RAPORT

**Imie i Nazwisko:** Adam liwecki

**Projekt:** Indiana Ford

**Projekt Indywidualny:** Drzewa Decyzyjne

Drzewa dezycyjne wykorzystywane są do uczenia maszynowego. Dając komputerowi Dając mu ''bagaż'' doświadczeń, czyli informację o przypadkach określonego typu, chcielibyśmy, aby podejmował racjonalną decyzję. Racjonalną znaczy najbliższą indukowanym regułom, które w jego mniemaniu zostały użyte do podjęcia tych dobrych decyzji i jednocześnie o niskim przewidywalnym poziomie pomyłek.

Algorytm budowy drzewa jaki użyłem jest rekurencyjny algorytm ID3

**Idea algorytmu :**

1. Drzewo zaczyna od pojedynczego węzła reprezentującego cały zbiór treningowy.
2. Jeżeli wszystkie przykłady należą do jednej klasy decyzyjnej, to zbadany węzeł staje się liściem i jest on etykietowany tą decyzją.
3. W przeciwnym przypadku algorytm wykorzystuje miarę entropii (funkcja przyrostu informacji) jako heurystyki do wyboru atrybutu, który najlepiej dzieli zbiór przykładów treningowych.
4. Dla każdego wyniku testu tworzy się jedno odgałęzienie i przykłady treningowe są odpowiednio rozdzielone do nowych węzłów (poddrzew).
5. Algorytm działa dalej w rekurencyjny sposób dla zbiorów przykładów przydzielonych do poddrzew.
6. Algorytm kończy się, gdy kryterium stopu jest spełnione.

**Kryteria stopu:**

1. Wszystkie przykłady przydzielone do danego węzła należą do jednej klasy decyzyjnej, lub
2. Nie istnieje atrybut, który może dalej podzielić zbiór przykładów. W tym przypadku, liść jest etykietowany nie jedną wartością decyzji, lecz wektorem wartości zwanym *rozkładem decyzji*, lub
3. Wszystkie liście mają założoną z góry przewagę jednej klasy decyzyjnej (np. w żadnym nie ma mniej, niż 1% obiektów z innych klas, niż dominująca).

# **Cechy algorytmu ID3**

#### **Zalety**

* prostota
* jeżeli w Training Set nie ma tzw. hałasu (ang. noise), czyli że nie ma rekordów które dla tych samych wartości atrybutów mają przypisaną różną kategorię, to ID3 daje poprawny wynik dla wszystkich rekordów z Training Set.

#### **Wady**

* nie radzi sobie z ciągłymi dziedzinami atrybutów (zakłada, że wartośći atrybutów są dyskretne)
* zakłada sztywno, że wszystkie rekordy w Training Set są wypełnione tzn. nie działa, jeżeli choć jeden rekord zawiera niepełne dane.
* duży rozmiar drzewa
* brak odporności na zjawisko zwane *overfitting*. Polega to na tym, że algorytm nie radzi sobie z danymi zaburzającymi ogólną ich informację. Może to w rezultacie prowadzić do wysokiego współczynnika błędów na danych testowych.