RAPORT KLASYFIKACJI CUKRZYKÓW

KRÓTKO O CUKRZYCY I ZASTOSOWANIU AI DO KLASYFIKACJI OSÓB

Z DANYCH UDOSTĘPNIONYCH PRZEZ STRONE PACJENT.GOV, W 2018 2,9 MLN OSÓB ZOSTAŁO ZDIAGNOZOWANYCH Z CHOROBĄ CUKRZYCY. OSOBY CHORE NA CUKRZYCE SĄ BARDZIEJ NARAŻONE NA RETINOPATIE, NIEWYDOLNOŚĆ NEREK, CZY NEUROPATIĘ. ZASTOSOWANIE KLASYFIKACJI Z POMOCĄ AI POMOGŁOBY OSOBOM NA WCZESNO-WSTĘPNĄ DIAGNOZĘ ORAZ WCZEŚNIEJSZE LECZENIE CUKRZYCY. CHOĆ CUKRZYCA MOŻE DOTKNĄĆ KAŻDEGO, TAK STATYSTYKI POKAZUJĄ, ŻE SĄ CECHY LUDZKIE, KTÓRE ZWIĘKSZAJĄ PRAWDOPODOBIEŃSTWO ZACHOROWANIA NA CUKRZYCE, JAK WIEK CZY ILOŚĆ SPOŻYCIA CUKRU.

O DATASECIE WYKORZYSTANYM W RAMACH KLASYFIKACJI

- HTTPS://WWW.KAGGLE.COM/DATASETS/ALEXTEBOUL/DIABETES-HEALTH-INDICATORS-DATASET;
- DATASET SKŁADA SIĘ Z 22 KOLUMN, W TYM:
    - 21 TO ODPOWIEDZI PACJENTÓW NA PYTANIA;
    - 1 TO KOLUMNA ETYKIETOWA, Z 3 MOŻLIWOŚCIAMI:
        - BRAK CUKRZYCY;
        - PRZED CUKRZYCĄ;
        - CUKRZYCA;
- STĄD TEŻ, PROJEKTEM JEST KLASYFIKACJA WIELOKLASOWA;
- DATASET POWSTAŁ Z ANKIET TELEFONICZNYCH PRZEPROWADZONYCH PRZEZ THE BEHAVIORAL RISK FACTOR SURVEILLANCE SYSTEM (BRFSS);
- DATASET ZOSTAŁ ZREDUKOWANY Z 330 ATRYBUTÓW ORAZ ZOSTAŁ OCZYSZCZONY PRZEZ AUTORA;
- PO WSTĘPNEJ ANALIZIE DYSTRYBUCJI KLAS, DATASET TEN MOŻNA OPISAĆ JAKO NIE ZBALANOSOWANY (0.82% : 0.02% : 0.16%).

CZYSZCZENIE I PRZYGOTOWANIE DANYCH

- USUNIĘCIE POWTARZAJĄCYCH SIĘ INSTANCJI WIĘCEJ NIŻ 2 RAZY;
- ZASTOSOWANIE TRAIN-TEST SPLIT 2-KROTNIE, W CELU STWORZENIA DANYCH TRENINGOWYCH, TESTOWYCH I WALIDACYJNYCH;
- STANDARDYZACJA DANYCH;
- OVERSAMPLING (BORDERLINE SMOTE) I UNDERSAMPLING (RANDOM UNDER SAMPLER) DANYCH TRENINGOWYCH, W CELU ZNIWELOWANIA NIEZRÓWNOWAŻENIU KLAS W DATASECIE ORAZ POPRAWIENIU JAKOŚCI NAUKI ALGORYTMÓW;

METRYKI

- F1 WEIGHTED;
- ORAZ DODATKOWO:
  - ROC;
  - CONFUSION MATRIX;
  - ACCURACY;
- ACCURACY ZOSTAŁO ZASTĄPIONE PRZEZ METRYKĘ F1 WEIGHTED ZE WZGLĘDU NA DATASET, KTÓRY JEST NIEZRÓWNOWAŻONY. PRZY OBRANIU METRYKI ACCURACY ZA GŁÓWNĄ, ALGORYTM PRIORETYZOWAŁ KLASE BEZ CUKRZYCY, CO JEST ODWROTNYM DO CZEGO AI POWINNO DĄŻYĆ. STĄD, ZOSTAŁO ZASTOSOWANE F1 WEIGHTED, KTÓRE PRIORETYZUJE F1.

MODELE KLASYFIKACJI PODJĘTE PODCZAS PROJEKTU

- RANDOM FOREST CLASSIFIER (SKLEARN);
- LINEAR REGRESSION (SKLEARN);
- NEURAL NETWORK (TENSORFLOW);

RANDOM FOREST CLASSIFIER - HYPERTUNING

1. N_ESTIMATORS (100);
2. MIN_SAMPLES_SPLIT (2);
3. MIN_SAMPLES_LEAF (1);
4. MAX_DEPTH (40);
5. MAX_FEATURES (6);

HYPERTUNING ZOSTAŁ WYKONANY W KOLEJNOŚCI WYŻEJ, DLA JEDNEGO PARAMETRU NA RAZ.

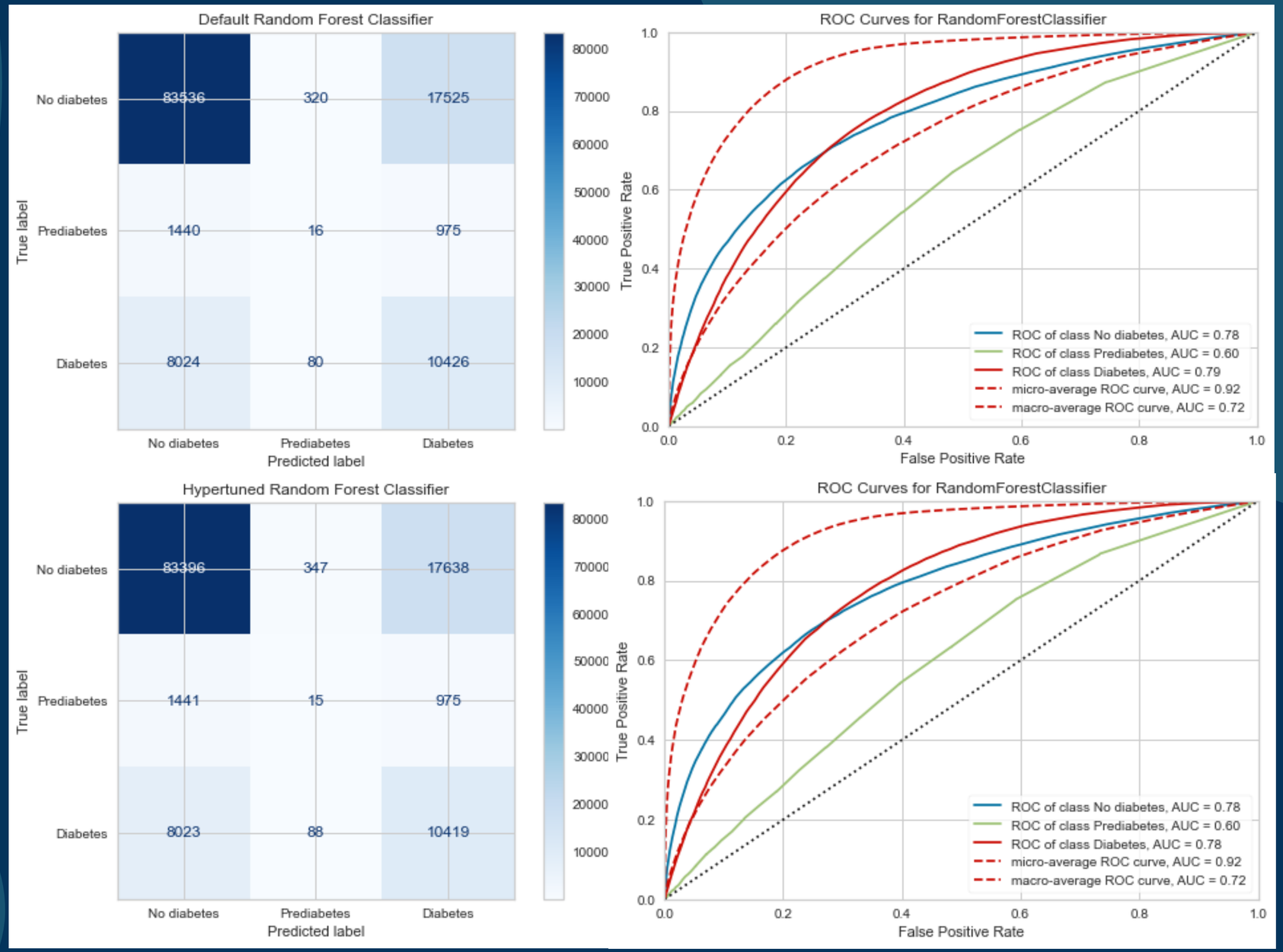
RANDOM FOREST CLASSIFIER - WYNIKI

MODEL BAZOWY

WYNIK F1:
0.779

MODEL PO HIPERPARAME-TRYZACJI

WYNIK F1:
0.7781



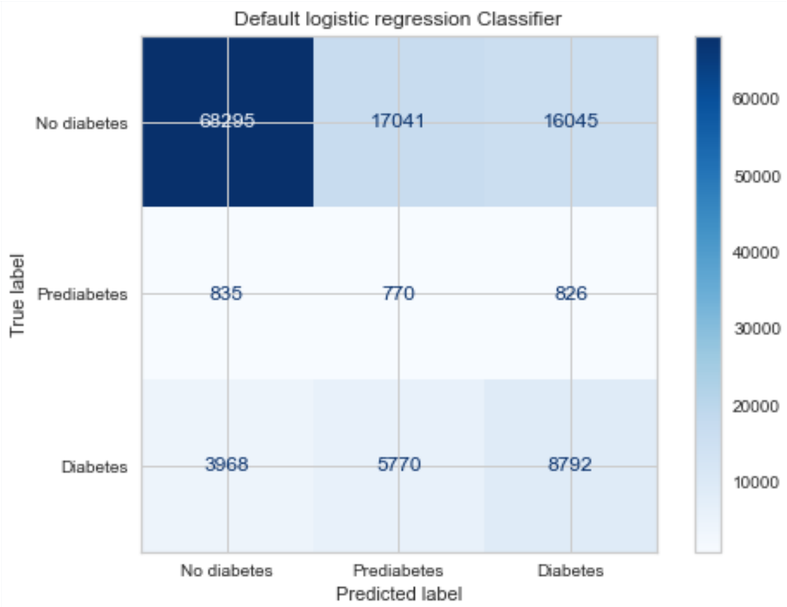
LOGICAL REGRESSION - HYPERTUNING

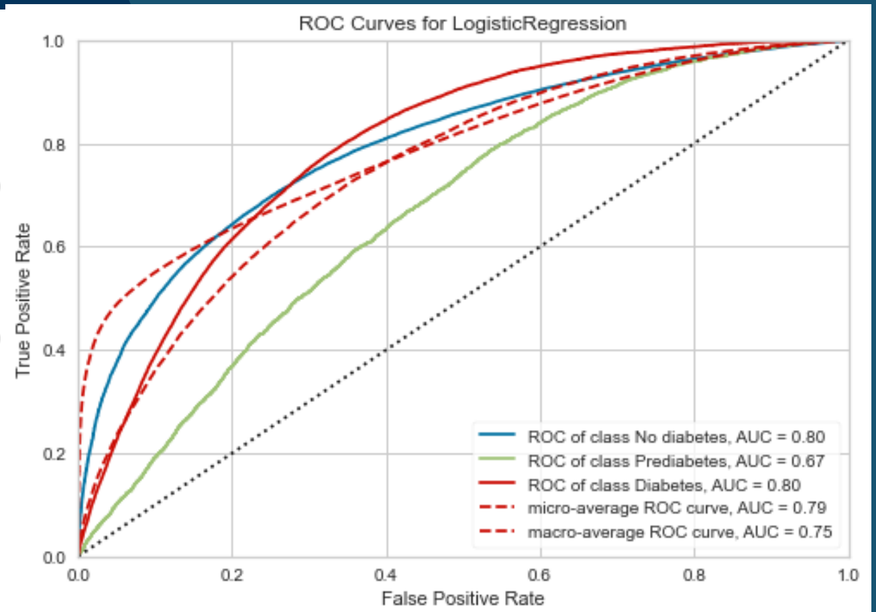
1. C (0.5);

LOGICAL REGRESSION - WYNIKI

MODEL BAZOWY

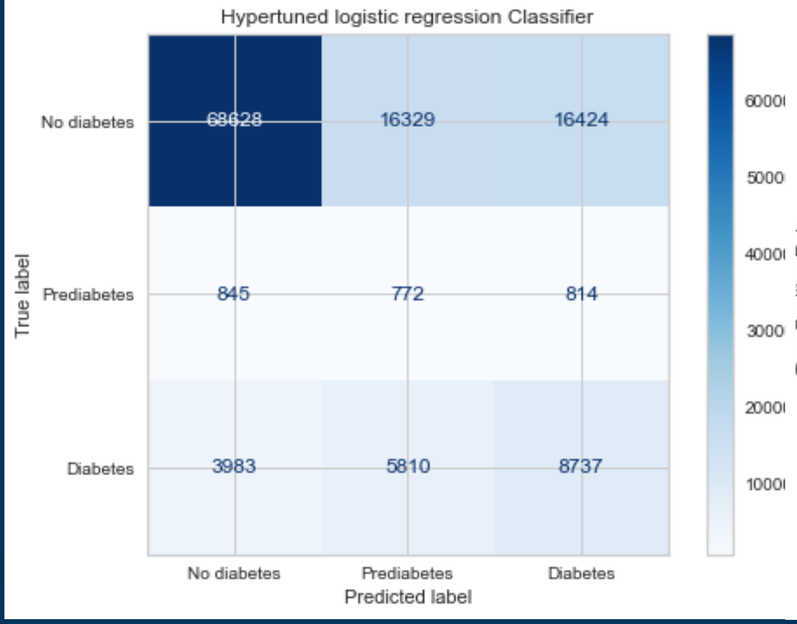
WYNIK F1:
0.7102

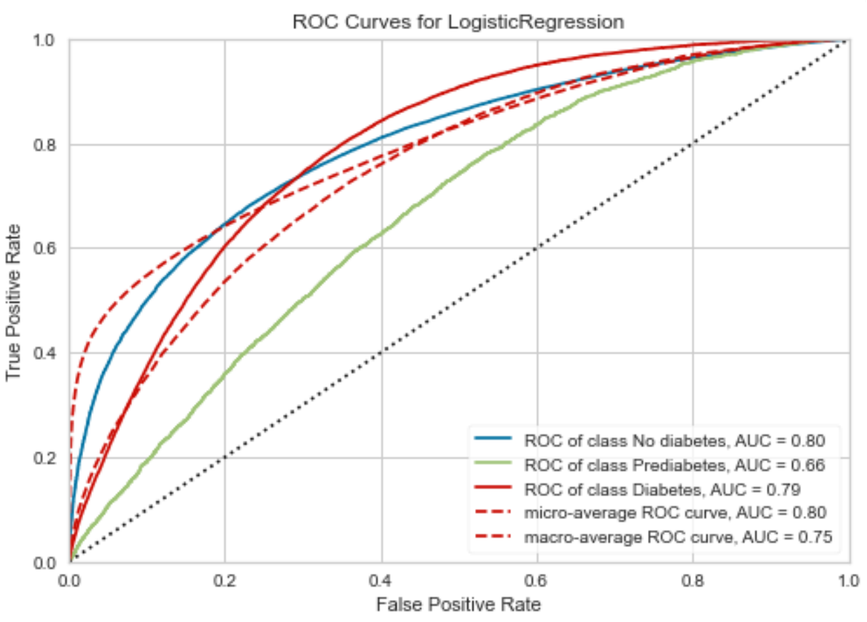




MODEL PO HIPERPARAME-TRYZACJI

WYNIK F1:
0.7112





NEURAL NETWORK - HYPERTUNING

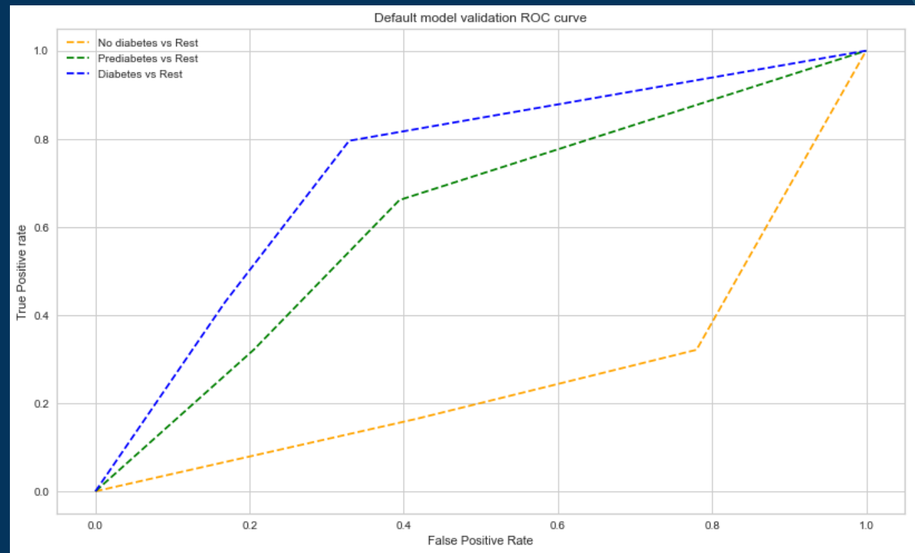
1. 1 HIDDEN LAYER'S SIZE (30);
2. 2 HIDDEN LAYERS' SIZES (1ST_HIDDEN: 60; 2ND_HIDDEN:40);

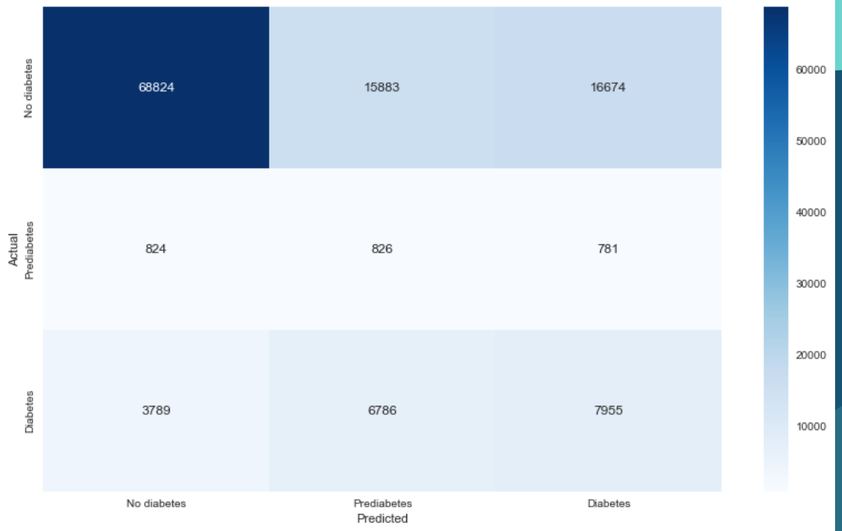
HYPERTUNING ZOSTAŁ WYKONANY W KOLEJNOŚCI WYŻEJ, DLA JEDNEGO PARAMETRU NA RAZ.

NEURAL NETWORK - WYNIKI

MODEL BAZOWY

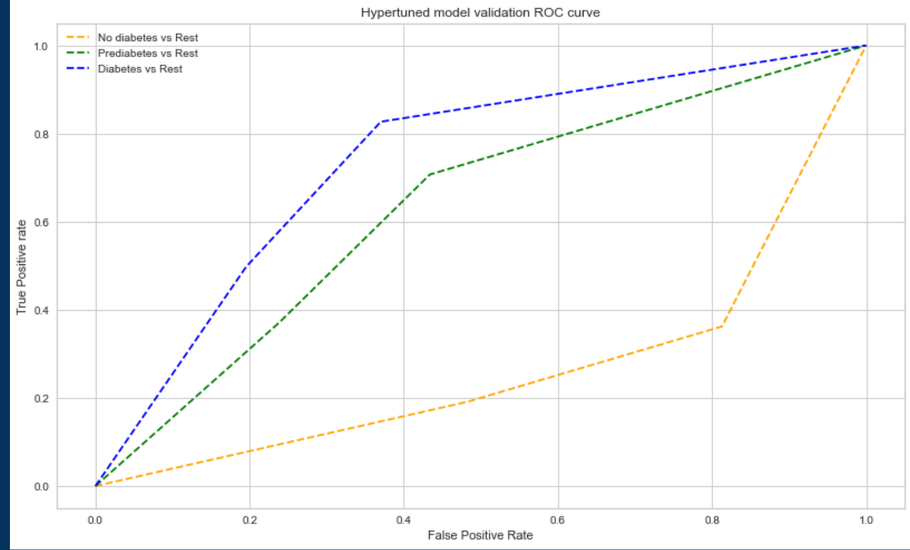
WYNIK F1: 0.7086

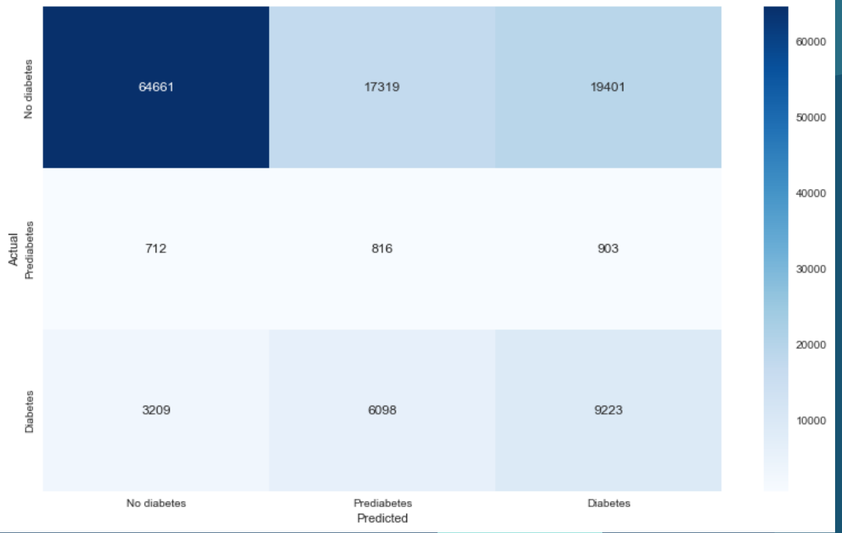




MODEL PO HIPERPARAME-TRYZACJI

WYNIK F1: 0.6899





WNIOSKI

- HYPERTUNING WYNIKÓW ZBYTNIO NIE POMÓGŁ, CO MOŻE BYĆ SPOWODOWANE RÓŻNYMI PRZYCZYNAMI, TAKIMI JAK:
  - NIEZRÓWNOWAŻONY DATASET;
  - ATRYBUTY, KTÓRE W ŻADEN SPOSÓB NIE PRZEWIDUJĄ CUKRZYCY;
- WIĘKSZOŚĆ ALGORYTMÓW MA PROBLEM Z DATASETAMI, KTÓRE SĄ NIEZRÓWNOWAŻONE. DO WALKI Z PROBLEM, ZOSTAŁY ZASTOSOWANE TECHNIKI OVERSAMPLING I UNDERSAMPLING, A WYMAZANE ZOSTAŁY PRÓBY MANIPULOWANIA CLASS_WEIGHTS, KTÓRE SĄ INNYM ROZWIĄZANIEM PROBLEMU.
- NAJLEPSZYM MODELEM SPOŚRÓD WSZYSTKICH BYŁ RANDOM_FOREST (WEDŁUG F1), KTÓRY MIAŁ WYSOKĄ WYKRYWALNOŚĆ CUKRZYKÓW MIMO NISKIEJ WYKRYWALNOŚĆ OSÓB PRZED CUKRZYCĄ;
- W PRZYSZŁOŚCI TRZEBA BY SIĘ POCHYLIĆ GŁĘBIEJ NAD DOKŁADNĄ ANALIZĄ DATASETU I SPRÓBOWAĆ RÓŻNYCH FORM JEGO AUGMENTACJI, JAK FEATURE SELECTION, FEATURE ENGINEERING, CZY DATA WRANGLING.