A Machine Learning Approach Using Interpretable Models for Predicting Success of NCAA Basketball Players to Reach NBA

Dante de Araújo Costa[1], Joseana Macêdo Fechine[1], José Rubens da Silva Brito[2], João Victor Ribeiro Ferro[2], Evandro de Barros Costa[2] and Roberta Vilhena Vieira Lopes[2]
[1]*Department of Systems and Computing, Federal University of Campina Grande, Campina Grande, Brazil*
[2]*Computing Institute, Federal University of Alagoas, Maceio, Brazil*

Keywords: Supervised Machine Learning, Feature Selection, Genetic Algorithms, Knowledge Discovery in Databases, White-Box Prediction, Data Mining.

Abstract: Predictive models in machine learning and knowledge discovery in databases have been used in various application domains, including sports and basketball, in the context of the National Basketball Association (NBA), where one can find relevant predictive issues. In this paper, we apply supervised machine learning to examine historical and statistical data and features from players in the NCAA basketball league, addressing the prediction problem of automatically identifying NCAA basketball players with an excellent chance of reaching the NBA and becoming successful. This problem is not easy to resolve; among other difficulties, many factors and high uncertainty can influence basketball players' success in the mentioned context. One of our main motivations for addressing this predicting problem is to provide decision-makers with relevant information, helping them to improve their hiring judgment. To this end, we aim to have the advantage of producing an interpretable prediction model representation and satisfactory accuracy levels, therefore, considering a trade-off between Interpretability and Predictive Accuracy, we have invested in white-box classification methods, such as induction of decision trees, as well as logistic regression. However, as a baseline, we have considered a relevant method as a reference for the black-box model. Furthermore, in our approach, we explored these methods combined with genetic algorithms to improve their predictive accuracy and promote feature reduction. The results have been thoroughly compared, and models exhibiting superior performance have been emphasized, revealing predictive accuracy differences between the best white box and black box models were very small. The pairing of the genetic algorithm and logistic regression was particularly noteworthy, outperforming other models' predictive accuracy and significant feature reduction, assisting the interpretability of the results. Furthermore, the analysis also highlighted which features were most important in the model.

1 INTRODUCTION

Artificial Intelligence applied to Sports has increased in importance in the last years. In particular, some studies indicate that data mining and machine learning algorithms have been a relevant option to automate part of the data analysis processes for understanding and improving sports performance(Clacy et al., 2017).

In this study, we have considered the knowledge discovery in database (KDD) process applied to American basketball, only focusing on exploring historical data and features from National Collegiate Athletic Association (NCAA) men's basketball datasets. The context of our study is on the following specific scenario, connected to National Basketball Association (NBA), as follows:

Annually, 60 players are selected sequentially to join one of the 30 teams in the NBA, with the majority coming from the basketball NCAA league. To be successful in the NBA, a team needs to select players efficiently. However, even teams with the best opportunities for recruiting often make poor choices. Sometimes, a player treated as a top priority does not perform well, while another player selected in a lower priority position ends up performing better. This makes it difficult to know which player to select based on the team's recruiting possibilities and order. (Alamar, 2013).

Given the mentioned context, identifying NCAA basketball players with an excellent chance to reach the NBA is still a challenging mining problem. We

Costa, D., Fechine, J., Brito, J., Ferro, J., Costa, E. and Lopes, R.
A Machine Learning Approach Using Interpretable Models for Predicting Success of NCAA Basketball Players to Reach NBA.
DOI: 10.5220/0012390000003636

In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 3*, pages 758-765
ISBN: 978-989-758-680-4; ISSN: 2184-433X

tackle this problem, assuming that obtaining quality information from statistical data collected in NCAA matches is less expensive when compared to sophisticated technologies regarding computer vision or even a human look, in addition to being possible to cover more teams and athletes from this type of approach, since it is only necessary to process the collected data, which are easily accessible. Therefore, in this paper, we explore historical data and features from NCAA men's basketball datasets to provide decision-makers with relevant information and improve their judgment when hiring. Hence, we address the relevant prediction problem of automatically identifying NCAA basketball players with a good chance of reaching the NBA. Thus, we propose to apply supervised machine learning techniques, to examine data from NCAA, considering white-box classification methods, such as inducer decision trees with C4.5, C5.0, and CART algorithms, as well as logistic regression. This choice related to white-box models is mainly because they are widely used in knowledge discovery and data mining, especially in scenarios where, not only high accuracy is required, but model comprehensibility is desired. In addition, however, we have adopted the Support Vector Machine and Multi-layer perceptron (MLP), as baseline algorithms and examples of a successful black-box model, serving to compare its predictive accuracies, with the accuracy of our white-box algorithms.

Concerning data used in our study, we have worked on datasets that contain redundant, noisy, and irrelevant features that could potentially have a negative influence on decision-making processes. Thus, we have especially invested in feature selection questions to identify player attributes that contribute the most to being chosen by an NBA team for a professional contract. Thus, in particular, to improve the predictive accuracy of the used models with feature selection, while producing comprehensible models, we explore the mentioned classifiers in an approach that combines them with genetic algorithms.

In summary, the main purpose of our study is to investigate which features and classification models would be most useful, considering these four following research questions:

- **RQ1 -** How to enhance the feature selection process by identifying the most relevant attributes?
- **RQ2 -** Which classification techniques, among those studied, exhibit the best predictive performance?
- **RQ3 -** What are the advantages and disadvantages of algorithms with and without the use of genetic algorithms?
- **RQ4 -** How to provide explanations for predictive models by analyzing the contribution of the attributes used in the prediction process?

We performed an experimental study to evaluate the performance of the proposed methods with datasets from the NCAA repository. This study brought some exciting results, for instance, showing that the proposed approach is feasible and relevant in its purpose of combining genetic algorithms with the used predictive models. The outcomes have been compared and the models with the best results are presented and discussed in the remaining part of the paper, adequately answering the three mentioned research questions.

The remainder of this paper is organized as follows. In Section 2, we provide some background and discuss some similar works. Section 3 describes our approach, explaining the proposed white-box classification models used for our prediction task, with its experimental evaluation. Section 4 reports the predictive accuracy results and performance of the algorithms we tested. Section 5 presents conclusions and future work.

2 BACKGROUND AND RELATED WORK

The use of machine learning in sports is becoming increasingly popular, whether it is to predict the winner of a football match or to train a player through virtual environments. Basketball also uses these tools to determine which team will win a tournament, as demonstrated by Gumm et al. (Gumm et al., 2015) in their article, where they discuss the statistical challenges in correctly predicting the winning teams in a tournament, to present a machine learning strategy to perform such prediction.

University teams participating in the draft seek to develop strategies against their competitors. It is through this aspect that Kannan et al. (Kannan et al., 2018) analyze the data of players who participated in the NCAA to predict success in the NBA. They used 28 variables as input for algorithms such as Random Forest, Support Vector Machine (SVM), and Logistic Regression, obtaining Precision, Recall, and F1-Score values at the end of the executions.

In North American professional leagues like the NBA, it is common for various types of data to be collected and used to understand the probabilities of success of individual players as well as teams. Houde and Matthew (Houde, 2021) developed a study in this line of research, where they made a comparison be-

tween different models and their metrics for predicting games of the season based on data collected from matches of all teams in previous years.

In the work of Mahmood et al. (Mahmood et al., 2021), an analysis is conducted to determine if a specific player is an up-and-coming star in the NBA. To perform this analysis, a term called *Co-player* was utilized, which essentially refers to a person who belongs to the same team or the opposing team and has played matches during a common period. According to the author, *Co-player* is an important factor for predicting up-and-coming stars. For the prediction, the following machine learning algorithms were used: Support Vector Machine (SVM), Decision Tree CART, Maximum Entropy Markov Model (MEM), Bayesian Network, and Naive Bayes. In the study, some new attributes were created based on the pre-existing attributes in the databases. For example, the average *Hollinger Score* of a player P's *Co-players* was computed. At the end of the study, the significance of this *Co-player* related attributes in aiding the prediction of rising stars was demonstrated.

Meanwhile, Albert et al. (Albert et al., 2021) proposed a hybrid approach called ANN (*Adaboost*, *Random Forest*, and Multi-Layer Perceptron - MLP) that feeds on the same dataset. According to the author, this weighted combination of the three conventional models has not been the subject of research, making it an innovative approach to the problem of predicting stars in the NBA. This combination was obtained from the individual results of a variety of tested machine learning models, and the mentioned three models yielded better metrics in terms of sensitivity and specificity.

This ANN was constructed to serve as a hidden layer of a Recurrent Neural Network (RNN). Upon retesting the proposed model, the authors achieved a specificity of 90% and a sensitivity of 80%. While the specificity decreased slightly compared to the individual models, there was a significant increase in sensitivity.

The work by Hsu et al. (Hsu et al., 2018) attempts to predict the top sixteen NBA teams by applying machine learning algorithms based on player characteristics. These characteristics are related to statistics such as points, blocks, offensive and defensive rebounds, and other game metrics. The applied models aim to calculate the players' winning contribution to the team. To achieve this result, the following models were employed: Polynomial Regression, Random Forest Regression, and Support Vector Regression. To compare the models, a player efficiency rating (PER) was used as a measure of player performance.

The hybrid intelligent system proposed by Ozkan (Ozkan, 2020) was a combination of an artificial neural network and fuzzy logic. This concurrent neuro-fuzzy system was established to determine which team would win, considering data on overall team success, performance in recent games, and player quality. The neural network was developed to predict which team (home team or visiting team) would win the game based on certain parameters. The fuzzy inference system developed was able to predict the game's favorite, aiming to enhance the precision and sensitivity generated by the neural network.

In the work by Geng and Hu (Geng and Hu, 2020), they proposed the use of genetic programming to model and predict the final results of the NBA playoffs. For this purpose, they used performance statistics from the regular season of each team, attempting to predict the possible final classification of these teams in the playoffs. Historical data were collected to train the proposed predictive model. The results obtained demonstrated that the algorithm achieved good prediction accuracy and provided a valuable assessment of various performance statistics that are crucial in determining the team's probability of winning the championship.

The utilization of genetic algorithms for attribute reduction in datasets has proven to be relevant in the field of data mining. Babatunde et al. (Babatunde et al., 2014) conducted an experiment using a genetic algorithm on a dataset with 100 attributes, successfully reducing the dimensionality to only 11 attributes. The author compared the genetic algorithm with two methods, namely, WEKA (Information Gain Ranking Filter) and WEKA (CFS Subset Evaluator), both of which managed to reduce the dataset to only 20 attributes. However, despite being a significant value, it did not come as close as the 11 attributes achieved by the genetic algorithm. The author also assessed the accuracy of several machine learning models using the features generated by the three models. The model that achieved the highest accuracy, at 94%, utilized the features generated by the genetic algorithm in a Multi-Layer Perceptron.

Observing the works in the literature, few of them portray the importance of selecting different types of features to measure the result, since each algorithm returns a different result. All of them can be explored to show the main features within a pre-selected set. The intersection of these features will demonstrate the characteristics that deserve attention in the addressed problem.

3 METHOD

This section is divided into four subsections that represent the stages of the well-known Knowledge Discovery in Databases (KDD) process that was addressed, namely: (i) Data Description, (ii) Preprocessing, (iii) Used Algorithms, and (iv) Evaluation Metrics.

3.1 Data Description

The selected dataset was from the period 2009 to 2021 of American university athletes who competed in the NCAA[1]. The database consists of 65 features and 65.039 instances that are related to the athletes and the matches played by them. Some of the attributes found are:

• Attributes: The minutes played per game (Min), position, field goals made (FGM), field goals attempted (FGA), 3-pointers made (3 PTM), 3-pointers attempted (3 PTA), free throws made (FTM), free throws attempted (FTA), offensive rebounds (OREB), defensive rebounds (DREB), rebounds in general (REB), assists (AST), steals (STL), blocks (BLK), personal fouls (PF), points (PTS), and starter status (Starter if true, reserve if false) are some of the attributes found in the database.

3.2 Pre-Processing

The dataset is unbalanced with:

Table 1: Amount of data in the dataset.

Label	Amount of data in the dataset
0	64437
1	320
2	282

Where the label 0 represents the number of people who were not called up to play in the NBA. While label 1 represents the people who were called in the ”first call”, they are the highest priority players, while the players who have label 2, were selected in the ”second call” and have a lower priority compared to the of label 1.

To achieve data balancing, both Undersampling and Oversampling techniques were initially tested. However, the application of these techniques did not yield satisfactory prediction indexes. To address this issue, an alternative approach for balancing the data was adopted, which involved the following steps:

- Separation of all players belonging to label 1;
- Separation of all players belonging to label 2;
- Random selection of 1000 instances from a total of 64,437.

As a result, a balanced dataset consisting of 1,602 instances was obtained for conducting further tests.

Pre-processing is a stage of KDD where several techniques are applied to the data to improve the learning rates of the models. Blum's work (Blum and Langley, 1997) shows the importance of feature selection for machine learning models. In Brito's work (Brito et al., 2023), four selection methods were employed: Embedded, Filter, Wrapper, and Genetic Algorithm.. However, in this paper, we only focus on the genetic algorithm combined with the used white-box predictive models.

We have passed test_size as 0.33 which means 33% of data will be in the test part and the rest will be in the train part.

3.3 Proposed Method

The proposed method can be visualized in Figure 1. It consists of an algorithmic framework that utilizes a genetic algorithm. In this framework, an individual is subjected to genetic operators, such as one-point crossover, inversion mutation, and elitist selection. After passing through these genetic operators, the chromosome is evaluated using a fitness function, which incorporates a machine-learning algorithm. Once the fitness is calculated, the termination condition is checked to determine if it has been met. If the condition is met, the output will be a subset containing the best-performing chromosomes; otherwise, the entire process is repeated.

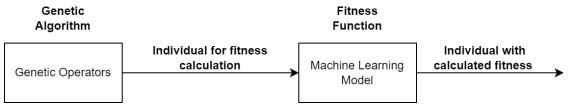

Figure 1: Proposed Method.

Various machine learning algorithms were tested in the fitness function of the proposed method, as illustrated in Figure 1. The primary objective was to assess the quality of the generated output. The fitness function incorporated the following machine learning models: Decision Trees CART, C4.5, and C5.0; Logistic Regression; Multi-layer Perceptron; and Support Vector Machines.

[1]https://www.kaggle.com/datasets/adityak2003/college-basketball-players-20092021

3.4 Used Algorithms[2]

The algorithms used were selected to diversify the approaches in machine learning models, including tree-based approaches such as C4.5, C5.0, and CART. Additionally, a statistical model, namely logistic regression, and a black-box kernel-based algorithm, the Support Vector Machine (SVM), and also the Neural Network, with the architecture Multi-Layer Perceptron were applied.

All of these algorithms served as a foundation for understanding the behavior of the dataset about performance metrics for each approach. Machine learning algorithms require the configuration of hyperparameters, and to identify the hyperparameters that best represented the dataset, a GridSearch was performed. GridSearch is a systematic approach to automate the parameter tuning process of an algorithm, generating and evaluating various combinations of parameters. The combination that best represents the dataset is selected as the most suitable (Liashchynskyi and Liashchynskyi, 2019).

Below, we present the best hyperparameter combinations obtained through GridSearch for the algorithms.

3.4.1 Decision Tree CART

The hyperparameters adopted for Cart algorithms are these:

- **Random_state:** 33
- **Criterion:** Gini
- **max_depth:** 4
- **max_features:** None
- **min_samples_leaf:** 1
- **splitter:** best

3.4.2 Decision Tree C4.5

The hyperparameters adopted for the feature selection criterion were entropy.

3.4.3 Decision Tree C5.0

For this decision tree, the hyperparameters are similar to the Cart decision tree, but we change the selection criterion and select the entropy

- **Criterion:** Entropy

3.4.4 Logistic Regression (LR)

In logistic regression, we adjusted the algorithm's hyperparameters, where we set **random_state = 33** and **max_iter = 200**

3.4.5 Support Vector Machine (SVM)

We used a black box algorithm as a baseline for our results. The selected algorithm was SVM. For more information on how it works, please refer to (Zhou, 2021), with the following hyperparameter settings:

- **C:** 1.0
- **gamma:** auto
- **kernel:** linear

3.4.6 Multi Layer Perceptron (MLP)

Another black box algorithm used with a baseline for our results was a neural network, as follows:

- **MLP with two hidden layers, each with 64 and 32 neurons, respectively.**
- **The activation function for each layer is the Rectified Linear Unit (ReLU).**
- **The model is trained with binary cross-entropy loss using the Adam optimizer with a learning rate of 0.01.**

3.4.7 Genetic Algorithm - GA

The settings adopted were:

- **individual's representation =** binary
- **length_population** = 50
- **length_chromosome** = 13
- **crossover_rate** = 75
- **mutation_rate** = 30

The stopping criterion used was a counter that records the number of generations in which there was no change in the best individual found. If this individual is not modified for 10 consecutive generations, the algorithm ends. Otherwise, the stop criterion counter will be reset and the process will be repeated.

3.5 Test Environment

The tests were performed on Google Computer Engine[3], in the following specifications:

- RAM: $\approx 12GB$
- Hard Drive: $\approx 108GB$

[2]https://anonymous.4open.science/r/Papers-04B6

[3]https://colab.research.google.com/?hl=pt_BR

3.6 Evaluation Metrics

The selected metrics were Accuracy, Precision, Recall, and F1_score as comparison measures between the algorithms, aiming to evaluate the performance of each algorithm according to the input parameters of each type of feature selection

4 RESULTS AND DISCUSSION

In this section, we delineate the outcomes and contributions stemming from the Machine Learning-based Approach posited within this investigation. The section is organized into three subsections to systematically articulate the results of the algorithms delineated in subsection 3.4. Herein, the outcomes will be expounded first without invoking the genetic algorithm, followed by an analysis of results when incorporating it. It is imperative to note that all findings are contingent upon the input of the feature subset ***['GP', 'Ortg', 'TPA', 'adjoe', 'rimmade/(rimmade+rimmiss)', 'dunksmade', 'dunksmiss+dunksmade', 'adrtg', 'dporpag', 'stops', 'gbpm', 'stl', 'blk', 'pts']***. Additionally, the concluding subsection presents a graphical representation elucidating the interpretability of the model that yielded the most accurate predictive outcomes.

4.1 Single Algorithm

In Table 2, the results obtained from the algorithms in single mode can be observed. Notably, the SVM achieved an accuracy metric of 79.58%, followed by the CART decision tree model with 79.01%.

Table 2: Performance Metrics for Single Algorithm.

Model	Accuracy	Precision	Recall	F1 Score
MLP	84.81%	77.21%	76.75%	76.97%
C4.5	77.90%	92.81%	47.29%	62.66%
SVM	79.58%	67.35%	67.01%	67.02%
LR	78.07%	64.23%	64.07%	64.14%
C5.0	78.26%	66.05%	64.64%	65.26%
CART	78.83%	68.58%	67.73%	66.87%

Another important point is the accuracy of the Decision Tree C 4.5 algorithm, which achieved a hit rate greater than 90%, indicating a high level of detection of true positives. However, when observing its Recall, a drop in the detection sensitivity of false positives is noticed. Therefore, a model that presented a balance between these aspects, measured by the F1_Score, was the CART decision tree, with a result of 66.87%. This result is just below the SVM algorithm, which obtained a rate of 67.02%. Balancing these metrics is crucial, as it will help scouts on basketball teams identify high-performing players.

In summary, achieving a balance among key evaluation metrics is imperative in the context of player scouting. This balance ensures that the model does not unduly prioritize a singular metric, such as high accuracy, to the detriment of others, such as recall. Consequently, this approach facilitates a more comprehensive and dependable assessment of player performance. The adoption of such a holistic methodology serves to assist scouts in rendering well-informed decisions regarding the identification of high-performing players within the basketball domain.

It is evident that the Multilayer Perceptron (MLP) neural network has demonstrated favorable outcomes in various metrics, particularly in terms of accuracy, recall, and F1-score. It consistently maintains a slight superiority over logistic regression, although occasionally with only marginal differences from the metrics obtained by the white-box predictor. Notably, within this specific domain of study, the implemented neural network did not yield results as remarkable when juxtaposed with the tested white-box algorithms.

4.2 Algorithms Combined with GA

In Table 3 we have the results obtained from the algorithms explained in the section 4.1 all combined with a genetic algorithm, in which we were able to obtain the metrics of accuracy, precision, recall, and F1-score, as well as the selection of features that the genetic algorithm was able to process during its execution.

It is noted that there was an increase in the accuracy of the models combined with the genetic algorithm, where the C4.5 decision tree obtained the result of 82.95% of accuracy, followed by the logistic regression with 82.23% and the SVM with 81.85%. However, it can be observed that the C4.5 model obtained the lowest indexes about precision, recall, and f1-score, that is, when this model was combined with GA, there was a significant loss in detecting true positives. However, it can be noticed that the recall, even being one of the lowest among the combined algorithms, was superior when compared to the C4.5 single.

It is noticed that the logistic regression, when combined with the GA, obtained the best metrics of accuracy, precision, recall, and f1-score, among the white box algorithms. That is, this model was able

to represent the data set well, as well as being able to match the SVM black box model, in relation to some metrics.

Table 3: Performance Metrics for Algorithms Combined with GA.

Model	Accuracy	Precision	Recall	F1 Score
C5.0	78.83%	67.65%	67.58%	67.09%
CART	79.02%	66.75%	62.75%	62.39%
LR	82.23%	71.50%	70.85%	70.94%
SVM	81.85%	71.18%	71.10%	71.11%
C4.5	82.95%	57.50%	57.32%	57.41%

To demonstrate the statistical significance of the results, the Student's t-test (Mishra et al., 2019) was applied to analyze whether there is a mean difference between the Logistic Regression algorithms with and without the use of Genetic Algorithm. This test was conducted using a sample of 30 accuracy results for each algorithm, aiming to determine if there is a statistically significant difference in predictive accuracy between the algorithms that utilized GA.

Independent Samples T-Test

		Statistic	df	p
Accuracy	Student's t	-10.9 ª	58.0	< .001

Note. H_a $\mu_{RL} \neq \mu_{RL+AG}$

ª Levene's test is significant (p < .05), suggesting a violation of the assumption of equal variances

Figure 2: Student's t-test.

As shown in Figure 2, according to the t-test for the accuracy variable, there was a statistically significant difference between the Logistic Regression algorithm that did not employ the genetic algorithm and the one that utilized GA.

Finally, as can be seen from Figure 3, the reduction in the number of features performed by the Genetic Algorithm (GA) in combination with the selected algorithms brought benefits, such as the increased performance of the evaluation metrics defined in Section 3.6. In addition, this process highlights the main features that should be observed by the scout during the evaluation. An example of this is the CART algorithm, which obtained the greatest feature reduction, using only three, and managed to maintain a performance comparable to the case in which the GA was not used.

In addition, the features that are present in most models are the **dunksmade, gbpm e adjoe**, which are represented by CART. That is, according to the executed models, these features stand out during the athletes' evaluation and are important for their work and performance. Consequently, performing well on these traits increases the likelihood of getting an NBA position.

Model	Selected Features	Count of features
CART + GA	adjoe, dunksmade, gbpm	3
C5.0 + GA	adrtg, gbpm, stl, blk	4
LR + GA	GP, dunksmade, adrtg, dporpag, gbpm, blk	6
C4.5 + GA	Ortg, TPA, adjoe, rimmade/(rimmade+rimmiss), dunksmiss+dunksmade, stops	6
SVM + GA	GP, TPA, adjoe, dunksmade, adrtg, dporpag, stops, gbpm, stl, blk	10

Figure 3: Feature selection of models with GA.

4.3 Explainability of the Model

The model demonstrating the most notable accuracy performance was logistic regression employed in conjunction with a genetic algorithm. Consequently, we opted for a subset generated through this amalgamation, encompassing the following features: ['GP', 'dunksmade', 'adrtg', 'dporpag', 'gbpm', 'blk']. Subsequently, this subset underwent analysis using the SHAP (SHapley Additive exPlanations) explainability tool.

The selection of the SHAP tool is motivated by its foundation as a game-theoretic methodology tailored for expounding upon the output of machine learning models. This framework establishes a nexus between optimal credit allocation and local explanations, drawing upon classical Shapley values from game theory and their relevant extensions (Schlömer et al., 2018).

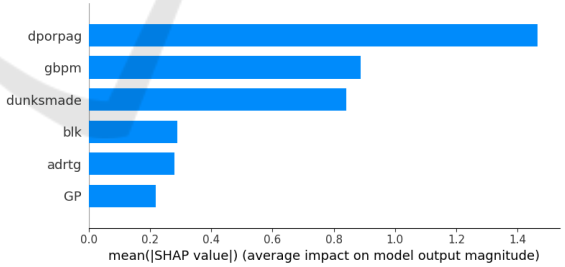

Figure 4: Result obtained from SHAP.

In Figure 4, it is possible to visualize the most contributing features and the magnitude of their impact on the model. Through this visualization, one can gain a global interpretation of how the features impact the model. As observed in the image, the factors that most influenced the model are 'dporpag' and 'gbpm'. Therefore, these attributes are of great relevance to the output generated by the models.

5 CONCLUSION AND FUTURE WORK

In this paper, we propose a supervised machine-learning approach with white-box models for predicting the chance of NCCA basketball players in the NBA. It is a novel approach with feature selection and predictive methods, whose experimental results indicate that the number of selected features is significantly reduced, besides achieving better predictive accuracy and offering a comprehensible predictive model. In our approach, therefore, we combine genetic algorithms with the white-box predictive models, allowing interpretability of the results, as well as knowing which features are more informative. The Logistic Regression classifier outperformed the use of white-box models and the two black-box models, that is, multi-layer perceptron and support vector machine. However, the predictive accuracy differences between the best white box and black box models were very small, the interpretability aspect is in favor of Logistic Regression model representation.

As an immediate future work, we aim to advance further in terms of prediction models by exploring ensemble methods, particularly Random Forest. Additionally, we intend to invest more in aspects of explainability for the algorithms employed.

REFERENCES

Alamar, B. C. (2013). *Sports analytics: A guide for coaches, managers, and other decision makers*. Columbia University Press.

Albert, A. A., de Mingo López, L. F., Allbright, K., and Gómez Blas, N. (2021). A hybrid machine learning model for predicting usa nba all-stars. *Electronics*, 11(1):97.

Babatunde, O. H., Armstrong, L., Leng, J., and Diepeveen, D. (2014). A genetic algorithm-based feature selection.

Blum, A. L. and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1-2):245–271.

Brito, J., Ferro, J., Costa, D., Costa, E., Lopes, R., and Fechine, J. (2023). A ranking between attributes selection models using data from ncaa basketball players to determine their tendency to reach the nba. In *2023 18th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6. IEEE.

Clacy, A., Goode, N., Sharman, R., Lovell, G. P., and Salmon, P. M. (2017). A knock to the system: A new sociotechnical systems approach to sport-related concussion. *Journal of sports sciences*, 35(22):2232–2239.

Geng, S. and Hu, T. (2020). Sports games modeling and prediction using genetic programming. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–6. IEEE.

Gumm, J., Barrett, A., and Hu, G. (2015). A machine learning strategy for predicting march madness winners. In *2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pages 1–6. IEEE.

Houde, M. (2021). Predicting the outcome of nba games.

Hsu, P.-H., Galsanbadam, S., Yang, J.-S., and Yang, C.-Y. (2018). Evaluating machine learning varieties for nba players' winning contribution. In *2018 International Conference on System Science and Engineering (ICSSE)*, pages 1–6.

Kannan, A., Kolovich, B., Lawrence, B., and Rafiqi, S. (2018). Predicting national basketball association success: A machine learning approach. *SMU Data Science Review*, 1(3):7.

Liashchynskyi, P. and Liashchynskyi, P. (2019). Grid search, random search, genetic algorithm: a big comparison for nas. *arXiv preprint arXiv:1912.06059*.

Mahmood, Z., Daud, A., and Abbasi, R. A. (2021). Using machine learning techniques for rising star prediction in basketball. *Knowledge-Based Systems*, 211:106506.

Mishra, P., Singh, U., Pandey, C. M., Mishra, P., and Pandey, G. (2019). Application of student's t-test, analysis of variance, and covariance. *Annals of cardiac anaesthesia*, 22(4):407.

Ozkan, I. A. (2020). A novel basketball result prediction model using a concurrent neuro-fuzzy system. *Applied Artificial Intelligence*, 34(13):1038–1054.

Schlömer, N., danielhkl, Wehrfritz, A., Berndt, H., Stathopoulos, S., Boeddeker, C., Edler, D., Spott, A., Gaul, A., Rossi, M., Vinot, B., Schürmann, D., Lipp, M., Dawson, D., mrtnschltr, pwohlhart, hgwd2, Koslowski, S., Lacasse, P., Verdier, O., Haberthür, D., and Kuzmin, A. (2018). nschloe/matplotlib2tikz v0.6.15.

Zhou, Z.-H. (2021). *Machine learning*. Springer Nature.