Compararea Rezultatelor Inițiale vs. Rezultatele cu Datele Preprocesate

Analiza exploratorie EDA

Stoian Andreea

1. Distribuțiile variabilelor numerice:

- Vârsta (age):
    - Rezultate inițiale: Distribuția vârstei era ușor asimetrică spre dreapta, dar aproape normală.
    - Rezultate după preprocesare: Distribuția vârstei a rămas similară, dar valorile au fost probabil scalate sau standardizate (ex. prin StandardScaler), ceea ce poate fi observat din boxploturi și analiza descrierei statistice. Aceasta nu schimbă semnificativ distribuția, dar face valorile mai ușor de comparat între variabile.
- Numărul de ore lucrate pe săptămână (hours-per-week):
    - Rezultate inițiale: Existau câțiva indivizi care lucrau mult mai puțin de 10 ore pe săptămână sau mult mai mult de 60 de ore, ceea ce sugerea posibile valori extreme (outlieri).
    - Rezultate după preprocesare: Distribuția a rămas similară, dar din cauza preprocesării, valorile extrem de mari și mici au fost poate modificate pentru a se încadra într-un interval mai standardizat sau mai ușor de interpretat. Aceste extreme pot fi mai ușor de gestionat în modelele viitoare dacă sunt gestionate ca outlieri.
- Diferența capitalului (capital_diff):
    - Rezultate inițiale: Această variabilă nu aducea informații relevante (erau toate valorile 0).
    - Rezultate după preprocesare: Aceasta a fost eliminată complet, ceea ce este un pas important pentru a preveni adăugarea de zgomot în analiza ulterioară. Preprocesarea a identificat corect că variabila este inutilă și a permis o curățare a datelor.

2. Distribuțiile variabilelor categorice:

- Sectorul de muncă (workclass_Private):
    - Rezultate inițiale: Se observa că majoritatea persoanelor active lucrează în sectorul privat, iar restul sunt într-un număr mai mic.
    - Rezultate după preprocesare: Nicio modificare semnificativă nu a avut loc aici. Variabila a rămas relativ similară, iar preprocesarea nu a afectat distribuția sa.

- Nivelul de educație (education_Bachelors):
    - Rezultate inițiale: Persoanele cu studii superioare de licență (True) reprezenta o proporție mică comparativ cu cele fără acest nivel de educație.
    - Rezultate după preprocesare: Distribuția a rămas aproximativ aceeași, dar preprocesarea ar putea include o eventuală transformare a variabilei pentru a facilita înțelegerea relațiilor ulterioare (de exemplu, binarizarea în variabile categorice).
- Starea civilă (marital-status_Married-civ-spouse):
    - Rezultate inițiale: Majoritatea persoanelor sunt căsătorite civil, iar un număr mic au alte statuturi matrimoniale.
    - Rezultate după preprocesare: Similar cu cele inițiale, această variabilă a rămas neschimbată în urma preprocesării. Variabila poate fi folosită direct pentru analiza mai detaliată a distribuției statutului marital.

3. Corelațiile între variabilele numerice:

- Educația (education-num) și venitul ridicat (high_income):
    - Rezultate inițiale: Corelația dintre aceste variabile era moderată, ceea ce sugerează că educația joacă un rol important în determinarea veniturilor.
    - Rezultate după preprocesare: Corelația a rămas similară (0.30), indicând faptul că preprocesarea nu a afectat semnificativ relația dintre educație și venit. Aceasta rămâne o corelație pozitivă importantă.
- Vârsta (age) și venitul ridicat (high_income):
    - Rezultate inițiale: Corelația era moderată (0.25), indicând o asociere între vârstă și venit, posibil datorită experienței de muncă.
    - Rezultate după preprocesare: Corelația a rămas similară, ceea ce sugerează că, indiferent de preprocesarea datelor, experiența acumulată în câțiva ani de muncă influențează venitul într-o măsură mai mică, dar semnificativă.
- Numărul de ore lucrate pe săptămână (hours-per-week) și venitul ridicat (high_income):
    - Rezultate inițiale: Corelația era slabă (0.17), ceea ce sugerează că numărul de ore lucrate nu este un factor determinant pentru venituri mari.
    - Rezultate după preprocesare: Această corelație a rămas la un nivel scăzut și după preprocesare, confirmând că alte variabile, cum ar fi educația sau experiența, au un impact mult mai mare asupra venitului decât numărul de ore lucrate.

4. Outlieri și variabile eliminate:

- Rezultate inițiale: Am identificat câțiva outlieri în variabila hours-per-week, care sugerează că există valori extreme ce ar putea afecta modelele ulterioare.

- Rezultate după preprocesare: În urma preprocesării, s-au eliminat variabilele cu informații insuficiente, cum ar fi capital_diff, iar outlierii din hours-per-week au fost, probabil, tratați prin diverse tehnici, cum ar fi transformarea sau normalizarea datelor.

Concluzie:

Comparând rezultatele inițiale și cele preprocesate, putem observa câteva diferențe semnificative:

- Preprocesarea a îmbunătățit curățarea datelor, eliminând variabilele inutile și ajutând la identificarea unor outlieri, care au fost tratate corespunzător.
- Scalarea sau standardizarea variabilelor numerice a făcut datele mai omogene, facilitând compararea lor între ele.
- Corelațiile și distribuțiile variabilelor nu s-au schimbat semnificativ, ceea ce sugerează că preprocesarea a fost adecvată și nu a afectat semnificativ informațiile esențiale din setul de date.

Aceste observații indică faptul că preprocesarea a contribuit la o curățare eficientă a datelor, ceea ce va facilita construirea unui model predictiv mai robust în etapele următoare ale analizei.