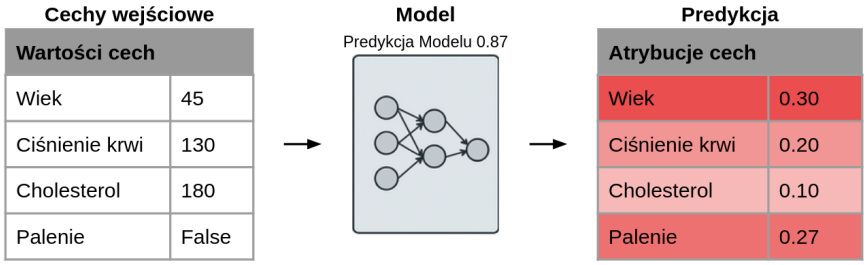
Lokalna Wyjaśnialność Modeli Uczenia Maszynowego z Uwzględnieniem Struktury Przyczynowości: Metody Teoriogrowe dla Niezawodnej Atrybucji Cech



Rysunek 1: Pipeline atrybucji cech

Wprowadzenie Systemy sztucznej inteligencji coraz częściej podejmują kluczowe decyzje, które wpływają na nasze codzienne życie—od diagnoz medycznych i decyzji o udzieleniu kredytu po rekomendacje w systemie sprawiedliwości karnej. Chociaż te systemy AI mogą być bardzo dokładne, często działają jako „czarne skrzynki", uniemożliwiając lekarzom, sędziom czy innym specjalistom zrozumienie, dlaczego podjęto konkretną decyzję. Ten brak przejrzystości staje się niebezpieczny, gdy systemy AI podejmują błędne lub stronnicze decyzje, potencjalnie szkodząc jednostkom i społecznościom.

Obecne metody wyjaśnianialności mogą dostarczać mylących wyjaśnień sugerujących fałszywe relacje między czynnikami, lub wyjaśnień, które zmieniają się nieprzewidywalnie, gdy system AI uczy się z nowych danych. Dla specjalistów, którzy muszą ufać i weryfikować decyzje AI w sytuacjach wysokiego ryzyka, te zawodne wyjaśnienia mogą prowadzić do złego podejmowania decyzji lub całkowitego odrzucenia potencjalnie korzystnych narzędzi AI.

Cele Badawcze i Motywacja Ten projekt ma na celu opracowanie nowego podejścia do wyjaśnialności AI, które zapewnia przyczynowo spójne i lokalnie wiarygodne wyjaśnienia. Zamiast pokazywać, które czynniki były statystycznie powiązane z decyzją, nasza metoda będzie uwzględniać rzeczywiste relacje przyczynowe, które kierują przewidywaniami AI—odpowiadając na kluczowe pytanie „dlaczego"zamiast tylko „co".

"Dlaczego"jest fundamentalnym pytaniem dla ludzkiego zrozumienia: Ludzie naturalnie myślą w kategoriach przyczyny i skutku. Gdy lekarz pyta „Dlaczego AI zaleciło to leczenie?"oczekuje wyjaśnień przyczynowych takich jak „Ponieważ objawy pacjenta wskazują na tę chorobę", a nie korelacji statystycznych. Prowadzone badania będą .

Prezentowane badania będą łączyły teorię gier – do sprawiedliwiego rozdzielania uznania między różne czynniki przyczyniające się do decyzji oraz przyczynowości – do identyfikowania prawdziwych relacje przyczynowo skutekowo zamiast korelacji.

W trakcie gdy systemy AI stają się coraz bardziej rozpowszechnione w opiece zdrowotnej, finansach i wymiarze sprawiedliwości, potrzebujemy metod wyjaśniania, które odpowiadają temu, jak ludzie naturalnie rozumują o przyczynie i skutku, pozostając jednocześnie stabilne i godne zaufania w różnych scenariuszach.

Oczekiwany Wpływ Badania wygenerują teoretycznie ugruntowane metody, które zapewnią matematycznie gwarantowaną spójność przyczynową, gwarantując, że wyjaśnienia AI uwzględniają prawdziwe relacje przyczyna-skutek, a nie pozorne korelacje. Te postępy przełożą się na praktyczne narzędzia, które działają z rzeczywistymi systemami AI w, czyniąc korzyści dostępnymi w różnorodnych zastosowaniach - od diagnozy medycznej po podejmowanie decyzji finansowych.