1. Przygotowanie i standaryzacja Danych

Dane zawierające ludzie i ssacze sekwencje białkowe o długości 27 aminokwasów zawierające miejsca karbonylacji zostały pobrane i podzielone na 4 zestawy train_pos, train_neg, test_pos, test_neg. Cechy zaproponowane przez autorów (9 cech) zostały wystandaryzowane. Sekwencje białkowe zostały przepisane na wektory o wymiarze 243(9*27) i w takiej postaci zostały użyte do dalszej część analizy.

2. AdaBoost

AdaBoost to algorytm uczenia maszynowego, który jest używany do rozwiązywania problemów z klasyfikacją i regresją. Jest to metaalgorytm, który służy do poprawy wydajności innych algorytmów uczenia maszynowego. AdaBoost działa poprzez łączenie wielu słabych klasyfikatorów, znanych również jako klasyfikatory podstawowe, w celu utworzenia silnego klasyfikatora. Algorytm iteracyjnie poprawia wydajność klasyfikatora, przywiązując większą wagę do błędnie sklasyfikowanych przykładów, dzięki czemu następny klasyfikator w zespole skupia się bardziej na trudnych przykładach. Ze względu na większą wagę błędnie sklasyfikowanych przykładów algorytm nauczył się rozpoznawać miejsca karbonylacji mimo 243 cech, natomiast algorytm Random Forest nie zadziałał w tym przypadku.

3. PCA

Służy do analizowania i wizualizacji wzorców w danych wielowymiarowych poprzez zmniejszenie wymiarowości danych przy jednoczesnym zachowaniu jak największej wariancji. Przy zredukowaniu cech z 243 do 100 poziom wyjaśnianej wariancji wynosi 86%.

4. SVM

Algorytm SVM szczególnie przydatny w sytuacjach, gdy dane są zaszumione, wielowymiarowe lub nieliniowo separowalne. W tym przypadku algorytm SVM dał wynik accuracy 0,84 natomiast F1_score = 0 może to być spowodowane tym że dane są niezbilansowane.

5. Element neuronowy

Stworzono prostą sieć neuronową z dwoma warstwami: warstwę wejściową z 32 neuronami i warstwę wyjściową z 1 neuronem. Funkcją aktywacji dla warstwy wejściowej jest „relu”, a funkcją aktywacji dla warstwy wyjściowej jest „sigmoid”, dobrze nadaje się do klasyfikacji binarnej. Stosowanym optymalizatorem jest „adam”, który jest ogólnie dobrym wyborem w przypadku wielu problemów, a loss function jest „binary_crossentropy”, powrzechnie stosowana w klasyfikacji binarnej. Model jest trenowany przy użyciu metody dopasowania na danych treningowych, gdzie X_train to cechy wejściowe, a Y_train to odpowiednie etykiety (0 lub 1) Na koniec model jest oceniany na podstawie danych testowych. Acurracy na poziomi 81,64% nieco lepiej niż Klasyfikator AdaBoost który dał acc = 81,15%