Raport z projektu -Dominik Strzałko - 434788

# Cel projektu

Celem projektu było stworzenie modelu, który przewiduje, czy na podstawie danych z dnia poprzedniego, da się przewidzieć deszcz.

# Dane

Dane pochodzą z wyzwania „Rain in Australia” na platformie Kaggle.com (link: <https://www.kaggle.com/jsphyg/weather-dataset-rattle-package>).

Po odrzuceniu obserwacji odstających, uzyskano 107868 przykładów, które podzielono na zbiór uczący (86294 przykłady) i zbiór testowy (21574 przykładów).

# Modele

W projekcie porównano działanie 5 modeli:

* Regresja logistyczna
* Drzewa decyzyjne
* Las losowy (n\_estimators=100, max\_depth=4)
* KNeighborsClassifier (n\_neighbors = 2)
* Sieć neuronowa zbudowana w pytorch (Binary Cross Entropy Loss, Adam, 1000 epochs, (3,32,16,8,1))

# Ewaluacja

Do ewaluacji wykorzystano metryki *accuracy, precision, recall* i *F1-score*. Wyniki ewaluacji przedstawia poniższa tabelka:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-score |
| Regresja logistyczna | 0.8414294984703811 | 0.79 | 0.67 | 0.70 |
| Drzewa decyzyjne | 0.8328080096412348 | 0.77 | 0.66 | 0.68 |
| Las losowy | 0.842263836099008 | 0.79 | 0.67 | 0.70 |
| KNeighborsClassifier | 0.8090757393158431 | 0.71 | 0.61 | 0.63 |
| Sieć neuronowa | 0.8424955965514045 | 0.79 | 0.67 | 0.70 |

# Wnioski

Praktycznie wszystkie modele poradziły sobie równie dobrze z zadaniem. Accuarcy wacha się pomiędzy 80% a 84%. Najwyższy wynik uzyskała sieć neuronowa (accuracy: 0.8424955965514045). Głównym problemem każdego modelu jest fakt, że dane są mało zbalansowane (dni deszczowych jest 3x mniej niż bez deszczu). W przypadku upsamplingu/downsamplingu dataframe-a, uzyskamy mniejsze wyniki dla „accuracy”, lecz model będzie lepiej przewidywał wyniki dla dni deszczowych.