Predicting Relegation Clubs in Italian Serie A with Method based C4.5 Decision Tree Algorithm

F Syukmana¹, E Wahyudi¹, W Gata², H Wahono¹, N I Febianto¹, A Y Kunto³, R O Nitra¹, L Effendi¹, D D Saputra¹ and O R Sulaeman¹

¹ Master of Computer Science Postgraduate Program, STMIK Nusa Mandiri, Indonesia
² Graduate School Master’s Degree Computer Science, STMIK Nusa Mandiri, Indonesia

Abstract. The purpose of this study is to help small clubs from Italian Serie A in finding the minimum targets to avoid relegation into Serie B competition (below Serie A league). Relegation will reduce the club’s income from TV revenues and the decline of enthusiastic supporters. Based on the data from the final standings (seasons 2006 until 2018), this can be explained by the Decision Tree method using the C4.5 algorithm. The methods used in this study are data collection, data pre-processing, model proposal, model testing, and model validation. In this study, it is expected that the value of accuracy exceeds 85% to achieve a proper classification.

Keywords: C4.5 Algorithm, Decision Tree, Relegation, Prediction, Football

1. Introduction
Football is the most popular sport in the world. Football is the most multicultural sport compared to other sports[1]. Besides, football has become a unifying icon of the countries of the world with the availability of many international competitions such as the World Cup, Asian Cup, European Cup, African Cup, Confederation Cup, and others. A large number of significant football fans means considerable business potential for football clubs. Clubs with a significant number of fans spend hundreds of millions of dollars [2] to buy good players. The value of the contract for TV rights is also enormous. For professional Italian Serie A clubs, competing in the major leagues is a mean to reap big profits. Therefore, small clubs whose revenue come mostly from TV contract will try hard to avoid relegation to the league below the major leagues (Serie B).

The previous study discussed the prediction simulation of victory in a football match with the C4.5 algorithm method with a gain of 0.543. Also, in another study of the C4.5 algorithm, the sample predictions of legislative elections results in the DKI Jakarta Regional House of Representatatives (DPRD) were 97.84% [3]. Another study of predictive results in football with a different method is the Bayesian hierarchical model for the prediction of football results [4] that uses attack and defense data to predict the results. Because predicting clubs' relegation has not yet published in national and international scientific journals, it is the authors' interest to analyse the problem with the C4.5 algorithm. The purpose of this research is to find out the main factors of surviving in Italian Serie A football competition, so that small clubs can have the minimum target that must be achieved to avoid relegation from the main competition which at the end of the day will affect the club's revenues.

2. Methods

2.1. Dataset
This research uses the standings data from the Italian Serie A League from 2006 to 2018. Because the purpose of this study is to find out the parameters to predict which Serie A clubs to be relegated, so that data utilized in the study is obtained from the club finished in 11-20 position.

2.2. Literature Study
Literature study aims to find information/data relevant to research. This research takes references from journals, articles, and books conversing data mining, decision trees, C4.5 algorithms, and football topics.

2.3. Data Collection
The research data was collected from the website www.worldfootball.net. It contains information and statistics on football throughout the world. The steps for collecting the data are as follows:
1. Visiting the website
2. Choose Italian competition
3. Select the schedule menu
4. Select the competition period (For example 2005/2006)
5. Copy the table to the text for each competition period
6. Copy the text to Microsoft Excel file

Data sample for this study are as follows:

![Data Sample](image)

The following are the descriptions of each column headers of the samples:
- Relegation: club status at the end of the competition (YES: relegated, NO: survive)
- Years: year of competition
- Positions: club rankings at the end of the competition
- Clubs: club names
- Match: number of matches from the competition
- Win: number of wins from the club
- Draw: number of series from the club
- Lose: number of defeats from the club
- Goals_For: number of goals scored by the club
- Goals_Against: number of goals conceded from the club
- Different_Goals: the difference between the number of goals scored and goals conceded from the club
- Points: total points of the club at the end of the competition

2.4. Pre-Processing Data
Preliminary data processing is performed by merging the data from the 2005/2006 season to 2017/2018 or thirteen seasons in total. Total records after the merging of data are 130 records.

2.5. Processing Data with C4.5 Algorithm
The model proposed in this study is the C4.5 algorithm. C4.5 algorithm is the result of the development of algorithm ID3 (Iterative Dichotomiser 3) by Quinlan [5]. C4.5 algorithm or decision tree is similar to a tree that has internal nodes (not leaves) to explain attributes, each branch describes the results of the attributes being tested, and each leaf explains its classes. The Decision Tree is easily converted to classification rules. Mostly, tree classification decisions have good accuracy, but the success of their use is influenced by the data to be processed.

The design of this research phase based on the C4.5 algorithm is as follows:

1. Prepare training data by taking from the information that has previously occurred and grouped into specific classes.
2. Determine the root of the tree by calculating the highest gain value of each attribute or based on the lowest entropy index value. Previously, the entropy index value was calculated first, using the following formula:

   \[ \text{Entropy} (i) = - \sum_{j=1}^{m} f(i,j) \cdot \log_2 f[(i,j)] \]

   Information:
   - \( i \) = set of cases
   - \( m \) = number of partitions \( i \)
   - \( f(i,j) \) = proportion \( j \) into \( i \)

3. Calculate the gain value with the formula as follows:

   \[ \text{Entropy Split} = - \sum_{i=1}^{n} \frac{n_i}{n} \cdot IE(i) \]

   Information:
   - \( p \) = number of partition attributes
   - \( n_i \) = current proportion to \( i \)
   - \( n \) = number of cases in \( n \)

4. Repeat 2\textsuperscript{nd} step to all partitioned records. The partitioning process will stop when:
   - All tuples in the record in node \( m \) get the same class.
   - There are no attributes in the partitioned record anymore.
   - There are no records in an empty branch.

2.6. Evaluation and Validation
The evaluation and validation of this study used Confusion Matrix and AUC (Area Under Curve) charts. Additionally, validation data is obtained by the Cross-Validation method.

3. Results and Analysis

3.1. Result
This study aims to identify the factors that caused Serie A clubs to be relegated from the competition. Inputs from this study are the results of the final competition standings of each year (2006-2018).

The target of the degraded club’s variable is YES and NO. While the input parameters are:
1. Number of goals scored (Goals For)
2. Number of goals conceded (Goals Against)
3. Amount of Points (Points)

The design of this research process in the Rapidminer application is as follows:
The design in figure 2 consists of operators as follows:

- **Relegation Data**: input process originating from Excel files and already imported into the Rapidminer application.
- **Select Attributes**: the selection of required attributes
- **Cross-Validation**: perform cross-validation to estimate the statistical value of the model. The operators in the Cross-Validation operator, are as follows:
  
  - **Decision Tree**: processing data into a tree-like form that has rules for classifying data from input. Because this study uses the C4.5 algorithm, the selected criteria are information gain.
  - **Apply Model**: operator to apply a model from the Example Set.
  - **Performance**: an evaluation operator to display the results of evaluation and model validation.

The results of the Example Set are as follows:

Based on figure 4, the relegation attribute is bi-nominal type label (YES / NO). While other attributes (goals_for, goals_against, and points) are regular attributes with numerical types.

### 3.2. Analysis

The results of this study are the output of the Rapidminer application with input from the final Serie A data from Italy. In the testing phase, the C4.5 algorithm produces the Decision Tree graph as follows:

Based on the Decision Tree graph above, it can be concluded that Serie A clubs will not be relegated if they have a point above 36.50, because there is no decimal value in football. Then the minimum point for surviving degradation is 37. A complete description of the decision tree is as follows:

- **points > 36.500**: NO {NO=86, YES=2}
- **points ≤ 36.500**: YES {NO=6, YES=36}
The results of the evaluation of this study used the following confusion matrix and ROC/AUC curves:

1. **Confusion Matrix**

   This method displays in the form of a matrix table consisting of:
   - Value accuracy: the percentage of data records that are correctly classified by an algorithm can make a classification after testing the results of the classification (Han & Kamber, 2006).
   - Value of precision: also known as the value of confidence is the proportion of the number of cases that are positively predicted, which are also positively correct in the actual data.
   - Whereas the value of recall or sensitivity is the proportion of the actual number of positive cases that are positively predicted (Powers, 2011).
   - True Positive is the number of positive records that are classified as positive.
   - False positives are the number of negative records classified as positive.
   - False Negative is the number of positive records that are classified as negative.
   - True Negative is the number of negative records that are classified as negative.

   **Table 1. Result of Confusion Matrix.**

   | Prediction | True NO | True YES | Precision Class |
   |------------|---------|----------|-----------------|
   | Prediction NO | 86      | 6        | 93.48%          |
   | Prediction YES | 6       | 32       | 84.21%          |

   Based on the results of the evaluation matrix confusion above, the accuracy value is 90.77%. The value of precision is 86.00% with the positive class being YES, while the recall value is 85.00%. True Positive amount is 86 records classified as NO (survived), and the number of False Negatives is 6 records classified as YES (relegated). The number of False Positive is 6 records classified as NO (survived) and the number of True Negatives is 32 records which are classified as YES (relegated).

2. **ROC/AUC Curve**

   Receiver Operating Characteristic (ROC) curve is a metric for comparing actual and target values in the classification model. The ROC/AUC curve consists of plots X and Y.

   - **Figure 6. AUC Optimistic Curve**
   - **Figure 7. AUC Curve**
   - **Figure 8. AUC Pessimistic Curve**
On the AUC Optimistic graph (Figure 6), the AUC value of 0.956 with a positive class is YES. The AUC graph (Figure 7), the AUC value of 0.912 with a positive class is YES. On the AUC Pessimistic graph (Figure 8), the AUC value of 0.879 with a positive class is YES.

The Rapidminer application displays a summary of the results of testing and evaluation in the Performance Vector format. Below are the results:

![Performance Vector](image)

### 4. Conclusions

Based on the results of the discussion in the previous chapters, it can be concluded that the C4.5 algorithm can predict the factors for the relegation of the Italian Serie A clubs. The main factor in determining a relegated club is the number of points (points). In the decision, tree decides the club can’t be relegated if above 36.50 points (minimum 37 points). The accuracy value of the confusion matrix is 90.77%, the precision value is 86%, and the recall value is 85%. With a total data of 130 records, the prediction value is Good Classification.

### 5. References

[1] Handoko A. Sepakbola Tanpa Batas: City of Tolerance. Yogyakarta: Kanisius Publishers; 2008.
[2] Swatman R. Confirmed: Neymar becomes the most expensive footballer in transfer history. Guinness World Rec, 2017:1. [http://www.guinnessworldrecords.com/news/2017/8/confirmed-neymar-becomes-most-expensive-footballer-in-transfer-history-484811](http://www.guinnessworldrecords.com/news/2017/8/confirmed-neymar-becomes-most-expensive-footballer-in-transfer-history-484811).
[3] Evicienna E, Amalia H. Algoritma C4.5 Untuk Prediksi Hasil Pemilihan Legislatif DPRD DKI Jakarta. J Techno Nusa Mandiri 2013; IX:48–56.
[4] Baio G, Blangiardo M. Bayesian hierarchical model for the prediction of football results. J Appl Stat 2010. doi:10.1080/02664760802684177.
[5] Han J, Kamber M. Data Mining Concept, and Techniques. Second. San Francisco: Morgan Kaufman; 2006.
[6] Hamil S, Michie J, Oughton C, Warby S. The Changing Face of the Football Business: Supporters Direct. London: Frank Cass Publishers; 2000.
[7] Muniyandi AP, Rajeswari R, Rajaram R. Network anomaly detection by cascading k-Means clustering and C4.5 decision tree algorithm. Procedia Eng., 2012. doi:10.1016/j.proeng.2012.01.849.
[8] Mantas CJ, Abellán J. Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data. Expert Syst Appl 2014. doi:10.1016/j.eswa.2014.01.017.
[9] Powers DMW, Ailab. EVALUATION: FROM PRECISION, RECALL, AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION. J Mach Learn Technol ISSN 2011.
[10] Singh S, Gupta P. Comparative Study Id3, Cart and C4.5 Decision Tree Algorithm: a Survey. Int J Adv Inf Sci Technol 2682.
[11] Jia W, Huang LJ. Improved C4.5 decision tree. Int. Conf. Internet Technol. Appl. ITAP 2010 Proc., 2010. doi:10.1109/ITAPP.2010.5566133.
[12] Zhou ZH, Jiang Y. NeC4.5: Neural ensemble-based C4.5. IEEE Trans Knowl Data Eng 2004. doi:10.1109/TKDE.2004.11.