Czy to cukrzyk?

Maciej Pawlikowski
Mariusz Słapek
Adam Stańdo

Dane

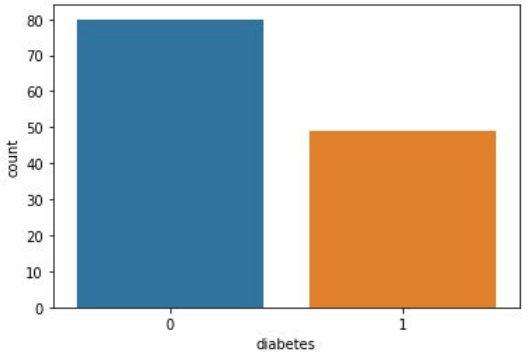
- dane pochodzą z bazy MIMIC-III-demo
- korzystaliśmy z tabel:
  D_ICD_DIAGNOSES - kody diagnoz,
  DIAGNOSES_ICD - zdiagnozowani pacjenci,
  PATIENTS - dane pacjentów,
  ADMISSIONS - dane dotyczące przyjęć,
  D_LABITEMS - rodzaje badań laboratoryjnych,
  LABEVENTS - wyniki badań laboratoryjnych

- dane “wyciągnięte” z bazy danych to:
  płeć, wiek, wyznanie, kolor skóry, rodzaj ubezpieczenia,
  ilość przyjęć nieplanowanych (emergency i urgent) oraz planowanych (elective),
  średnia wartość z wyników badań krwi; na podstawie artykułu: hemoglobina,
  kreatynina oraz glukoza wraz z liczbą tych badań i liczbą wyników oznaczonych
  jako ‘abnormal’, informacja, czy dany pacjent jest cukrzykiem

Link do artukułu: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5961793/

Preprocessing danych

- usunięcie kolumn, w których jest ponad 70% wartości NULL - wyniki laboratoryjne dt. poziomu hemoglobiny
- zastąpienie wartości NULL średnią
- pod uwagę wzięliśmy tylko średnią wartość wyników badań
- one-hot encoding dla zmiennych kategorycznych
- MinMaxScaler() do standaryzacji zmiennych



Modelowanie

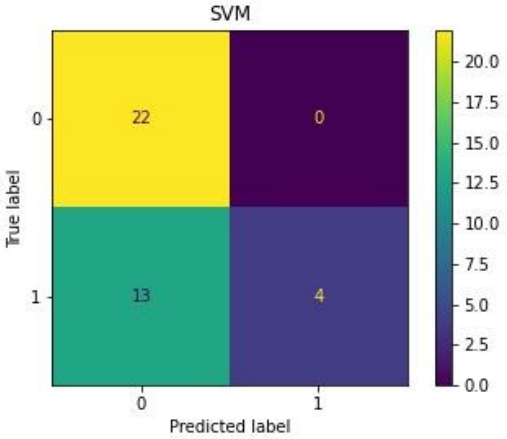
Dane podzieliliśmy w stosunku 70:30 (dane uczące i testowe).

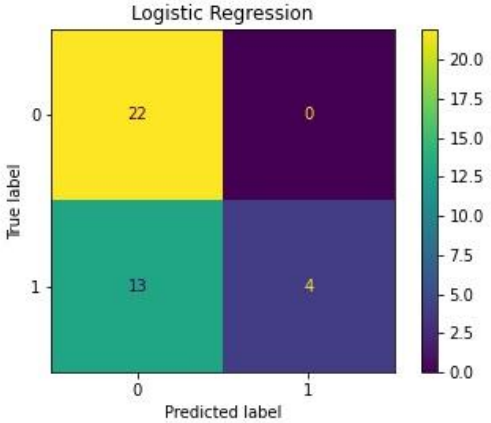
Do modelowania użyliśmy modeli:

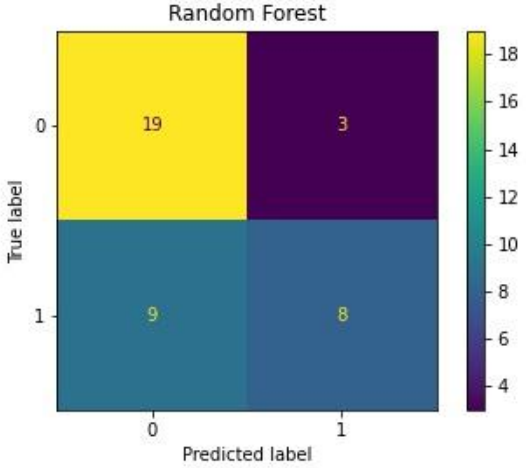
- SVM
- Random Forest
- Logistic Regression
- Decision Tree

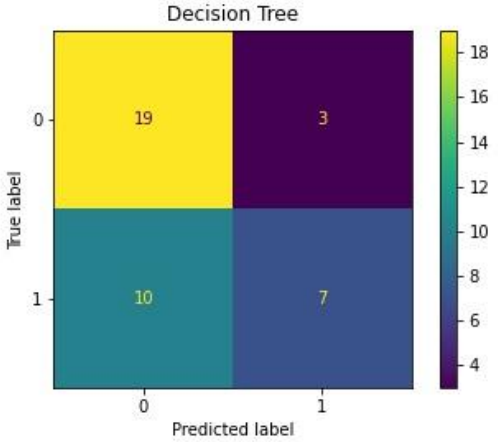
W celu znalezienia optymalnych hiperparametrów zastosowaliśmy metodę grid search.

Wyniki modeli









DODATEK
Wyniki modeli - dane z całej bazy MIMIC-III