|  |  |
| --- | --- |
| Paweł Galewicz 234053  Karol Podlewski 234106 | Rok akademicki 2019/20  Poniedziałek, 12:30 |

JĘZYKI PROGRAMOWANIA W ANALIZIE  
DANYCH – LABORATORIUM

Zadanie 3

# Opis implementacji

Stworzone rozwiązanie to program konsolowy, zaimplementowany w języku Python 3.7.5. W projekcie wykorzystano następujące biblioteki: Matplotlib, NumPy, Pandas oraz Scikit-learn.

# Zbiór danych

W zadaniu wykorzystano zbiór danych zawierający informacje o zdiagnozowanych przypadkach raka piersi w stanie Wisconsin – klasyfikacja polega na przypisaniu rekordów do nowotworów złośliwych lub łagodnych. Zbiór danych zawiera 30 kolumn, wśród których znajdziemy 10 cech opisanych przez trzy miary: średnią, odchylenie standardowe oraz wartości najgorsze zebrane dla każdego jądra komórkowego.

Zbiór został załadowany z paczki zbiorów biblioteki Scikit-learn, jest on możliwy do odnalezienia także w najpopularniejszych repozytoriach: <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>.

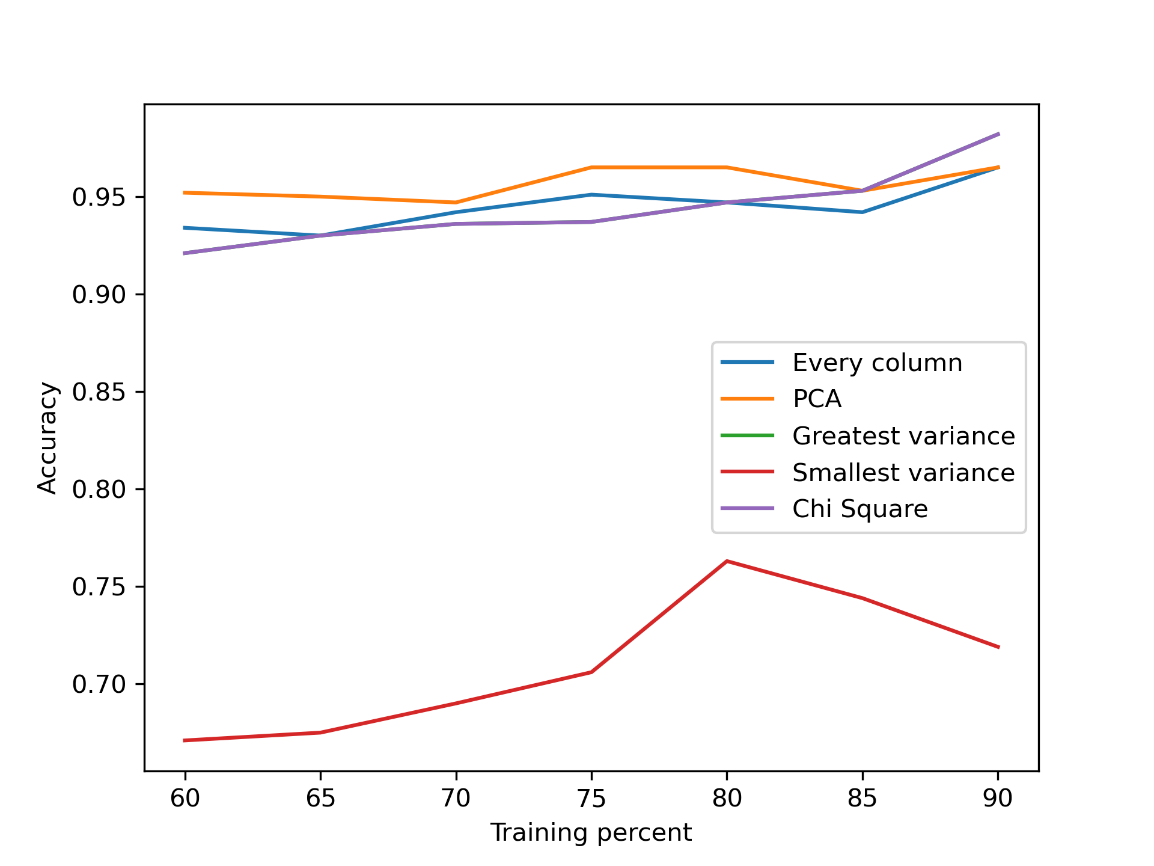
# Klasyfikacja

Pierwszą częścią zadania było przeprowadzenie klasyfikacji na zbiorze danym za pomocą wybranej techniki klasyfikacji – wykorzystaliśmy klasyfikator Maszyny wektorów nośnych. Do klasyfikacji wykorzystaliśmy pełny zbiór, a także zbiory zredukowane do dwóch cech z wykorzystaniem analizy głównych składowych, wyboru największej i najmniejszej wariancji oraz selekcji testem niezależności chi-kwadrat. Przy sprawdzaniu skuteczności przypisania posłużyliśmy się miarą dokładności. Klasyfikacja została przeprowadzona dla zbiorów danych, w których część treningowa stanowiła 60%, 65%, 70%, 75%, 80% 85% lub 90% całego zbioru.

W Tabeli 1 przedstawiono cechy, do których został zredukowany zbiór przy wykorzystaniu konkretnej metody. Redukcja cech przy metodzie opartej na wyborze największej wariancji oraz teście niezależności chi-kwadrat wykazały identyczne cechy, dlatego uzyskana dokładność dla każdej z metod jest  
identyczna – w obu wypadkach klasyfikator pracuje na identycznym zbiorze.

**Tabela 1.** Cechy wybrane przy redukcji cech

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metoda** | **Pierwsza cecha** | **Wartość pierwszej cechy** | **Druga cecha** | **Wartość drugiej cechy** |
| Największa wariancja | worst area | 324167,39 | mean area | 123843,55 |
| Najmniejsza wariancja | fractal dimension error | 0,0000070016 | smoothness error | 0.0000090151 |
| Test niezależności Chi-kwadrat | worst area | 112598,43 | mean area | 53991,66 |



**Rysunek 1.** Dokładność dla klasyfikatora SVM dla pełnego oraz zredukowanych zbiorów danych

Uzyskane rezultaty pokazują, że dla zbioru opisującego zdiagnozowane przypadki raka piersi redukcja cech z wykorzystaniem najmniejszej wariancji znacząco pogarsza rezultaty klasyfikacji. Redukcja  
metodą analizy głównych składowych zawsze osiąga przynajmniej identyczną dokładność jak klasyfikacja na pełnym zbiorze – zazwyczaj wyniki są lepsze. Wybór cech z wykorzystaniem testu niezależności  
chi-kwadrat oraz wyboru cech o największej wariancji uzyskały bardzo podobną dokładność jak analiza pełnego zbioru. Przy tych metodach redukcji dokładność rosła wraz ze wzrostem zbioru treningowego, co nie charakteryzowało pracy zbiorach zredukowanych za pomocą innych metod, a także pełnym zbiorze.

# Analiza skupień

Pierwszy eksperyment polegał na przetestowanie różnych metod uzupełniania brakujących danych i sprawdzenie jak wpływają one na cechy charakterystyczne zbioru danych. Braki danych w rozważanych kolumnach ustawiono na poziom **~8%**. Dane z zastosowanymi metodami wypełniania porównane zostały z danymi, gdzie rekordy z występującymi brakami zostały usunięte. Dla każdego zbioru wyliczono wartości średniej, odchylenia standardowego oraz trzech kwartyli (Q1, Q2, Q3), a także wyznaczono krzywą regresji.

**Tabela 1.** Cechy charakterystyczne wyliczone dla kolumny **sqft\_living** dla 8% brakujących danych

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metoda** | **Średnia** | **Odchylenie standardowe** | **Pierwszy kwartyl** | **Drugi kwartyl** | **Trzeci kwartyl** |
| Usuwanie rzędów | 2078.89 | 913.23 | 1420 | 1914 | 2550 |
| Mean imputation | 2079.98 | 881.81 | 1470 | 2010 | 2480 |
| Interpolacja | 2082.02 | 902.46 | 1440 | 1920 | 2540 |
| Hot-Deck | 2081.9 | 922.68 | 1420 | 1910 | 2550 |
| Wartości z krzywej regresji | 2079.98 | 920.23 | 1420 | 1910 | 2550 |

**Tabela 2.** Cechy charakterystyczne wyliczone dla kolumny **price** dla 8% brakujących danych

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metoda** | **Średnia** | **Odchylenie standardowe** | **Pierwszy kwartyl** | **Drugi kwartyl** | **Trzeci kwartyl** |
| Usuwanie rzędów | 540729.68 | 364617.1 | 321500 | 450000 | 645500 |
| Mean imputation | 539749.05 | 346961.99 | 330000 | 479500 | 622500 |
| Interpolacja | 539662.35 | 354272.61 | 325000 | 454975 | 644750 |
| Hot-Deck | 540247.04 | 361080.86 | 322000 | 450000 | 648000 |
| Wartości z krzywej regresji | 541034.87 | 358477.11 | 325000 | 454925 | 649950 |

|  |  |  |
| --- | --- | --- |
| **Rysunek 1.** Krzywa regresji dla danych z usuniętymi rzędami, dla 8% braków |  | **Rysunek 2.** Krzywa regresji dla danych wypełnionych metodą mean imputation, dla 8% braków |
|  | | |
| **Rysunek 3.** Krzywa regresji dla danych wypełnionych metodą interpolacji, dla 8% braków |  | **Rysunek 4.** Krzywa regresji dla danych wypełnionych metodą hot-deck, dla 8% braków |
|  | | |
| **Rysunek 5.** Krzywa regresji dla danych wypełnionych wartościami uzyskanymi z tej krzywej, dla 8% braków | | |

**Tabela 3.** Współczynniki regresji dla 8% braków

|  |  |  |
| --- | --- | --- |
| **Metoda** | **Współczynnik kierunkowy** | **Wyraz wolny** |
| Usuwanie rzędów | 280.49 | -42376.33 |
| Mean imputation | 253.89 | 11671.62 |
| Interpolacja | 244.2 | 31228.63 |
| Hot-Deck | 234.2 | 52661.13 |
| Wartości z krzywej regresji | 280.49 | -42376.33 |

# Wszystkie testowane metody wykazały porównywalnie dobre wyniki i nieznacznie zmieniały charakterystyki zbioru. Każda metoda, ze względu na swą implementację, cechowała się zmianami innego rodzaju i tak na przykład zbiór wykorzystujący metodę *mean imputation*, która zakłada wypełnianie brakujących danych średnią wartością danej kolumny w zbiorze nie wpływała zauważalnie na wartość średniej, zaś mocno zmniejszała odchylenie standardowe. Wszystkie metody zmniejszają kąt krzywej regresji – oczywistym wyjątkiem od tego jest metoda wypełniająca dane wartościami z krzywej.

# Wpływ dodatkowych braków na uzyskane wyniki

Drugim eksperymentem było sprawdzenie jak wypełnianie danych wpływa na zmianę parametrów zbioru przy różnych procentowych progach brakujących danych. Dla każdego z następujących progów: **15%**, **30%**, **45%**, porównane wyznaczone cechy charakterystyczne zbiorów z:

* Usuniętymi brakującymi danymi
* Brakującymi danymi wypełnionymi metodą Hot-Deck, która wybrana została na podstawie zadowalających wyników w pierwszym eksperymencie

**Tabela 4.** Cechy charakterystyczne wyliczone dla kolumny **sqft\_living** dla 8% metody Hot-Deck

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **% brakujących danych** | **Imputacja** | **Średnia** | **Odchylenie standardowe** | **Pierwszy kwartyl** | **Drugi kwartyl** | **Trzeci kwartyl** |
| 15 | Nie | 2077.44 | 906.11 | 1430 | 1910 | 2550 |
| Tak | 2075.44 | 910,3 | 1430 | 1910 | 2550 |
| 30 | Nie | 2080.8 | 917.51 | 1430 | 1920 | 2550 |
| Tak | 2068.4 | 904.5 | 1420 | 1900 | 2530 |
| 45 | Nie | 2081.58 | 913.72 | 1430 | 1910 | 2550 |
| Tak | 2088.12 | 920.66 | 1430 | 1920 | 2550 |

**Tabela 5.** Cechy charakterystyczne wyliczone dla kolumny **price** dla 8% metody Hot-Deck

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **% brakujących danych** | **Imputacja** | **Średnia** | **Odchylenie standardowe** | **Pierwszy kwartyl** | **Drugi kwartyl** | **Trzeci kwartyl** |
| 15 | Nie | 538033.16 | 356032.88 | 322000 | 450000 | 642000 |
| Tak | 539294.79 | 361392.88 | 322500 | 450000 | 641000 |
| 30 | Nie | 542757.81 | 372321.61 | 323000 | 450700 | 649950 |
| Tak | 538597.33 | 368482.93 | 320000 | 450000 | 643002 |
| 45 | Nie | 539222.08 | 357915.16 | 321013.5 | 450000 | 649950 |
| Tak | 536711.58 | 352735.35 | 320000 | 450000 | 643000 |

|  |  |  |
| --- | --- | --- |
| **Rysunek 6.** Krzywa regresji dla danych z usuniętymi rzędami, dla 15% braków |  | **Rysunek 7.** Krzywa regresji dla danych wypełnionych metodą hot-deck, dla 15% braków |
|  | | |
| **Rysunek 8.** Krzywa regresji dla danych z usuniętymi rzędami, dla 30% braków |  | **Rysunek 9.** Krzywa regresji dla danych wypełnionych metodą hot-deck, dla 30% braków |
|  | | |
| **Rysunek 10.** Krzywa regresji dla danych z usuniętymi rzędami, dla 45% braków |  | **Rysunek 11.** Krzywa regresji dla danych wypełnionych metodą hot-deck, dla 45% braków |

**Tabela 6.** Współczynniki regresji dla metody hot-deck

|  |  |  |  |
| --- | --- | --- | --- |
| **% brakujących danych** | **Imputacja** | **Współczynnik kierunkowy** | **Wyraz wolny** |
| 15% | Nie | 272.76 | -30610.46 |
| Tak | 201.54 | 119854.04 |
| 30% | Nie | 283.82 | -41621.11 |
| Tak | 159.76 | 220236.49 |
| 45% | Nie | 273.26 | -36070.69 |
| Tak | 106.08 | 347025.71 |

Wraz ze wzrostem brakujących danych imputacja miała coraz większy wpływ na cechy zbioru, co widoczne jest przede wszystkim, kiedy porównujemy współczynniki regresji. Im więcej danych brakowało, tym bardziej malał współczynnik kierunkowy – krzywa regresji była jeszcze bardziej wypłaszczona. Natomiast różnice te nadal nie są tak duże jak można by się było tego spodziewać. Prawdopodobnie jest to spowodowane bardzo dużą liczbą danych, przez co nawet 50% obserwacji wystarczająco dobrze opisuje zbiór. Dzięki temu metody imputacji dobrze się sprawdzały.