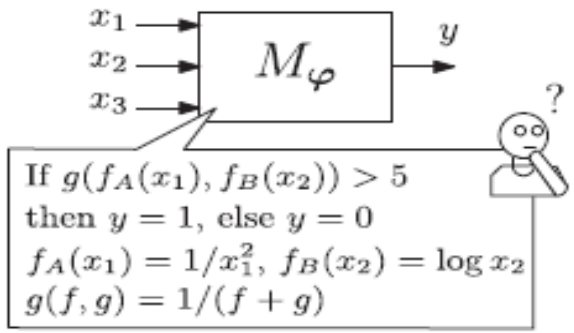


SIMULE EDILEBILIR ŞEFFAFLIK:

- Bir modelin tam olarak bir insan tarafından simüle edilebilme veya üzerinde düşünülebilme yeteneğidir.

- Basit ama kapsamlı kural tabanlı sistemler bu özelliğin dışında kalırken, tek bir algılayıcı sinir ağı bunun içine girer.

- Ayrıştırılabilir bir modele simule edilebilme özelliği kazandırmak, ayrıştırılan parçaların yeterince bağımsız olmasını gerektirir.



ŞEFFAF MAKINE ÖĞRENMESI MODELLERI

Linear/Logistic regression

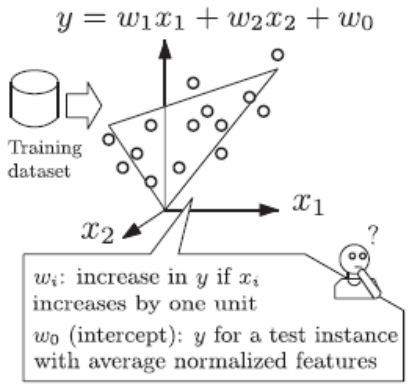
- İkili (ikili) olan bir bağımlı değişkeni (kategori) tahmin etmeye yönelik bir sınıflandırma modelidir. Bağımlı değişken sürekli olduğunda, doğrusal regresyon onun eş anlamlısı olacaktır.

- Bu model, verilere esnek bir uyum sağlar.

- Model onu kimin yorumlayacağına bağlı olarak her iki kategoriye de girmektedir. (POST-HOC ve şeffaflık)

ŞEFFAF MAKINE ÖĞRENMESI MODELLERI

- Lojistik regresyon, denetimli öğrenmede en basit sınıflandırma modelleri arasında yer alsa da dikkat edilmesi gereken kavramlar vardır.
- Ayrıştırılabilirliğin ve simule edilebilirliğin sağlanması için lojistik veya doğrusal regresyon gibi bir modelin boyutunun sınırlı olması ve kullanılan değişkenlerin kullanıcıları tarafından anlaşılabilir olması gerekir.
- Modelin girdileri, karmaşık veya anlaşılması zor, yüksek düzeyde tasarlanmış özellikler ise, eldeki modelin ayrıştırılabilmesini zorlaştırır.



ŞEFFAF MAKINE ÖĞRENMESI MODELLERI

Karar ağaçları

- Şeffaflık için her türlü kısıtlamayı kolaylıkla yerine getirebilecek bir model örneğidir.

- Karar ağaçları, regresyon ve sınıflandırma problemlerini desteklemek amacıyla karar vermek için hiyerarşik yapılardır.

- En basit haliyle, karar ağaçları simüle edilebilir modellerdir.

ŞEFFAF MAKINE ÖĞRENMESI MODELLERI

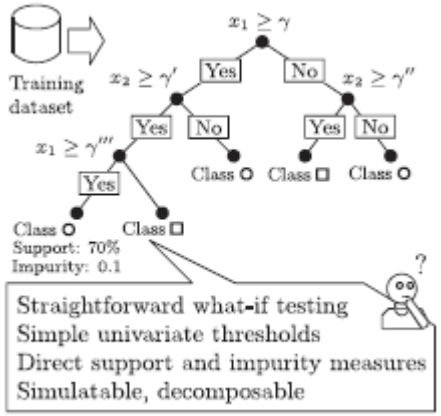
- Ancak, özellikleri onları ayrıştırılabilir veya algoritmik olarak şeffaf hale getirebilir.
- Karar ağaçları her zaman şeffaf modellerin farklı kategorileri arasında kalmıştır.
- Şeffaf modeller içerisinde her kategoriye uyum sağlayabilmesine rağmen sahip oldukları özellikleri onları algoritmik şeffaf modeller kategorisine sokabilmektedir.
- Karar ağaçları, hazır şeffaflıkları nedeniyle uzun süredir karar destek bağlamlarında kullanılmaktadır.
- Bu modellerin pek çok uygulaması geniş bir kullanım alanına sahiptir.

ŞEFFAF MAKINE ÖĞRENMESI MODELLERI

- Genelleme özelliklerinin zayıf olması, öngörü gerektiren uygulamalarda kullanılabilirliğini azaltmaktadır.

- Ağaç toplulukları, alt dallarındaki farklı ağaçların elde ettikleri tahminleri bir araya getirerek düşük performans sorununun önüne geçmeyi hedefler.

- Ancak bu durum karar ağaçlarının şeffaflığını kaybetmesine yol açar ve bu da POST-HOC analizi daha avantajlı hale getirir.



ŞEFFAF MAKINE ÖĞRENMESI MODELLERI

K- En yakın Komşular

- Sınıflandırma problemlerini metodolojik olarak basit bir şekilde ele alır

- K, bilinmeyen noktanın en yakın komşularının miktarını temsil eder.

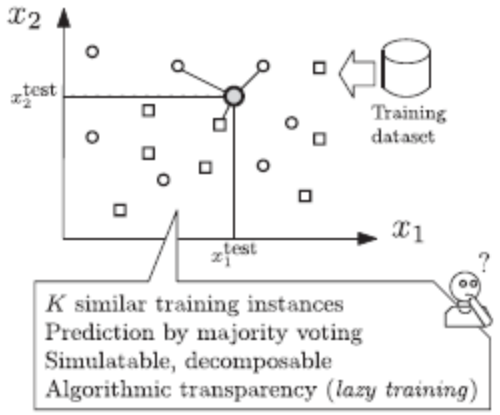
- K en yakın komşusunun sınıflarını oylayarak bir test örneğinin sınıfını tahmin eder. İlişki, veriler arasındaki mesafe ölçüleri ile belirlenir.

- Regresyon problemleri bağlamında kullanıldığında, oylamanın yerini en yakın komşularla ilişkili hedef değerlerin bir toplamı (örneğin ortalama) alır.

ŞEFFAF MAKINE ÖĞRENMESI MODELLERI

- Bu tahmin yaklaşımı, geçmişteki benzer vakaların sonucuna göre karar veren, insanların, deneyime dayalı karar verme yaklaşımına benzemektedir

- Çok yüksek bir K, kullanıcı tarafından model performansının tam simülasyonunu engeller. Benzer şekilde, karmaşık özelliklerin veya uzaklık işlevlerinin kullanılması, modelin ayrıştırılabilirliğini engelleyecek ve yorumlanabilirliğini yalnızca algoritmik işlemlerinin şeffaflığıyla sınırlayacaktır.



ŞEFFAF MAKINE ÖĞRENMESI MODELLERI

Kural'a dayalı öğrenme

- Öğrenilmesi amaçlanan verileri karakterize etmek için kurallar üreten her modeli ifade eder.

- Kurallar, basit koşullu if-then kuralları olabileceği gibi, bilgilerini oluşturmak için basit kuralların daha karmaşık kombinasyonları şeklinde de olabilir.

- Tahminlerini açıklayan kurallar üreterek karmaşık modelleri açıklamak için sıklıkla kullanılmış olan açıkça şeffaf modellerdir.

ŞEFFAF MAKINE ÖĞRENMESI MODELLERI

- Bu genel model ailesiyle de bağlantılı olarak, bulanık kural tabanlı sistemler, daha geniş bir eylem kapsamı için tasarlanır ve kesin olmayan alanlar üzerinde sözlü olarak formüle edilmiş kuralların tanımlanmasına izin verir.

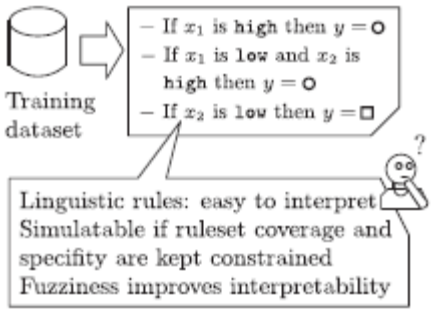
- **Bulanık sistemler**;
  1. Dilsel terimlerle çalıştıkları için daha anlaşılır modelleri güçlendirirler.
  2. Belirli derecelerde belirsizliğe sahip bağlamlarda klasik kural sistemlerinden daha iyi performans gösterirler.

- Temel iki sorun mevcuttur;
  - oluşturulan kuralların kapsamı (miktarı)
  - özgüllüğü (uzunluğu)

ŞEFFAF MAKINE ÖĞRENMESI MODELLERI

- Kuralların miktarı arttıkça anlaşılabilirlik azalmakta, ancak performans artmaktadır.

- Çok sayıda girdi veya sonuç bir kuralın yorumlanmasını zorlaştırır.



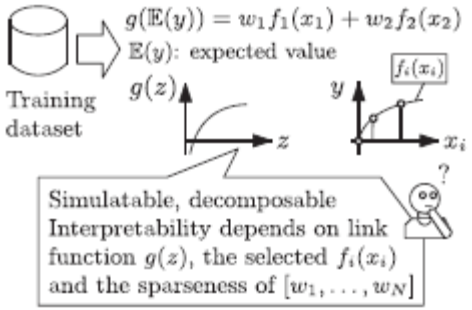
ŞEFFAF MAKINE ÖĞRENMESI MODELLERI

Genel Eklemeli Modeller

- İstatistikte kullanılan doğrusal bir modeldir.
- Bu modelin amacı, toplam kompozisyonu, tahmin edilen değişkene yaklaşan düzgün fonksiyonları çıkarmaktır.
- Finans, çevre çalışmaları, jeoloji, sağlık, biyoloji ve enerji alanlarındaki pratik uygulamalar için kullanıcılara anlaşılabilirlik sağlamaktadır.
- Modelin yorumlanmasını daha da kolaylaştırmak için görselleştirme yöntemleri kullanılır.

ŞEFFAF MAKINE ÖĞRENMESI MODELLERI

- Modelin yorumlanmasını daha da kolaylaştırmak için görselleştirme yöntemleri kullanılır.



ŞEFFAF MAKINE ÖĞRENMESI MODELLERI

Bayesian Modeller

- Bir Bayes modeli genellikle, bağlantıları bir dizi değişken arasındaki koşullu bağımlılıkları temsil eden, olasılıksal yönlendirilmiş döngüsel olmayan grafik modellerdir

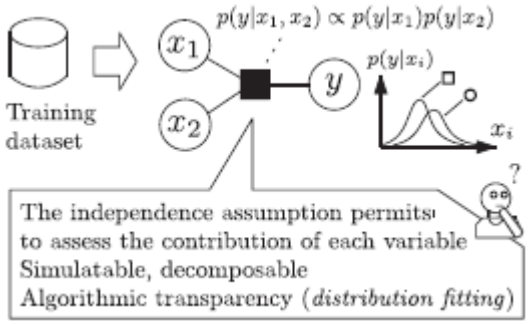
- Örneğin, bir Bayes ağı, hastalıklar ve semptomlar arasındaki olasılıksal ilişkileri temsil edebilir. Verilen semptomlar, ağ, çeşitli hastalıkların varlığının olasılıklarını hesaplamak için kullanılabilir.

- Bu modeller aynı zamanda özellikler ve hedef arasındaki ilişkilerin açık bir temsilini de iletir, bu durumda değişkenleri birbirine bağlayan bağlantılar açıkça verilir.

ŞEFFAF MAKINE ÖĞRENMESI MODELLERI

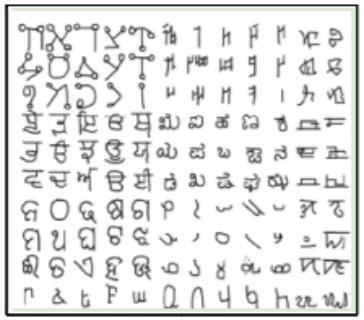
- Bayes modellerinin bilişsel modelleme, balıkçılık, oyun, iklim, ekonometri veya robotik gibi çeşitli uygulamalarda büyük içgörülere yol açtığı gösterilmiştir.

- Ayrıca, ağaç topluluklarının ortalamasının alınması gibi diğer modelleri açıklamak için de kullanılmıştır.

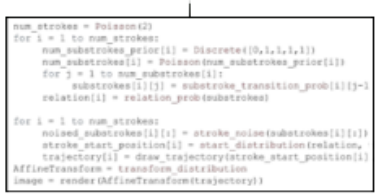


ŞEFFAF MAKINE ÖĞRENMESI MODELLERI

Training Data
1623 Characters

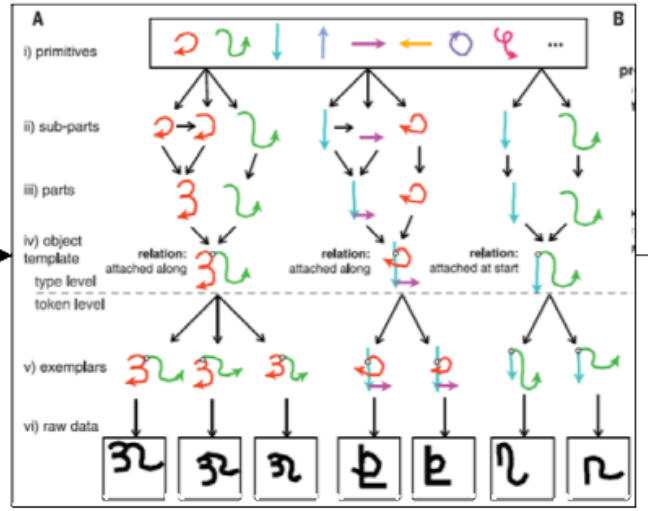


Bayesian Program Learning

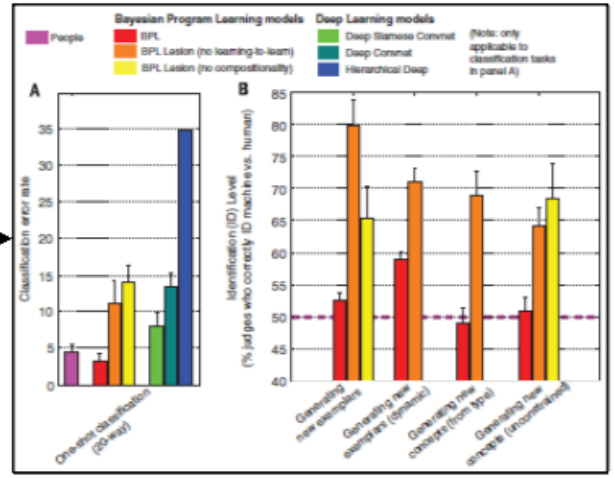


Seed Model
A simple Probabilistic Program that describes the parameters of character generation

Concept Learning Through Probabilistic Program Induction



Generative Model
Recognizes characters by generating an explanation of how a new test character might be created (i.e., the most probable sequence of strokes that would create that character)



Performance
This model matches human performance and out performs deep learning

POST-HOC ANALİZİ

- Post-hoc açıklanabilirlik, yorumlanabilirliklerini geliştirmek için, çeşitli araçlara başvurarak kolayca yorumlanamayan modelleri hedefler. Bu araçları iki kısımda inceliyoruz;
  - Model Agnostic
  - Model Spesifik

- Bunun yanı sıra başvurduğumuz araçlar şu şekilde sıralanabilir;
  - Metin açıklamaları
  - Görsel açıklamalar
  - Yerel açıklamalar
  - Örnek açıklamalar
  - Basitleştirme yoluyla açıklamalar
  - Özellik alaka açıklama teknikleri

POST-HOC ANALİZİ

- **Metin Açıklamaları;**
  - Metin açıklamaları, modelden elde edilen sonuçları açıklamak için metinsel açıklamalar üretir.
  - Modelin işleyişini temsil eden semboller üreten her yöntemi içerir.
  - Bu semboller, modelden sembollere semantik bir haritalama vasıtasıyla algoritmanın mantığını göstermeyi hedefler.

- **Görsel Açıklama;**
  - Modelin davranışını görselleştirmeyi amaçlar
  - Görselleştirme yöntemlerinin birçoğu, boyutsallık azaltma teknikleri ile birlikte gelir.
  - Karmaşık etkileşimleri tanıtmanın en uygun yoludur.
  - Diğer tekniklerle birlikte kullanılabilir

POST-HOC ANALİZİ

- **Yerel Açıklamalar**
  - Çözüm uzayını bölümlere ayırır.
  - Tüm model için uygun olan bu, daha az karmaşık çözüm alt uzaylarına açıklamalar vererek açıklanabilirliği ele alır.
  - Sistem, açıklanabilirliği sağlamak amacıyla sistemin sadece bir kısmı anlatılacak şekilde indirgenebilir.

- **Örneklendirme ile Açıklama**
  - Belirli bir model tarafından oluşturulan sonuçla ilgili veri örneklerinin çıkarılması ile kullanıcıya açıklamayı hedefler
  - İnsanların açıklama yaparken kullanmış oldukları örneklemeye benzetilebilecek olan bu durum analizi yapılan modelin sahip olduğu iç ilişkileri ve korelasyonları açıklayan örneklere odaklanır.

POST-HOC ANALİZİ

- **Basitleştirme ile Açıklama**
  - Modelin orjinaline dayanan tamamen yeni bir modelin oluşturulmasına dayanır.
  - Bu basitleştirilmiş model, orjinal modelin karmaşıklığını azaltırken, performans durumunu olabildiğince korumaya ve optimize etmeye çalışır.
  - Yeni modelin kullanımı, orjinal modelden daha kolaydır.

- **Özellik Alaka Açıklama**
  - Yönetilen değişkenleri için bir uygunluk puanı hesaplayarak bir modelin iç işleyişini netleştirir.
  - Modelin sahip olduğu özelliklerin bu uygunluk puanı üzerindeki etkisini inceler.
  - Farklı değişkenler arasındaki puanlar karşılaştırılması sonucu, değişkenlerin, çıktı için olan etkisi (önemi) ortaya çıkarılır.

POST-HOC ANALİZİ

- ML modelleri, şeffaf olmak için gerekli kriterlerden herhangi birini karşılamadığında, kararlarını açıklamak için modele ayrı bir yöntem tasarlanmalı ve uygulanmalıdır.

- POST-HOC tekniklerinin esas amacı;

“Bir modelin, herhangi bir girdi için tahminlerini nasıl ürettiği konusunda anlaşılır bilgiler iletmektir.”

- Modelleme sonrası açıklanabilirlik de denilebilmektedir.

POST-HOC ANALİZİ

MODEL AGNOSTİK

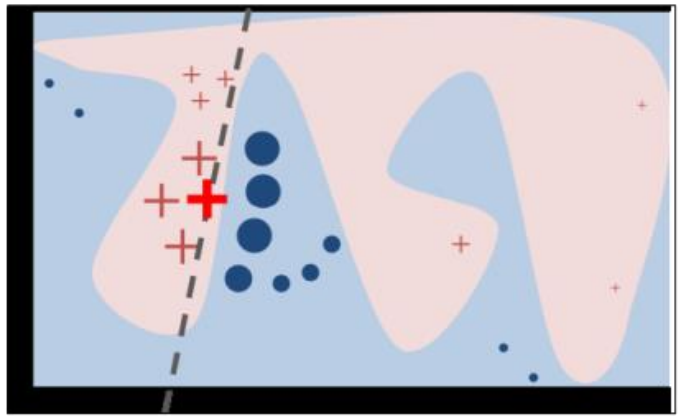
Maalesef çoğu makine öğrenimi modelini doğrudan yorumlamak mümkün değil. Rastgele ormanlar, gradyan destekli makineler ve sinir ağları gibi popüler modeller için modelden bağımsız yöntemlere ihtiyacınız vardır;

- Basitleştirme ile Açıklama
- Özellik Alaka Açıklama
- Yerel Açıklamalar
- Görsel Açıklama

POST-HOC ANALİZİ

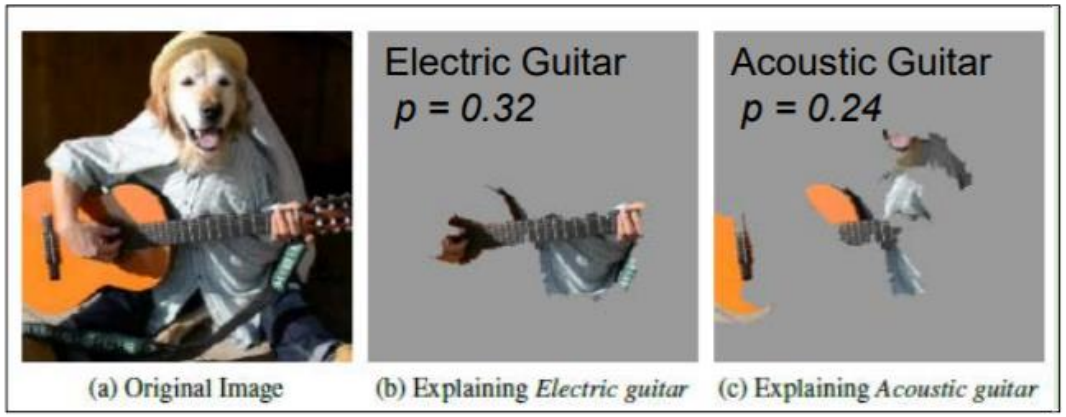
Local Interpretable Model-agnostic Explanations (LIME)

Black-box Induction



The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background. The bright bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful. .

Example Explanation



- **LIME** is an algorithm that can explain the predictions of any classifier in a faithful way, by approximating it locally with an interpretable model.

- **SP-LIME is** a method that selects a set of representative instances with explanations as a way to characterize the entire model.

POST-HOC ANALİZİ

MODEL SPESIFIK

- **Sığ ML modellerinde post-hoc açıklanabilirlik:**
  - Karar ağaçları, ağaç toplulukları, çoklu sınıflandırıcı sistemler
  - Vektör makinelerin desteklenmesi

- **Derin öğrenmede açıklanabilirlik:**
  - Çok katmanlı Sinir Ağları
  - Evrişimli Sinir Ağları
  - Tekrarlayan Sinir Ağları
  - Hybrid transparent and black-box methods

AÇIKLANABİLİR YAPAY ZEKA'NIN KARŞILAŞTIĞI KRİTER VE SORUNLAR

AÇIKLANABILIR YAPAY ZEKA'NIN KARŞILAŞTIĞI KRİTER VE SORUNLAR

Yorumlanabilirlik ve Performans İlişkisi

- Daha karmaşık olan modellerin daha doğru olduğuna dair net bir şey söylenemez.

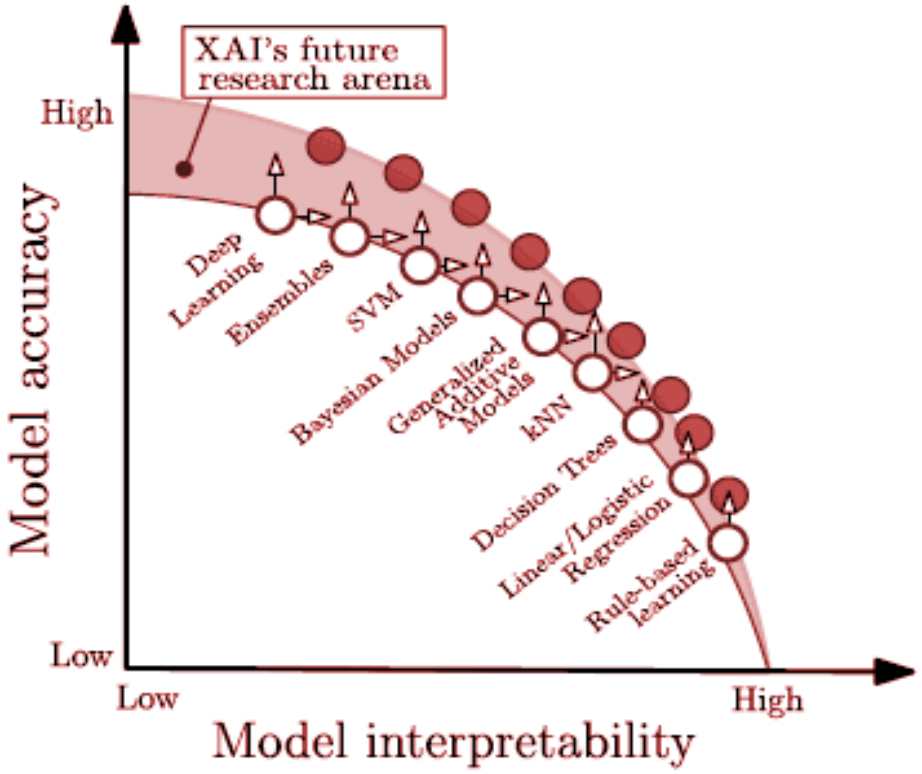
- Verilerin bir modelin çalışmasında büyük etkisinin olduğu durumlarda modelin sağlıklı olup olmadığını anlayamayız. Özellikle endüstri gibi alanlarda bu durumla çok sık karşılaşılır.

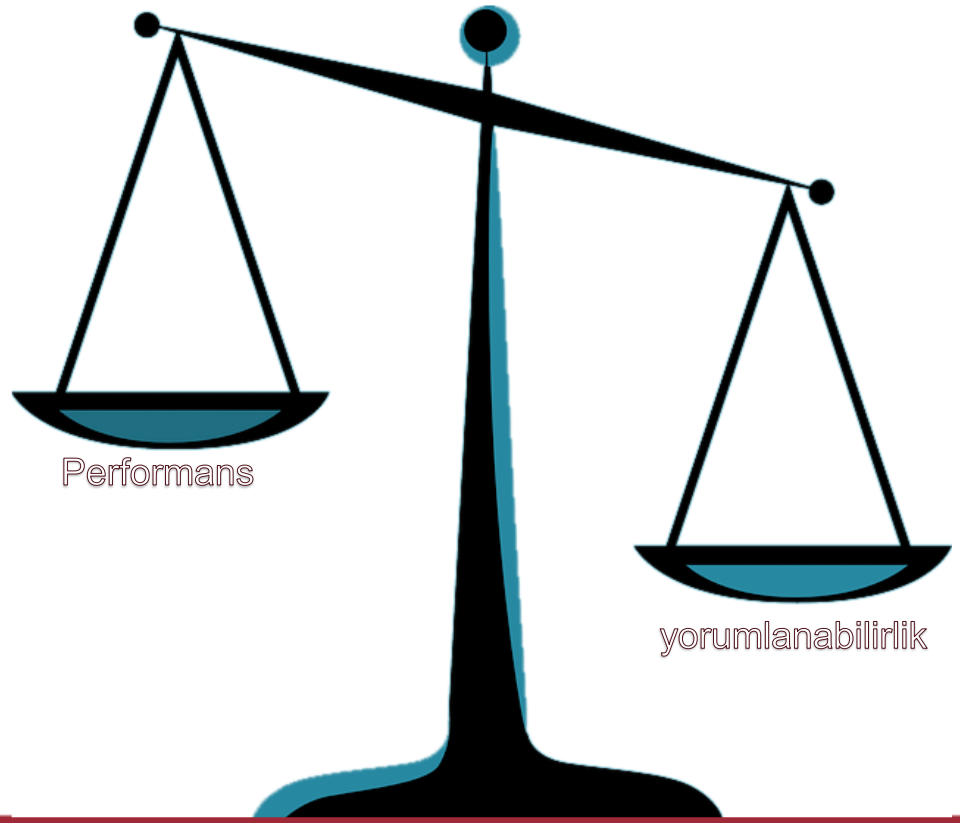
- Daha karmaşık modeller, daha basit benzerlerine göre çok daha fazla esnekliğe sahiptir.

- Ayrıca model ne kadar karmaşık ve iyi yapılandırılmış gibi olursa olsun, verilerin eksik ya da hatalı girilmesi sonucunda bir anlamı kalmayacak ve hatalı sonuç verecektir.

AÇIKLANABILIR YAPAY ZEKA'NIN KARŞILAŞTIĞI KRITER VE SORUNLAR

- Performansa önem verilmesi, yorumlanabilirlikte kayba yol açarken aksi durumda performansta bir düşüş olduğu gözlemlenmektedir.





AÇIKLANABILIR YAPAY ZEKA'NIN KARŞILAŞTIĞI KRİTER VE SORUNLAR

Kavramların etkisi

- Modelin ya da modellerin açıklanabilir olması için literatüre uyan ortak ve benimsenmiş kavramlar kullanılmalıdır.

- Modellerin açıklanabilirliğinde somutlaştırma büyük önem arz eder.

AÇIKLANABILIR YAPAY ZEKA'NIN KARŞILAŞTIĞI KRİTER VE SORUNLAR

- Henüz yeni yeni ortaya çıkan bir alan olduğundan global çapta kabul edilmiş bir teminolojisi bulunmamaktadır.

- Bu nedenle topluluk tarafından önerilen mevcut ölçüm prosedürlerini ve araçlarını desteklemek için daha niceliksel, genel XAI kavramlarına ihtiyaç vardır.

Nedensellik:

- Açıklanabilirlik sonuçtan ziyade nedene odaklanmalıdır.

2 + 5 = 7

Ama neden?

AÇIKLANABILIR YAPAY ZEKA'NIN KARŞILAŞTIĞI KRİTER VE SORUNLAR

Verilerdeki Tutarlılık ve Genel Kabuller

- Açıklanabilir yapay zekada nesnelliğin sağlanması ve doğruluk payının artmasının yanı sıra elde edilen verilerin resmileşmesi için sosyal bilimlerden çokça yararlanılır.

Gizlilik

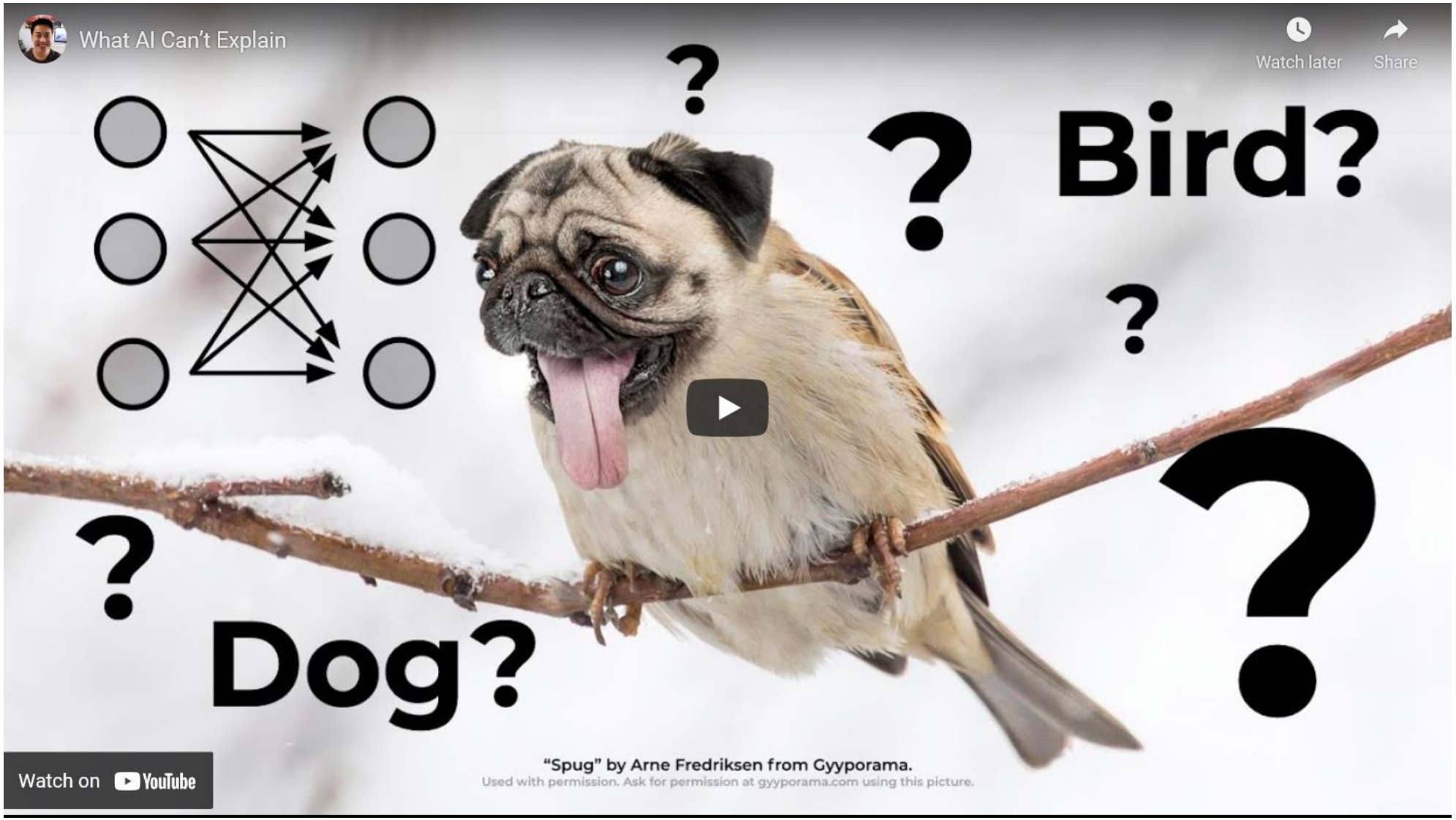
- Açıklanabilir yapay Zeka, modellerin kullanıcı veya gözlemciye aktarılmasını sağlarken gizliliği göz önünde bulundurmalıdır.

- Üretken modeller, kendilerine öğretildiği takdirde ve gizliliğin ihlali söz konusu olduğunda elde ettikleri verilerin benzerlerini üreterek modellerin manipüle edilmesine yol açabilir ve hatta bu verileri kullanarak öngörülerini geliştirerek orjinal modelin alacağı kararları verebili yapacağı tahminleri öngörebilir.

AÇIKLANABILIR YAPAY ZEKA'NIN KARŞILAŞTIĞI KRITER VE SORUNLAR

Öğrenim Modelleri

- Model tipi karşılaşılması beklenen ya da düşünülen ilişki tipine göre seçilmelidir.
- Önceki deneyimlerin de takip edilmesi büyük önem arz etmektedir.
- Modelin çıktısı, modelin öğrendiği her şey hakkında bilgi vermeli ve kümülatif deneyime izin vermelidir.


What AI Can't Explain
Watch later
Share
?
?
Bird?
?
?
Dog?
?
"Spug" by Arne Fredriksen from Gyyporama.
Used with permission. Ask for permission at gyyporama.com using this picture.
Watch on YouTube

KAYNAKLAR

- https://www.analyticsvidhya.com
  - https://www.analyticsvidhya.com/blog/2021/01/explain-how-your-model-works-using-explainable-ai/
- https://pythonrepo.com
  - https://pythonrepo.com/repo/microsoft-interpret-python-machine-learning
  - https://pythonrepo.com/repo/pbiecek-DALEX-python-machine-learning
- https://www.youtube.com
  - https://www.youtube.com/watch?v=fQ2eNFCSRiA
  - https://www.youtube.com/watch?v=0np9N0C5ukc

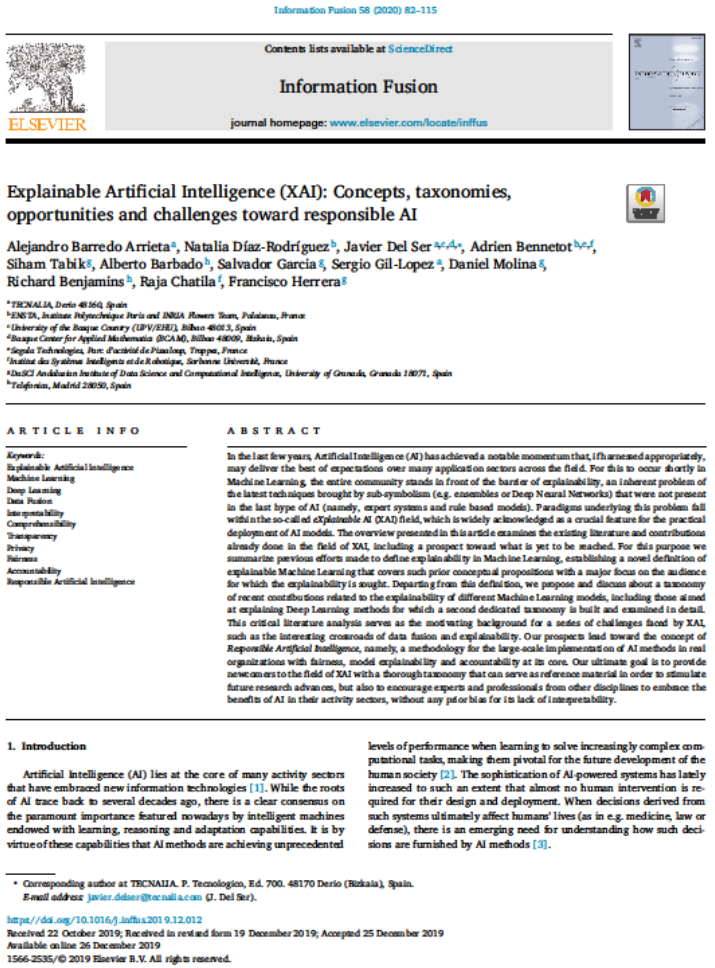
KAYNAKLAR

- https://www.cc.gatech.edu
  - https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf
- https://ambiata.com
  - https://ambiata.com/blog/2021-04-12-xai-part-1/
- https://veribilimcisi.com
  - https://veribilimcisi.com/2017/07/20/k-en-yakin-komsu-k-nearest-neighborsknn/

REFERANS MAKALE

Explainable Artificial Intelligence (XAI):

Concepts, taxonomies, opportunities and challenges toward responsible AI



Information Fusion 58 (2020) 82–115

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI

Alejandro Barredo Arrieta[a], Natalia Díaz-Rodríguez[b], Javier Del Ser[a,c,d,*], Adrien Bennetot[b,e,f], Siham Tabik[g], Alberto Barbado[b], Salvador Garcia[g], Sergio Gil-Lopez[a], Daniel Molina[g], Richard Benjamins[h], Raja Chatila[f], Francisco Herrera[g]

[a] TECNALIA, Derio 48160, Spain
[b] ENSTA, Institute Polytechnique Paris and INRIA Flowers Team, Palaiseau, France
[c] University of the Basque Country (UPV/EHU), Bilbao 48013, Spain
[d] Basque Center for Applied Mathematics (BCAM), Bilbao 48009, Bizkaia, Spain
[e] Segula Technologies, Parc d'activité de Pissaloup, Trappes, France
[f] Institut des Systèmes Intelligents et de Robotique, Sorbonne Université, France
[g] DaSCI Andalusian Institute of Data Science and Computational Intelligence, University of Granada, Granada 18071, Spain
[h] Telefonica, Madrid 28050, Spain

ARTICLE INFO

Keywords:
Explainable Artificial Intelligence
Machine Learning
Deep Learning
Data Fusion
Interpretability
Comprehensibility
Transparency
Privacy
Fairness
Accountability
Responsible Artificial Intelligence

ABSTRACT

In the last few years, Artificial Intelligence (AI) has achieved a notable momentum that, if harnessed appropriately, may deliver the best of expectations over many application sectors across the field. For this to occur shortly in Machine Learning, the entire community stands in front of the barrier of explainability, an inherent problem of the latest techniques brought by sub-symbolism (e.g. ensembles or Deep Neural Networks) that were not present in the last hype of AI (namely, expert systems and rule based models). Paradigms underlying this problem fall within the so-called *eXplainable AI* (XAI) field, which is widely acknowledged as a crucial feature for the practical deployment of AI models. The overview presented in this article examines the existing literature and contributions already done in the field of XAI, including a prospect toward what is yet to be reached. For this purpose we summarize previous efforts made to define explainability in Machine Learning, establishing a novel definition of explainable Machine Learning that covers such prior conceptual propositions with a major focus on the audience for which the explainability is sought. Departing from this definition, we propose and discuss about a taxonomy of recent contributions related to the explainability of different Machine Learning models, including those aimed at explaining Deep Learning methods for which a second dedicated taxonomy is built and examined in detail. This critical literature analysis serves as the motivating background for a series of challenges faced by XAI, such as the interesting crossroads of data fusion and explainability. Our prospects lead toward the concept of *Responsible Artificial Intelligence*, namely, a methodology for the large-scale implementation of AI methods in real organizations with fairness, model explainability and accountability at its core. Our ultimate goal is to provide newcomers to the field of XAI with a thorough taxonomy that can serve as reference material in order to stimulate future research advances, but also to encourage experts and professionals from other disciplines to embrace the benefits of AI in their activity sectors, without any prior bias for its lack of interpretability.

1. Introduction

Artificial Intelligence (AI) lies at the core of many activity sectors that have embraced new information technologies [1]. While the roots of AI trace back to several decades ago, there is a clear consensus on the paramount importance featured nowadays by intelligent machines endowed with learning, reasoning and adaptation capabilities. It is by virtue of these capabilities that AI methods are achieving unprecedented levels of performance when learning to solve increasingly complex computational tasks, making them pivotal for the future development of the human society [2]. The sophistication of AI-powered systems has lately increased to such an extent that almost no human intervention is required for their design and deployment. When decisions derived from such systems ultimately affect humans' lives (as in e.g. medicine, law or defense), there is an emerging need for understanding how such decisions are furnished by AI methods [3].

* Corresponding author at TECNALIA. P. Tecnologico, Ed. 700. 48170 Derio (Bizkaia), Spain.
E-mail address: javier.delser@tecnalia.com (J. Del Ser).

https://doi.org/10.1016/j.inffus.2019.12.012
Received 22 October 2019; Received in revised form 19 December 2019; Accepted 25 December 2019
Available online 26 December 2019
1566-2535/