

# ZPOiF - Konspekt Projektu

Paulina Przybyłek  
Jakub Wiśniewski  
Dawid Przybyliński

9 grudnia 2019

## 1 Cel projektu i jego krótka charakterystyka

Projekt ma na celu stworzenie implementacji *lasu losowego (Random forest)* w Javie, działającego na zasadzie biblioteki, który będzie możliwy do wykorzystania na ogólnych danych.

Realizowany jest przez grupę trzyosobową a wyniki postępu prac umieszczane są w repozytorium na Githubie. Wybrany środowiskiem jest *IntelliJ IDEA*. Czas na jego wykonanie kończy się w dniu 27.01.2020. Sposób realizacji projektu jest zgodny z wytycznymi podanymi przez prowadzącego przedmiot a podział obowiązków wewnątrz grupy będzie ustalany dynamicznie, zależnie od funkcji, które akurat będą do zaimplementowania.

Zakładamy, że każdy członek grupy będzie wystarczająco dobrze zaznajomiony z całym projektem, aby mógł on poprawiać błędy i implementować nowe funkcjonalności na własną rękę.

## 2 Zadania realizowane w ramach projektu

Poniżej wymieniono funkcje, które zamierzone są w projekcie. Podział jest rozłożony na 10 punktów, które w przybliżeniu powinny stanowić po 10% pracy. Jednak te założenia są ustalane przed tworzeniem tych funkcji, więc możliwym jest, że niektóre z nich mogą być bardziej lub mniej obszerne niż zakładano.

### *Lista zadań:*

1. Klasy do wprowadzania i odczytywania danych do zastosowania przez algorytm.

Zakładamy, że dane są podawane przez użytkownika w pliku .csv. Przerabiamy je na ramki danych - na kształt listy list.

2. Znalezienie odpowiednich danych do wykorzystywania przy tworzeniu algorytmu. Podział zbioru na trenujący i testowy.

3. Indeks Giniego/Entropii - algorytmy klasyfikacji.
4. Klasa Tree (prawdopodobnie na zasadzie klasy wydmuszki z uzupełnianiem jej powoli innymi klasami).
5. Struktura drzewa. Można założyć, iż jest to kontynuacja powyższego punktu.

Dwa powyższe punkty dotyczą tworzenia podstaw drzew decyzyjnych - przed rozpoczęciem pracy nad tymi klasami i metodami nie jesteśmy świadomi ile może to pochłąć czasu. W tym momencie sugerujemy, że jest to około 20% do 30%.

6. Dokonywanie podziału zmiennych.
7. Metoda tworzenia predycji i budowania drzewa.
8. Testowanie drzew. Poprawianie i ulepszanie algorytmu.

W tym momencie powinniśmy zakończyć tworzenie drzew decyzyjnych, dlatego poświęcimy więcej czasu na testowanie algorytmu i udoskonalanie go zanim przejdziemy do budowania lasów losowych.

9. Struktura lasu losowego. Bagging.
10. Testowanie i analiza wyników. Ostatnie szlify nad biblioteką.

Dodatkowo algorytm będzie na bieżąco testowany, aby wyłapać błędy i móc je poprawić przed przejściem do kolejnego punktu.

**UWAGA:** Wydaje nam się, że projekt może pochłonąć zbyt dużo czasu i pracy nad dopracowaniem wszystkiego - dodatkowo uczymy się pojęcia drzew decyzyjnych, dlatego projekt nie zawiera w zakładanych zadaniach stworzenia intersejsu graficznego.