Projekt 1

Analiza danych

Gabriela Matuszewska, Mateusz Wojtulewicz

1 Opis datasetu

Do analizy wybrano dataset "Adult income", który zawiera klasyfikacje przychodu rocznego ($< 50k$, $>= 50k$) w zależności, od miedzy innymi pochodzenia, wykształcenia, miejsca zamieszkania.

Liczba cech i ich typy oraz rozmiar:

```
df.shape
```

```
(48842, 15)
```

```
df.head()
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country	income
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	0	40	United-States	0
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0	50	United-States	0
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0	40	United-States	1
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	0	40	United-States	1
4	18	Private	103497	Some-college	10	Never-married	Prof-specialty	Own-child	White	Female	0	0	30	United-States	0

2 Czyszczenie danych

Ze wzgledu na specyfikacje zestawu danych konieczna była pewna modyfikacja tych danych. Brakujace wartości zostały zastapione przez najcześciej wystepujace wartości. Ze wzgledu na przeważajaca liczbe native-country="United-States", która mogłaby zakłócić wyniki postanowiono usunać kolumne native-country. Kolejnym spostrzeżeniem jest podobieństwo kolumn education i education-num oraz marital-status orazz relationship dlatego zostały one zmapowane do jednej kolumny. Atrybut "workclass" zastepuje nam atrybut "occupation" dlatego możemy pozbyć sie tej kolumny. Po usunieciu zbednych kolumn oraz wypełnieniu brakujacych wartości, a także zmapowaniu kategorii wypełnionych typem "object" na int, nasze dane wygladaja nastepujaco:

3 Random Forest Classification

TO DO: OPIS

3.1 Parametry

TO DO: Dlaczego takie wartości a nie inne?

4 Macierz pomyłek

5 Walidacja Krzyżowa

TO DO: NA CZYM POLEGA (OPIS DZIAŁANIA METODY WYBORU HIPER-PARAMETRÓW)

5.1 Grid Search

TO DO: NA CZYM POLEGA

6 Normalizacja

7 Standarycja

8 PCA

References