# PREDICTION OF SUCCESSFULNESS OF YOUNG TENNIS PLAYERS

[1,2]Andrej Panjan, [3]Aleš Filipcic, [4]Nejc Šarabon,

(1) Wise Technologies Ltd., Laboratory for Advanced Measurement Technologies, Ljubljana, Slovenia
(2) Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, Koper, Slovenia
(3) Faculty of Sport, University of Ljubljana, Ljubljana, Slovenia
(4) University of Primorska, Science and Research Centre, Institute for Kinesiological Research, Koper, Slovenia

## Introduction

The purpose of this study was to assess the possibilities for predicting playing successfulness of young tennis players in competitive tennis, thereby, using machine learning methods applied to the players` motor abilities and morphological tests results. Machine learning is an artificial intelligence field which deals with discovering knowledge in data by data analysis and by the automatic generation of knowledge databases for expert systems, for the construction of numeric and qualitative models using classification and regression analyses, etc. In recent years, we have witnessed a rapid increase in the volume of data in digital form. Machine learning is becoming an important tool for transforming these data into useful information, since manual processing of such a vast quantity of data has become impossible.

Predictions of competitive performance can be made using either classification or regression methods. Both approaches have in common that out of a multitude of independent variables they can construct a model, whose output is an observed variable [1]. It is possible to improve the reliability of individual methods by selecting only the most promising attributes [2, 3].

The aim of this study was to assess the possibilities for making predictions of performance in competitive tennis by employing classification machine learning methods on the basis of motor and morphological tests of young tennis players. The efficiency of predicting competitive performance was studied with regard to two age groups (under 12 years and between 12 and 16 years) in male and female categories. Our final aim was to identify those attributes which proved as the most useful in making predictions. For this purpose we used two methods for attribute selection. Additionally, we were interested in correlation between all tests (attributes) with each other, and between all tests and ranking (class), that might also be useful for selection of the most promising attributes.

## Methods

- 511 Slovene male tennis players and 372 female tennis players, i.e. 883 individual tennis players in total.
- Subjects were divided into age categories:
  - **U12/U12:** 170 male tennis players (age 12.14±1.02 years, body height 153.53±7.95 cm, body weight 42.85±7.58 kg) and 157 female tennis players (11.85±.75 years, 155.74±8.16 cm, 44.02±8.33 kg).
  - **12-16/ 12-16:** 341 male tennis players (14.88±1.20 years, 170.35±10.08 cm, 58.43±11.49 kg) and 215 female tennis players (14.80±1.19 years, 166.65±6.18 cm, 55.93±7.46 kg).
- A test battery contained tests of anthropometrical characteristics and motor abilities, whose usefulness in predicting competitive performance in tennis was already identified.
- Several machine learning methods were used to predict competitive performance.
  - Two machine learning methods for the identification of the most promising predictors were included.
  - Test of the suitability of correlation analysis for selection of the most promising predictors.
- The possibility of automatically identifying the most promising attributes was tested using both the ReliefF method and the wrapper approach.
- Correlation analysis was carried out between all tests and between all attributes and class. Pearson's correlation coefficient was used.

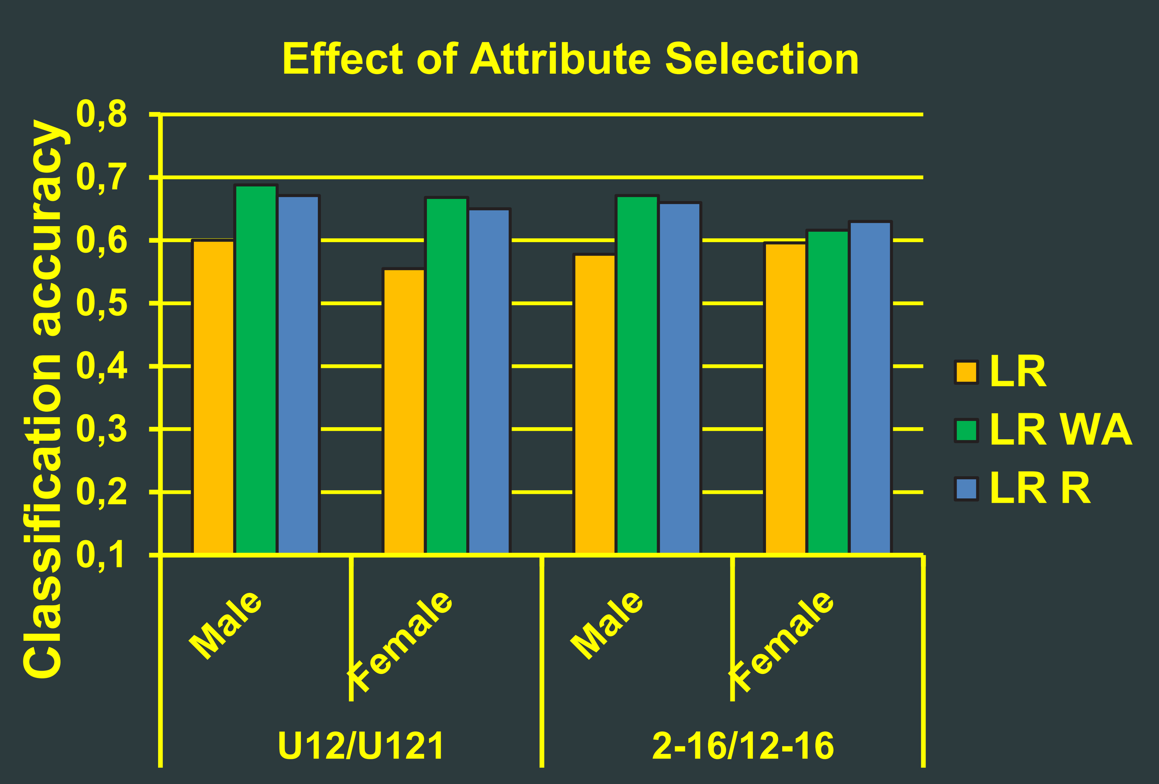


Figure 1. Effect of attribute selection for the logistic regression method.

## References

1. Hand, D. J., Cambridge: The MIT Press, 2001.
2. Kohavi, R., Artificial Intelligence, 97(1-2), 273-324, 1997.
3. Kononenko, I., Chichester: Horwood Publishing Ltd., 2007.
4. Caruana, R., Pittsburgh, Pennsylvania: ACM, 2006.

Acknowledgements

*Operation part financed by the European Union, European Social Fund. Operation implemented in the framework of the Operational Programme for Human Resources Development for the Period 2007-2013, Priority axis 1: Promoting entrepreneurship and adaptability, Main type of activity 1.1.: Experts and researchers for competitive enterprises.*

## Results

- Both the naive Bayes method with ReliefF and logistic regression with the wrapper approach, proved to be the most accurate in predicting competitive performance in U12/U12 and in 12-16/12-16.
- Comparing the prediction in U12/U12 and in 12-16/12-16 regarding gender there were no larger deviations, except for female in U12/U12 for naive Bayes with ReliefF the classification accuracy was 0.73 (Table 1). The naive Bayes with ReliefF was on average better method then the logistic regression with wrapper approach, but results of both of them differed by 0.06 at most.
- In U12/U12 and in 12-16/12-16, the wrapper approach on average improved accuracy by 0.08, whereas the ReliefF method improved accuracy by 0.07. All other methods improved in accuracy by ≤ 0.04.
- C-3 and A-2 were the most commonly selected attributes by means of the logistic regression method in combination with the wrapper approach for predicting competitive performance in U12/U12 and in 12-16/12-16, whereas the ReliefF method selected C-3 and A-1 most commonly.
- Regarding morphological dimensions high positive correlation coefficient (r ≥ 0.8) was observed between M-3 and M-2; and high negative correlation coefficients (r ≤ -0.8) were observed between M-2 and M-6, as well as between M-3 and M-6.
- Only one correlation coefficient, between A-1 and A-2, regarding motor tests stood out with r ≥ 0.55 for all groups except for 12-16 females. Among correlation coefficients between attributes and class C-3 correlation coefficient was the highest (r around 0.40 for all groups), while all other correlation coefficients were low (r ≤ 0.35 and r ≥ -0.35).

| | U12/U12 | | 12-16/12-16 | |
|---|---|---|---|---|
| **Method** | **Male** | **Female** | **Male** | **Female** |
| **NB** | 0.67 | 0.66 | 0.67 | 0.62 |
| **NB WA** | 0.62 | 0.70 | 0.68 | 0.61 |
| **NB R** | 0.67 | 0.73 | 0.66 | 0.67 |
| **DT** | 0.59 | 0.62 | 0.65 | 0.56 |
| **DT WA** | 0.63 | 0.65 | 0.51 | 0.62 |
| **DT R** | 0.66 | 0.63 | 0.53 | 0.66 |
| **C4.5** | 0.58 | 0.62 | 0.60 | 0.58 |
| **C4.5 WA** | 0.66 | 0.63 | 0.59 | 0.64 |
| **C4.5 R** | 0.65 | 0.61 | 0.60 | 0.66 |
| **kNN** | 0.64 | 0.58 | 0.55 | 0.61 |
| **kNN WA** | 0.53 | 0.63 | 0.62 | 0.55 |
| **kNN R** | 0.62 | 0.56 | 0.55 | 0.63 |
| **SVM** | 0.57 | 0.64 | 0.61 | 0.55 |
| **SVM WA** | 0.65 | 0.61 | 0.55 | 0.54 |
| **SVM R** | 0.60 | 0.59 | 0.60 | 0.58 |
| **LR** | 0.60 | 0.56 | 0.58 | 0.60 |
| **LR WA** | 0.69 | 0.67 | 0.67 | 0.62 |
| **LR R** | 0.67 | 0.65 | 0.66 | 0.63 |

Table 1. Classification accuracies of all models for studied samples of tennis players. (NB – naive Bayes, DT – decision tree, C4.5 – C4.5 algorithm, kNN – k-nearest neighbor, SVM – support vector machine, LR – logistic regression, WA – wrapper approach, R - ReliefF).

## Discussion

Regarding classification analysis, a classification accuracy value above 0.60 was considered a satisfactorily accurate result. With a well selected sample, classification accuracy around 0.50 can be achieved with a random classifier [3] alone, which makes models with classification accuracy under 0.60 unsuitable for the issues in question. The most accurate classification models are the naive Bayes method with ReliefF and the logistic regression with the wrapper approach. Their classification accuracies differ by 0.06 at most, which is a negligibly small difference. These two methods are therefore equally suitable for predicting competitive performance of the young tennis players. While the SVM method is usually considered to be one of the most reliable ones when it comes to complex actual issues [3, 4]. A conclusion to be drawn from this is that the machine learning methods tested (the naive Bayes and the logistic regression) was the most suitable for predicting competitive performance in U12/U12 and 12-16/12-16 for males and females. However, they can be hardly considered as suitable for the use in practice. We speculate that for practical purposes classification accuracy should be above 0.80. In this study, predictions of competitive performance of tennis players turned out to be a highly complex issue, as the accuracy of predicted models, based on morphological and motor factors, was relatively poor for practical use. Reasons for this lie in the fact that the competitive performance was predicted only on the basis of estimates of potential performance in the fields of morphological, motor, and functional dimensions, and in doing so the players' personality traits, mental and competitive abilities, technical and tactical competencies, and experience were not taken into account.