Projekt Preference learning

1 Założenia ogólne

Ten projekt polega na przeprowadzeniu eksperymentów z użyciem kilku metod i przeprowadzeniu analizy uzyskanych wyników. Znajdź zbiór danych, który zawiera kilka kryteriów monotonicznych (4-9), ponad 100 wariantów i co najmniej 2 klasy (jeśli problem ma więcej klas, możesz je zbinaryzować). Można wykorzystać jeden z następujących zbiorów danych link.

- Projekt może być wykonywany indywidualnie lub w parach.
- Raport może być wykonany w jupyter notebook (**.ipynb** + **PDF**/**HTML**) lub **plik z kodem w pythonie** + **raport w PDF**.
- Kod z raportem należy przesłać przed **8/06/2025 23.59**

Krótko opisz zbiór danych, w tym opisy kryteriów. Dla tego zbioru danych wytrenuj następujące modele:

- Jeden prosty, interpretowalny model ML (np. XGBoost, rankSVM lub regresja logistyczna)
- Jeden interpretowalny model ANN-MCDA przedstawiony na zajęciach (ANN-UTADIS)
- Złożony model sici neuronowej zawierającej kilka warstw ukrytych i nieliniową funkcję aktywacji.

2 Eksperymenty

Dla każdego modelu:

1. Przedstawi miary jakości Accuracy, F1 oraz AUC
2. Modele powinny być zaprezentowane wraz ze wizualizacjami w celu ułatwienia interpretacji (np funkcje cząstkowe, schemat drzewa ...).
3. Wszystkie prezentowane wartości powinny być zaokrąglone do maksymalnie 4 miejsc po przecinku.

2.1 Wyjaśnienie wybranych decyzji

1. Dla 3 wybranych wariantów dokonaj wyjaśnienia decyzji. Dlaczego wariant został przydzielony do danej klasy. Jakie oceny na kryteriach na to wpłynęły. (Do odpowiedzenia na to pytanie w przypadku sieci neuronowych można wykorzystać na przykład guided gradient).
2. Znajdź minimalną zmianę w ocenach tych 3 wariantów tak, aby zostały one przypisane do innej klasy. Zmianie powinno podlegać wyłącznie jedno kryterium.
   - Spróbuj odpowiedzieć na to pytanie w sposób analityczny, opierając się tylko na wartościach parametrów modelu i wyjaśnij, dlaczego taka zmiana jest minimalna (bez próbkowania).
   - Wykonaj próbkowanie przestrzeni, nieznacznie zmieniając oceny, tak, aby uzyskać inną klasę. Czy wyniki zgadzają się z przewidywaniami teoretycznymi?
3. Dokonaj wyjaśnienia predykcji za pomocą conajmniej jednej techniki (Anchors LIME, SHAP, ...)

2.2 Interpretacja modelu

Zinterpretuj modele oraz odpowiedz na następujące pytania:

- Czy na podstawie uzyskanych parametrów możemy powiedzieć coś o preferencjach użytkowników?
- Jaki jest wpływ każdego z kryteriów. Czy są jakieś kryteria, które nie mają żadnego znaczenia, czy też mają wpływ decydujący.
- Jaki jest charakter danego kryterium: zysk, koszt, niemonotoniczne?
- Czy istnieją jakieś progi preferencji? Czy istnieją oceny kryteriów, które są nierozróżnialne z punktu widzenia preferencji?
- Wykonaj interpretację modelu korzystając z co najmniej jednej techniki(Global Surrogate, Partial Dependence Plot Permutation Feature Importance ...)
  - Czy wyniki uzyskane z tych technik pokrywają się się z analizą wykonaną w poprzednim punkcie?

Lista narzędzi, które zawierają różne techniki wyjaśniania predykcji i interpretacji modelu:

- Shapash
- Alibi
- Explainerdashboard
- DALEX
- eli5
- aix360

3 Ocenianie

- 3 - Interpretowalny model ML wraz z całym eksperymentem
- 4 - To co na 3 + jeden interpretowalny model ANN-MCDA
- 5 - To co na 4 + sieć neuronowa z nieliniową funkcja aktywacji.

Uwaga: stwierdzenie, że sieć neuronowa jest czarną skrzynką i nie da się jej zinterpretować i wyjaśnić decyzji przez nią podjętych nie jest wystarczająca.