VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

http://fit.vut.cz

Projekt - ZZN 2019/2020

Tenisové zápasy

Jakub Svoboda – xsvobo0z@stud.fit.vut.cz
Vadym Hladyuk – xhlady01@stud.fit.vut.cz

1. Popis datasetu

Následující sekce popisuje všechna data vyskytující se v datasetu a možné hodnoty kterých mohou nabývat. Samotný dataset se skládá z osmi csv souborů, z celkem čtyřech turnajů. Jeden řádek datasetu popisuje jeden zápas. V případě, že údaj je obsažen dvakrát s jiným číslem, tak to znamená, že je to údaj pro hráče číslo 1 nebo pro hráče číslo 2. Vysvětlení tenisových pojmů naleznete zde[1] a význam zkratek čerpán zde[2].

- PLAYER 1 – jméno a příjmení hráče číslo 1 (`str`)
- PLAYER 2 – jméno a příjmení hráče číslo 2 (`str`)
- ROUND – kolo turnaje, v jakém se zápas odehrál (`int`)
- RESULT – referuje výhru nebo prohru hráče číslo 1, v případě výhry 1, v případě prohry 0 (`bool`)
- FNL.1/FNL.2 – finální počet vyhraných setů (`int`)
- FSP.1/FSP.2 – procentuální podíl prvních podání (`int`)
- FSW.1/FSW.2 – počet vyhraných prvních podání (`int`)
- SSP.1/SSP.2 – procentuální podíl druhých podání (`int`)
- SSW.1/SSW.2 – počet vyhraných druhých podání (`int`)
- ACE.1/ACE.2 – počet podaných es (`int`)
- DBF.1/DBF.2 – počet zahraných dvojchyb (`int`)
- WNR.1/WNR.2 – počet vítězných úderů (`int`)
- UFE.1/UFE.2 – počet nevynucených chyb (`int`)
- BPC.1/BPC.2 – počet vytvořených break-pointů (`int`)
- BPW.1/BPW.2 – počet vyhraných break-pointů (`int`)
- NPA.1/NPA.2 – počet pokusů získat net-pointů (`int`)
- NPW.1/NPW.2 – počet vyhraných net-pointů (`int`)
- TPW.1/TPW.2 – celkový počet vyhraných bodů (`int`)
- ST1.1/ST1.2 – počet vyhraných gamů v setu číslo 1 (`int` nebo `NA` v případě, že set neodehrál)
- ST2.1/ST1.2 – počet vyhraných gamů v setu číslo 2 (`int` nebo `NA` v případě, že set neodehrál)
- ST3.1/ST1.2 – počet vyhraných gamů v setu číslo 3 (`int` nebo `NA` v případě, že set neodehrál)
- ST4.1/ST1.2 – počet vyhraných gamů v setu číslo 4 (`int` nebo `NA` v případě, že set neodehrál)
- ST5.1/ST1.2 – počet vyhraných gamů v setu číslo 5 (`int` nebo `NA` v případě, že set neodehrál)

2. Formulace úlohy

Samotný projekt se bude skládat ze dvou podpříkladů. V první úloze bude naším cílem vytvořit prediktivní model, schopný predikovat výsledek pátého setu (u zápasů žen třetího setu) dle dat z předchozích setů daného zápasu.

Druhá úloha se bude zabývat shlukováním. Cílem bude aplikovat shlukovací algoritmy na data popisující podání jednotlivých hráčů. Očekávaný výsledek by měl hráče rozřadit do shluků dle agresivity prvního podání, chybovosti, počtu zahraných es a podobně.

V obou úlohách plánujeme použít více algoritmů a nalézt nejpřesnější model pro naši úlohu.

[1] https://en.wikipedia.org/wiki/Glossary_of_tennis_terms

[2] https://archive.ics.uci.edu/ml/datasets/Tennis+Major+Tournament+Match+Statistics