Metody Analizy Dużych Zbiorów Danych

PROBLEM KLASYFIKACJI - CUKRZYCA

24 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:

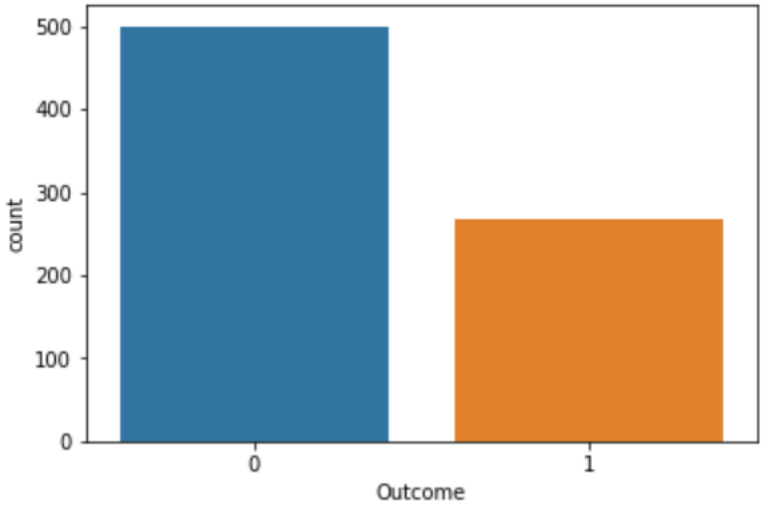
- Pregnancies - ilość ciąż,
- Glucose - koncentracja glukozy wg 2-godzinnego testu,
- BloodPressure - rozkurczowe ciśnienie krwi (mm Hg),
- SkinThickness - grubość fałdu skóry na tricepsie (mm),
- Insulin - poziom insuliny mierzony (mu U/ml)
- BMI - index BMI (waga w kg/(wzrost w m^2)
- DiabetesPedigreeFunction - 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

Tabela 1: Opis zbioru danych, cz. 1

	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

Tabela 2: Opis zbioru danych, cz. 2

	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

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 celu efetywnej implementacji kolejnych etapów projektu wykorzystano środowisko Jupiter Notebook, a współpracę zespołu umożliwiło repozytorium na serwisie GitHub oraz Overleaf do współdzielenia dokumentacji w LaTeX.. W projekcie posłużono się językiem Python oraz bibliotekami poświęconymi analizie danych i uczeniu maszynowemu, takimi jak: NumPy, Pandas, SciKit-learn. Do wytworzenia wykresów zastosowano biblioteki Matplotlib oraz Seaborn.

4 Analiza zbioru

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

4.1 Wartości zerowe

Z tabel **??** oraz **??** 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**. Wszystkie trzy przypadki posłużą jako dane testowe.

4.2 Wartości odstające

Pomimo, że w zbiorze dla niektórch cech występują wartości odstające nie zostały one usunięte. Dla niektórych zmiennych z powodów medycznych nie został zdefiniowany górny lub dolny zakres mierzalny. W niektórych przypadkach wartości odstające mogą również świadczyć o objawach choroby, a więc usunięcie ich ze zbioru wpłynęłoby niekorzystnie na dopasowanie modelu.

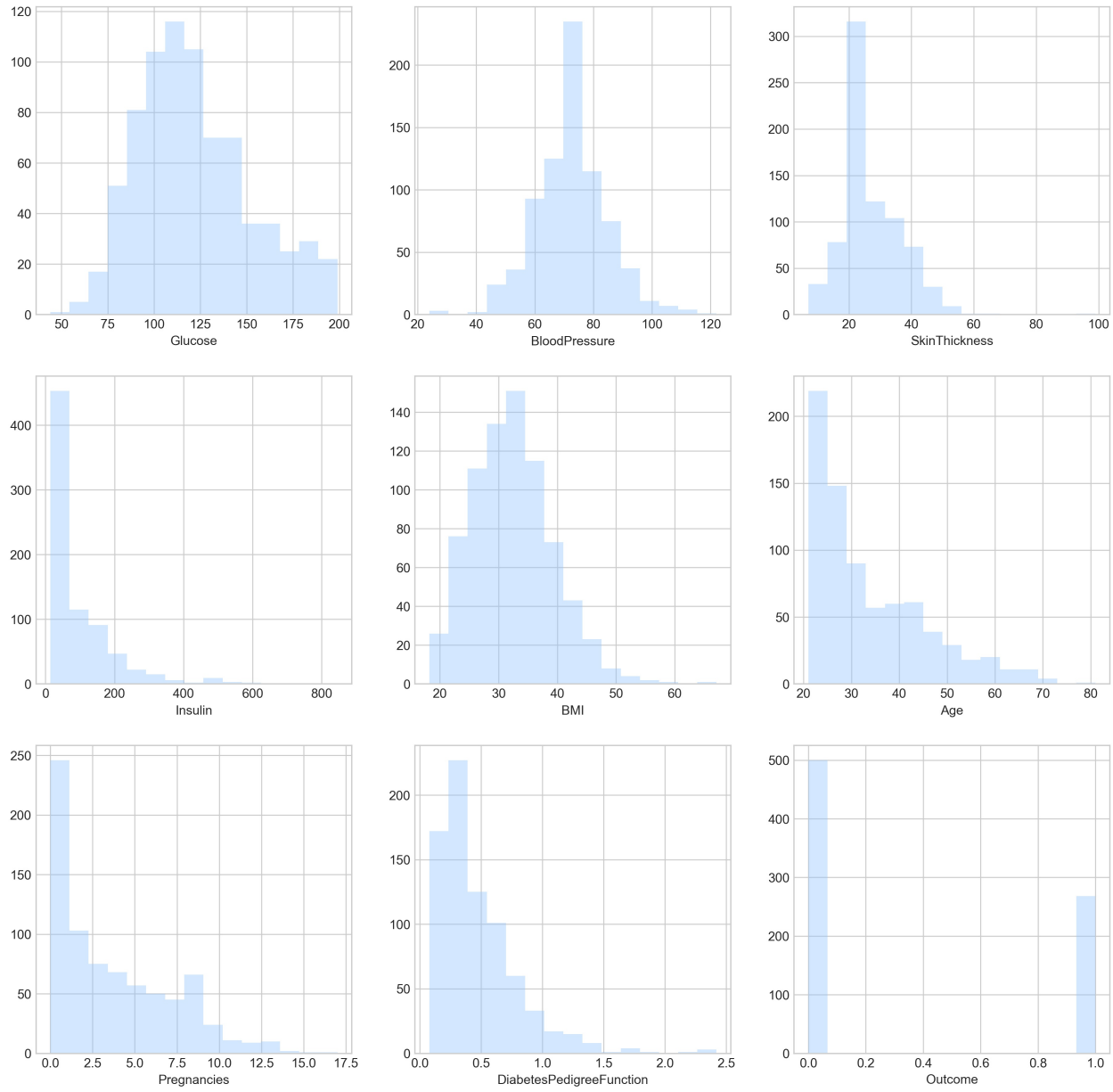
4.3 Selekcja cech

Przy użyciu *ExtraTreesClassifier* z modelu zostały usunięte cechy, które nie zostały uznane za istotne. Dzięki tej redukcji, trening modelu może być przeprowadzony w krótszym czasie bez straty na jakości. Cechy z najniższym wynikiem zostają odrzucone. Wyniki istotności wg. *ExtraTreesClassifier* zostały przedstawione w tabeli **??**.

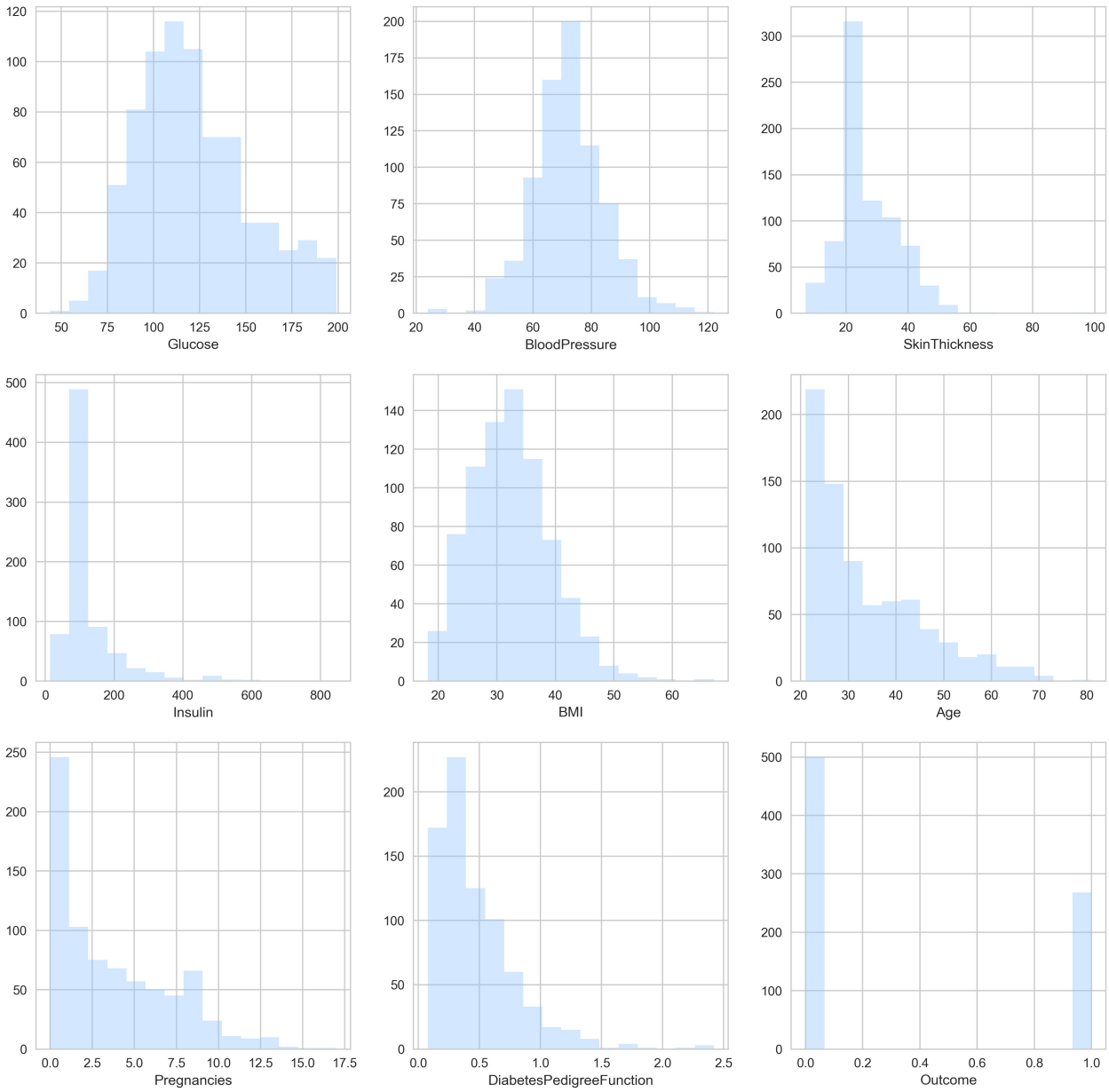
Tabela 3: Wyniki przeprowadzenia eksperymentu selekcji cech

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPdigFun	Age
0.108638	0.245338	0.083665	0.084513	0.081918	0.15441	0.1096581	0.127862

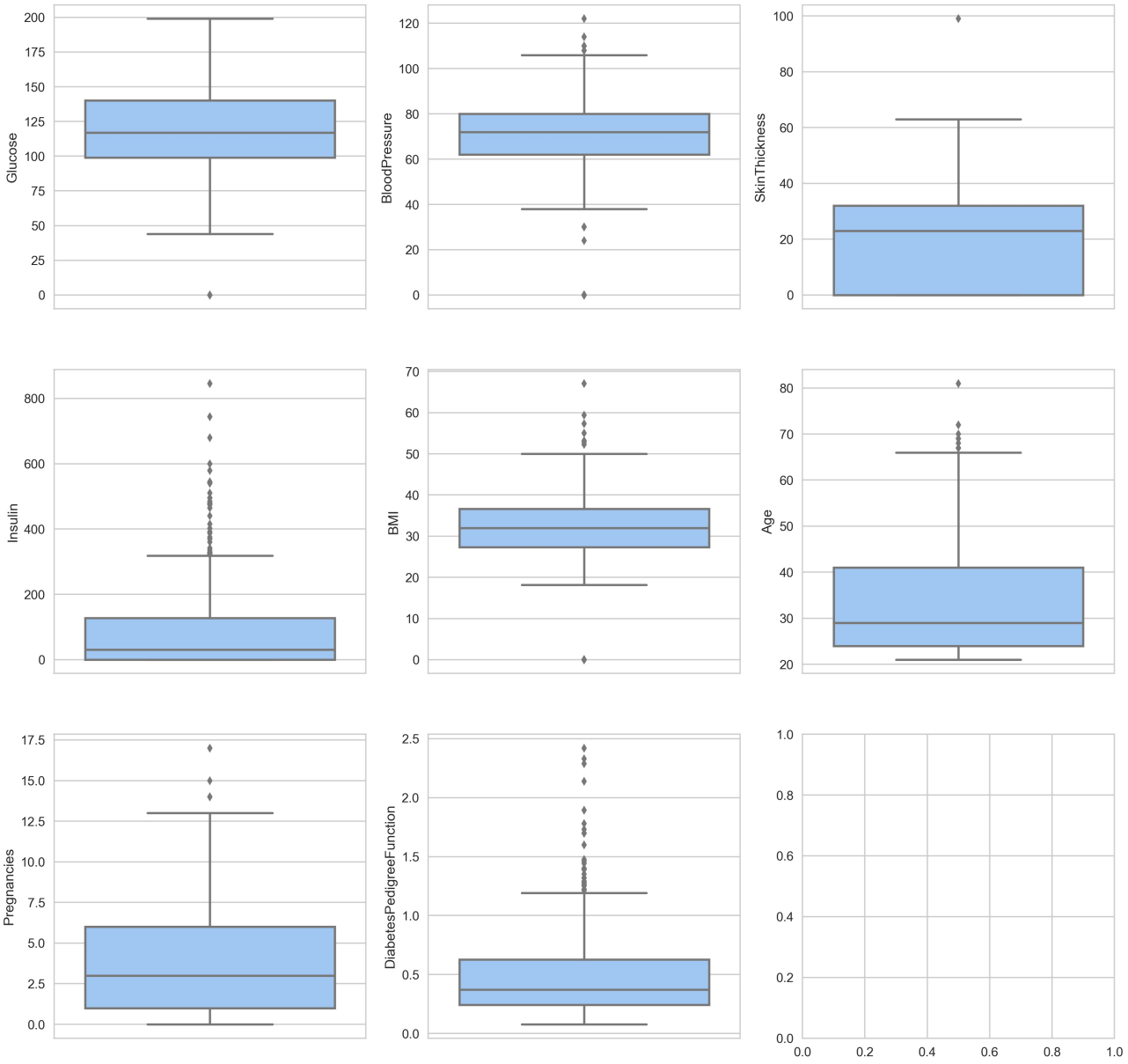
Zgodnie z powyższą tabelą postanowiono odrzucić cechy: BloodPressure oraz Insullin.



Rys. 2: Histogramy po zastąpieniu wartości 0 medianą



Rys. 3: Histogramy po zastąpieniu wartości 0 średnią



Rys. 4: Wartości odstające

4.4 Normalizacja

Normalizacja jest jednym z najważniejszych przekształceń dokonywanych na danych. W przypadku kiedy zakresy wartości różnych cech znacznie się różnią, klasyfikator może usnać wyższe wartości za bardziej wpływające na model [**?**]. Do normalizacji użyto:

- *MinMaxScaler*

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{1}$$

- StandardScaler

$$x' = \frac{x - \overline{x}}{\sigma} \tag{2}$$

Najlepsze wyniki zostały uzyskane w przypadku *StandardScaler*.

5 Walidacja krzyżowa

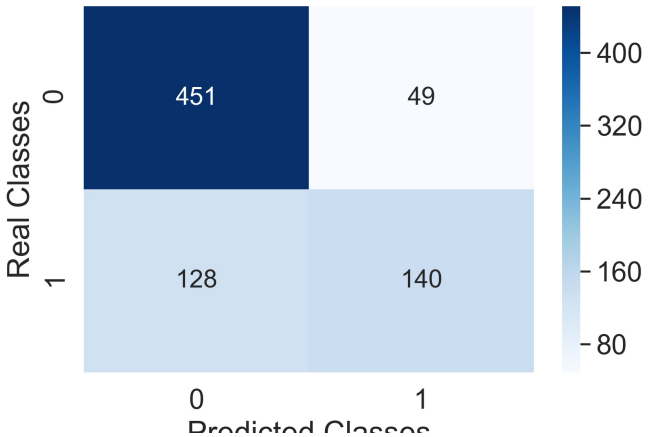
Przed przystąpieniem do klasyfikacji zastosowano podział zbioru danych na dane treningowe oraz testowe, a następnie wykorzystano 5-krotną walidację krzyżową. Proces ten stosuje się w celu minimalizacji problemu nadmiernego dopasowania (*overfitting*). Dzięki niemu można uzyskać informacje takie jak dokładność modelu (*accuracy*) czy macierz pomyłek, które umożliwiają ocenę jakości modelu.

6 Wyniki

Do klasyfikacji zostało użytych siedem różnych klasyfikatorów w celu porównania wyników. Dla każdego klasyfikatora zastosowano metodę *GridSearch*, w celu znalezienia najlepszych parametrów modelu. Zostały wykonanane łącznie 334 porównania dla 5-krotnej walidacji krzyżowej. Dzięki wyspecyfikowaniu parametru *n_jobs* obliczenia wykonywane były szybciej przy użyciu kilku rdzeni procesora. Poniżej przedstawiono parametry modeli, którymi posłużono się w dalszej części badań.

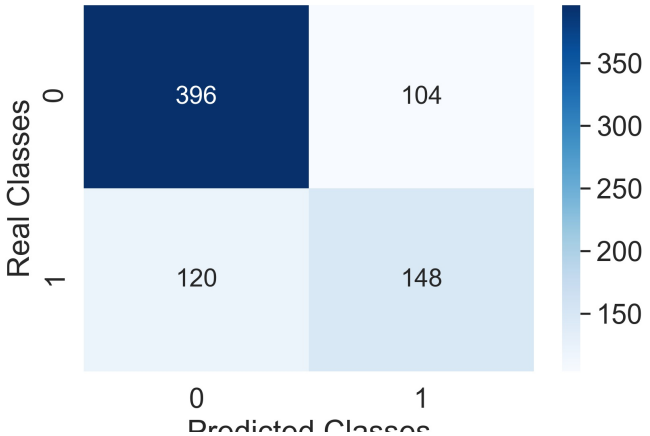
- Maszyna wektorów nośnych (SVM)
  - C = 1 (parametr kary)

- gamma = 0.01 (współczynnik jądra)
- kernel = rbf (typ jądra)



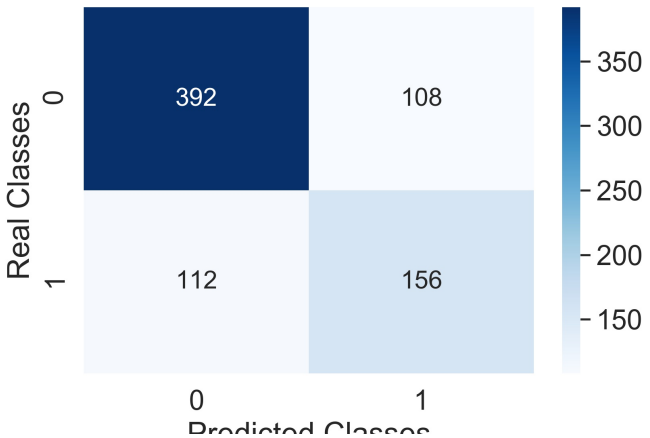
Rys. 5: Macierz pomyłek dla klasyfikatora SVC

- K najbliższych sąsiadów (KNN)
  - n_neighbors = 3 (liczba sąsiadów)
  - weights = uniform (funkcja wagowa)



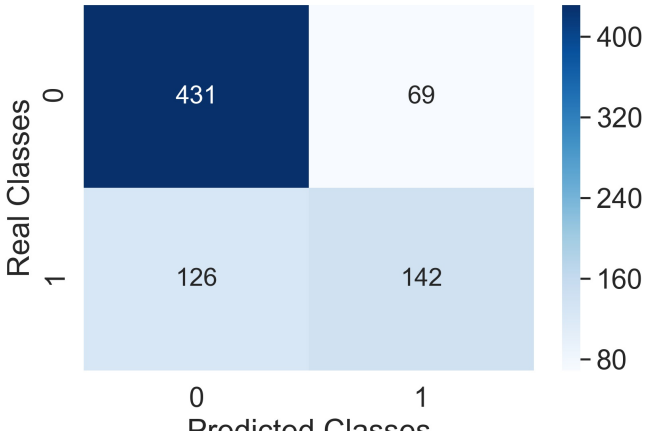
Rys. 6: Macierz pomyłek dla klasyfikatora KNN

- Drzewo decyzyjne
  - max_depth = 6 (maksymalna głębokość drzewa)



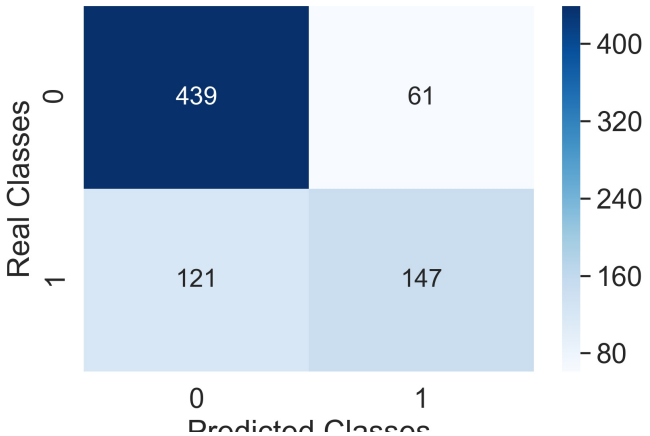
Rys. 7: Macierz pomyłek dla klasyfikatora Decision Tree

- Las losowy
  - max_depth = 3 (maksymalna głębokość drzewa)
  - max_features = 4 (liczba zmiennych rozpatrywanych przy budowie drzewa)
  - min_samples_split = 3 (minimalna liczba próbek potrzebna to rozgałęzienia)
  - bootstrap = True (czy wykorzystywane są próbki typu bootstrap)
  - criterion = gini (funkcja mierząca jakość rozgalęzienia)
  - n_estimators = 10 (liczba drzew w lesie)



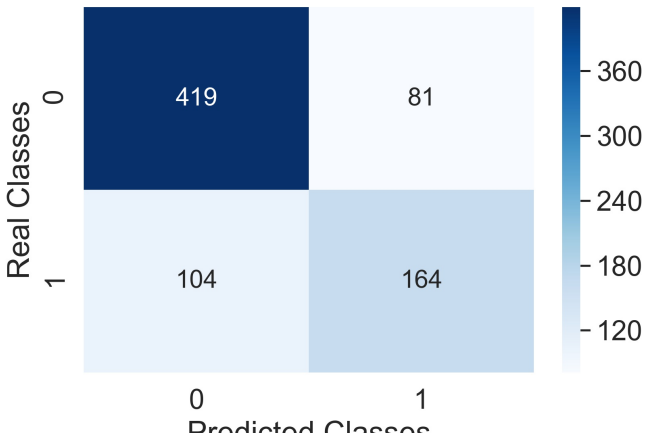
Rys. 8: Macierz pomyłek dla klasyfikatora Las Losowy

- Regresja logistyczna
  - C = 0.1 (paramter kary)
  - penalty = l1 (norma wykorzystywana w procesie karania)



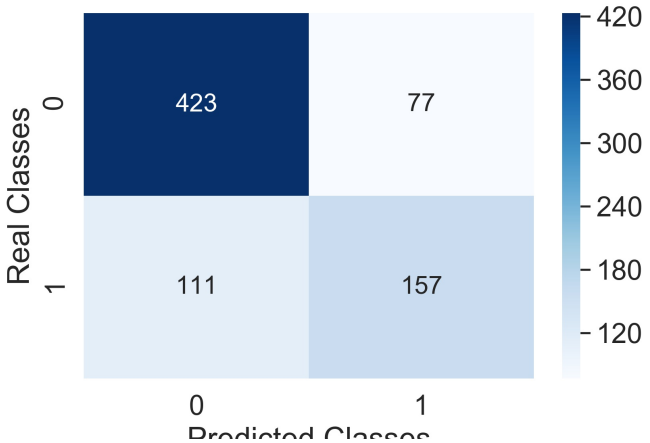
Rys. 9: Macierz pomyłek dla klasyfikatora Regresji Logistycznej

- Naiwny klasyfikator bayesowski



Rys. 10: Macierz pomyłek dla Naiwny klasyfikatora bayesowskiego

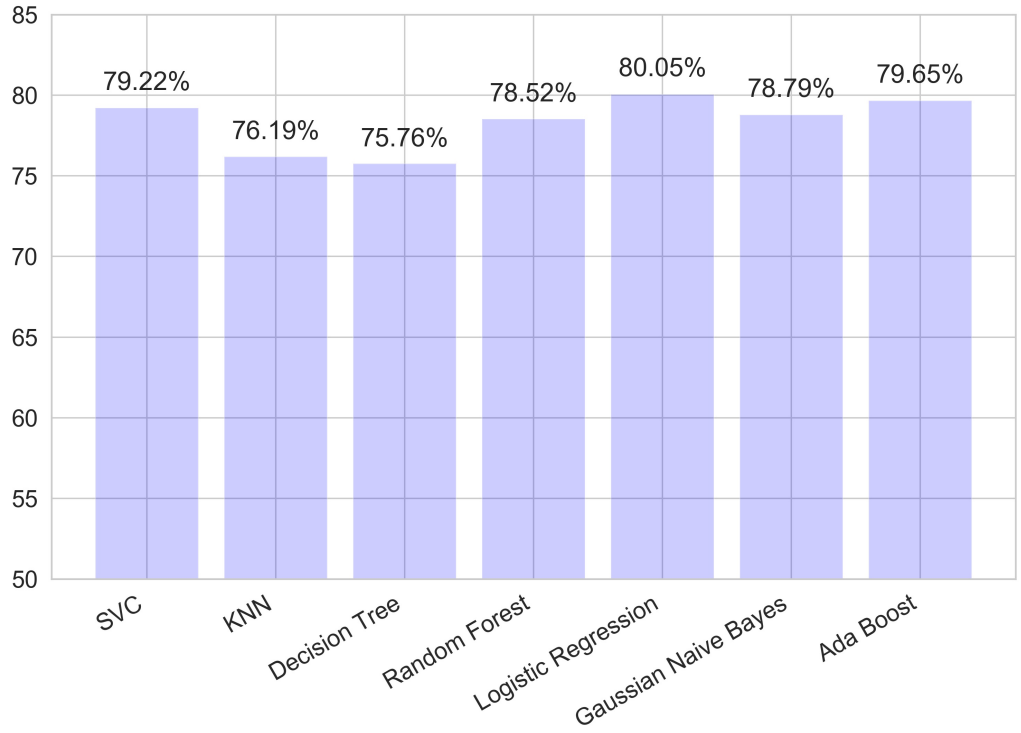
- Ada Boost



Rys. 11: Macierz pomyłek dla klasyfikatora Ada Boost

Tabela 4: Ostateczne wyniki dokładności klasyfikatorów dla najlepiej dopasowanych parametrów

	SVC	KNN	Dec. Tree	Rnd Forest	Log Reg	Gaussian naive	Ada Boost
Std Scaller	0.7922	0.7619	0.7706	0.7706	0.8052	0.78	0.7965
Min-Max	0.7965	0.7449	0.7619	0.7749	0.7965	0.78	0.7965
Brak	0.7965	0.7449	0.7446	0.7965	0.7965	0.78	0.7965



Rys. 12: Dokładność klasyfikatorów

7 Wnioski

Zdecydowanie najbardziej pracochłonną częścią projektu była sama analiza i przygotowanie danych wykorzystanych w późniejszej klasyfikacji. Jest też to najistotniejszy element tego typu projektów. Niepoprawne próbki, czyli te

które zawierały wartości zerowe w niektórych kolumnach stanowiły większość zbioru i bez zastosowania odpowiednich metod wynik klasyfikacji byłby dużo słabszy. Kolejnym wnioskiem, który nasuwa się po zakończeniu prac jest to, że wynik rzędu 80% skuteczności klasyfikatora dla zastosowań medycznych nie jest wynikiem zadowalającym. W rzeczywistym przypadku nie można pozwolić sobie na tak duży błąd. Być może udało by się usyzkać lepszy efekt posiadając wyniki większej ilości badań, bądź dane uzupełnione o brakujące wartości.

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

[4] https://en.wikipedia.org/wiki/Cross-validation_(statistics)

[5] https://en.wikipedia.org/wiki/Hyperparameter_optimization#Grid_search

[6] Aurélien Géron, „Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow."