Data Foundations

Hoofdstuk 6

Lab: bias detecteren in een dataset

Hassan Haddouchi

Na dit lab kan je:

- Bias in een dataset identificeren en visualiseren.
- De invloed van bias op modelprestaties analyseren.
- Reflecteren op de ethische implicaties van bias in data.
- Fairness-metrics berekenen en analyseren.
- Gegevens anonimiseren en de impact hiervan op data en modellen evalueren.

Opdracht

Je werkt met de **Adult Income Dataset**, een dataset die informatie bevat over leeftijd, geslacht, ras, opleiding, beroep en inkomen. Het doel van deze opdracht is om:

- Bias in de dataset te onderzoeken.
- Fairness-metrics te berekenen en te interpreteren.
- Een eenvoudig classificatiemodel te trainen en de invloed van bias op de modelprestaties te analyseren.
- Gegevens te anonimiseren en de impact hiervan te evalueren.

Stappen voor de opdracht

Stap 1: Dataset laden en verkennen

1. Importeer de **Adult Income Dataset** via `sklearn.datasets`.
2. Verken de dataset om de kenmerken en verdeling van gegevens te begrijpen.
3. Identificeer welke kenmerken mogelijk bias kunnen bevatten (bijv. geslacht, ras).

Stap 2: Exploratory Data Analysis (EDA)

1. Analyseer de verdeling van kenmerken zoals geslacht, ras en inkomen.
2. Gebruik grafieken zoals barplots en histogrammen om patronen van bias inzichtelijk te maken.
3. Beantwoord vragen zoals:
    - Hoe varieert het inkomen tussen mannen en vrouwen?
    - Zijn bepaalde rassen onder- of oververtegenwoordigd in de dataset?

Stap 3: bias visualiseren

1. Visualiseer de verdeling van inkomen per geslacht of ras.
2. Gebruik geschikte grafieken (zoals stacked bar charts of boxplots) om verschillen te tonen.

Stap 4: classificatiemodel bouwen

1. Splits de dataset in een trainings- en testset.
2. Train een eenvoudig classificatiemodel (bijv. Logistic Regression) om te voorspellen of iemand meer of minder dan $50K verdient.
3. Evalueer de prestaties van het model met een confusion matrix en een classificatierapport.

Stap 5: evalueren van bias in het model

1. Analyseer hoe goed het model presteert voor verschillende groepen (bijv. mannen versus vrouwen).
2. Bereken prestaties zoals precision, recall, en F1-score per groep.
3. Reflecteer op:
    - Worden bepaalde groepen slechter geclassificeerd dan andere?
    - Hoe beïnvloedt de verdeling van data de prestaties van het model?

Stap 6: fairness-metrics berekenen

1. Bereken fairness-metrics zoals:
    - **Demographic Parity**: zijn de voorspellingen gelijk verdeeld over mannen en vrouwen?
    - **Equal Opportunity**: is de True Positive Rate (TPR) gelijk tussen mannen en vrouwen?
2. Gebruik Python om de metrics te berekenen en te visualiseren.
3. Reflecteer op de resultaten en bespreek hoe fairness-metrics kunnen bijdragen aan ethisch datagebruik.

Stap 7: data anonimisatie

1. Implementeer anonimisatie-technieken zoals:
    - Het groeperen van leeftijden in categorieën (bijv. 18-25, 26-35).
    - Het vervangen van geslacht en ras door generieke labels (bijv. `M/V` of `Group A/B`).
    - Het toepassen van hashing op specifieke kenmerken zoals namen of landen.
2. Analyseer hoe anonimisatie de verdeling van gegevens en modelprestaties beïnvloedt.
3. Reflecteer op de balans tussen privacy en bruikbaarheid van gegevens.

Vereisten voor inlevering

1. **Code**:
    - Een werkend Python-script dat alle stappen uitvoert.
    - Correcte implementatie van EDA, visualisaties, fairness-metrics en anonimisatie.
2. **Reflectie**:
    - Een korte reflectie in je portfolio waarin je:
        - De gevonden bias beschrijft.
        - De prestaties van het model analyseert.
        - Fairness-metrics interpreteert.
        - De impact van anonimisatie op data en modellen evalueert.