Journal of Modern Technology and Engineering
Vol.10, No.1, 2025, pp.63-69
https://doi.org/10.62476/jmte.10163

ANALYZING MOBA GAME DATA USING FEATURE SELECTION AND CLASSIFICATION

Nur Uylaş Sati*

Muğla Sıtkı Koçman University, Bodrum Maritime Vocational School, Department of Motor Vehicles and Transportation Technologies, Bodrum, Muğla, Turkiye

Abstract. This work presents a research designed to predict the winner outcomes in video games and analyzing the game events' effect on victory. The approach uses various methods from data mining, preprocessing of the acquired data, feature selection, and classification implementations. Different classification algorithms in Weka (Waikato Environment for Knowledge Analysis) are applied on both the original and the reduced new dataset. A comparative performance analysis on accuracy and running time are presented both for the two dataset and for the used classification algorithms. The study offers insights into the role of game event data in outcome prediction and contributes to the broader understanding of data-driven decision-making in video game analytics.

Keywords: Data mining, classification, feature selection, video game.
AMS Subject Classification: 68T20, 68T30, 68W40.
Corresponding author: Nur, Uylaş Sati, Muğla Sıtkı Koçman University, Bodrum Maritime Vocational School,Department of Motor Vehicles and Transportation Technologies, Bodrum, Muğla, Turkiye,
e-mail: *nursati@mu.edu.tr*
Received: 5 February 2025; Revised: 22 March 2025; Accepted: 13 April 2025; Published: 30 April 2025.

1 Introduction

With the advancement of technology, vast amounts of data are collected in video games, including player, team, game, and season-related records. Analyzing such large datasets manually is impractical. Data mining has been extensively researched across various domains, and one of its emerging application areas is video games (Zdravevski & Kulakov, 2010). This motivated us to explore and test different data mining algorithms on datasets containing video game records.

Video games are a type of digital entertainment that involves the use of electronic devices, such as personal computers or gaming consoles, to play games.The first examples of video games emerged in the 1960s, and they became commercially successful and widely popularised in the 1970s (Jones, 2008). Today, video games have become a multi-billion dollar industry, used for both entertainment and educational purposes. This industry has also contributed to the development of virtual world economies (Altinpulluk, 2021).

MOBA (Multiplayer Online Battle Arena) games, a subgenre of MMO (Massively Multiplayer Online) games, are played over the Internet, where players compete to take control of key objectives on a map and ultimately destroy the opposing team's base. These games typically feature two teams of five players each. One of the most well-known examples of this genre is League of Legends (LoL). In recent years, predicting match outcomes in online games has become a significant area of study for machine learning models (Arık, 2024).

In the scholarly work conducted by Maymin in the year 2020, a sophisticated amalgamation of computer vision, dynamic client hooks, machine learning, visualization, logistic regression, extensive cloud computing resources, and rapid and efficient tree algorithms is employed to generate novel high-frequency data pertaining to millions of ranked League of Legends (LoL) matches. The researchers meticulously calibrate an in-game win probability model, formulate refined definitions for conventional metrics, introduce a multitude of advanced metrics, automate the analysis of player improvement, and implement an innovative player-evaluation framework based on both fundamental and advanced statistical measures (Maymin, 2020).

A multi-modal approach towards stress response modeling in competitive LoL games is proposed by Blom and his friends. They collect wearable physiological sensor data, mouse and keyboard logs and in-game data in order to study the relationship between player stress responses and in-game behaviour (Blom et. al., 2019).

In the research of Ani and his friends, with our same purpose, good forecasting rate for a popular electronic sport called League of Legends is aimed. Prediction is done by using ensemble models of classification algorithms and the performance was evaluated. The results show that the reliable match result prediction is possible in the League of Legends game (Ani et. al., 2019).

Also in the research of Hitar-García and his friends in 2023, a similar study was presented by adding new features and results with accuracy values above 0.70 were obtained. The difference in this study lies in the dataset used and the classification methods applied (Hitar-García et. al., 2023).

In the light of these studies, in this article, it is aimed to find the method that predicts the winner of the game with the highest accuracy value by using the classification methods, which is frequently used in machine learning algorithms, in the League of Legends video game. Before starting the classification studies, feature selection operations are performed as well as data editing for the big dataset found on the internet in the preprocessing section and the algorithm running- time is aimed to be reduced by decreasing the size of the problem. The work carried out for this purpose also enabled the most effective features to emerge in case of winning in this particular game.

In section 2 of the article, the methods used are presented in detail. In Section 3, information about the dataset used is given. In Section 4, the results of the implementation are presented in tables and comparisons are made and finally, the study is concluded in Section 5.

2 Materials and Methods

In this section, the classification algorithms and feature selection methods used in the study are explained in detail, along with the performance metrics employed for comparison. The implementations are carried out using the open-source software WEKA, meaning that the methods applied are limited to those available within this program.

Used classification algorithms:

Naive Bayes: Naive Bayesian classifiers assume that the effect of an attribute value on a particular class is independent of the values of the other attributes. This is termed class conditional independence. It simplifies calculations, hence its ”naive” designation. Bayesian belief networks, unlike naive Bayes classifiers, model dependencies between attributes using graphical models. Bayesian networks are applicable to classification problems.(Han et. al., 2011). For more information on Naive Bayes classifiers, see (John & Langley, 1995).

Logistics: Logistic regression is basically used as a binary classification technique based on supervised learning algorithm. It predicts the probability of a target variable using statistical techniques. Linear regression is unsuitable for binary classification as its unbounded output and differing baselines result in poor performance. Logistic regression employs the logistic or sigmoid function, an S-shaped curve. Logistic regression's best-fit line is updated using this formula.

$$y = \max \sum_{i=1}^{n} f(y_i w_i^T x_i)$$

Here y_i and x_i are the known values, only weight parameter w_i^T needs to be updated (Hussain & Gogoi, 2022).

Classification Via Regression: This method uses regression methods for classification. Class is binarized and one regression model is built for each class value. For more information, see (Frank et. al., 1998).

J48: The J48 classifier is an implementation of the C4.5 decision tree algorithm (Tripathi et. al., 2011). J48 classifies a new instance by creating a decision tree from the attribute values of the given training set. The moment it come across the training set, it recognizes the attribute that is responsible for categorizing the various instances most accurately. The possible feature values with no ambiguity are assigned to the concern branch by terminating it (Panigrahi & Borah, 2018).

Used feature selection methods:

In the preprocessing phase, in addition to the arrangements made in the data, feature selection is performed using the InfoGainAttributeEval function from the Attribute Evaluator section in the Select Attributes tab in Weka in order to reduce the problem size. This method ranks attributes by their individual evaluations.

Used performance metrics:

For comparison of the used classification algorithms, running time and k-fold cross validation performance metrics are used.

The running time of an algorithm refers to the length of time it takes for it to run as a function. An algorithm's running time for a given input depends on the number of operations executed. The applications in this study are performed on the same dataset on the same PC using the same programme. Therefore, the running times (sec.) given for each algorithm are comparable.

To determine the accuracy value, 'use full training set' is used in WEKA. The formula applied here is presented below.

wcp: number of well classified points

tsp: number of training set points

$$\text{Accuracy} = \frac{\text{wcp} \times 100}{\text{tsp}}$$

In the accuracy value calculation, in addition to the calculation made with the 'use full training set' in the Attribute Selection Mode tab in WEKA, the 'cross validation' calculation with higher reliability is also used.

In k fold cross validation, the dataset is divided into k equally (or nearly equally) sized segments, commonly referred to as folds. Following this partitioning, k iterations of the training and validation process are executed, in which, for each iteration, a distinct fold of the dataset is withheld for validation purposes while the remaining k-1 folds are utilized for the training process (Irizarry, 2019). In the context of this research, the parameter "k" is specified to be 10.

3 Dataset Description

As mentioned in introduction for this study the mostly used LoL game dataset in data mining studies is preferred This dataset is obtained from Kaggle (Mitchell, 2017). A single LoL game generates a large amount of data. We choose this to predict LoL game outcomes and analyze the in-game events most likely to cause victory.

The original dataset that was collected using the Riot Games API, is a collection of 51.490 ranked EUW games from the game League of Legends. For each game, there are 20 attributes. These attribute names are given below (Mitchell, 2017):

- ID of the game
- Timestamp of Creation (expressed in Epoch format)
- Duration of the Game (measured in seconds)
- Season Identification Number
- Champion of Victory (1 = team1, 2 = team2)
- Initial Baron, dragon, tower, blood, inhibitor, and Rift Herald (1 = team1, 2 = team2, 0 = none)
- Champions and summoner spells utilized by each team (Documented as Riot's champion and summoner spell identifiers)
- The tally of tower, inhibitor, Baron, dragon, and Rift Herald eliminations attributed to each team
- The five bans imposed by each team (Once more, champion identifiers are employed)

4 Numerical Experiments

In the experiments, firstly decreasing the problem size by choosing the most effective attributes on victory is aimed. InfoGainAttribute Eval is used in WEKA. Winner attribute is selected as the class and Ranker is selected for the method. Obtained results are given in Figure 1.

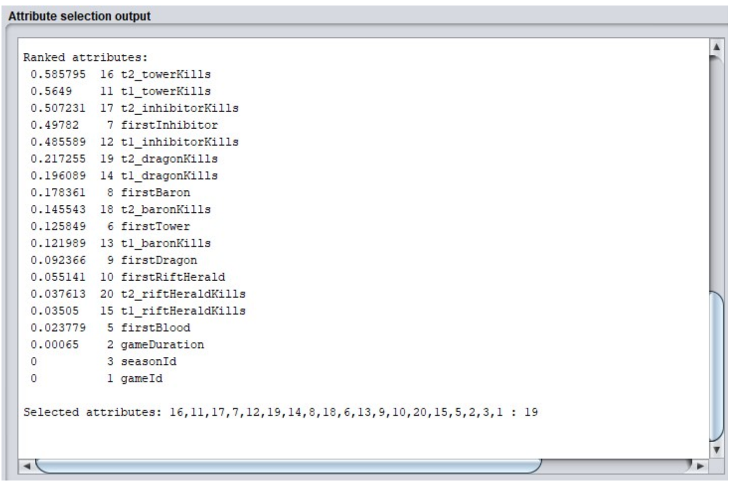

Figure 1: InfoGainAttribute Eval Outcomes

As can be seen from the results and as might be expected the gameid and seasonid have no effect on victory with rank value of 0. Also we see that the most effective ones are the number of the towerkills, inhibitorkills and the first team of the inhibitorkill. For classification applications, we deleted the attributes that have the rank of less than 0.4 since they have less effect on the victory. So in classification a new dataset that consists of 51.490 instances and 6 attributes with the class attribute(winner)is used. The obtained performance results on the new reduced dataset are given in Table 1.

Table 1: Performance Results of Classification Algorithms on the new dataset

Algorithms	**Accuracy (%)**	**Cross Validation (%)**	**Running Time (sec)**
Naive Bayes	94.31	94.30	Accuracy: 0.23 Cross-Vald.: 0.07
Logistics	95.93	95.92	Accuracy: 0.18 Cross-Vald.: 0.46
Classification Via Regression	96.78	96.69	Accuracy: 0.15 Cross-Vald.: 1.67
J48	96.84	96.64	Accuracy: 0.04 Cross-Vald.: 0.5

In order to see the effect of the attribute selection and change of the problem size, we conducted an experiment in addition to the new dataset (with reduced 6 attributes) we performed a new experiment involving the classification algorithms on the original dataset (with 20 attributes). The results obtained are shown in Table 2.

Table 2: Performance Results of Classification Algorithms on the original dataset

Algorithms	**Accuracy (%)**	**Cross Validation (%)**	**Running Time (sec)**
Naive Bayes	94.08	94.07	Accuracy: 1.08 Cross-Vald.: 0.38
Logistics	96.15	96.14	Accuracy: 0.04 Cross-Vald.: 1.08
Classification Via Regression	97.20	97.01	Accuracy: 0.09 Cross-Vald.: 4.78
J48	98.00	96.85	Accuracy: 0.05 Cross-Vald.: 2.64

As shown in Table 1 and Table 2, the original dataset generally yields higher classification accuracy across all algorithms. For instance, the J48 classifier achieved an accuracy of 98.00% on the original dataset, compared to 96.84% on the new dataset. Similarly, ClassificationViaRegression showed an increase from 96.78% to 97.20%. However, this accuracy gain comes at the cost of significantly higher computational time. For example, J48's cross-validation time increased from 0.50 seconds on the new dataset to 2.64 seconds on the original one. This trend is consistent across all classifiers, indicating that while the original dataset offers slightly better performance, the new dataset significantly reduces processing time. Therefore, feature selection can be a valuable preprocessing step when efficiency is a priority without incurring substantial loss in accuracy.

5 Conclusion

In this study, classification algorithms used in machine learning are used for predicting the winner of a game, League of Legends (LoL) on WEKA (Waikato Environment Knowledge Analysis) an open-source data mining software developed by the University of Waikato. Furthermore, the influence of game events on the outcome of the game is examined through an attribute selection algorithm. Ineffective attributes are eliminated according to the obtained ranking values, thereby reducing the dimensionality of the problem. The implication results of the classification algorithms both for original and new dataset is presented in tables. The most significant difference is encountered in the running time results since the problem size is really so effective on it. No significant difference is seen in accuracy values. It is foreseen that machine learning studies can be carried out on more comprehensive game data in future studies.

References

Altinpulluk, H. (2021). Video oyunların eğitim araştırmalarında kullanımı: Bir sistematik tarama. *Anadolu University Journal of Social Sciences, 21*, 185–212. https://doi.org/10.18037/ausbd.902597

Ani, R., Harikumar, V., Devan, A. K., & Deepa, O. S. (2019). Victory prediction in League of Legends using feature selection and ensemble methods. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)* (pp. 74–77). IEEE. https://doi.org/10.1109/ICCS45141.2019.9065758

Arık, K. (2024). Machine learning models on MOBA gaming: League of Legends winner prediction. *Acta Infologica, 7*(1), 139–151. https://doi.org/10.26650/acin.1180583

Blom, P. M., Bakkes, S. C., & Spronck, P. (2019). Towards multi-modal stress response modelling in competitive League of Legends. In *2019 IEEE Conference on Games (CoG)* (pp. 1–4). IEEE. https://doi.org/10.1109/CIG.2019.8847986

Frank, E., Wang, Y., Inglis, S., Holmes, G., & Witten, I. H. (1998). Using model trees for classification. *Machine Learning, 32*, 63–76. https://doi.org/10.1023/A:1007421302149

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann. https://books.google.com.tr/books?id=pQws07tdpjoC

Hitar-García, J.A., Morán-Fernández, L., & Bolón-Canedo, V. (2023). Machine learning methods for predicting League of Legends game outcome. *IEEE Transactions on Games, 15*, 171–181. https://doi.org/10.1109/TG.2022.3153906

Hussain, M.A., Gogoi, L. (2022). Performance analyses of five neural network classifiers on nodule classification in lung CT images using WEKA: A comparative study. *Physical and Engineering Sciences in Medicine, 45*(4), 1193–1204. https://doi.org/10.1007/s13246-022-01187-3

Irizarry, R.A. (2019). Cross validation. In *Introduction to data science*. https://rafalab.github.io/dsbook/cross-validation.html

John, G.H., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (pp. 338–345). https://doi.org/10.48550/arXiv.1302.4964

Jones, S.E. (2008). *The meaning of video games: Gaming and textual strategies* (1st ed.). Routledge. https://doi.org/10.4324/9780203929926

Maymin, P.Z. (2020). Smart kills and worthless deaths: eSports analytics for League of Legends. *Journal of Quantitative Analysis in Sports, 17*(1), 11–27. https://doi.org/10.1515/jqas-2020-0002

Mitchell, J. (2017). League of Legends dataset. *Kaggle*. https://www.kaggle.com/datasets/datasnaek/league-of-legends

Panigrahi, R., Borah, S. (2018). Rank allocation to J48 group of decision tree classifiers using binary and multiclass intrusion detection datasets. *Procedia Computer Science, 132*, 323–332. https://doi.org/10.1016/j.procs.2018.05.195

Tripathi, N., Oakes, M., & Wermter, S. (2011). Hybrid parallel classifiers for semantic subspace learning. In T. Honkela, W. Duch, M. Girolami, & S. Kaski (Eds.), *Artificial neural networks and machine learning – ICANN 2011* (Vol. 6792, pp. 73–80). Springer. https://doi.org/10.1007/978-3-642-21738-8_9

Zdravevski, E., & Kulakov, A. (2010). System for prediction of the winner in a sports game. In D. Davcev & J. M. Gómez (Eds.), *ICT Innovations 2009* (pp. 231–240). Springer. https://doi.org/10.1007/978-3-642-10781-8_7