Big data

PROBLEM KLASYFIKACJI - CUKRZYCA

14 stycznia 2019

Emilia Lubos
Daria Pacewicz
Michał Gandor

Spis treści

1 Opis zbioru danych

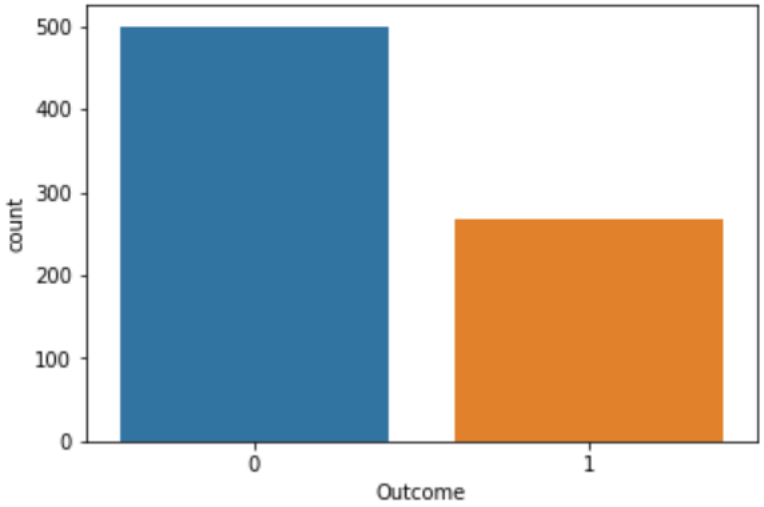
Zbiór zawiera informacje czy u danego pacjenta występuje cukrzyca czy też nie. Pacjentami są kobiety w wieku 21 lat lub starszych pochodzących z Indii. Opis dokonany jest za pomocą zmiennych:

- 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

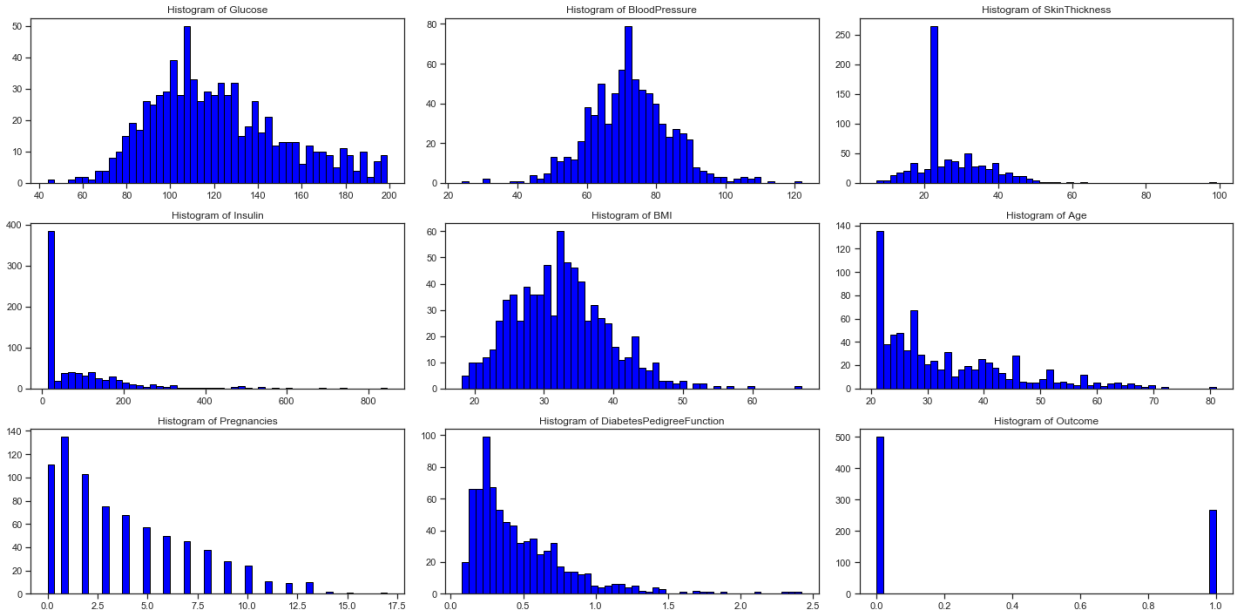
Celem projektu jest dokonanie klasyfikacji oraz zbadanie czy u danego pacjenta wystąpi cukrzyca czy nie. Zbadane zostanie także czy dane zawarte w zbiorze są wystarczające do decyzji o prawdopodobieństwie wystąpienia choroby oraz czy wszystkie z nich wpływają znacząco na wystąpienie choroby. W projekcie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

3 Narzędzia

W tym projekcie posłużono się językiem Python oraz bibliotekom poświęconym analizie danych i uczeniu maszynowym, takimi jak: NumPy, Pandas, SciKit-learn. Do wytworzenia wykresów zastosowano biblioteki Matplotlib oraz Seaborn.

4 Analiza zbioru

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



Rys. 2: Histogramy po zastąpieniu wartości 0 medianą

wartości odstające.

4.1 Wartości zerowe

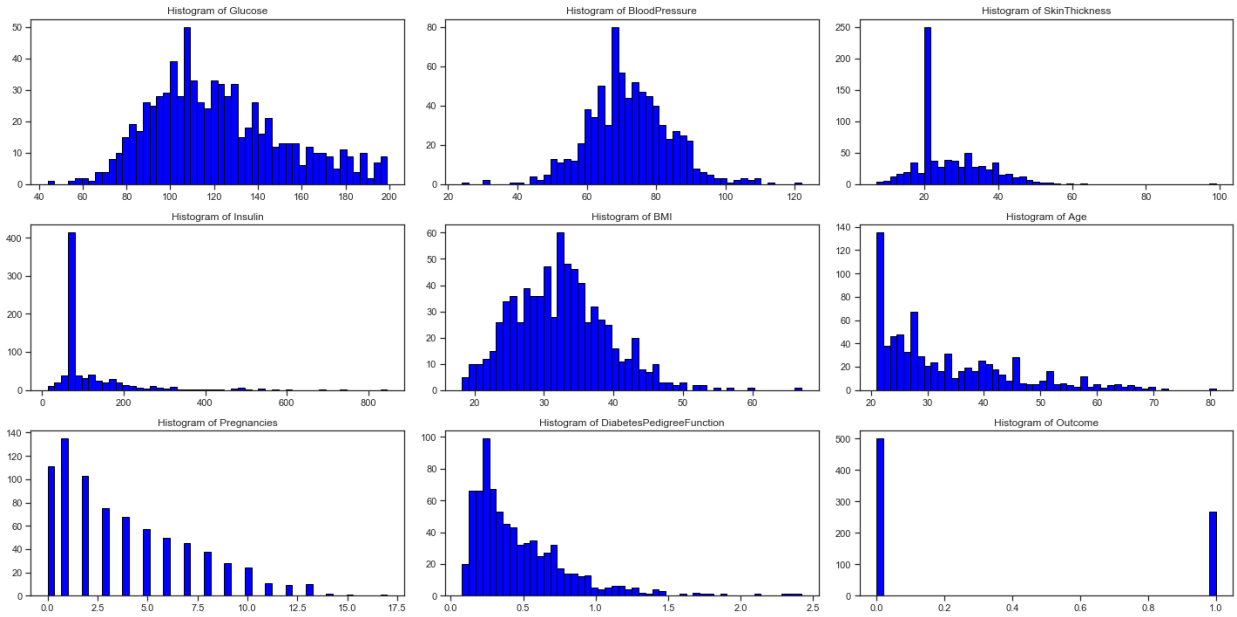
Z tabel wynika, że istnieją wartości zerowe w kolumnach *Pregnancies, Glucose, BoloodPressuure, SkinThickness, Insulin, BMI*. Z medycznego punktu widzenia cechy w badanym zbiorze nie mogą być równe zeru, świadczy to o braku poprawności danych danych. W przypadku pierwszej kolumny wartości zerowe są poprawne - jest to informacja, że dana kobieta nie była w ciąży. W przypadku reszty obserwacji dane zostają zamienione na **średnią wartość** w kolumnie, **medianę** oraz zostają całkowicie **usunięte**. Oba trzy przypadki posłużą jako dane testowe.

4.2 Wartości odstające

- Pregnancies

  Brak wartości odstających.

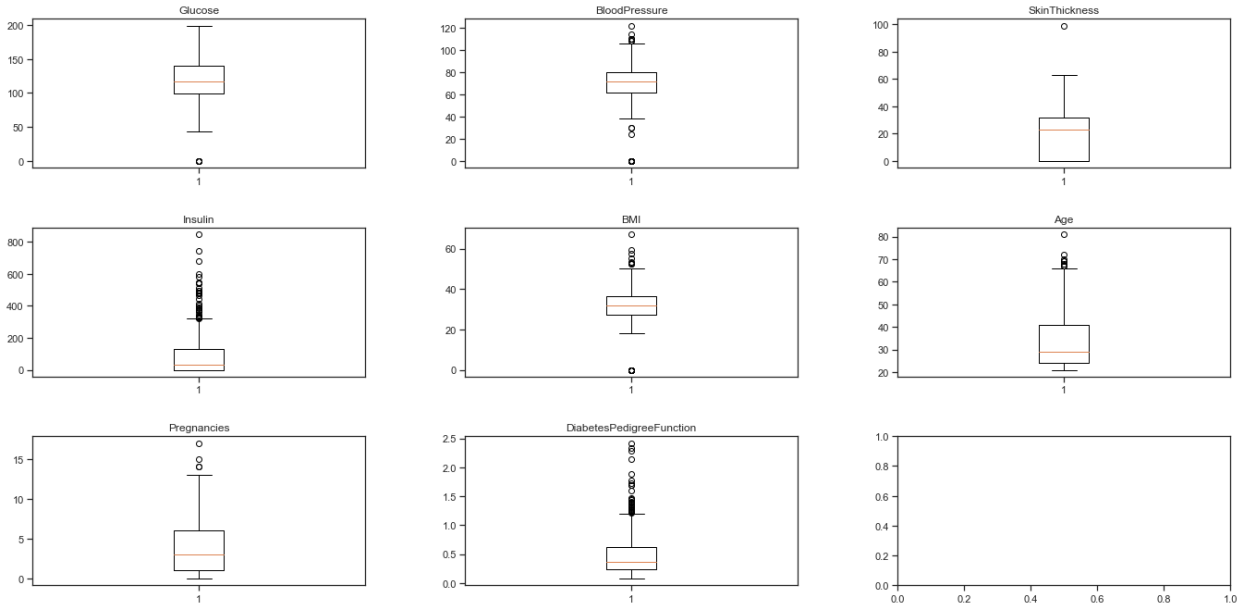
- Glucose



Rys. 3: Histogramy po zastąpieniu wartości 0 średnią

	Pregnancies	**Glucose**	**BloodPressure**	**SkinThickness**	**Insulin**
count	768	768	768	768	768
mean	3.845052	120.8945	69.10547	20.53646	79.79948
std	3.369578	31.97262	19.35581	15.95222	115.244
min	0	0	0	0	0
25%	1	99	62	0	0
50%	3	117	72	23	30.5
75%	6	140.25	80	32	127.25
max	17	199	122	99	846

	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768	768	768	768
mean	31.99258	0.471876	33.24089	0.348958
std	7.88416	0.331329	11.76023	0.476951
min	0	0.078	21	0
25%	27.3	0.24375	24	0
50%	32	0.3725	29	0
75%	36.6	0.62625	41	1
max	67.1	2.42	81	1



Rys. 4: Wartości odstające

Według badań medycznych poziom cukru poniżej 140mg/dl jest wynikiem prawidłowym. Ponad 200mg/dl świadczy o wystąpieniu cukrzycy [**?**].

- BloodPressure

  Rozkurczowe ciśnienie tętnicze uznawane jest za poprawne w granicach 60-80 mm/Hg. Poniżej tego progu zostaje stwierdzona chorba. Ciśnienie nie może być mniejsze niż 40 [**?**].

- 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 [**?**]. Wiersz wykraczające poza ten zakres powinny zostać usunięte ze zbioru.

- DiabetesPedigreeFunction - funkcja rodowodu cukrzycy,
- Age

  Brak wartości odstających.

4.3 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.

4.4 Normalizacja

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

5 Walidacja krzyżowa

Przed przystąpieniem do klasyfikacji zastosowano prosty podział zbioru danych na dane treningowe oraz testowe oraz 5-krotną walidację krzyżową.

Zdecydowano się na podział zbioru w następującej proporcji: 70% - dane treningowe, 30% - dane testowe. Podział ten zastosowano do wyboru najlepszych parametrów modeli klasyfikacji. Walidację krzyżową stosuje się w celu minimalizacji problemu nadmiernego dopasowania (*overfitting*). Dzięki niej można uzyskać informacje takie jak dokładność modelu (accuracy) czy macierz pomyłek, które umożliwiają ocenę jakości modelu.

6 Klasyfikacja

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

Tu można dodać coś o wybranych parametrach i dlaczego takie....

- Maszyna wektorów nośnych (SVN)
- K najbliżych sąsiadów (KNN)
- Drzewo decyzyjne
- Las losowy
- Regresja logistyczna
- Naiwny klasyfikator bayesowski

7 Wyniki

Literatura

[1] https://www.diabetes.co.uk/diabetescare/blood-sugar-level-ranges.html

[2] http://www.bloodpressureuk.org/BloodPressureandyou/Thebasics/Bloodpressurechart)

[3] https://pl.wikipedia.org/wiki/Wskaźnikmasyciała