POLITECHNIKA WARSZAWSKA

WYDZIAŁ MECHATRONIKI

Data Mining

projekt grupowy  
„Regresja logistyczna”

Wykonali:  
Adrian Szymański  
Michał Kalski



Dane zostały pobrane z zbioru dostępnego pod linkiem: [UCI Machine Learning Repository: Estimation of obesity levels based on eating habits and physical condition Data Set](https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+).   
Pochodzą z badań prowadzonych nad wagą ciała osób charakteryzowanych 17-stoma zmiennymi: [Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S2352340919306985?via%3Dihub).

Obraz zawierający tekst, Czcionka, zrzut ekranu

Opis wygenerowany automatycznie



Dane zawarte w pliku zawierają 17 zmiennych:

* Gender -> płeć osoby badanej [mężczyzna/kobieta]
* Age -> wiek [lata]
* Height -> wzrost [m]
* Weight -> masa ciała [kg]
* Family\_history\_with\_overveight -> problemy z masą ciała w rodzinie [tak/nie]
* FAVC -> częstotliwość spożywania wysokokalorycznych posiłków [tak/nie]
* NCP -> liczność głównych posiłków [liczba]
* CAEC -> spożycie pomiędzy posiłkami [4 poziomy]
* SMOKE -> palenie papierosów [tak/nie]
* CH2O -> dzienne spożycie wody [l]
* SCC -> świadome obliczanie liczby kalorii [tak/nie]
* FAF -> częstotliwość wysiłku fizycznego [4 poziomy]
* TUE -> czas użytkowania sprzętu elektronicznego dziennie [3 poziomy]
* CALC -> częstotliwość spożywania alkoholu [3 poziomy]
* MTRANS -> środek transportu/ sposób dojeżdżania do pracy [4 typy]
* NObeyesdad -> oznaczenie poziomu wagi [6 poziomów]

OPIS ZAGADNIENIA REGRESJI LOGISTYCZNEJ

Dane zostały przekształcone tak, aby można było przeprowadzić regresję logistyczną:

* Zmienna wyjściowa *NObeyesdad* została przekształcona na binarną postaci *Proper\_weight,* która przyjmuje wartość 1 gdy waga jest „Normal\_weight” w przeciwnym wypadku 0
* Pozostałe zmienne przekształcone zostały z typu *character* na *numeric* z odpowiednią rozpiętością wartości typu *intiger*

Obraz zawierający tekst, Czcionka, dokument, zrzut ekranu

Opis wygenerowany automatycznie



Sporządzono podsumowanie przekształconych/przygotowanych danych:

Obraz zawierający tekst, biały, czarne i białe, paragon

Opis wygenerowany automatycznie

Obraz zawierający tekst, diagram, linia, Wykres

Opis wygenerowany automatycznie

Sprawdzono korelacje danych wejściowych.

Aby dokonać poprawnego podziału danych na uczące i testowe, najpierw rozdzielono dane względem zmiennej wyjściowej. Następnie podzielono w stosunku 70:30 i scalono wyniki podziału.

Obraz zawierający tekst, zrzut ekranu, Czcionka, dokument

Opis wygenerowany automatycznie



Obraz zawierający tekst, zrzut ekranu, Czcionka, algebra

Opis wygenerowany automatycznie

Pierwszy model pełny, *formuła* uwzględnia wszystkie zmienne wejściowe. Dopasowanie zostało utworzone dla rodziny *quasibinominal*. Jako próg odcięcia przyjęto *0.5.*

Obraz zawierający tekst, zrzut ekranu, design

Opis wygenerowany automatycznie

Dopasowanie sprawdzono na danych testowych. Model poprawnie opisuje przyporządkowanie do grupy z wagą niepoprawną, jednak popełnia duży błąd przyporządkowując osoby z poprawną masą ciała do grupy przeciwnej.

Obraz zawierający tekst, zrzut ekranu, menu, dokument

Opis wygenerowany automatycznie

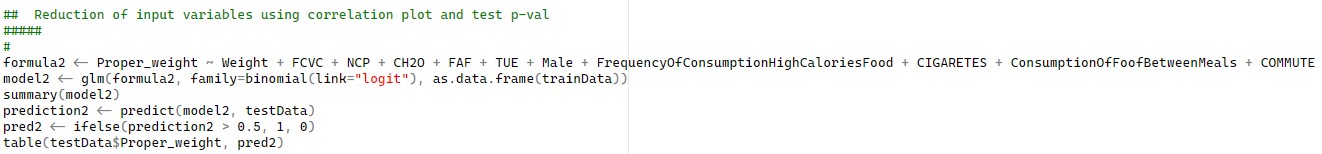
Na podsumowaniu powstałego modelu można zauważyć, że ze względu na zbyt duże wartości testu p-val wiele zmiennych nie jest istotnych.

Obraz zawierający tekst, diagram, zrzut ekranu, linia

Opis wygenerowany automatycznie

OPIS DO WYKRESU





Na podstawie correlogramu i wykonanych testów p-val w modelu pierwszym zredukowano ilość zmiennych wejściowych.

Obraz zawierający tekst, zrzut ekranu, menu, dokument

Opis wygenerowany automatycznie

Obraz zawierający tekst, zrzut ekranu, Czcionka, typografia

Opis wygenerowany automatycznie

Jak widać, stworzony model zawiera nadal kilka zmiennych o niewielkim znaczeniu oraz popełnia nieco mniejszą choć podobną liczbę błędów.

Obraz zawierający tekst, zrzut ekranu, Czcionka, algebra

Opis wygenerowany automatycznie

Aby poprawić jakość modelu usunięto z zbioru uczącego obserwacje o dużej odległości Cooke’a. Dodatkowo ograniczono ponownie liczbę zmiennych wejściowych do modelu.

Obraz zawierający tekst, zrzut ekranu, dokument, Czcionka

Opis wygenerowany automatycznie

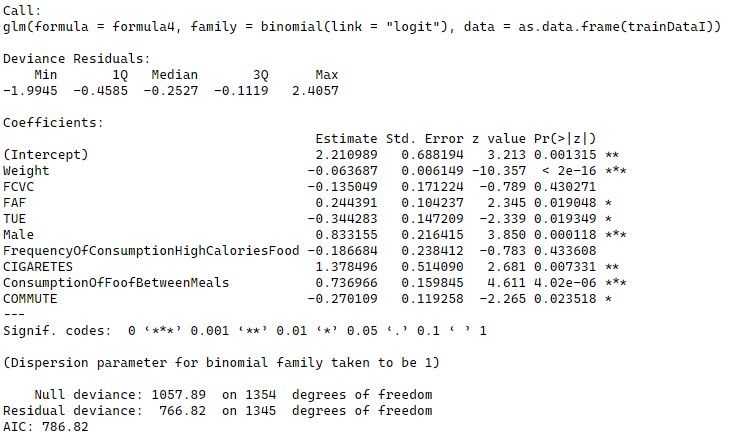


Dopasowany model mimo prostszej budowy nie popełnia większej ilości błędów.



Obraz zawierający tekst, zrzut ekranu, Czcionka, algebra

Opis wygenerowany automatycznie





Ostateczny model ograniczony został do 9 zmiennych wejściowych. Próg odcięcia zmieniono na 0.1. Model popełnia znacząco mniej błędów w przypadku przyporządkowania osób z poprawna masą ciała jako niepoprawną. Popełnia przy tym niewiele więcej błędów przyporządkowania w odwrotnej sytuacji.

1. TEN NO I TU