# Cel projektu

Celem projektu było stworzenie modelu, który przewiduje rodzaj drzewa rosnącego na danym obszarze na podstawie otaczających go cech.

# Dane

Dane pochodzą z Forest Cover Type Dataset znajdującego się na platformie kaggle

https://www.kaggle.com/datasets/uciml/forest-cover-type-dataset?select=covtype.csv

Na początku część danych była zakodowanych w formacie one-hot, a dokładnie dane o Wilderness\_Area oraz Soil\_Type, po przekształceniu ich do danych kategorycznych otrzymujemy 581012 przykładów z czego każdy przykład ma 12 cech oraz wartość przewidywaną Cover\_Type. Po podzieleniu danych na zbiór uczących oraz testowy: zbiór uczący – 464809 przykładów, zbiór testowy 116203 przykładów (20%). Dane nie zawierają żadnych wartości null oraz wszystkie mają typ int64, co ułatwia prace z nimi. Wartości jednak różnią się znacznie wielkością między sobą, więc aby polepszyć działanie modeli standaryzuje dane używając StandardScaler oraz w przypadku sieci neuronowej zmieniam ich typ na float32.

# Modele

W projekcie porównano działanie 3 modeli:

* Regresja logistyczna wielomianowa 7. stopnia. Jako optymalizatora użyto lbfgs. Zastosowano regularyzację L2.
* Klasyfikator K najbliższych sąsiadów, najbardziej optymalna wartość k po analizie wyniosła 3.
* Sieć neuronowa ANN, składająca się z 10 warst, wykorzystałem warstwę Dropout, aby zapobiegać overfittowaniu, warstwe BatchNorm1d, która znajduje się między każdą warstwą ukrytą, aby normalizować dane płynące z jednej warstwy do drugiej. Jako funckja aktywacji została użyta funckja ReLU.

# 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 | 71% | 0.70 | 0.47 | 0.49 |
| KNN | 97% | 0.95 | 0.93 | 0.94 |
| ANN | 85.13% | - | - | - |

# Wnioski

Najlepsze wyniki w każdej kategorii osiąga algorytm KNN, można wysunąć wnioski, że jest on najlepszy dla tego problemu, ponieważ drzewa podobnego rodzaju w lesie najczęściej sąsiadują ze sobą. Sieć ANN osiągnęła by prawdopodobnie lepszy wynik niż 85%, po dodaniu większej ilośći neuronów w każdej z warstw oraz po pogłębieniu jej, lecz generalnie nie jest ona najlepsza metoda przy danych tabelarycznych.