www.aileaders-project.eu

DEMO

- Narzędzie do audytu predyktorów

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

DEMO – narzędzie do audytu predyktorów

01 Streszczenie

Rodzaj OER

Prezentacja wykorzystująca AEQUITAS w środowisku Google Colab

Cel/przeznaczenie

Zapewnienie studentom narzędzia do kontroli predyktorów uczenia maszynowego pod kątem stronniczości i sprawiedliwości. Koncentrując się na scenariuszu wykrywania oszustw, studenci uczą się krytycznie oceniać wydajność algorytmów nie tylko pod kątem dokładności, ale także sprawiedliwego traktowania różnych grup demograficznych.

Oczekiwane efekty kształcenia

Po zakończeniu demonstracji studenci będą potrafili:

01 Zastosować Aequitas do kontroli modeli klasyfikacyjnych (drzewo decyzyjne, las losowy, FairGBM) pod kątem sprawiedliwości;

02 Interpretować kluczowe wskaźniki sprawiedliwości, takie jak wskaźnik fałszywych alarmów (FPR), wskaźnik fałszywych odkryć (FDR) i parytet statystyczny;

03 Identyfikować i wyjaśniać dysproporcje na poziomie grup w wynikach prognoz;

04 Podjąć refleksje nad rolą audytu stronniczości i sprawiedliwych technik uczenia maszynowego w kontekście podejmowania decyzji wysokiego ryzyka.

Słowa kluczowe

- Uczenie maszynowe
- Aequitas
- Audyt
- Stronniczość
- Sprawiedliwość

Sugerowane podejście metodologiczne

Nauka oparta na problemach

UWAGA

Aby zrozumieć i pracować z treścią tego OER, wymagana jest średniozaawansowana znajomość **programowania** w języku **Python.**

02 Wprowadzenie

W miarę jak uczenie maszynowe staje się coraz bardziej zintegrowane z procesami decyzyjnymi – szczególnie w dziedzinach o wysokiej stawce, takich jak finanse – niezbędna jest krytyczna ocena etycznych implikacji wdrażania modeli. Chociaż modele predykcyjne są zazwyczaj oceniane na podstawie wskaźników wydajności, takich jak dokładność, same one nie wystarczają do zagwarantowania sprawiedliwości, zwłaszcza gdy dane bazowe zawierają wrażliwe atrybuty, takie jak płeć, pochodzenie etniczne lub status społeczno-ekonomiczny (Jesus et al., 2024).

To ćwiczenie oparte na demonstracji zachęca uczniów do zapoznania się z Aequitas, zestawem narzędzi audytowych typu open source, poprzez praktyczne zbadanie realistycznego scenariusza wykrywania oszustw. Cel jest dwojaki: pogłębienie zrozumienia przez uczniów algorytmicznych uprzedzeń oraz zapoznanie ich z praktycznymi narzędziami i technikami, które wspierają rozwój bardziej przejrzystych i odpowiedzialnych systemów uczenia maszynowego.

Aequitas zapewnia kompleksowe ramy oceny sprawiedliwości modeli klasyfikacyjnych poprzez badanie rozkładu wyników prognozowanych w różnych grupach demograficznych. Oferuje szereg wskaźników sprawiedliwości, w tym wskaźnik fałszywych alarmów (FPR), wskaźnik fałszywych odkryć (FDR) i parytet statystyczny, które pomagają wykryć dysproporcje w wydajności modeli, które mogą być niewidoczne przy użyciu konwencjonalnych metod oceny. Nawet w przypadku modeli szkolonych na niezrównoważonych lub stronniczych zbiorach danych można przeprowadzić audyt w celu oceny, czy traktują one różne podgrupy w sposób sprawiedliwy.

Dzięki bezpośredniemu zaangażowaniu w Aequitas w tym kontekście zastosowań studenci zdobędą zarówno umiejętności techniczne w zakresie kontroli sprawiedliwości, jak i szerszą świadomość etyczną wyzwań związanych z wdrażaniem uczenia maszynowego w dziedzinach, w których konsekwencje stronniczości mogą być szczególnie poważne.

03 Prezentacja narzędzi

Scenariusz obejmuje audyt **modelu klasyfikacji binarnej**, który został przeszkolony w celu wykrywania **oszustw związanych z kontami bankowymi.**

Systemy wykrywania oszustw zazwyczaj borykają się z następującymi problemami

- **Poważna nierównowaga klas** (przypadki oszustw są rzadkie);
- **Wysoka stawka** (błędna klasyfikacja może zaszkodzić osobom fizycznym lub instytucjom);
- **Ukryte uprzedzenia** (wrażliwe cechy mogą korelować z wynikami etykiet).

*W symulacji wykorzystano **zbiór danych Bank Account Fraud (BAF)**, czyli duży, syntetyczny zbiór danych chroniący prywatność, który odzwierciedla rzeczywiste wzorce oszustw bankowych.*

Najważniejsze cechy to

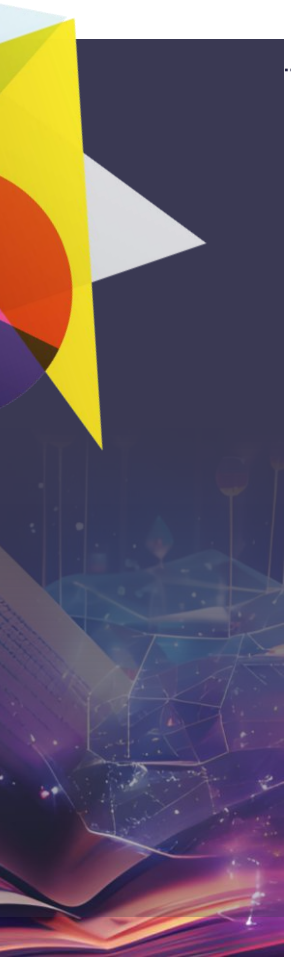
01 Warianty symulujące tendencyjność próbkowania, dryf czasowy i nierównowagę cech

02 Atrybuty demograficzne umożliwiające analizę sprawiedliwości;

03 Silna nierównowaga klas dla realizmu.

Warunki te pozwalają na przeprowadzenie znaczących eksperymentów dotyczących zachowania wskaźników sprawiedliwości w różnych założeniach modelowych i konfiguracjach danych.

04 Wykonanie symulacji



01 Dostęp do zeszytu symulacji

Przejdź do strony https://tinyurl.com/4b9u9hun

02 Uruchom cały kod

Uruchom cały notatnik, aby załadować zbiór danych, wyszkolić model klasyfikacji binarnej i wygenerować prognozy. W tym celu kliknij przycisk odtwarzania w lewym górnym rogu każdej komórki.

03 Przeprowadź audyt sprawiedliwości za pomocą Aequitas

- Użyj Aequitas, aby wygenerować raport sprawiedliwości
- Skoncentruj się na wskaźnikach takich jak FPR, FDR i parytet statystyczny
- Zidentyfikuj, które grupy są traktowane niesprawiedliwie w prognozach modelu

04 Porównaj wyniki i zinterpretuj wskaźniki

Przejrzyj różnice w sprawiedliwości między grupami. Porównaj wskaźniki wydajności (np. dokładność) ze wskaźnikami sprawiedliwości, aby ocenić kompromisy.

05 Zastanów się i przedyskutuj

- Jakie wzorce niesprawiedliwości zaobserwowano?
- Jak te wyniki mogą wpłynąć na rzeczywiste osoby?
- W jaki sposób można włączyć takie narzędzia do procesu uczenia maszynowego, aby poprawić wyniki pod względem etycznym?

Szczegóły procesu

Proces ten składa się z kilku kolejnych etapów mających na celu przygotowanie danych, stworzenie modeli predykcyjnych oraz ocenę ich wydajności i sprawiedliwości.

Kluczowe etapy są następujące:

01 **Ładowanie danych** — zestaw danych jest najpierw ładowany z określonego źródła (np. pliku CSV, bazy danych). Oczekuje się, że będzie on zawierał zarówno zmienne cech, jak i jeden lub więcej atrybutów wrażliwych (np. płeć, rasa) niezbędnych do oceny sprawiedliwości.

02 **Wstępne przetwarzanie** — dane przechodzą kilka etapów wstępnego przetwarzania w celu zapewnienia jakości i spójności (imputacja i normalizacja).

03 **Modelowanie —** trenowanych i ocenianych jest wiele modeli uczenia maszynowego. Modele te mogą obejmować zarówno algorytmy nie uwzględniające sprawiedliwości (standardowe), jak i podejścia uwzględniające sprawiedliwość, które wykorzystują techniki ograniczania stronniczości podczas trenowania lub przetwarzania końcowego.

04 **Ocena sprawiedliwości za pomocą Aequitas** – wyszkolone modele są oceniane nie tylko pod kątem dokładności i innych wskaźników wydajności, ale także pod kątem sprawiedliwości przy użyciu zestawu narzędzi Aequitas.

05 Wnioski

To demo pokazuje, że nawet dokładne modele mogą prowadzić do **niesprawiedliwych wyników,** jeśli są wdrażane bez kontroli sprawiedliwości.

Korzystając z Aequitas, studenci odkrywają, w jaki sposób **uprzedzenia mogą utrzymywać się** w ustawieniach klasyfikacji binarnej i jak różne grupy demograficzne mogą doświadczać różnych wskaźników błędnej klasyfikacji. Angażując się bezpośrednio w pomiary sprawiedliwości i przeprowadzając rzeczywiste audyty, studenci rozwijają kompetencje techniczne i świadomość etyczną. Co ważniejsze, uczą się**, że wykrywanie uprzedzeń nie jest dodatkiem,** ale istotną częścią budowania **godnych zaufania systemów sztucznej inteligencji** — szczególnie w dziedzinach takich jak finanse, gdzie prognozy modeli mają poważne konsekwencje.

06 Bibliografia

- NJesus, S., Saleiro, P., e Silva, I. O., Jorge, B. M., Ribeiro, R. P., Gama, J., ... & Ghani, R. (2024). Aequitas flow: Streamlining fair ml experimentation. Journal of Machine Learning Research, 25(354), 1-7.

Śledź naszą podróż

www.aileaders-project.eu

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.