Analiza ryzyka śmierci w przypadku Covid-19

06-DUMAUI0 2023/SZ

Cel projektu

Celem projektu było stworzenie modelu, który przewiduje, czy pacjent jest zagrożony śmiercią w przypadku Covid-19, na podstawie różnych chorób przewlekłych, wieku płci itp.

Dane

Dane pochodzą z serwisu Kaggle (https://www.kaggle.com/datasets/meirnizri/covid19-dataset). Dostarczył je Meksykański rząd.

Danych o pacjentach początkowo było 1 048 576, po ich oczyszczeniu zostało 1 025 152, które zostały podzielone na zbiory treningowy (820121 przykładów), testowy (102515 przykładów), walidcyjny(102516 przykładów).

Modele

Porównano działanie 7 modeli:

- Logistic Regression,
- Ridge Classifier, z parametrem $\lambda = 100000$,
- Gaussian Naive Bayes,
- Bernoulli Naive Bayes,
- Decision Tree,
- Random Forest, z wykorzystaniem 100 estymatorów,
- Gradient Boosting, z wykorzystaniem 100 estymatorów.

Ewaluacja

Do ewaluacji wykorzystano metryki precision, recall i Fscore. Wyniki ewaluacji przedstawia poniższa tabelka:

Model	Precission	Recall	Fscore
Logistic Regression	0.84	0.75	0.78
Ridge Classifier	0.87	0.67	0.73
Gaussian Naive Bayes	0.67	0.84	0.71
Bernoulli Naive Bayes	0.73	**0.87**	0.77

Decision Tree	0.78	0.73	0.76
Random Forest	0.84	0.75	0.77
Gradient Boosting	0.84	0.76	0.80

Wnioski

Metryką, która powinna być głównie brana pod uwagę jest recall (czułość), może być to przydatne do ostrzegania pacjentów przed ryzykiem śmierci. Zdecydowanie najlepiej poradziły sobie naiwne klasyfikatory Bayes'owskie. Najlepszy okazał się model w wersji Bernoulliego, jest to prawdpopodobnie spowodowane sporą ilość danych boolowskich w zbiorze.