# Raport z projektu

06-DUMAUI0 2021/SL

Sebastian Wałęsa 478839
https://git.wmi.amu.edu.pl/s478839/uma_s478839

## Cel projektu

Celem projektu było stworzenie modelu, który przewiduje poziom wykształcenia osoby na podstawie: płci, posiadania samochodu, nieruchomości, liczby dzieci, dochodów, liczby osób w rodzinie, źródła przychodów, oraz statusu rodzinnego.

## Dane

Dane pochodzą z datasetu „Credit Card Approval Prediction" na platformie Kaggle.com (link: https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction?select=application_record.csv).

Po odrzuceniu 159 obserwacji odstających, uzyskano 438398 przykładów, które podzielono na zbiór uczący (350718 przykładów) i zbiór testowy (87680 przykładów).

## Modele

W projekcie porównano działanie 3 modeli:

- Regresja logistyczna. Zastosowano regularyzację L2.
- Klasyfikator SGD – stochastyczny spadek gradientowy. Jako optymalizatora użyto mini-batch gradient descent z wielkością batcha równą 100. Brak regularyzacji.
- Sieć neuronowa z liniową transformacją danych, optymalizatorem Adam oraz entropią krzyżową jako funkcja straty. Wykorzystano bibliotekę PyTorch.

## Ewaluacja

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

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Regresja logistyczna z regularyzacją | 0.7030 | 0.6406 | 0.7030 | 0.6269 |
| SGD | 0.6919 | 0.7409 | 0.6919 | 0.5659 |
| Sieć neuronowa | 0.6998 | 0.6370 | 0.6998 | 0.6250 |

# Wnioski

Najlepsze wyniki pod względem F1-score, Accuracy oraz Recall uzyskano dla modelu regresji logistycznej z regularyzacją. Model SGD osiągnął najwyższą precyzję, jednak pozostałe metryki obliczone dla niego osiągnęły najniższe wartości. Wyniki dla sieci neuronowej były bardzo zbliżone do regresji logistycznej, jednak różniły się one o około 0,5 punktu procentowego. Uzyskane wyniki wskazują, że poruszany problem nie jest łatwy do rozwiązania, a przewidywana cecha może nie być mocno skorelowana z cechami wykorzystanymi do procesu uczenia.