|  |  |
| --- | --- |
| Zbigniew Nowacki 234102  Bartosz Jurczewski 234067 | Rok akademicki 2019/20  Poniedziałek, 12:30 |

JĘZYKI PROGRAMOWANIA W ANALIZIE  
DANYCH – LABORATORIUM

Zadanie 2

# Opis implementacji

Zadanie zostało zrealizowane przy użyciu języka Python w wersji 3.7, z wykorzystaniem bibliotek: *pandas*, *matplotlib*, *numpy* oraz *sklearn*.

# Zbiór danych

Wybranym przez nas zbiorem zadań był zbiór „Horse Colic”. Zawiera on dane dotyczące chorób wśród koni. Spośród 28 kolumn wybraliśmy dwie, charakteryzujące się dużą wariacją:

* total protein
* packed cell volume

Ze zbioru usunięto część kolumn tak, aby uzyskać % brakujących danych mieszczący się w zakładanym przedziale 5-10%. Otrzymany przez nas zbiór ma **8.2%** brakujących danych.

Zbiór dostępny jest pod następującym adresem: <https://sci2s.ugr.es/keel/dataset.php?cod=180>

# Wpływ metody uzupełniającej dane na uzyskane wyniki

Przeprowadzony przez nas eksperyment testował różne metody uzupełniania braków w zbiorach danych i sprawdzenie ich wpływu na cechy charakterystyczne.

Dla każdego zbioru wyliczono średnią, odchylenie standardowe oraz wartość trzech kwartyli (Q1,Q2, Q3).

**Tabela 1.** Cechy charakterystyczne wyliczone dla kolumny **total\_protein**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Średnia** | **Odchylenie standardowe** | **Pierwszy kwartyl** | **Drugi kwartyl** | **Trzeci kwartyl** |
| **Usuwanie rzędów** | 24,27 | 27,36 | 6,50 | 7,50 | 56,75 |
| **Mean imputation** | 24,27 | 25,80 | 6,60 | 7,70 | 52,00 |
| **Interpolacja** | 24,36 | 26,69 | 6,60 | 7,50 | 55,00 |
| **Hot Deck** | 24,31 | 27,28 | 6,50 | 7,50 | 56,50 |
| **Wartość z krzywej regresji** | 24,28 | 25,81 | 6,60 | 7,70 | 52,00 |

**Tabela 2.** Cechy charakterystyczne wyliczone dla kolumny **packed\_cell\_volume**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Średnia** | **Odchylenie standardowe** | **Pierwszy kwartyl** | **Drugi kwartyl** | **Trzeci kwartyl** |
| **Usuwanie rzędów** | 46.31 | 10.44 | 38.00 | 45.00 | 52.00 |
| **Mean imputation** | 46,31 | 9,92 | 39,00 | 46,00 | 50,00 |
| **Interpolacja** | 46,34 | 10,13 | 39,00 | 45,00 | 52,00 |
| **Hot Deck** | 46,40 | 10,61 | 38,00 | 45,00 | 52,00 |
| **Wartość z krzywej regresji** | 46,31 | 9,92 | 39,00 | 45,31 | 50,00 |

|  |  |
| --- | --- |
| **Rysunek 1.** Krzywa regresji dla usuniętych rzędów | |
| **Rysunek 2.** Krzywa regresji dla danych wypełnionych metodą **mean imputation** | **Rysunek 3.** Krzywa regresji dla danych wypełnionych metodą **interpolacji** |
| **Rysunek 4.** Krzywa regresji dla danych wypełnionych metodą **hot deck** | **Rysunek 5.** Krzywa regresji dla danych wypełnionych wartościami z tej krzywej |

**Tabela 3.** Współczynnik regresji

|  |  |  |
| --- | --- | --- |
| **Metoda** | **Współczynnik kierunkowy** | **Wyraz wolny** |
| **Usuwanie rzędów** | 0.00223108 | 24.55769689 |
| **Mean imputation** | 0.00780387 | 23.11165952 |
| **Interpolacja** | 0.00863741 | 23.07372798 |
| **Hot Deck** | 0.01046172 | 22.74923077 |
| **Wartość z krzywej regresji** | 0.00861749 | 22.99113488 |

Przeprowadzone przez nas eksperymenty pozwoliły nam dojść do następujących wniosków:

* Uzupełnienie braków średnią nie wpływa na zmianę średniej.
* Uzupełnianie braków z krzywej regresji i średnią daje najbardziej zbliżone do siebie rezultaty w statystykach.
* Linia regresji po uzupełnianiu z krzywej regresji jest najbardziej podobna do linii po uzupełnianiu z interpolacji.
* Rezultaty w statystykach przy metodzie hotdeck są najbardziej zbliżone do statystyk przed imputacją.