baseball dataset

Matematički fakultet Univerziteta u Beogradu

Relja Pešić

August 2023

Abstract

blablabla sumirano sta radimo

Contents

Chapter 1

Uvod

Ovaj rad je predvidjen za demonstriranje rada brojnih tehnika istraživanja podataka. Baza podataka koja se koristi sadrži podatke o bejzbol igračima i njihovim statistikama koje su postigli.

1.1 Analiza skupa podataka

Ukupan broj instanci je 1340 od kojih njih 20 ima nedostsajuće vrednosti. Podaci su opisani narednim atributima:

- 'Games_played': podatak o broju utakcima koje je igrač odigrao
- 'At_bats':
- 'Runs'
- 'Hits'
- 'Doubles'
- 'Triples'
- 'Home_runs'
- 'RBIs'
- 'Walks'
- 'Strikeouts'
- 'Batting_average'
- 'On_base_pct'

- 'Slugging_pct'
- 'Fielding_ave'

Iz skupa podataka su uklonjeni karakteri navodnika, apostrofa i obnutih kosih crta, a razmaci su zamenjeni karakterom donja crta. Napomenuto je da je kolona 'Player' identifikator koji je potrebno ignorisati prilikom kreiranja modela.

1.2 Identifikacija elemenata van granica

1.3 Rad sa nedostajućim vrenostima

1.4 Pretprocesiranje

1.4.1 Priprema za klasifikaciju

1.4.2 Priprema za klasterovanje

1.4.3 Priprema za pravila pridruživanja

Chapter 2

Klasifikacija

2.1 DTC

2.2 Naive Bayes

2.3 KNN

Chapter 3

Klasterovanje

3.1 Algoritam K-sredina

3.2 Hijerarhijsko klasterovanje

3.2.1 Vizuelizacija podataka

Chapter 4

Pravila pridruživanja

4.1 Apriori algoritam

Chapter 5

Zakljucak