

Optimization of the C4.5 Algorithm Using Particle Swarm Optimization and Discretization in Predicting the Results of English Premier League Football Matches

Muhammad Bahyul Anwar Fuadi ^{1*}, Alamsyah ¹

^{1,1} Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang, Indonesia
*Corresponding author: bahyulm@students.unnes.ac.id

ARTICLE INFO

Article history

Received 5 October 2022
Revised 13 October 2022
Accepted 27 October 2022

Keywords

Data Mining
C4.5 Algorithm
Discretization
Particle Swarm Optimization
Football Match Prediction

ABSTRACT

Football is one of the most popular sports. One of the most competitive football competitions is the English Premier League. This study aims to determine the prediction of the results of the football match in English Premier League. The prediction results in the form of home win, away win, and draw. This prediction uses data mining techniques, namely using the C4.5 algorithm as a classification algorithm with Particle Swarm Optimization as a feature selection method and Discretization as a preprocessing method. The dataset used was obtained from the football-data.co.uk website for four league seasons from the 2017/2018 season to the 2020/2021 season with a total of 1,520 instances. In this study, a comparison was made to the methods used to determine the increase in accuracy obtained. Based on ten times the data mining process, the final result of the best accuracy from using the C4.5 algorithm is 57.24%, then the C4.5 algorithm with Discretization gets an accuracy of 65.13%, and the C4.5 algorithm with Discretization and Particle Swarm Optimization gets accuracy of 71.05%. The conclusion is that the use of Discretization and Particle Swarm Optimization can improve the performance of the C4.5 algorithm in predicting the results of English Premier League matches with an increase in accuracy of 13.81%.

This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



1 Introduction

The English Premier League is the top level of the men's English football league system held by the football association in England, namely The Football Association (Prabowo, 2020). The English Premier League is one of the leagues with the most fans in the world. This can be seen from the number of followers on the English Premier League club's Instagram account. Based on Instagram data, in April 2022, Manchester United had 57.7 million followers, then Liverpool had 38 million followers, and Chelsea had 33.9 million followers. So that the results of the English league matches are very interesting to make a prediction to find out the results of the match. Research in predicting the results of football matches has been done previously using Machine Learning (Alfredo & Isa, 2019; Baboota & Kaur, 2019; Prabowo, 2020; Razali et al., 2017; Zhang et al., 2021).

Machine Learning can be used to make predictions on a dataset using data mining classification techniques (Yuliani, 2021). Classification is a technique that can be used to predict data or describe data classes (Alamsyah & Fadila, 2021). In data mining, a popular classification algorithm for analyzing Decision Tree-based statistical data is the C4.5 algorithm (Muslim et al., 2018). The C4.5 algorithm has been used to predict data (Prihanditya & Alamsyah, 2020). The algorithm is used to classify the results of the English Premier League matches based on the dataset obtained from the statistics of the results of the English Premier League matches. The game results are home wins,

away wins and draws. Based on research that has been done previously, there are obstacles in the selection of attributes and the data used has a continuous value so that the algorithm used for the classification process does not run optimally (Baboota & Kaur, 2019).

In this study, the classification process will be carried out using a Decision Tree-based algorithm, namely the C4.5 algorithm. The C4.5 algorithm is one of the Decision Tree algorithms that has the ability to produce the best prediction accuracy and requires a minimum execution time (Saputra, 2020). Then this research performs a preprocessing process to get data that is easily accessible by the system and more concisely using the Discretization method. The Discretization method is able to divide the data into a range so that the data used becomes more concise (Dash et al., 2011). After the preprocessing process, a feature selection was applied to reduce attributes that did not increase the results of classification accuracy (Lestari & Alamsyah, 2020). A feature selection process is carried out to overcome attributes that do not have an important influence on the classification results. The feature selection process uses Particle Swarm Optimization because this method is a more optimal method compared to genetic algorithms, especially in the field of optimization (Muslim et al., 2018). Particle Swarm Optimization is also able to overcome the problem of imbalance in the dataset. The imbalance of the dataset can be derived from the algorithm used in the classification process (Fanny & Cenggoro, 2018).

2 The Proposed Method

2.1 C4.5 Algorithm

The C4.5 algorithm is one of the Decision Tree algorithms in data mining techniques, namely a classification that is interpreted into a decision tree and decision rules (Hermanto & Azhari, 2017). The C4.5 algorithm includes a continuation of the ID3 algorithm with several developments such as processing continuous data and being able to overcome missing values (Irena & Setiawan, 2020). The C4.5 algorithm is able to handle data that has numeric and discrete values (Irawan, 2021).

The performance of the C4.5 algorithm in building a decision tree is to choose the attribute to be used as the root, then create branches until all attributes have the same class (Muzakir & Wulandari, 2016). Determination of the attributes used is based on the highest gain value of all the attributes in the dataset (Sulistyo, 2020).

2.2 Discretization

Discretization is used when preprocessing data by changing the value of a continuous-valued attribute into an attribute that has an interval value with a discrete numeric value (Mirqotussa'adah et al., 2017). Discretization has an important role in the data preprocessing stage in data mining, especially in classifying data (Kapoor et al., 2017).

The purpose of the discretization process is to determine a set of cut points to divide the range into a small number of intervals. There are two main functions of discretization. First, find the number of discrete intervals. Second, look for the width or boundaries of the interval given the range of values of the continuous-valued attribute (Kapoor et al., 2017).

2.3 Particle Swarm Optimization

Particle Swarm Optimization is a search method that has a simple structure, strong operating capabilities, and is easy to implement (Sulistyo, 2020). The function of Particle Swarm Optimization is to find a set of particles that is updated in all iterations. To achieve the optimal solution, each particle moves towards the personal best and global best position of a group (Sundaramurthy & Jayavel, 2020).

Particle Swarm Optimization is one of the techniques in the optimization field that can handle optimization problems on data that has continuous or discrete values (Sengupta et al., 2018). Differences in Particle Swarm Optimization with other optimization techniques is that each Particle Swarm Optimization particle is associated with a velocity. Particle Swarm Optimization is almost the same as Genetic Algorithm, but Particle Swarm Optimization does not use evolution operators such as mutation and recombination or crossover (Muslim et al., 2019). The Particle Swarm Optimization process begins with the initialization of the Particle Swarm Optimization parameters, initializing the initial particle position and velocity, calculating the fitness value, updating the

personal best and global best, updating the particle position and velocity, completing all iterations, then displaying the optimal solution.

3 Method

In this research, prediction of the results of football matches is carried out by applying the C4.5 algorithm as a classification method for data mining, then discretization is used at the data preprocessing stage and Particle Swarm Optimization is used in the feature selection process to obtain the selected attribute. The result of this research is the accuracy resulting from the proposed method. The flowchart of the method proposed in this research can be seen in Figure 1.

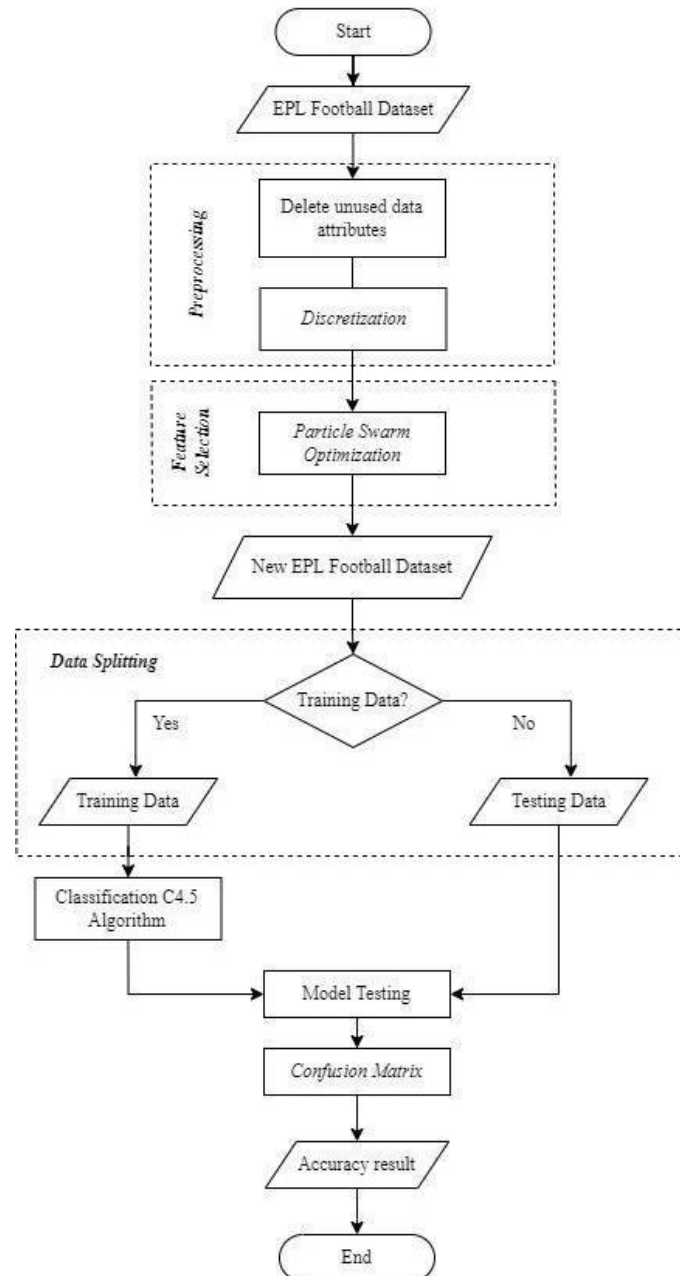


Figure 1. Flowchart of proposed method

Based on Figure 1 before the classification process is carried out, the dataset of the statistics results of the English Premier League matches used is first processed at the preprocessing stage by applying Discretization. The results of the discretization data are then used for the Particle Swarm Optimization feature selection process. The flowchart of Particle Swarm Optimization can be seen

in Figure 2. Selected attributes obtained from the feature selection process are used in the classification process by dividing them into training data and testing data. The classification process is applied using the C4.5 algorithm. The flowchart of the C4.5 algorithm can be seen in Figure 3.

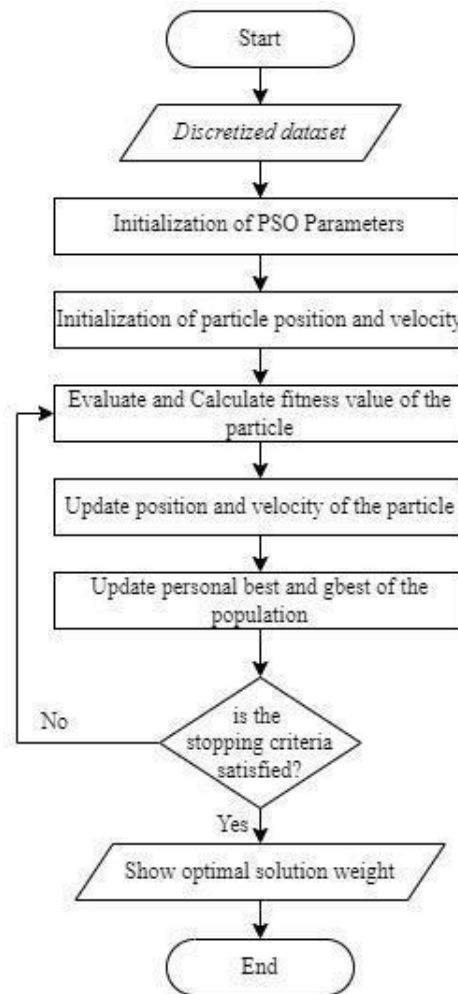


Figure 2. Flowchart of Particle Swarm Optimization

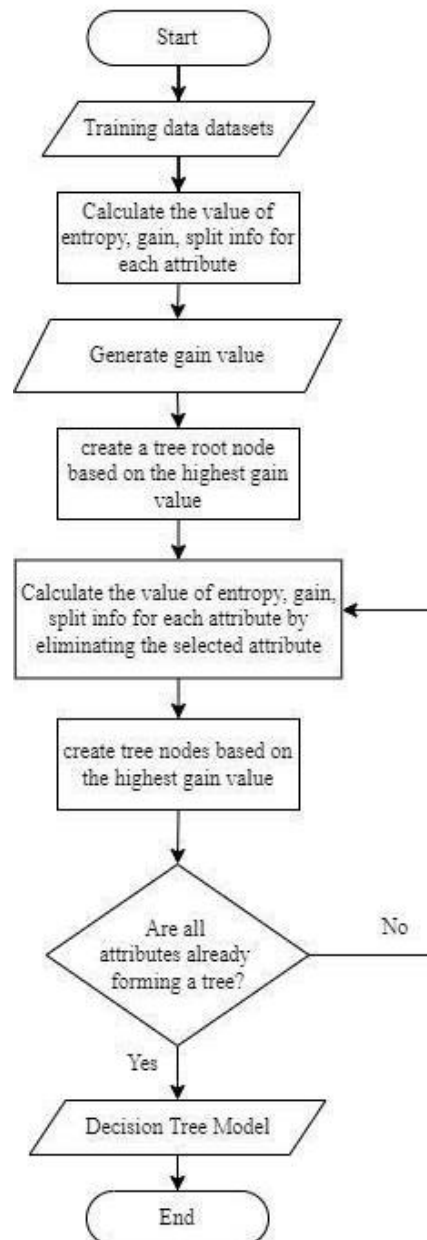


Figure 3. Flowchart of C4.5 algorithm

4 Results and Discussion

This section is divided into 2 parts, results and discussion. The results are a description of the data and findings obtained using the methods and procedures described in the data collection method. Discussion is a review of the results that answer research questions more comprehensively.

4.1 Result

4.1.1 Data Collection

The results of data collection in the form of datasets that have been obtained from the football-data.co.uk site is then divided into two types of data, namely football match statistics and bookmaker odds prediction. The data taken is in the form of football match statistics, while the bookmaker odds prediction data is deleted. The data used has a total of 1,520 instances. The football match statistics dataset can be seen in Table 1.

Table 1. Football match statistics dataset

No.	HTHG	HTAG	HS	AS	HST	AST	HF	AF	HC	HY	FTR
1	2	2	27	6	10	3	9	12	9	0	H
2	0	0	6	14	2	4	6	9	3	0	A
3	0	3	19	10	6	5	16	11	8	3	A
4	0	2	14	8	4	6	7	19	12	1	A
5	1	0	9	9	4	1	13	10	6	1	H
6	0	0	29	4	2	0	10	13	13	2	D
7	2	1	9	14	4	5	14	8	3	0	D
...
1516	1	0	19	5	5	4	10	8	14	2	H
1517	2	0	21	8	11	3	8	10	7	2	H
1518	1	0	12	10	3	3	11	1	8	3	H
1519	2	0	14	17	7	5	5	9	2	0	H
1520	1	2	14	9	4	4	14	3	6	4	A

Based on Table 1, the dataset will then be processed at the preprocessing stage using the Discretization method and feature selection using Particle Swarm Optimization.

4.1.2 Data Preprocessing

The result of data preprocessing is to change the value in the FTR attribute which was originally a polynomial value to be numeric. In the discretization process, not all attributes will be applied to the discretization process. Attributes that are not applied by the discretization process are HTHG, HTAG, and FTR. The HTHG and HTAG attributes are attributes for the number of goals entered by the home and away teams. If the attribute is applied to the discretization process, it will affect the classification results and reduce the accuracy value obtained. The discretization results can be seen in Table 2.

Table 2. Discretized Dataset

No.	HTHG	HTAG	HS	AS	HST	AST	HF	AF	HC	HY	FTR
1	2	2	1	0	1	0	0	1	1	0	0
2	0	0	0	1	0	1	0	0	0	0	1
3	0	3	1	0	1	1	1	1	1	1	1
4	0	2	1	0	1	1	0	1	1	1	1
5	1	0	0	0	1	0	1	1	1	1	0
6	0	0	1	0	0	0	1	1	1	1	2
7	2	1	0	1	1	1	1	0	0	0	2
...
1515	1	1	0	1	1	1	0	0	1	0	1
1516	1	0	1	0	1	1	1	0	1	1	0
1517	2	0	1	0	1	0	0	1	1	1	0
1518	1	0	0	0	0	0	1	0	1	1	0
1519	2	0	1	1	1	1	0	0	0	0	0
1520	1	2	1	0	1	1	1	0	1	1	1

The attributes applied by the discretization process are, HS, AS, HST, AST, HF, AF, HC, and HY. The attribute initially has a continuous value, then a Discretization process is applied which is divided into two intervals with equal frequency division. So that the discretization results data changes to values 0 and 1.

4.1.3 Feature Selection

The results of the feature selection process using Particle Swarm Optimization will be used for the classification process in data mining. Where the attribute that produces a value of 0 will not be used in the data mining process. While the attribute that produces a value of 1 will be used in the data mining process. The results of the feature selection process can be seen in Table 3.

Table 3. Result of Feature Selection Particle Swarm Optimization

HTHG	HTAG	HS	AS	HST	AST	HF	AF	HC	HY
1	1	0	1	1	1	0	0	1	1

The results obtained from the feature selection process with Particle Swarm Optimization using the initialization of several parameters, namely, $C1 = 0.7$, $C2 = 0.7$, $w = 0.72$, number of particles = 50, and iteration = 100. Where the selected attribute is HTHG, HTAG, AS, HST, AST, HC and HY. While the attributes that are not used are HS, HF, and AF because they have a less important influence in the data mining process.

4.1.4 Data Splitting

Results The data separation process uses a ratio of 90:10. Because this number is the best data split (Prabowo, 2020). The more training, the better the learning model obtained by the system. The results of the data split can be seen in Table 4 and Table 5.

Table 4. Training Data

No.	HTHG	HTAG	HS	AS	HST	AST	HF	AF	HC	HY	FTR
1	0	0	7	11	4	2	8	9	0	2	2
2	0	0	19	6	6	0	13	13	12	0	0
3	1	0	11	7	6	2	5	12	7	2	0
4	0	0	14	6	2	0	13	20	7	5	0
5	0	2	21	14	3	5	11	10	8	2	1
6	1	0	24	10	4	5	11	12	13	1	1
7	0	1	5	18	2	8	10	10	4	1	1
...
1363	0	1	12	6	6	3	7	7	6	2	1
1364	0	0	10	11	4	3	6	12	2	0	0
1365	0	0	9	7	4	4	11	7	2	3	1
1366	1	1	7	10	2	3	19	11	1	3	2
1367	1	3	10	17	3	8	13	9	2	0	1
1368	0	0	14	8	1	2	8	10	6	1	1

Based on Table 4 and Table 5, the training data is obtained from 90% of the total number of 1,520 instances, so that 1,368 rows are obtained for training data. While testing data obtained 10% of the total number of 1,520 instances resulting in 152 rows of data.

Table 5. Testing Data

No.	HTHG	HTAG	HS	AS	HST	AST	HF	AF	HC	HY	FTR
1	1	0	10	15	4	5	13	9	4	2	0
2	0	0	8	13	3	2	8	10	1	3	1
3	0	0	6	18	3	6	6	10	5	1	1
4	0	0	18	4	3	0	9	10	3	1	0
5	0	1	3	17	0	6	6	8	3	1	1
6	0	1	8	16	2	6	11	8	2	1	1
...
147	2	0	17	7	9	2	14	10	10	0	0
148	0	0	9	9	2	0	11	14	8	3	2
149	1	1	14	12	3	8	10	7	4	1	1
150	0	0	15	17	2	5	13	13	3	2	2
151	0	1	6	7	3	4	14	22	1	2	2
152	1	1	13	13	5	3	5	11	6	2	2

4.1.5 Data Mining

The data mining process is carried out into 3 testing processes. The first process is prediction of the results of the English Premier League match using the C4.5 algorithm without preprocessing and feature selection. The results were evaluated using a confusion matrix. Table 6 shows the results of the classification calculated using the confusion matrix.

Table 6. Confusion Matrix using C4.5 Algorithm

Actual	Predicted		
	Class 0	Class 1	Class 2
Class 0	44	10	12
Class 1	6	33	15
Class 2	17	5	10

$$Accuracy = \frac{\text{number of correct predictions}}{\text{total predictions}} \times 100\% \quad (1)$$

Based on Equation 1, the accuracy resulting from the confusion matrix in the Table 6 is:

$$Accuracy = \frac{44+33+10}{44+10+12+6+33+15+17+5+10} = \frac{87}{152} = 57.24\%$$

The accuracy obtained from the use of the C4.5 algorithm without data preprocessing and feature selection in the prediction of English Premier League match results is 57.24%. The accuracy obtained is considered too low, so the next testing process is carried out.

The second process is prediction of English Premier League match results using the C4.5 algorithm with preprocessing Discretization without feature selection. The dataset resulting from discretization is carried out by a data mining process and then the results are calculated using a confusion matrix. Table 7 shows the results of the confusion matrix in the second testing process.

Table 7. Confusion Matrix using C4.5 Algorithm with Discretization

Actual	Predicted		
	Class 0	Class 1	Class 2
Class 0	51	5	10
Class 1	9	40	5
Class 2	16	8	8

Based on Equation 1, the accuracy resulting from the confusion matrix in the Table 7 is:

$$Accuracy = \frac{51+40+8}{51+5+10+9+40+5+16+8+8} = \frac{99}{152} = 65.13\%$$

The accuracy obtained from the use of the C4.5 algorithm with preprocessing discretization without feature selection in the prediction of English Premier League match results is 65.13%. The accuracy obtained still needs to be improved, so further testing is carried out.

The third testing process is prediction of English Premier League match results using the C4.5 algorithm with preprocessing Discretization and feature selection Particle Swarm Optimization. The discretized dataset with 7 selected attributes will be used in the data mining process. The results of this process are calculated using a confusion matrix. Table 8 shows the results of the confusion matrix in the third test.

Table 8. Confusion Matrix using C4.5 Algorithm with Discretization and Particle Swarm Optimization

Actual	Predicted		
	Class 0	Class 1	Class 2
Class 0	53	8	5
Class 1	7	44	3
Class 2	13	8	11

Based on Equation 1, the accuracy resulting from the confusion matrix in the Table 8 is:

$$Accuracy = \frac{53+43+11}{53+8+5+7+44+3+13+8+11} = \frac{107}{152} = 71.05\%$$

The accuracy obtained from the use of the C4.5 algorithm with preprocessing Discretization and feature selection Particle Swarm Optimization in the prediction of English Premier League match results is 71.05%. Based on this accuracy, the Discretization and Particle Swarm Optimization processes are able to improve the accuracy of the use of the C4.5 algorithm in predicting the results of English Premier League matches.

4.2 Discussion

In this research, a comparison of accuracy using the C4.5 algorithm was carried out before and after the application of preprocessing and feature selection. At the data preprocessing stage, the Discretization method and feature selection using Particle Swarm Optimization are used in predicting the results of English Premier League football matches. Based on the application of the proposed method, there is an increase in the accuracy of the use of the C4.5 algorithm. the following comparison of accuracy obtained from the use of the proposed method can be seen in Table 9.

Table 9. Accuracy Comparison

Method	Accuracy
C4.5	57.24%
C4.5 + Discretization	65.13%
C4.5 + Discretization + Particle Swarm Optimization	71.05%

Based on the application of the proposed method, the accuracy of the implementation of the Discretization preprocessing and Particle Swarm Optimization feature selection on the C4.5 algorithm increased by 13.81%. The increase in accuracy is obtained from the accuracy of C4.5 without preprocessing and feature selection getting an accuracy of 57.24% compared to the C4.5 algorithm with preprocessing and feature selection getting an accuracy of 71.05%. The increase in accuracy is due to the discretization process making datasets have data that is more concise and easily understood by the system. The value of the discretized dataset is in the range of 0 and 1. And in the feature selection process, the selected attributes that have the most influence on the classification results are obtained. The selected attributes obtained from the selection of the Particle Swarm Optimization features are HTHG, HTAG, AS, HST, AST, HC, and HY.

The application of preprocessing Discretization and feature selection using Particle Swarm Optimization on the C4.5 algorithm obtained a higher accuracy value than the previous research conducted by Prabowo in 2020, where the K-Nearest Neighbor and Naïve Bayes Classifier methods were used on the same dataset as this research. accuracy as follows, K-Nearest Neighbor 62.7% and Naïve Bayes Classifier 63.5%. In that study, the use of K-Nearest Neighbor and Naïve Bayes Classifier was not suitable when encountering irrelevant attributes, such as attributes that did not have an important influence on the classification results (eg HS, HF and AF attributes). So in this study, a feature selection process was carried out to get the selected attributes. And in the K-Nearest Neighbor and Naïve Bayes Classifier methods require large memory usage, then this study overcomes this problem by conducting the Discretization preprocessing process. This is done to reduce memory usage on the system by making the values in the dataset a value that has two strings of 0 and 1. Then, the research conducted by Baboota & Kaur in 2019 using the Gradient Boost method resulted in an accuracy of 59%. The Gradient Boost method takes a long time to create a classification model, because it must complete one tree first to build the next tree by reducing errors. As well as research conducted by Alfredo & Isa in 2019, where the selection feature used is Wrapper Model and then classification using Random Forest produces an accuracy of 68.55% and Extreme Gradient Boosting 67.89%. Research using Random Forest and Extreme Gradient Boosting has weaknesses in terms of making classification models, learning cannot be repeated and runs slowly. In this study, the C4.5 algorithm is used because it has better performance and runs optimally. Comparison of the accuracy used by previous studies can be seen in Table 10.

Table 10. Comparison of the Accuracy by Previous Studies

Research	Method	Accuracy
Prabowo (2020)	K-Nearest Neighbors	62.7%
	Naïve Bayes Classifier	63.5%
Baboota & Kaur (2019)	Gradient Boost	59%
	Random Forest + Wrapper Model	68.55%
Alfredo & Isa (2019)	Extreme Gradient Boosting + Wrapper Model	67.89%
	Proposed Method	71.05%

Based on Table 10 the proposed method to get better accuracy results than previous studies. The proposed method gets better accuracy because in the proposed method the data preprocessing process uses Discretization and the feature selection process uses Particle Swarm Optimization. While the

research method carried out by previous researchers did not carry out the preprocessing and feature selection process.

5 Conclusion

Based on the results of research and discussion related to the application of Discretization and Particle Swarm Optimization methods on the C4.5 algorithm to predict the results of English Premier League football matches with datasets obtained from football-data.co.uk, the application of the proposed method can produce better accuracy. The data processing process starts from data preprocessing by applying the Discretization method with the number of $n_bins = 2$ with an equal frequency strategy. Then the discretized data is carried out by a feature selection process using Particle Swarm Optimization so that 7 selected attributes are obtained. Then the discretized data with selected features is carried out by a data mining process with split data distribution, 90% for training data and 10% for testing data used for the C4.5 algorithm classification process.

The highest accuracy results from the C4.5 algorithm without Particle Swarm Optimization and Discretization are 57.24%, the C4.5 algorithm with Discretization is 65.13%, and the C4.5 algorithm with Discretization and Particle Swarm Optimization is 71.05% in performing English Premier League match predictions. Increased accuracy resulting from the C4.5 Algorithm before applying Discretization and Particle Swarm Optimization with C4.5 after applying Discretization and Particle Swarm Optimization by 13.81%.

6 References

- Alamsyah, A., & Fadila, T. (2021). Increased accuracy of prediction hepatitis disease using the application of principal component analysis on a support vector machine. *Journal of Physics: Conference Series*, 1968(1).
- Alfredo, Y. F., & Isa, S. M. (2019). Football match prediction with tree based model classification. *International Journal of Intelligent Systems and Applications*, 11(7), 20-28.
- Baboota, R., & Kaur, H. (2019). Predictive analysis and modelling football results using machine learning approach for English Premier League. *International Journal of Forecasting*, 35(2), 741-755.
- Dash, R., & Paramguru, R. L. (2011). Comparative analysis of supervised and unsupervised discretization techniques. *International Journal of Advances in Science and Technology*, 2(3), 29-37.
- Fanny, & Cenggoro, T. W. (2018). Deep Learning for imbalance data classification using class expert generative adversarial network. *Procedia Computer Science*, 135, 60-67.
- Hermanto, B., & Azhari, S. N. (2017). Klasifikasi Nilai Kelayakan Calon Debitur Baru Menggunakan Decision Tree C4.5. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, 11(1), 43.
- Irawan, Y. (2021). Application of the C4 . 5 Decision Tree Algorithm to Predict the Eligibility of Prospective Blood Donors by Classification Data Mining. *Jurnal Teknologi Informasi Dan Multimedia*, 2(4), 181-189.
- Irena, B., & Setiawan, E. B. (2020). Fake News (Hoax) Identification on Social Media Twitter using Decision Tree C4.5 Method. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 4(4), 711-716. <https://doi.org/10.29207/resti.v4i4.2125>
- Kapoor, P., Arora, D., & Kumar, A. (2017). Implications of discretization towards improving classification accuracy for software defect data. *Journal of Theoretical and Applied Information Technology*, 95(24), 6893-6901.
- Lestari, A., & Alamsyah. (2020). Increasing Accuracy of C4 . 5 Algorithm Using Information Gain Ratio and Adaboost for Classification of Chronic Kidney Disease. *Journal of Soft Computing Exploration*, 1(1), 32-38.
- Mirqotussa'adah, M., Muslim, M. A., Sugiharti, E., Prasetyo, B., & Alimah, S. (2017). Penerapan dizcretization dan teknik bagging untuk meningkatkan akurasi klasifikasi berbasis ensemble pada algoritma c4.5 dalam mendiagnosa diabetes. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, 135.
- Muslim, M. A., Rukmana, S. H., Sugiharti, E., Prasetyo, B., & Alimah, S. (2018). Optimization of

- C4.5 algorithm-based particle swarm optimization for breast cancer diagnosis. *Journal of Physics: Conference Series*, 983(1).
- Muslim, M. A., Prasetyo, B., Harum, E. L., Juli, A., Mirqotussa'adah, Hardiyanti, S., & Nurzahputra, A. (2019). *Data Mining Algoritma C4.5 Disertai contoh kasus dan penerapannya dengan program computer*.
- Muzakir, A., & Wulandari, R. A. (2016). Model Data Mining sebagai Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree. *Scientific Journal of Informatics*, 3(1), 19–26.
- Prabowo, D. (2020). Prediksi hasil pertandingan sepakbola english premier league dengan menggunakan algoritma k-nearest neighbors dan naive bayes classifier. *Skripsi*. Fakultas Teknologi Industri. Universitas Islam Indonesia: Yogyakarta
- Prihanditya, H. A., & Alamsyah, A. (2020). The Implementation of Z-Score Normalization and Boosting Techniques to Increase Accuracy of C4.5 Algorithm in Diagnosing Chronic Kidney Disease. *Journal of Soft Computing Exploration*, 1(1), 63–69.
- Razali, N., Mustapha, A., & Yatim, F. A. (2017). Predicting football matches results using bayesian networks for English Premier League (EPL). *Top conference series: Materials science and engineering*, 22(1).
- Saputra, R. H. (2020). Optimasi Algoritma C4.5 Menggunakan Seleksi Fitur Particle Swarm Optimization (Pso) Dan Teknik Bagging Pada Diagnosis Penyakit Kanker Payudara. *Skripsi*. Fakultas Matematika dan Ilmu Pengetahuan Alam. Universitas Negeri Semarang: Semarang
- Sengupta, S., Basak, S., & Peters, R. (2018). Particle Swarm Optimization: A survey of historical and recent developments with hybridization perspectives. *Machine Learning and Knowledge Extraction*, 1(1), 157–191.
- Sulistyo, G. B. (2020). Penerapan Decision Tree C.45 Berbasis Particle Swarm Optimization untuk Prediksi Hasil Pemilu Legislatif. *Journal Speed – Sentra Penelitian Engineering dan Edukasi*, 12(1), 21-26.
- Sundaramurthy, S., & Jayavel, P. (2020). A hybrid Grey Wolf Optimization and Particle Swarm Optimization with C4.5 approach for prediction of Rheumatoid Arthritis. *Applied Soft Computing Journal*, 94, 106500.
- Yuliani, D. D. (2021). Implementasi machine learning untuk melakukan prediksi hasil panen tebu (Studi kasus di KUD Subur, Malang). (*Doctoral dissertation, Universitas Brawijaya*). Brawijaya Knowledge Garden. 181481.
- Zhang, Q., Zhang, X., Hu, H., Li, C., Lin, Y., & Ma, R. (2021). Sports match prediction model for training and exercise using attention-based LSTM network. *Digital Communications and Networks* 8(4), 508-515 .