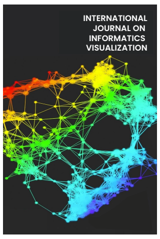
INTERNATIONAL JOURNAL
ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv



# A Comparative Study of Feature Selection Technique for Predicting the Professional Tennis Matches Outcome in a Grand Slam Tournament

Nur Amira Sariaty Ruslan [a], Zuraini Zainol [b,*], Ummul Fahri Abdul Rauf [c]

[a] *Department of Defence Science, Universiti Pertahanan Nasional Malaysia, Sungai Besi Camp 57000 Kuala Lumpur, Malaysia*
[b] *Department of Computer Science, Universiti Pertahanan Nasional Malaysia, Sungai Besi Camp 57000 Kuala Lumpur, Malaysia*
[c] *Centre for Defence Foundation Studies, Universiti Pertahanan Nasional Malaysia, Sungai Besi Camp 57000 Kuala Lumpur, Malaysia*
*Corresponding author: [*]zuraini@upnm.edu.my*

***Abstract*— Tennis is one of the world's most played sports, attracting many spectators to participate in the game. One of the most essential strokes in a tennis match is serve performance. This research is intended to determine the most critical strokes in tennis serve performance in predicting the tennis match outcome. This research focuses on the Grand Slam Tournaments of the Australian Open, French Open, Wimbledon, and US Open. The data are collected on the tennis serve performances such as Percentage First Serve In (PFSI), Percentage First Serve Won (PFSW), Percentage First Serve Return Won (PFSRW), Aces, and many more. For one tournament, it consists of 254 observations. This study applied feature selection methods available in R programming, such as Correlation Matrix, Relative Importance Metrics, Boruta, MARS, and cForest. Selecting the most essential and correlated variables with the match status can improve the model and help produce better results. This might help the practitioners to apply this method to obtain the closest result to the actual outcome when we include the most correlated variables in the model. From the result obtained, variables of first and second serve, either win on serve or return serve, are identified as the most critical attributes in the tennis match. As a future implication, we suggest that these are all the factors the players need to pay extra attention to in winning the tennis match.**

***Keywords*— Serve; correlation matrix; Boruta; MARS; feature selection.**

*Manuscript received 5 Oct. 2023; revised 31 Jan. 2024; accepted 23 Feb. 2024. Date of publication 31 Mar. 2024.*

## I. INTRODUCTION

Many well-known problems revolve around dependency analysis, which is the central task in statistics. While the relationship between contingency tables and independence tasks is obvious, other techniques, such as regression and variable selection, can also be viewed as dependence issues. Qualitative sports analysis is one of the most attractive and growing research areas. Tennis is the fifth most popular sport worldwide and has become one of the most played sports in recent years [1]. Tennis is a highly technical sport, and any error in its player's strength training can significantly impact the competition [2]. Every year, four Grand Slam tournaments are held across the world by the official tennis association. The tennis tournaments are the Australian Open (hosted in Melbourne, Australia, in January), the French Open (held in Paris, France, in June), Wimbledon (held in Wimbledon, United Kingdom, in July) and the U.S. Open (held in New York, United States, in late August). Tennis has become significantly more dynamic, powerful, and fast-paced in the last decade. The data on the website made it much easier for researchers to collect it for analysis. With the advancement of technology, the amount of data available on the performance of tennis players on the court has increased over the years, which can be of extreme significance for analyzing statistical parameters and sports performance [3].

In the tennis dataset, numerous variables may influence the outcome of the tennis match. For instance, either player wins or loses the match. [4] justified tennis serve performance is vital because high-quality serves enhance the likelihood of winning points by shortening the opponent's time to return accurately. The variables that may influence the outcome are 'Percentage first serve in,' 'Percentage first serve won,' 'Percentage second serve won,' 'Percentage first serve return won,' 'Percentage second serve return won,' 'Aces,' 'Double faults' and many more. Among all these variables, it is crucial to determine the variables that contribute most to the outcome of the match.

To develop an efficient model for tennis sports data, researchers must select the most essential attributes for constructing the model. This is useful when we have a dataset with multiple attributes and must decide which ones are most important [5]. Feature selection is one of the most critical steps before model fitting. It is used to avoid overfitting. The model will not target irrelevant features when building the model. Furthermore, feature selection may enhance model accuracy by removing irrelevant features [6]. Since only the relevant features are selected in the model, the runtime can be reduced by decreasing the dimension of the data and running the models more quickly [7].

Data Mining (DM) is the process of discovering exploitable patterns in complex data using sophisticated analytical and computer techniques [8], [9], [10]. With little or no human involvement, exploratory data analysis uses computationally feasible methodologies, such as searching for unknown exciting structures. In many circumstances, the DM process provides valuable information and insights. Examples of the application of DM are education [11], [12], insurance and healthcare [13], [14], [15], finance and banking sector [16], [17], social media analytics [18], [19] and others. Various DM techniques have been applied in predicting tennis match outcomes [20], [21], [22], [23], [24], [25]. Among these DM techniques, classification is often used as the outcome in the form of a binary response. Classification uses the training set to predict categorical class labels and classify data, which involves a two-step procedure [26].

The paper justified classification is used to predict the labels of categorical classes and assign labels to newly available data. Methods of classification can handle both numerical and categorical attributes. For example, a study by [27] applied the classification methods in their analysis of tennis stroke classification. Developing fast and accurate classifiers for big data sets is a critical challenge in mining and knowledge discovery. The primary goal of a classification algorithm is to maximize the predicted accuracy of the classification model. When the label/class variables are discrete or categorical, the classification technique is identified as the most popular and effective DM method for classifying data in the prediction model [28]. Tennis match outcomes can be predicted using a variety of methods under the classification technique, including Logistic Regression (LR), Naïve Bayes (NB), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Random Forest (RF) [22], [24], [29], [30]. Work by [23] applied three various techniques such as LR, RF, and SVM, in determining serves performance using different indicators such as Percentage of Ace over Double Faults, Previous Percentage of games owned, Number of championships, Percentage of games won in the same tournament, Games won before in the same round and many more. Based on their study, they found that RF is the best model with an accuracy greater than 80%. Another study by [31] compared different classifiers for predicting tennis match results. The decision tree is the best model compared to the learning vector quantization (LVQ) and SVM. However, none of the feature selection techniques were used before creating the models. The LR model has been applied in [20] to predict the result of Novak Djokovic's matches in the Australian Open tennis match based on the predictor listed in the model. A previous study by [29] conducted the feature selection to win the points of ATP tennis players using rally information. They deleted the feature with a zero variance, indicating they have the same value.

Work in [21] stated that training and fine-tuning the nonlinear SVM, LR, and ANN models was challenging for the local computers due to the amount of the dataset (approximately 45000 rows with 80 features). They stated that they were able to train these models with their dataset. However, iterating over the hyperparameters of the models proved difficult. If they had more computing power available, they could improve the issue of hyperparameters. At this point, selecting features can assist in overcoming this issue.

Through the previous articles search, we found that most researchers tend to include all the variables in the model formulation, such as [20], [25], [31], [32]. Reducing the number of variables in the model may lower the computation cost and accelerate model construction. Additionally, feature selection promotes generalization, which reduces model overfitting. Many variables in the model with little or no predictive value are often noisy. The data mining models learn from this noise, reducing generalization and causing overfitting. We can significantly increase generalization and decrease overfitting by eliminating this noise. Reducing the number of variables also lowers the possibility of data-gathering errors. By deleting the highly correlated characteristic and keeping only the relevant features, variable redundancy can be reduced without losing crucial data [33].

Most of the prior research on different areas of interest has proven that feature selection before model construction can aid in facilitating the selection of more relevant and correlated features to minimize the overfitting of the model with excessively irrelevant features [5], [6], [34], [35], [36]. Hence, we are applying feature selection techniques on this tennis match status to determine which factors are most significant in winning the tennis match, and we are using only the essential variables for our model formulation rather than incorporating all variables.

This paper explores several approaches for evaluating the most critical attributes in the tennis sports database. In any statistical analysis, the existence of factors (features) that do not contribute significantly to identifying the dependent variable (DV) while creating an effective prediction model can be a severe issue. The primary goal of this study is to establish what factors (attributes) are the most accurate predictors for predicting tennis match outcomes. The following indicates how the paper is organized. Section II provides an overview of the methodology used. Section III presents the results and discussion, and Section IV gives the conclusion of the research presented.

## II. Materials and Method

Table 1 describes the variables that would be used in analyzing a tennis match. The dependent variable (DV), or a response (e.g., 'status'), describes whether a player wins or loses the match when competing against an opponent. The remaining variables are treated as independent variables (IVs) or predictors.

TABLE I
DATASET DESCRIPTION OF THE TENNIS DATASET

| Variable | Description |
|---|---|
| **Tournament** | French Open, Australian Open, Wimbledon and U.S. Open |
| **Round** | 1,2,3,4, Semifinal, Quarterfinal and Final |
| **Type** | Left-handed and Right-handed |
| **Surface** | Hard court, Grass court, and Clay court |
| **Height** | Height of player (cm) |
| **Weight** | Weight of player (kg) |
| **Age** | Age of player |
| **PFSI** | Percentage of first serve in (%) |
| **PFSW** | Percentage of first serve won (%) |
| **PSSW** | Percentage of second serve won (%) |
| **PFSRW** | Percentage of first-serve return won (%) |
| **PSSRW** | Percentage of second serve return won (%) |
| **FSS** | First serve speed (kmh) |
| **SSS** | Second serve speed (kmh) |
| **DF** | Double faults |
| **Aces** | Aces |
| **UE** | Unforced error |
| **Status** | Win or lose |

To develop a model that can predict the outcome of the tennis match, it is essential to determine which attributes are most significant. We employed a set of algorithms already implemented in the R language to achieve this. Many packages and methods are available in the R programming language. In the R language, numerous packages and methods can be used to identify the most influential factors in explaining the DV. Figure 1 depicts the workflow of this study. It consists of several steps/stages, including data collection, checking missing values, detecting outliers, applying feature selection techniques, formulating a model, comparing, and selecting the best model.

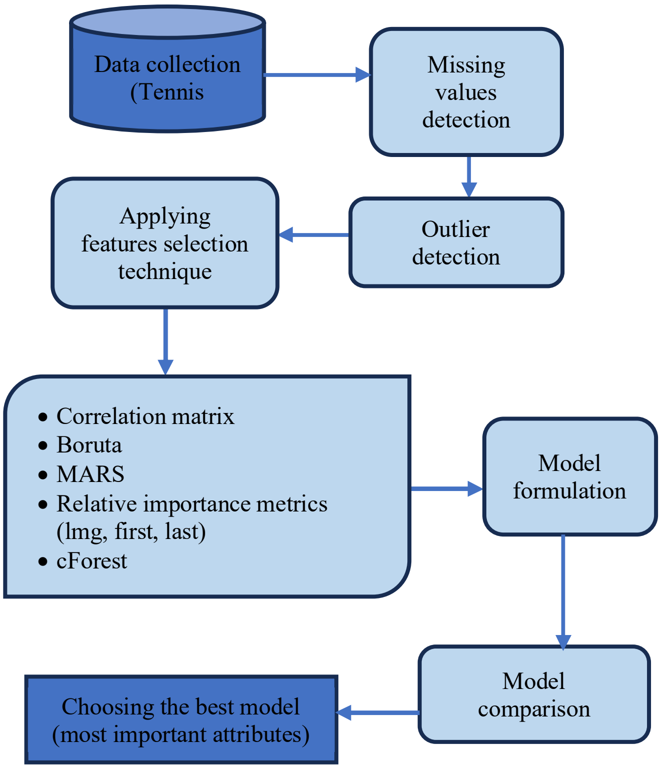


Fig. 1 Workflow of the features selection process

We extract all datasets from the official tennis website for the four Grand Slam tournaments: French Open, Wimbledon, Australian Open, and U.S. Open. Next, we conduct an initial data analysis to verify the missing values. This investigation uses the MICE imputation on R packages to manage missing values. We then move on to the detection of outliers. For outlier detection, we address the outlier rather than removing it from all server performance indicators. We decided to replace the outlier with missing values and implement the iteration of the mean using MICE imputation to replace the missing values.

After completing the preliminary data analysis, we used feature selection techniques to choose the most essential attributes. In this study, we employed the correlation matrix and relative importance metrics such as lmg, first and last, Boruta, MARS, and the cForest feature selection technique. From this point onwards, we will create a few models and conduct some model comparisons to choose the best model with the most essential attributes.

### A. *Imputation*

Almost every dataset contains missing data, which can result in severe problems such as biased estimates or decreased efficiency due to a smaller dataset. Missing data can be replaced with new values using imputation methods to alleviate these issues. [37] defined missing data imputation as a typical method for dealing with missing values in which the missing values' substitutes are identified using a statistical technique.

Multivariate imputation by chained equation (MICE) has been developed as a principled method of dealing with missing databases [38]. MICE is a popular method available in several statistical software programs, including R and SAS. The algorithm generates multiple complete datasets, each imputed with plausible values drawn from a distribution fitted to each incomplete variable [39], [40].

According to [38], the MICE procedure involves running a series of regression models in which each variable with missing data is conditionally modeled based on the other variables in the data. This means that each variable can be modeled based on its distribution; for example, binary variables can be modeled using logistic regression, and continuous variables can be modeled using linear regression. The implementation of MICE varies slightly across software packages, with some using a multinomial logit model for categorical variables and a Poisson model for count variables. In the MICE procedure, every time a value in the dataset is missing, a simple imputation is carried out, such as imputing the mean. These mean imputations may be considered "placeholders." We noticed missing values for certain variables in the tennis dataset. To overcome this problem, we have used the MICE procedure to replace the missing values in those variables.

### B. *Outliers*

After checking the boxplot for outlier detection, we discovered the presence of outliers among the variables. Instead of removing them from the model, we decided to replace them using the MICE imputation. However, the outliers must be treated as missing values before applying the imputation process [5].

We used a boxplot to visualize quantitative variables by displaying the minimum, median, first and third quartiles, maximum, and any observation appointed as an outlier based on the interquartile range (IQR) criterion [41]. According to the IQR criterion, possible outliers are those observations that fall outside of the first and third quartiles, respectively, and whose IQR is calculated as the difference between those two quartiles. Alternatively, all data outside the subsequent range will be considered potential outliers.

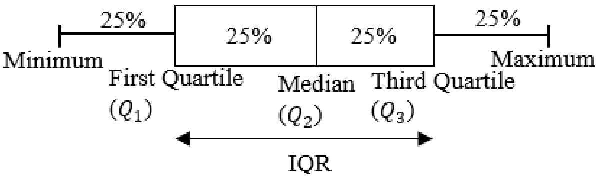


Fig. 2 Boxplot criteria

$$I = q_{0.25} - 1.5(IQR);\ q_{0.75} + 1.5(IQR) \qquad (1)$$

A boxplot, one of the graphical tools for displaying the locality, spread, and skewness group of numerical data through their quartiles, was employed in Figure 2. A plot may have lines, also called whiskers extending from the box to indicate variability outside the top and lower quartiles. Consequently, the plot is also known as the box-and-whisker plot. Minimum and maximum are the lowest and highest data points in the data set, excluding any outliers, the median is the middle value in the data set, and the first quartile ($Q_1$) is the median of the lower half of the dataset, the third quartile ($Q_3$) is the median of the upper half of the dataset and interquartile range (IQR) is the distance between upper and lower quartiles. They are using the boxplot.stats()$out function available in R, extracting the values of possible outliers based on the IQR criterion is also possible.

### C. Feature Selection

Prediction is the process of approximating the value of an unknown output variable's value based on its input variables' values. Previous research has shown that incorporating all the variables into a model can contribute to model inaccuracy, and in some instances, overfitting might occur. Hence, to avoid this problem occurring, it is important to apply the features selection technique in order to include only the variables that contribute the most in the model formulation. This ensures that only the variables with the highest correlation to the outcome are included in the model analysis.

*1) Correlation Matrix:* A correlation matrix is a table presenting correlation coefficients between the variables. The correlation between the two variables is displayed in each table cell. Data are summarized using correlation matrices, also used as inputs for advanced studies and as diagnostics for such analyses. The value of the matrix lies between -1 and 1.

*2) Using the Relative Importance R Package:* For the linear model, calc.relimp computes a number of relative importance metrics. The three different metrics employed in the model development are lmg, first, and last.

- Lmg: it converts $R^2$ into non-negative contributions that sum up to the total $R^2$ that require more computational work. The lmg measure is based on the sequential $R^2$, but it handles the dependence on orderings by averaging over orderings using simple unweighted averages.
- First, the regressors' importance is ranked based on their univariate cap R squared values, which correspond to the squared correlations between the regressors and the response. If the regressors are correlated, the sum of these individual contributions is frequently much higher than the overall cap R squared of the model when all regressors are considered.
- Last: evaluates the importance of a regressor by measuring the increase in the total $R^2$ when this regressor is included as the last one. It assigns the irisein $R^2$ to each regressor when this regressor is included as the last of the p regressors. When regressors are linked, their contributions do not add up to the total $R^2$, but are often significantly less than the overall $R^2$.

*3) Using Boruta:* The Boruta approach can select the most essential variables from a set of variables. This technique, known as Boruta in R, takes the dependent and independent variables as inputs and outputs a set of significant variables. In detail, the Boruta algorithm, which employs Random Forest by default, is an all-relevant feature selection wrapper algorithm compatible with any classification approach that produces a variable importance measure (VIM)—differentiating the importance of the original attributes with the significance that can be randomly achieved, which will be estimated using their permuted duplicates, and gradually removing irrelevant features to stabilize the test. The approach performs a top-down search for significant features.

*4) Using MARS:* The MARS technique, a component of the Earth package, uses the generalized cross-validation (GCV) statistic to calculate each predictor's contribution or variable importance score. The earth() methods in the Earth package and *evimp()* are used to construct the model and determine the significance of the various variables, respectively.

*5) Using cForest:* The package party comprises the cForest method to build the model and the varimp() function to evaluate the relative importance of the predictors used in constructing the model.

### D. Model Comparison based on Model Development

*1) Logistic Regression (LR):* LR is a supervised modeling technique commonly used to explain the outcomes when it is in definite form. This method is often used when the outcome variable is binary or dichotomous. On the other hand, the features (predictors) can be categorical or numerical values. In this paper, the outcome variable is the match status in the tennis match, which is either won or lost the tennis match. LR is used to build classification rules for a given dataset based on the information on predictors of tennis serve performance. The LR algorithm also uses a linear equation with independent predictors to explain the match status. The predicted value could range from negative infinity to positive infinity.

There are some assumptions of LR that need to be fulfilled: The DV must be binary.

- LR requires the observations to be independent of each other.

- LR requires no multicollinearity to exist in the predictors.
- LR assumes linearity of independent variables and log odds.
- LR assumes the largest sample size.

The logistic regression equation is:

$$ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (2)$$

where:

$ln\frac{p}{1-p}$ *is the DV*

$x$ *is the IV*

$\beta$ *is the coefficient*

Based on the equation stated, it could be used to calculate the probability that the competing player will win the tennis match based on the predictors included. Therefore, the output of the linear equation needs to be compressed between 0 to 1.

### E. Model Performance Comparison

Accuracy is defined as the ratio of the number of correct predictions to the total sample size. Sensitivity measures the proportion of true positives that are correctly identified, and specificity measures the proportion of true negatives that are correctly identified. Kappa is used to test the model's interrater reliability.

- TP: True positive; values of precisely predicted event values.
- TN: True negative; values of wrongly predicted values.
- FP: False positive; values of no-event that were successfully calculated.
- FN: False negative; values of no-event that were mistakenly calculated.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (4)$$

$$Specificity = \frac{TN}{TN+FP} \quad (5)$$

$$Kappa = \frac{2\ x\ (TP\ x\ TN - FN\ x\ FP)}{(TP+FP)\ x\ (FP+TN)\ x\ (TP+FN)\ x\ (FN+TN)} \quad (6)$$

## III. Results and Discussion

This section describes the results and discussion based on the analysis obtained from Section II.

### A. Descriptive Analysis

The descriptive statistics of the tennis dataset are displayed in Table 2. The table shows 12 attributes with missing values, with Weight, FSS, and SSS the highest missing attributes. We decided to use the MICE procedure in the R package to replace the missing values without omitting the variables from the list. After performing the MICE procedure, we discovered that none of those variables had any missing values. Multiple imputation steps that apply mean imputation were employed to replace all missing variables.

TABLE II
Descriptive Statistics of Tennis Dataset

| Variables | Mean | Missing value (n) |
|---|---|---|
| **Height** | 174.4 | 23 |
| **Weight** | 64.04 | 106 |
| **Age** | 26.57 | 0 |
| **PFSI** | 62.59 | 4 |
| **PFSW** | 65.02 | 4 |
| **PSSW** | 46.02 | 4 |
| **FSS** | 157.1 | 166 |
| **SSS** | 131.6 | 166 |
| **DF** | 3.505 | 3 |
| **Aces** | 2.817 | 2 |
| **PFSRW** | 35.29 | 4 |
| **PSSRW** | 50.57 | 4 |
| **UE** | 26.58 | 2 |

### B. Outlier Detection

After performing the imputation procedure, we must examine the boxplot for the presence of outliers. The existence of outliers can increase the dataset's variability, reducing the statistical power. Hence, removing or replacing the outlier with specific values can cause the result to become statistically more significant. After performing boxplots for each independent variable, we could see that most tennis serves performances contain outliers. Hence, we decided not to remove the outliers but to consider them missing values [5] and use MICE imputation to replace them. After completing this phase, we rechecked the boxplot for each variable and determined that all outliers had been addressed.

### C. Feature Selection Technique

This subsection explains all the features and essential techniques applied in this study. Figures 3 to 5 show the plot of crucial values for each variable using the Correlation matrix, Boruta, and MARS package. In Figure 3, the correlation matrix, or heatmap, represents the strength of the correlation between tennis serve performance and match status indicated by color depth. The stronger the color shades, the larger the correlation magnitude. Figure 4 shows the boxplots of all the tennis serve attributes. In determining the variables using Boruta packages, the system uses 4 different default color indicators. The green boxplot represents the confirmed attributes, the red boxplot is confirmed to be unimportant, the blue boxplot corresponds to the shadow attribute, and the yellow boxplot is tentative. Figure 5 shows the variable importance of the tennis serve using the generalized cross-validation (GCV) statistic to calculate the contribution also called as variable importance score of each predictor. The highest gcv values indicate the most correlated variables with match outcomes.

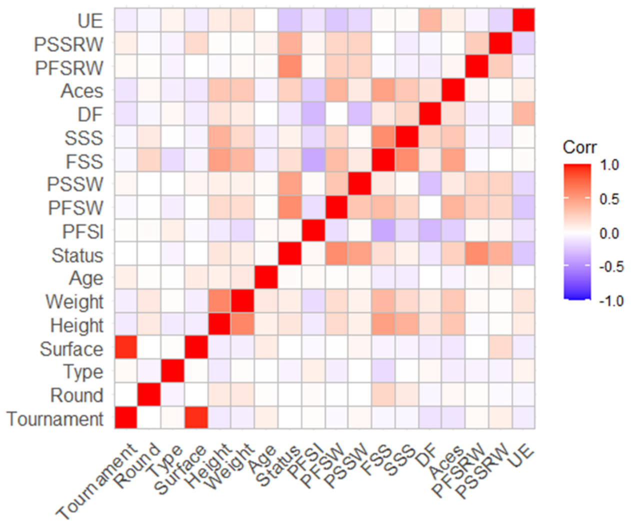


Fig. 3 Plot of correlation matrix.

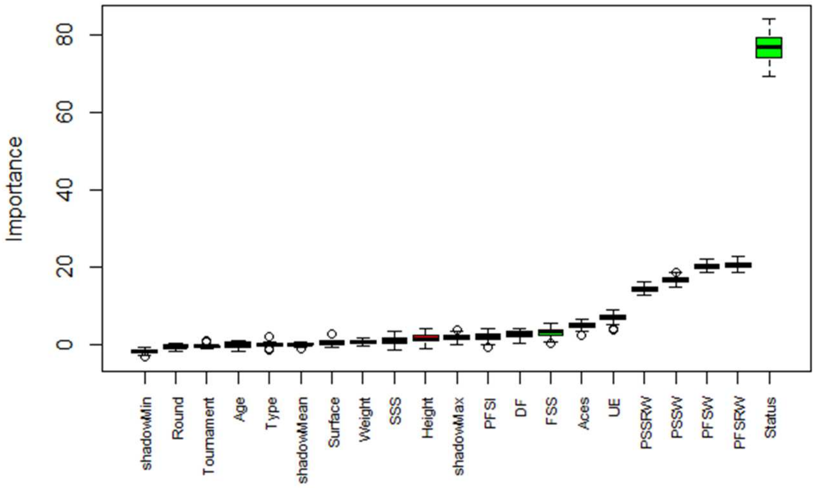


Fig. 4 Plot of variable importance using the Boruta package.

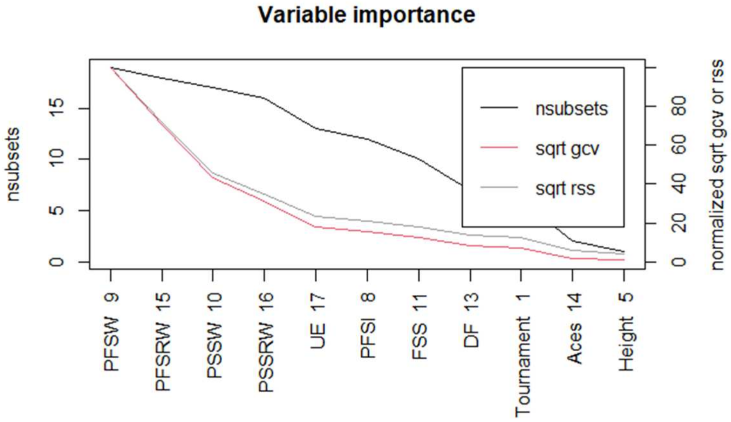


Fig. 5 Plot of variable importance using the MARS method.

Table 3 presents the summary, in descending order, of each predictor variable's contribution to the tennis match's outcome, which is either a win or a loss. We have simplified the fifth most correlated variables from the table with the match outcome. From there, we could say that the first serve correlates with winning the tennis match indicated by PFSW and PFSRW. As observed in Table 3, the four topmost importance values are similar for all the techniques used. The only difference is on the fifth variable, in which correlation matrix and relative importance (first) are given to the Aces. In contrast, relative importance (lmg), relative importance (last), Boruta, MARS and cForest stated UE as their fifth variable. We built five different models to determine the most important variables for comparison.

TABLE III
LIST OF IMPORTANCE VALUES FOR ALL FEATURE IMPORTANCE TECHNIQUES

| Technique | Variable | Importance value |
|---|---|---|
| Correlation Matrix | PFSW | 0.58295 |
| | PFSRW | 0.57896 |
| | PSSW | 0.47988 |
| | PSSRW | 0.40978 |
| | Aces | 0.24456 |
| Relative importance (lmg) | PFSRW | 0.32219 |
| | PFSW | 0.26855 |
| | PSSW | 0.16599 |
| | PSSRW | 0.12213 |
| | UE | 0.03639 |
| Relative importance (first) | PFSW | 0.26854 |
| | PFSRW | 0.26488 |
| | PSSW | 0.18198 |
| | PSSRW | 0.13270 |
| | Aces | 0.04726 |
| Relative importance (last) | PFSRW | 0.42241 |
| | PFSW | 0.25603 |
| | PSSW | 0.14560 |
| | PSSRW | 0.09325 |
| | UE | 0.01900 |
| Boruta | PFSRW | 22.57640 |
| | PFSW | 22.13580 |
| | PSSW | 18.53746 |
| | PSSRW | 15.97057 |
| | UE | 8.73340 |
| MARS | PFSW | 100.0 |
| | PFSRW | 70.7 |
| | PSSW | 43.4 |
| | PSSRW | 31.3 |
| | UE | 18.3 |
| cForest | PFSW | 0.05670 |
| | PFSRW | 0.05452 |
| | PSSW | 0.02555 |
| | PSSRW | 0.02349 |
| | UE | 0.00709 |

### *D. Model Comparison*

We develop LR models to compare the model's performance based on the previously discussed list of essential features. LR is well known to be used when the outcome is binary and is one of the methods used in classification techniques. To compare the performance of all modes, we compare AIC, Accuracy, Sensitivity, Specificity, and Kappa values. All the descriptions are stated in Tables 4 and 5, respectively.

TABLE IV
LOGISTIC EQUATION MODELS

| Model | Logistic Equation |
|---|---|
| **Model 1** | $ln\frac{p}{1-p} = -21.57748 + 0.21599PFSW + 0.21430PFSRW$ |
| **Model 2** | $ln\frac{p}{1-p} = -32.63030 + 0.25394PFSW + 0.26502PFSRW + 0.14853PSSW$ |
| **Model 3** | $ln\frac{p}{1-p} = -41.54155 + 0.27039PFSW + 0.29378PFSRW + 0.15900PSSW + 0.12284PSSRW$ |
| **Model 4** | $ln\frac{p}{1-p} = -41.02411 + 0.25548PFSW + 0.29451PFSRW + 0.16271PSSW + 0.11996PSSRW + 0.14058Aces$ |
| **Model 5** | $ln\frac{p}{1-p} = -38.25474 + 0.27882PFSW + 0.25943PFSRW + 0.16618PSSW + 0.08401PSSRW - 0.03051UE$ |

TABLE V
COMPARISON OF MODEL PERFORMANCE

| | AIC | Accuracy | Sensitivity | Specificity | Kappa |
|---|---|---|---|---|---|
| **Model 1** | 444.42 | 0.8684 | 0.8684 | 0.8684 | 0.7368 |
| **Model 2** | 319.38 | **0.9178** | **0.9276** | 0.9079 | **0.8355** |
| **Model 3** | 260.25 | **0.9178** | 0.9013 | **0.9342** | **0.8355** |
| **Model 4** | **257.52** | 0.9046 | 0.9079 | 0.9013 | 0.8092 |
| **Model 5** | 268.85 | 0.9112 | 0.8947 | 0.9276 | 0.8224 |

Comparing all the models developed, we chose Model 3 as the best model with variables: PFSW, PFSRW, PSSW, and PSSRW in the model. This is because Model 3 has the highest accuracy, specificity, and kappa values, and it has lower AIC values than Model 2. In comparison evaluation, we can see that Models 2 and 3 have the same performance indicators but adding more variables in Model 3 can lower AIC values. All of the methods used in the feature selection process have proven that these selected variables are the most critical variables, which can support strengthening our findings.

## IV. CONCLUSION

In previous work, most research done on tennis datasets did not use the features selection procedure, which can help improve the model's accuracy and reduce the tendency of researchers to produce an overfit model. Determining the attributes that are mostly related to the outcome of match status is crucial because it can help the coaches or any sports agencies to predict the status of a player in the middle of the tournament and gain early insight into the attributes or serve performance that must be improved for the next round to win the match.

In this paper, we have presented a few methods for the feature selection process by using an R package that can be utilized to select the most critical attributes in the tennis dataset before building the prediction model. Our methodology applies different techniques for feature selection to choose the most significant variables in tennis matches and use the selected variables to build different logistic models for model performance comparison.

After all the processes presented in this paper, we found out that the most essential variables to score winning in tennis matches are first serve and second serve, either win on serve or return serve indicated by PFSW, PFSRW, PSSW, and PSSRW. This implies that both serve strokes are crucial for the player to pay extra attention to if they want to score a win. However, to highlight more, the first serve is the most serve stroke that tennis players need to take extra care of to win. For future work, we would like to compare the selected attributes with another classification technique to demonstrate their consistency. Other techniques include decision tree, naïve bayes and SVM.

## References

[1] S. Das, "Top 10 Most Popular Sports in the World July 2022. Sports Browser."

[2] J. Wang and Y. Li, "Strength Training Method for Tennis Players," *Revista Brasileira de Medicina do Esporte*, vol. 29, 2023, doi:10.1590/1517-8692202329012022_0632.

[3] Z. Bilić, V. Dukarić, S. Šanjug, P. Barbaros, and D. Knjaz, "The Concurrent Validity of Mobile Application for Tracking Tennis Performance," *Applied Sciences (Switzerland)*, vol. 13, no. 10, May 2023, doi: 10.3390/app13106195.

[4] K. Jung and H. Kim, "Comparison of the Tennis Serve Performance: A Case Study of an Elite Korean Tennis Player," *International Journal of Human Movement Science*, vol. 16, no. 1, pp. 77–85, Apr. 2022, doi: 10.23949/ijhms.2022.04.16.1.6.

[5] O. El Aissaoui, Y. El Alami El Madani, L. Oughdir, A. Dakkak, and Y. El Allioui, "A Multiple Linear Regression-Based Approach to Predict Student Performance," in *Advances in Intelligent Systems and Computing*, Springer Science and Business Media Deutschland GmbH, 2020, pp. 9–23. doi: 10.1007/978-3-030-36653-7_2.

[6] S. B. Sakri and Z. Ali, "Analysis of the Dimensionality Issues in House Price Forecasting Modeling," in *Proceedings - 2022 5th International Conference of Women in Data Science at Prince Sultan University, WiDS-PSU 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 13–19. doi: 10.1109/WiDS-PSU54548.2022.00015.

[7] P. Khumprom, D. Grewell, and N. Yodo, "Deep neural network feature selection approaches for data-driven prognostic model of aircraft engines," *Aerospace*, vol. 7, no. 9, Sep. 2020, doi:10.3390/aerospace7090132.

[8] N. S. Harzevili, A. B. Belle, J. Wang, S. Wang, Z. M. Jiang, and N. Nagappan, "A Survey on Automated Software Vulnerability Detection Using Machine Learning and Deep Learning," Jun. 2023, [Online]. Available: http://arxiv.org/abs/2306.11673

[9] K.-L. Tsui, V. C. P. Chen, W. Jiang, and Y. A. Aslandogan, "Data Mining Methods and Applications," 2023. [Online]. Available: www.selectron.com

[10] Z. Zainol, M. T. H. Jaymes, and P. N. E. Nohuddin, "VisualUrText: A Text Analytics Tool for Unstructured Textual Data," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Jun. 2018. doi: 10.1088/1742-6596/1018/1/012011.

[11] M. Y. Abdul Mutalib, Z. Zainol, U. F. Abdul Rauf, and P. N. Nohuddin, "Prediction Analysis of Student Academic Performance Using MyCGPA Application," *Journal of Information System and Technology Management*, vol. 8, no. 31, pp. 80–94, Jun. 2023, doi:10.35631/JISTM.831006.

[12] A. M. Abdo, N. M. Ahmad Rasid, N. A. H. Mohd Badli, S. N. A. Sulaiman, S. Wani, and Z. Zainol, "Student's Performance Based on E-Learning Platform Behaviour using K-means Clustering," 2021.

[13] N. K. W. Chan, A. S. H. Lee, and Z. Zainol, "Predicting Employee Health Risks using Classification Ensemble Model," in *Proceedings - CAMP 2021: 2021 5th International Conference on Information Retrieval and Knowledge Management: Digital Technology for IR 4.0 and Beyond*, Institute of Electrical and Electronics Engineers Inc., Jun. 2021, pp. 52–58. doi: 10.1109/CAMP51653.2021.9498106.

[14] U. Ahmed *et al.*, "Prediction of Diabetes Empowered with Fused Machine Learning," *IEEE Access*, vol. 10, pp. 8529–8538, 2022, doi:10.1109/access.2022.3142097.

[15] A. A. Ab Zayin, Z. Zainol, P. N. E. Nohuddin, and H. Mohamed, "Peramalan Risiko Diabetes Mengunakan Aplikasi Mydiabeticrisk Article Info Abstract," *Journal of Defence Science, Engineering & Technology Journal homepage*, vol. 6, pp. 137–148, 2023, doi:10.58247/jdset-2023-0602-13.

[16] M. E. Lokanan and K. Sharma, "Fraud prediction using machine learning: The case of investment advisors in Canada," *Machine Learning with Applications*, vol. 8, p. 100269, Jun. 2022, doi:10.1016/j.mlwa.2022.100269.

[17] X. Xu, F. Xiong, and Z. An, "Using Machine Learning to Predict Corporate Fraud: Evidence Based on the GONE Framework," *Journal of Business Ethics*, Aug. 2022, doi: 10.1007/s10551-022-05120-2.

[18] Z. Zainol, P. N. E. Nohuddin, A. S.-H. Lee, N. F. Ibrahim, L. H. Yee, and K. Abd Majid, "Analysing political candidates' popularity on social media using POPularity MONitoring (POPMON)," *SEARCH Journal of Media and Communication Research*, no. GRACE 2020 Conference, pp. 39–55, 2021.

[19] A. Hassani and E. Mosconi, "Social media analytics, competitive intelligence, and dynamic capabilities in manufacturing SMEs," *Technol Forecast Soc Change*, vol. 175, Feb. 2022, doi:10.1016/j.techfore.2021.121416.

[20] P. K. Choudhary, S. Dubey, D. Brijwal, and R. Paswan, "A statistical model to predict the results of Novak Djokovic's matches in the Australian open tennis event using the binary logistic regression," *International Journal of Statistics and Applied Mathematics*, vol. 8, no. 1, pp. 17–21, Jan. 2023, doi: 10.22271/maths.2023.v8.i1a.921.

[21] A. Cornman, G. Spellman, and D. Wright, "Machine Learning for Professional Tennis Match Prediction and Betting."

[22] G. C. Domínguez, E. F. Álvarez, A. T. Córdoba, and D. G. Reina, "A comparative study of machine learning and deep learning algorithms for padel tennis shot classification," *Soft comput*, 2023, doi:10.1007/s00500-023-07874-x.

[23] Z. Gao and A. Kowalczyk, "Random forest model identifies serve strength as a key predictor of tennis match outcome," *Journal of Sports Analytics*, vol. 7, no. 4, pp. 255–262, Jul. 2021, doi:10.3233/jsa-200515.

[24] S. Solanki, V. Jakir, A. Jatav, and D. Sharma, "Prediction Of Tennis Match Using Machine Learning," *International Journal of Progressive Research In Engineering Management And Science (IJPREMS)*, vol. 02, no. 06, 2022, [Online]. Available: www.ijprems.com

[25] Vincenzo Candila and L. Palazzo, "Neural networks and betting strategies for tennis," *Risks*, vol. 8, no. 3, pp. 1–19, Sep. 2020, doi:10.3390/risks8030068.

[26] M. Sudhir, M. Gorade, A. Deo, and P. Purohit, "A Study of Some Data Mining Classification Techniques," *International Research Journal of Engineering and Technology*, 2017, [Online]. Available: www.irjet.net

[27] M. Skublewska-Paszkowska and P. Powroznik, "Temporal Pattern Attention for Multivariate Time Series of Tennis Strokes Classification," *Sensors*, vol. 23, no. 5, Mar. 2023, doi:10.3390/s23052422.

[28] F. Shahrabi Farahani, M. Alavi, M. Ghasemi, and B. Teimourpour, "Scientific Map of Papers Related to Data Mining in Civilica Database Based on Co-Word Analysis."

[29] M. Makino, T. Odaka, J. Kuroiwa, I. Suwa, and H. Shirai, "Feature Selection to Win the Point of ATP Tennis Players Using Rally Information," *Int J Comput Sci Sport*, vol. 19, no. 1, pp. 37–50, Jul. 2020, doi: 10.2478/ijcss-2020-0003.

[30] J. C. Yue, E. P. Chou, M. H. Hsieh, and L. C. Hsiao, "A study of forecasting tennis matches via the Glicko model," *PLoS One*, vol. 17, no. 4 April, Apr. 2022, doi: 10.1371/journal.pone.0266838.

[31] S. Ghosh, S. Sadhu, S. Biswas, D. Sarkar, and P. P. Sarkar, "A comparison between different classifiers for tennis match result prediction," *Malaysian Journal of Computer Science*, vol. 32, no. 2, pp. 97–111, 2019, doi: 10.22452/mjcs.vol32no2.2.

[32] E. E. Ogheneovo and P. A. Nlerum, "Iterative Dichotomizer 3 (ID3) Decision Tree: A Machine Learning Algorithm for Data Classification and Predictive Analysis," *International Journal of Advanced Engineering Research and Science*, vol. 7, no. 4, pp. 514–521, 2020, doi: 10.22161/ijaers.74.60.

[33] S. L. Nesamani, S. N. S. Rajini, I. C. Figueroa Sánchez, M. D. P. M. Figueroa, D. A. Manrique De Lara Suárez, and O. F. C. Fuentes, "Predictive Modeling for Classification Of Breast Cancer Data Set Using Feature Selection Techniques," 2021. [Online]. Available: https://orcid.org/0000-0003-

[34] A. Juárez-López, J. Hernández-Torruco, B. Hernández-Ocaña, and O. Chávez-Bosquez, "Comparison of classification algorithms using feature selection," in *2021 Mexican International Conference on Computer Science, ENC 2021*, Institute of Electrical and Electronics Engineers Inc., Aug. 2021. doi: 10.1109/ENC53357.2021.9534831.

[35] O. A. Montesinos-López *et al.*, "Do feature selection methods for selecting environmental covariables enhance genomic prediction accuracy?," *Front Genet*, vol. 14, Jul. 2023, doi:10.3389/fgene.2023.1209275.

[36] O. O. Oladimeji, A. Oladimeji, and O. Oladimeji, "Classification models for likelihood prediction of diabetes at early stage using feature selection," *Applied Computing and Informatics*, May 2021, doi:10.1108/aci-01-2021-0022.

[37] N. Z. Abidin, A. R. Ismail, and N. A. Emran, "Performance analysis of machine learning algorithms for missing value imputation," *International Journal of Advanced Computer Science and*

*Applications*, vol. 9, no. 6, pp. 442–447, 2018, doi:10.14569/ijacsa.2018.090660.

[38] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, “Multiple imputation by chained equations: What is it and how does it work?,” *Int J Methods Psychiatr Res*, vol. 20, no. 1, pp. 40–49, Mar. 2011, doi:10.1002/mpr.329.

[39] A. Khademi, “Flexible imputation of missing data 2nd edition,” *J Stat Softw*, vol. 93, pp. 1–4, 2020, doi: 10.18637/jss.v093.b01.

[40] S. Pan and S. Chen, “Empirical Comparison of Imputation Methods for Multivariate Missing Data in Public Health,” *Int J Environ Res Public Health*, vol. 20, no. 2, Jan. 2023, doi: 10.3390/ijerph20021524.

[41] T. Falahi, G. Nassreddine, and J. Younis, “Detecting Data Outliers with Machine Learning,” *Al-Salam Journal for Engineering and Technology*, pp. 152–164, May 2023, doi:10.55145/ajest.2023.02.02.018.