**Sztuczna inteligencja i inżynieria wiedzy**

**Lista 4**

Michał Pesta, 266899

1. **Eksploracja danych**
   1. **Podstawowe informacje**

Obraz zawierający tekst, Czcionka, zrzut ekranu, numer

Opis wygenerowany automatycznie

* 1. **Liczności różnych wartości kolumn**

1. Rozmiar

Obraz zawierający tekst, zrzut ekranu, diagram, Wykres

Opis wygenerowany automatycznie

1. Materiał

Obraz zawierający tekst, zrzut ekranu, diagram, Wykres

Opis wygenerowany automatycznie

1. Kolor

Obraz zawierający tekst, zrzut ekranu, Wykres, diagram

Opis wygenerowany automatycznie

1. Rękawy

Obraz zawierający tekst, zrzut ekranu, diagram, Prostokąt

Opis wygenerowany automatycznie

1. Zapotrzebowanie

Obraz zawierający tekst, zrzut ekranu, diagram, Prostokąt

Opis wygenerowany automatycznie

* 1. **Macierz korelacji**

Obraz zawierający tekst, zrzut ekranu, numer, Czcionka

Opis wygenerowany automatycznie

* 1. **Spostrzeżenia**

- W żadnej z kolumn nie ma brakujących danych

- Brak literówek, wszystkie dane są wprowadzone poprawnie

- Rozkład różnych wartości kolumn jest w miarę różnorodny

- Brak danych numerycznych – potrzebna będzie obróbka przed użyciem modelu do klasyfikacji

- Rozkład kolumny etykiet nie jest zbalansowany – znacznie mniej wartości „low” w porównaniu do „high” oraz „medium”

- Analizując macierz korelacji można stwierdzić, że cechy nie są ze sobą powiązane. Dane zatem nie wymagają aby dokonać selekcję cech.

1. **Przetworzenie danych i jego wpływ na wyniki**

Aby zamienić dane kategoryczne na numeryczne wykonano następujące operacje:

Obraz zawierający zrzut ekranu, tekst, oprogramowanie, Oprogramowanie multimedialne

Opis wygenerowany automatycznie

Otrzymano łącznie 18 kolumn:

Obraz zawierający tekst, zrzut ekranu, design

Opis wygenerowany automatycznie

* 1. **Klasyfikacja bez dodatkowego przetwarzania danych**
     1. **Naive Bayes**

Wyniki klasyfikacji:

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Confusion matrix:

Obraz zawierający tekst, zrzut ekranu, numer, diagram

Opis wygenerowany automatycznie

* + 1. **Decision Tree**

Wyniki klasyfikacji:

**Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie**

Confusion Matrix:

Obraz zawierający tekst, zrzut ekranu, numer, diagram

Opis wygenerowany automatycznie

* 1. **Klasyfikacja z normalizacją danych**
     1. **Naive Bayes**

Wyniki klasyfikacji:

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Confusion Matrix:

Obraz zawierający tekst, zrzut ekranu, numer, diagram

Opis wygenerowany automatycznie

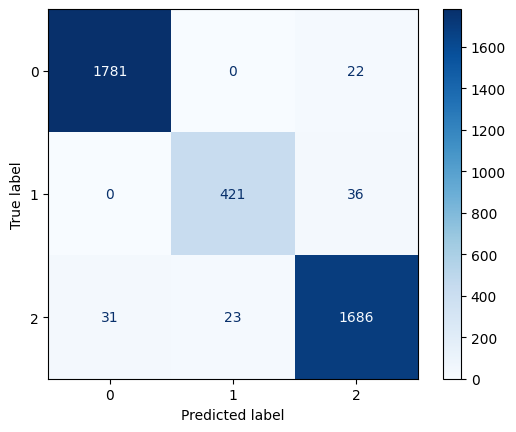
* + 1. **Decision Tree**

Wyniki klasyfikacji:

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Confusion Matrix:



* 1. **Klasyfikacja z użyciem PCA** 
     1. **Naive Bayes**

Wyniki klasyfikacji:

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Confusion Matrix:

Obraz zawierający tekst, zrzut ekranu, numer, diagram

Opis wygenerowany automatycznie

* + 1. **Decision Tree**

Wyniki klasyfikacji:

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Confusion Matrix

Obraz zawierający tekst, zrzut ekranu, numer, diagram

Opis wygenerowany automatycznie

* 1. **Podsumowanie**

**Naive Bayes**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Avg precision** | **Avg recall** | **Avg F1-score** | **accuracy** |
| **No preprocessing** | 0.76 | 0.56 | 0.57 | 0.62 |
| **Normalization** | 0.66 | 0.54 | 0.52 | 0.59 |
| **PCA** | 0.61 | 0.63 | 0.62 | 0.65 |

**Decision Tree**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Avg precision** | **Avg recall** | **Avg F1-score** | **accuracy** |
| **No preprocessing** | 0.97 | 0.96 | 0.96 | 0.97 |
| **Normalization** | 0.97 | 0.96 | 0.96 | 0.97 |
| **PCA** | 0.97 | 0.96 | 0.96 | 0.97 |

* 1. **Wnioski**

- Decision Tree sprawdza się znacznie lepiej niż Naive Bayes

- Dla Decision Tree przeprowadzony preprocessing nie miał znaczenia. Wyniki były wysokie dla wszystkich trzech przypadków

- Dla Naive Bayes motodą dającą najbardziej dokładne wyniki okazała się metoda PCA

- Z tego względu porównanie 3 zestawów hiperparametrów odbędzie się korzystając z danych przeprocesowanych metodą PCA

1. **Dopasowanie hiperparametrów**

Aby sprawdzić, jaka wartość parametrów jest optymalna użyto GridSearchCV z biblioteki sklearn

* 1. **Naive Bayes**

Testowane parametry:

- var\_smoothing

**Obraz zawierający tekst, zrzut ekranu, oprogramowanie, wyświetlacz

Opis wygenerowany automatycznie**

Znalezione optymalne wartości:

- Var\_smoothing = 0.0533

* 1. **Decision Tree**

Testowane parametry:

- max\_depth

- min\_samples\_split

- min\_saples\_leaf

**Obraz zawierający tekst, zrzut ekranu, Czcionka, oprogramowanie

Opis wygenerowany automatycznie**

Znalezione optymalne wartości:

- max\_depth = None

- min\_samples\_split = 2

- min\_saples\_leaf = 2