# Języki programowania Python i R

## Raport z modułu II

## Justyna Mackoś

1. **Tytuł**

Przewidywanie wyniku gry Dota2 na podstawie wybranych na początku rozgrywki bohaterów.

1. **Motywacja, cele**

Pomimo, że nie grałam konkretnie w tę grę, której dotyczą dane (Dota2), to miałam okazję spróbować innych gier typu MOBA (Multiplayer Online Battle Arena) i jestem ciekawa, czy sam wybór bohaterów ma duży wpływ na wynik rozgrywki. Pewnym jest, że niektóre postacie lepiej współpracują z innymi, co może się przełożyć na większą szansę wygranej. Dodatkowo podczas wyboru bohaterów tworzy się tak zwane „kontry”, czyli wybiera postać, której umiejętności (np. czary) blokują przeciwnika, bądź nie mogą być zablokowane przez gracza drużyny przeciwnej – co daje przewagę. Oczywiście sam wynik rozgrywki zależy przede wszystkim od umiejętności graczy, ale mam nadzieję, że na podstawie analizy przebiegu gier uda się z pewnym prawdopodobieństwem określić która drużyna wygra, nie znając pozostałych zmiennych.

1. **Opis danych**

Dane zostały pozyskane od:

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.   
Dokładny link do bazy: <https://archive.ics.uci.edu/ml/datasets/Dota2+Games+Results>

Łącznie w bazie znajduje się 102942 rekordów. Wszystkie zostały pozyskane w trakcie 2-godzinnego zapisu (rozgrywki odbywały się równolegle) w dniu 13.08.2016.

Dane zostały podzielone przez udostępniającego na dwa pliki – zbiór uczący i testowy w stosunku 90% – 10%. Jeśli używając tego podziału wyniki nie wyjdą obiecująco, będę je dzielić samodzielnie.

Każdy rekord składa się z 117 atrybutów oznaczających kolejno:

* 1. Kto wygrał grę - wartość 1 lub -1 oznaczającą drużynę.
  2. Lokalizacja gry – liczba całkowita oznaczająca region gry np. Europa wschodnia, Chiny – każdy region w zależności od wielkości może się składać z kilku podregionów   
     (np. Rosja to numery od 181 do 188).
  3. Mapa gry – liczba całkowita
  4. Typ tworzenia drużyn – liczba całkowita (np. gra samemu z losowym doborem drużyny, albo z drużyną dobraną przed rozpoczęciem rozgrywki)
  5. 113 atrybutów opisujących kolejnych bohaterów oznaczonych jako
* 1 gdy dana postać została wybrana przez pierwszą drużynę
* -1 gdy postać należała do drużyny przeciwnej
* 0 gdy bohater nie został wybrany do tego meczu

Pliki (nie mojego autorstwa) opisujące jacy bohaterowie/regiony/tryby kryją się pod ich ID.

Plik tłumaczący tryby rozgrywki:  
<https://github.com/kronusme/dota2-api/blob/master/data/mods.json>

Plik łączący ID bohatera z jego imieniem:   
<https://github.com/kronusme/dota2-api/blob/master/data/heroes.json>

Plik opisujący regiony:

<https://github.com/kronusme/dota2-api/blob/master/data/regions.json>

1. **Opis procesu przygotowywania danych do analizy**

Po sprawdzeniu w Pythonie wszystkie dane były już typu liczb całkowitych int64, więc nie była wymagana zmiana na zmienną kategoryczną.

Kolumny nie miały swoich nazw, więc przyjęły pewne automatyczne nazwy – próby wpisania ręcznie do pliku .csv w Excelu powodowały błędy i w rezultacie powstawała tylko jedna kolumna. Próby wpisania czegokolwiek do csv użytego na zajęciach i ponownego zapisania również powodowały zmianę odczytywania kolumn z działających wcześniej kilku, na jeden długi string. Problem udało się rozwiązać wpisując nazwy kolumn do pliku .csv używając notatnika, zamiast Excela.

Do analizy nie powinny być brane pod uwagę dane w których rozgrywka była inna niż 5 graczy na 5 graczy, oraz te, w których pojawiają się nieznane wartości. Zbadałam, że:

* W trybie tworzenia drużyn nie pojawia się wartość 8, czyli gra 1vs1
* W trybie tworzenia drużyn nie pojawia się wartość -1, czyli „Invalid”
* W wyborze mapy nie pojawia się wartość 0 czyli „Unknown”
* Przy wyborze mapy nie pojawia się wartość 21, czyli mapa 1vs1

1. **Analiza danych - przyjęte założenia, krótki opis metod i obranej metodologii analizy**

Podczas analizy danych odkryłam, że jeden z bohaterów o numerze ID 108 nie został ani razu wybrany podczas wszystkich gier. Taka sytuacja jest bardzo mało prawdopodobna, ponieważ w zbiorze testowym jest ok. 90 tys. rekordów, a każda gra polega na wyborze 10 (różnych) bohaterów. Dlatego też było to zaskakujące, że podczas wybierania prawie milion razy ze zbioru zaledwie 113 postaci jakaś postać została całkowicie ominięta. Gdyby tą analizę przeprowadzić wtedy, kiedy te dane zostały pobrane, można by ten fakt zgłosić do twórców gry, zwracając uwagę, że jedna z ich postaci prawdopodobnie nie spełnia wymagań użytkowników (możliwe, że użytkownicy doświadczalnie zauważyli, że wybierając tego bohatera częściej przegrywają i dlatego tego nie robili).

Po rozwinięciu listy okazało się, że jeszcze jeden bohater (o ID 24) również nie został wybrany ani razu.

Zdecydowałam się połączyć zbiory podzielone przez udostępniającego, żeby móc sprawdzić czy te wartości (i inne) nie pojawiają się też w zbiorze testowym.

Podane wcześniej ID dwóch bohaterów również nie pojawiły się w drugim zbiorze, więc nie będzie to generowało możliwych problemów.

1. Modelowanie danych - przyjęte założenia, krótki opis metod i obranej metodologii budowania modeli

Wykorzystałam metodę k-najbliższych sąsiadów. Tak naprawdę zagadnienie, które jest tutaj badane jest zależne od wielu czynników, które nie są tutaj zawarte – choćby ilość zagranych przez gracza meczy jako ta postać mogłaby przybliżyć wynik. Nie spodziewam się modelu o wysokiej skuteczności, jednak mam nadzieję, że będzie on wyższy niż 50% (w innym przypadku równie dobrze można by rzucić kością.

1. Rezultaty, wnioski i ich dyskusja

II) Moduł 2 (dr inż. Patryk Jasik) - Raport z analizy i modelowania wybranego zbioru danych (22 punkty)

Raport, wraz ze wszystkimi kodami, należy umieścić w wybranym repozytorium (np. GitLab, GitHub).

"Wynik","Region","Mapa","Typ\_gry","1","2","3","4","5","6","7","8","9","10","11","12","13","14","15","16","17","18","19","20","21","22","23","24","25","26","27","28","29","30","31","32","33","34","35","36","37","38","39","40","41","42","43","44","45","46","47","48","49","50","51","52","53","54","55","56","57","58","59","60","61","62","63","64","65","66","67",68","69","70","71","72","73","74","75","76","77","78","79","80","81","82","83","84","85","86","87","88","89","90","91","92","93","94","95","96","97","98","99","100","101","102","103","104","105","106,"107","108","109","110","111","112","113"