

BIG DATA

PROBLEM KLASYFIKACJI - CUKRZYCA

13 stycznia 2019

Emilia Lubos
Daria Pacewicz
Michał Gandor

1 Opis zbioru danych

Zbiór zawiera informacje czy u danego pacjenta występuje cukrzyca czy też nie. Pacjentami są kobiety w wieku 21 lat lub starszych pochodzących z Indii. Opis dokonany jest za pomocą zmiennych:

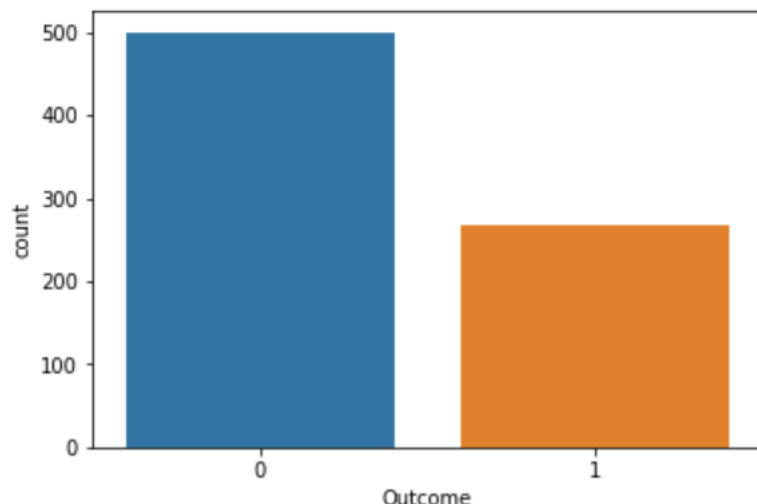
- Pregnancies - ilość ciąż,
- Glucose - koncentracja glukozy wg 2-godzinnego testu,
- BloodPressure - rozkurczowe ciśnienie krwi (mm Hg),
- SkinThickness - grubość fałdu skóry na tricepsie (mm),
- Insulin - poziom insuliny mierzony (mu U/ml)
- BMI - index BMI (waga w kg/(wzrost w m^2)
- DiabetesPedigreeFunction - funkcja rodowodu cukrzycy,
- Age - wiek w latach,
- Outcome - 0 = wynik negatywny (brak cukrzycy), 1 = wynik pozytywny (cukrzyca).

Rozkład klas:

- 0 - 500 próbek
- 1 - 268 próbek

Całkowita liczba obserwacji wynosi 786.

Źródło: <https://www.kaggle.com/uciml/pima-indians-diabetes-database>



Rys. 1: Outcome

2 Cel projektu

Celem projektu jest dokonanie klasyfikacji oraz zbadanie czy u danego pacjenta wystąpi cukrzyca czy nie. Zbadane zostanie także czy dane zawarte w zbiorze są wystarczające do decyzji o prawdopodobieństwie wystąpienia choroby oraz czy wszystkie z nich wpływają znacząco na wystąpienie choroby. W projekcie porównane zostaną wyniki skuteczności różnych klasyfikatorów.

3 Analiza zbioru

Zanim przystąpimy do próby klasyfikacji dane muszą zostać odpowiednio przygotowane. Sprawdzone zostają wiersze w których występują zera oraz wartości odstające.

3.1 Wartości zerowe

Z medycznego punktu widzenia cechy w badanym zbiorze nie mogą być równe zero, świadczy to więc o braku poprawnych danych. Wartości zerowe we wszystkich kolumnach zostają zastąpione przez średnią wartość w każdej z nich.

3.2 Wartości odstające

- Pregnancies

Brak wartości odstających.

- Glucose

Według badań medycznych poziom cukru poniżej 140mg/dl jest wynikiem prawidłowym. Ponad 200mg/dl świadczy o wystąpieniu cukrzycy. (<https://www.diabetes.co.uk/diabetescare/blood-sugar-level-ranges.html>)

- BloodPressure

Rozkurczowe ciśnienie tętnicze uznawane jest za poprawne w granicach 60-80 mm/Hg. Poniżej tego progu zostaje stwierdzona choroba. Ciśnienie nie może być mniejsze niż 40. (<http://www.bloodpressureuk.org/BloodPressureandyou/Theba>)

- SkinThickness

- Insulin - poziom insuliny mierzony (mu U/ml)

- BMI

Zakres BMI waha się pomiędzy 16 a 40. Gdzie 16 to wygłodzenie, a wartości powyżej 40 to 3 stopień otyłości. (<https://pl.wikipedia.org/wiki/Wskaźnikmasyciała>). Wiersz wykraczające poza ten zakres powinny zostać usunięte ze zbioru.

- DiabetesPedigreeFunction - funkcja rodowodu cukrzycy,

- Age

Brak wartości odstających.

3.3 Selekcja cech

Z modelu zostały usunięte cechy które nie zostały uznane za istotne. Kolumny które pozostaną w modelu zostały wybrane z użyciem ExtraTreesClassifier. Cechy z najniższym wynikiem zostają odrzucone.

3.4 Normalizacja

Dane zostały znormalizowane. Wszystkie wartości są teraz z zakresu 0-1. Do normalizacji użyto MinMaxScaler i StandardScaler.

4 Klasyfikacja

Do klasyfikacji zostało użytych kilka różnych klasyfikatorów w celu porównania wyników.

4.1 SVN

4.2