Zadanie 1. [Dane ALLAML]

Oceń ile zmiennych w tym zbiorze danych jest istotnych, tzn. ile zmiennych wystarczy, aby przewidywać zmienną celu z podobną dokładnością, jak przy wykorzystaniu wszystkich zmiennych. W analizie rozpatrz różne podejścia. Jakość predykcji oceniaj kroswalidacyjnie przy użyciu funkcji cross_val_score.

Zadanie 2.

Zaimplementuj algorytm selekcji zmiennych przy użyciu metody losowych podzbiorów.
Algorytm:

1. Powtórz N razy:
    1. Wylosuj k zmiennych,
    2. Spośród wybranych zmiennych wybierz zmienne istotne przy użyciu *selector* – obiekt z pakietu sklearn dokonujący selekcji. Zapisz informację o tym, które zmienne zostały wybrane.
2. Podaj zmienne, które w powyższym kroku zostały przynajmniej raz uznane za istotne.

Wersja 1
Napisz dowolny kod, który zrealizuje ten algorytm.

Wersja 2
Zdefiniuj funkcję, która realizuje ten algorytm i przyjmuje jako argumenty *N, k i selector* oraz dane *X* i *y*.

Wersja 3
Zdefiniuj klasę, która realizuje ten algorytm i ma interfejs analogiczny do selektorów z sklearn'a.

Zadanie 3. [Dane Adult]

1. Przygotuj do modelowania dane:
    1. Wyświetl informacje o typach zmiennych.
    2. Połącz zbiory (testowy z treningowym) w jeden zbiór.
    3. Zakoduj zmienną objaśnianą jak 0-1.
    4. Usuń obserwacje z brakami danych (wykryj jak oznaczone są braki).
    5. Rozważ zmienną fnlwgt – oznacza ona liczbę osób w populacji o podobnej charakterystyce. Zastanów się czy coś z nią trzeba robić?
    6. Zaproponuj sposób obsłużenia informacji dotyczących edukacji.
    7. Wypisz podstawowe informacje o rozkładach wszystkich zmiennych (nie tylko numerycznych).
    8. Wypisz liczności wystąpień poszczególnych narodowości.
    9. Przeanalizuj rozkład y w zależności od narodowości (np. rysując wykres słupkowy procentu obserwacji z y=1 względem narodowości). Czy kraj pochodzenia ma wpływ na zarobki?

10. Zastanów się czy zmienna dotycząca narodowości wymaga jakiegoś przekształcania.

2. Przetestuj wybrane modele (z uwzględnieniem optymalizacji).

Zadanie dodatkowe: podaj najważniejsze zmienne wpływające na wysokość zarobków.

Zadanie 4. [Dane creditcard]

Zaimplementuj następujący algorytm klasyfikacji oparty na komitecie klasyfikatorów. Uczymy N klasyfikatorów (np. drzew decyzyjnych) na losowych próbkach danych powstałych poprzez wzięcie wszystkich obserwacji klasy 1 oraz losowych obserwacji klasy 0 o liczności równej liczności obserwacji klasy 1. Predykcji dokonujemy w następujący sposób: dla danej obserwacji dokonujemy predykcji przy użyciu każdego z nauczonych wcześniej klasyfikatorów i przypisujemy obserwacji klasę, którą wskazało większość z nich. Przetestuj warianty, gdy stosunek obserwacji klasy 0 i 1 w pojedynczy podzbiorze jest inny niż 1:1.

Wersja 1:
Napisz dowolny kod, który to zrealizuje.

Wersja 2:
Zdefiniuj funkcję, która zwróci listę nauczonych modeli (ustal jakie argumenty powinno przyjmować funkcja). Następnie zdefiniuj funkcję, która zwraca predykcję, dla podanych danych przy użyciu podanej listy modeli.

Wersja 3
Zdefiniuj klasę realizującą ten algorytm, która będzie miała interfejs analogiczny do modeli z sklearn (będzie posiadać metody fit i predict).

Zadanie 5. [Dane HappyCustomerBank]

1. Przygotuj dane do modelowania.
2. Zoptymalizuj i przetestuj dwa lub trzy modele.

Elementy dodatkowe:
3. Powtórz analizę oceniając modele według następującego schematu: przypisanie jednej obserwacji klasy 1 „kosztuje” nas 100 zł, a trafienie z predykcją w klasę 1 przynosi nam 1000 zł przychodu. Spróbuj wygenerować jak największy zysk.
4. Przetestuj wykorzystanie technik dla niezbalansowanych danych.
5. Przetestuj klasyfikator XGBoost z optymalizacją parametrów według następującego schematu:
- ustal małą liczbę drzew i duży learning_rate i zoptymalizuj inne parametry random_search’em
- dla najlepszego zestawu parametrów z poprzedniego kroku zoptymalizuj liczbę drzew i współczynnik uczenia przy użyciu grid_search’a.