

Predicting the Success of Baseball Pitchers

Yilmaz Machine Learning 1 - 2022-2023

Wilson Chen and Bradley Cao

1 Introduction

1.1 Overview

For our project, we decided to look at a dataset on baseball pitchers. This dataset with 547 instances and 31 attributes contained information on different statistics recorded for pitchers. We wanted to see if we could use these statistics to predict whether or not a pitcher had a winning record.

1.2 Data Set

We got our original dataset from REPLACE. It contained 31 attributes:

- last_name
- first_name
- player_id
- year
- p_game
- p_total_hits
- p_home_run
- p_strikeout
- p_walk
- p_k_percent
- p_bb_percent
- batting_avg
- slg_percent
- on_base_percent
- on_base_plus_slg
- p_earned_run
- p_win
- p_loss
- p_era
- p_rbi

- `p_called_strike`
- `p_unearned_run`
- `exit_velocity_avg`
- `launch_angle_avg`
- `sweet_spot_percent`
- `barrel_batted_rate`
- `hard_hit_percent`
- `meatball_percent`
- `pitch_hand`
- `n_fastball_formatted`
- `fastball_avg_speed`
- `n_offspeed_formatted`

These attributes were collected by MLB or MLBStatCast, a partner company that specializes in collecting data from MLB games.

2 Preprocessing

2.1 Missing and Redundant Values

We first started by removing redundant values, such as OPS. The OPS attribute is just the sum of the On Base Percentage (**OBP**) and Slugging (**SLG**), so we can safely remove it without losing any information. Furthermore, there were missing values for the attributes Offspeed% and Offspeed Average MPH. We decided to remove these attributes altogether, as these values are missing because not all pitchers have an offspeed pitch; filling this data with an average would not be representative, and removing these attributes helps with dimensionality reduction. We also removed all attributes that did not relate to pitcher's performance (**name**, **year**, **player ID**, etc.).

2.2 Dimensionality Reduction

We used WEKA's **WrapperSubsetEval** technique with search method **BestFit** to perform further dimensionality reduction. As shown in the screenshot below, the attributes that were kept were **p_home_run** (number of home runs given up), **p_strikeout** (number of strikeouts), **p_walk** (number of walks), **batting_avg** (batting average of opposing batters), **slg_percent** (slugging percent of opposing batters), and **p_era** (earned run average).

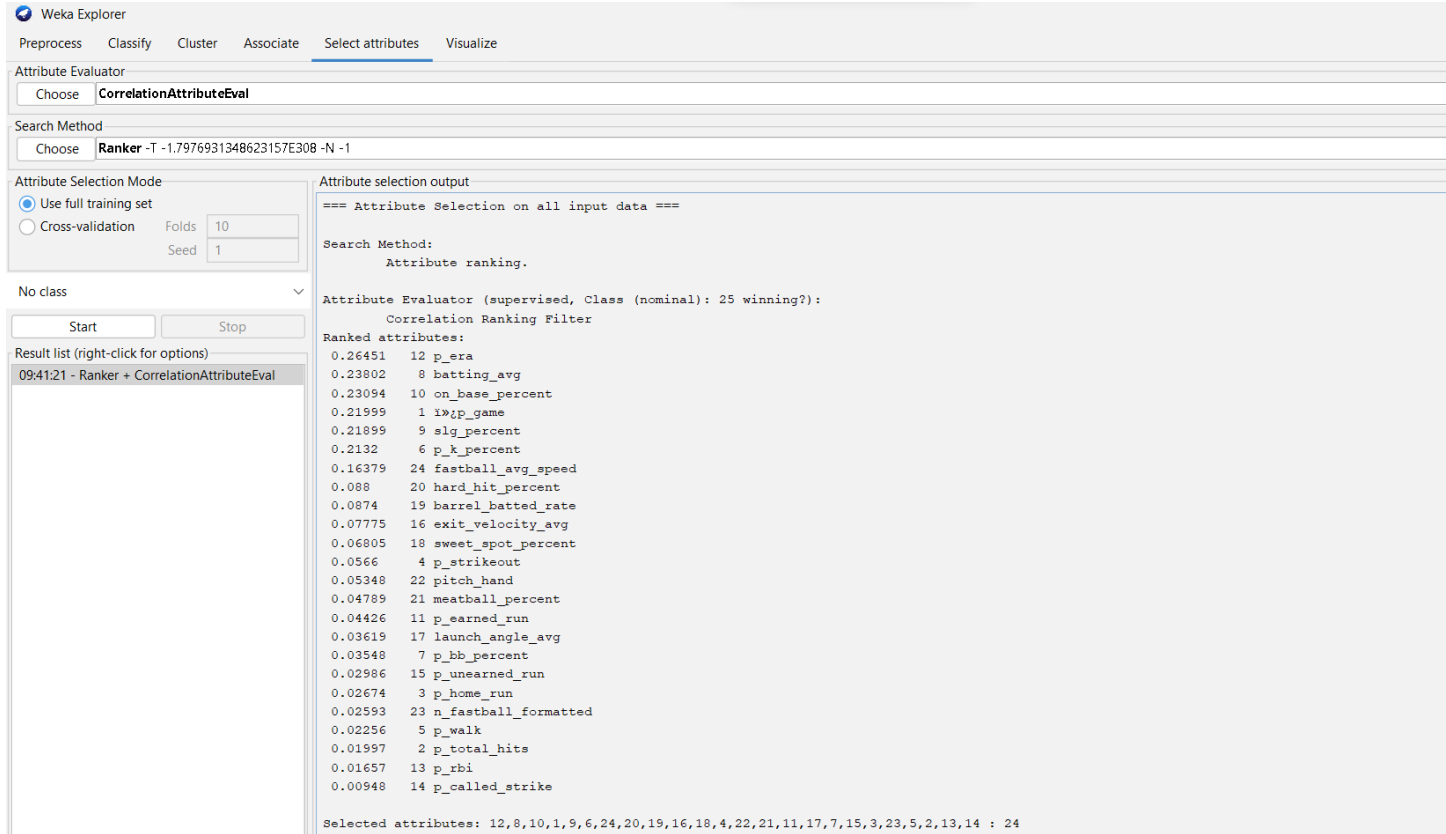


Figure 1: WEKA screenshot showing Dimensionality Reduction results

3 Splitting into Testing and Training

after reducing dimensionality to these 7 attributes, we used weka to split our data into training and testing sets. we used weka's supervised method **resample**, with sample size percent of 33% and **noreplacement** set to **true**. we ran it twice, once with **invertselection** set to **false** to obtain our testing set. The second time, we ran it with **invertSelection** set to **True** to obtain our training set.

The class distribution of the testing and training sets can be seen in Figures 2 and 3. As seen by the class distributions, the WEKA **Resample** method has maintained the non-uniform class distribution in the training and testing sets. This means that the

4 Classification

5 Discussion

6 Conclusion

7 Sources

We got our data from <https://www.mlb.com> and used WEKA to preprocess and do classification work.