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

Las losowy to zbiór drzew klasyfikacyjnych o podziałach binarnych. Dla konkretnej obserwacji (wyrażonej jako wektor wejściowy), każde z drzew zwraca decyzje lub krtoke prawdopodobieństw klasyfikacji, Prawdopodobieństwa z drzew wchodzacych w skład lasu sa traktowane jako głosy $->$ jako wynik zwracana jest decyzja która otrzymała najwiecej głosów (której średnie prawdopodobieństwo jest najwyższe).

3.1 Parametry

```
class sklearn.ensemble.RandomForestClassifier(n_estimators=100,
*, criterion='gini', max_depth=None, min_samples_split=2,
min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto',
max_leaf_nodes=None, min_impurity_decrease=0.0,
min_impurity_split=None, bootstrap=True, oob_score=False,
n_jobs=None, random_state=None, verbose=0, warm_start=False,
class_weight=None, ccp_alpha=0.0, max_samples=None)
```

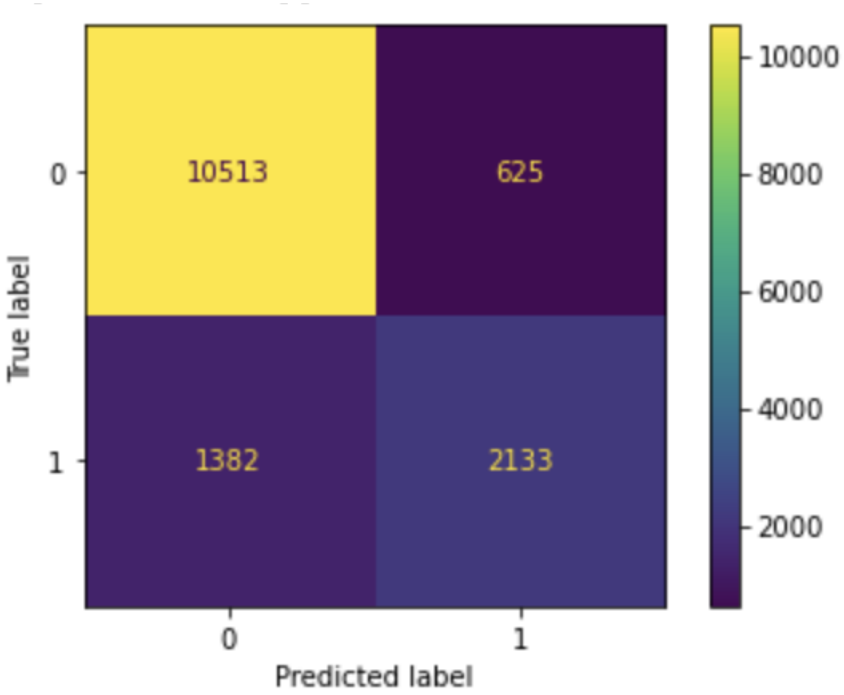
Opis najwazniejszych parametrów:

1. n_estimators
2. criterion"gini", "entropy"
3. max_depth
4. min_samples_split
5. min_samples_leaf
6. min_weight_fraction_leaf
7. max_features"auto", "sqrt", "log2"
8. max_leaf_nodes
9. min_impurity_decrease

4 Sposób wyboru zbioru testowego

5 Macierz pomyłek

Macierz pomyłek (confusion matrix) to narzedzie do oceny jakosci klasyfikacji. Składa sie z nastepujacych klas: true-positive, true-negative, false-positive, false-negative. W naszym przypadku wyglada tak:



6 Walidacja Krzyżowa

Walidacja krzyżowa służy do testowania klasyfikatora i polega na wybraniu parametru n (zazwyczaj 10) i podzieleniu danych na n równych podzbiorów, z których każdy nastepnie przyjmujemy jako testowy, a pozostałe tworza próbke treningowa. Wyniki klasyfikacji sa sumowane.

6.1 Grid Search

TO DO: NA CZYM POLEGA

7 Normalizacja

Normalizacja polega na przeskalowaniu wartości do zakresu [0,1]:

$$X^* = \frac{X - min(X)}{(max(x) - min(X}$$

8 Standarycja

Wartości mniejsze od średniej wartości po standaryzacji beda miały ujemne wartości, analogicznie wartości wieksze od średniej po standaryzacji beda miały

dodatnie wartości:

$X^* = \frac{X - \mu(X)}{\sigma(x)}$

9 PCA

PCA, czyli analiza głównych składowych, służy do redukcji liczby zmiennych (wymiarowości), tak aby najlepiej zachować strukture danych lub do odkrycia prawidłowości miedzy cechami. Opiera sie o wyznaczenie osi zachowujacej najwieksza wartość wariancji zbioru uczacego. Polega na wyznaczeniu składowych bedacych kombinacja liniowa badanych zmiennych.