Esports Analytics on PlayerUnknown's Battlegrounds Player Placement Prediction using Machine Learning Approach

N.F. Ghazali[1,2], N. Sanat[1], M.A. As'ari1[1,2]

[1]School of Biomedical Engineering and Health Sciences, Faculty Engineering, Universiti Teknologi Malaysia, Malaysia
[2]Sport Innovation and Technology Center, Institute of Human Centered Engineering, Universiti Teknologi Malaysia, Malaysia

amir-asari@biomedical.utm.my

Abstract — PUBG (PlayerUnknown's Battlegrounds) is a video game that has become popular in the past year. This paper aims to predict the placement of PUBG players during the match by detecting the influential features set that can impact the outcome of the PUBG game and build the best prediction model using a machine learning approach. In this study, the dataset is taken from Kaggle, which has 29 attributes that are categorized into one label (winPlacePerc). The training set has divided into five sets with each set has 6000 instances. The decision tree regression model was applied to find the optimum prediction. Other regression models such as Linear Regression and Support Vector Machine are also utilized to compare with the decision tree model's result. Based on the result analysis, the walkDistance feature was deemed as the most significant factor influencing the results of a PUBG game. Furthermore, there are other common features obtained from the five datasets that represent the crucial factors which are boosts, DBNOs, killPlace, kills, rideDistance and matchDuration. From the three regression models, the Support Vector Machine model built on the significant features has the best performance in terms of RMSE value while the Decision Tree Regression model has the fastest prediction speed among these regression models.

Keywords—Machine learning; regression model; decision tree, support vector machine, PUBG

Article history: Manuscript received dd Month 201X; received in revised form dd Month 201X; Accepted dd Month 201X.

I. INTRODUCTION

THE PlayerUnknown's Battlegrounds or more known as PUBG is a famous online multiplayer battle royal game, which is one of the creations of Brendan Greene [1]. The concept of this survival shooting game was formulated from previous mods he made for the ARMA game, which is a video game series of first-person tactical military shooters motivated by the Japanese film called Battle Royale [2]. This battleground game is played by up to one hundred players, where they can choose to play solo, duo, or with a small team of up to four people. They can also choose to play either from the first-person perspective (FPP) or third-person perspective (TPP), each having its advantages and disadvantages in combat and situational awareness. Players need to survive by killing other players of different teams until they become the last man standing and wins the match [3]. A full round averagely takes less than 30 minutes with each match starts with players parachuting from a plane onto any ground areas of the map. Once they landed, players need to search for weapons, armour, and other equipment in any buildings, ghost towns and other sites that have been procedurally distributed throughout the map. Figure 1 below

shows the PUBG gameplay after the player has landed on the ground.

Fig. 1. PUBG Gameplay

Esports, also known as electronic sports is a type of organized sports competition format of video games [4]. Esports is particularly participated by professional players, individually or as teams with the champion prize as their goal. The most common video games genres in esports are multiplayer online battle arena (MOBA), first-person shooter (FPS), fighting, card games, battle royals, and real-time strategy (RTS) [5]. Esports tournaments had always been between amateurs. However, in the late 2000s, its popularity surges when there is increasing in participation by professional gamers and spectatorship through live streaming [6]. In 2018, tournaments of esports games reached 380 million audiences, mostly made up of esports enthusiasts as reported in [7]. The report also stated that PUBG as the newly developed game contributed a lot to rising global growth in esports.

A corresponding need to evaluate tactics and anticipate behaviours has arisen in parallel with the rise of esports. This led to the emergence of esports analytics. Esports analytics can be defined as the process of using data relating to esports, primarily in order to find meaningful patterns and trends in behavioural data and use visualization tools to assist with the decision-making process [8]. In esports analytics, win prediction has become the main focus, hence many analytical techniques have been developed to forecast the results [9].

Machine learning is an application from Artificial Intelligence (AI) that has the capability of learning a set of data by itself and make models from the data sample to make a decision [10]. Technologies such as machine learning have become favourable in predicting the reliable outcome of esports games as it can provide useful information to plan strategies and tactics. Supervised learning which is a technique from the machine learning can trains a known set of input and output data to develop a predictive model which can be evaluated during testing [11], [12]. Supervised learning using a regression approach is used when the response data consists of continuous or real values such as for forecasting, time series modelling and finding the causal effect [12], [13]. There has been a lot of regression model that has been introduced such as regression tree (random forest), linear regression and support vector machine [14]–[16].

There has been a lot of studies conducted using supervised learning on esports analytics as it is very useful in handling a complex task that involves large amounts of variables. However, the majority of the previous studies were focusing on other esports games such as Dota with different machine learning approach [17]–[19]. Other studies related to PUBG analytics have used a few different models such as light gradient boost machine, multilayer perceptron regression, and random forest [20], [21]. There has been a lack of research conducted on PUBG game analytics using the decision tree regression model technique.

In this study, a large dataset that consists of 29 variables including one continuous response requires a supervised learning approach using a regression model that can help in finding the patterns between input and output variables. Thus, this study aims to use the Decision Tree Regression model that could predict the PUBG player placement in matches based on several attributes of player's performance. After that, the Regression Tree model will be analysed using the RMSE value which determines the most important attributes of a player's performance in PUBG matches. Lastly, the performance of the modelled regression tree will be compared with other supervised learning based on the most important attributes of a player's performance in PUBG matches.

This study uses the PUBG game stats from the Kaggle website [22], where it has already separated into a training set and testing set. As this study mainly focuses on identifying the most important features, only the training set will be used throughout this study. The data from the training set which consist of about 4.5 million instances will be randomly selected into five sets, each has 6000 instances for faster computation. Each set consists of 29 features which include winning placement percentile as the target of prediction. Overall, the analytic procedures will be carried out using Regression Learner Apps which is available in MATLAB. Feature selection is a critical process during the analytical procedures to eliminate the least significant or insignificant attributes. The remaining features will be trained and continue to remove each attribute to see the performance's changes until it produces the best performance model in term of RMSE value, containing a set of significant features. From the features set of the best model, the training process repeats using other regression models which are linear regression and support vector machine (SVM) to compare their results of RMSE value and training time with the regression tree model.

II. RELATED WORK

A. Sports Analytics

Sports analytics is the research and evaluation of competitive sports performance and tournaments using statistical methods [23]. There are three basic elements in sport analytic which are data mining, predictive models, and information systems. [24]. The purpose of sports analytics is to give information to the decision-makers which are the coaches, trainers, and personal executives for them to develop tactics to gain advantages in sports competitions [24]. There are thousands of websites devoted to sports statistics' research and analysis and how these contribute to forecast outcome. FiveThirtyEight.com is one example that provides information about the upcoming game, series or season sports [25].

There are several studies related to sport and game analytics that utilize the machine learning approach. For example, the work in [26] used few predictive machine learning models such as Gaussian naïve Bayes, support vector machine, random forest, and gradient boosting to determine the important factors that contribute to English Premier League Football match result. The data from 11 seasons of United Kingdom Football match results were used in this study and found that all the models obtained less predictive accuracy from bookmaker's predictions.

Another study of sports analytics has conducted to predict the outcomes of the Indian Premier League on cricket game using multivariate regression and neural networks [27]. Overall, the recent studies focused on machine learning implementation instead of conventional statistical approaches in formulating and finding the pattern which is useful in increasing the athlete performance though sport analytic concept.

B. Esports Analytics

Esports can be different from sports in term of the outcome and interface [28]. Esports found to give an outcome in the virtual world and mediated by the human-computer interface, which is the electronic systems such as keyboard, mouse, sensors. In these recent years, esports has become a major international sport with millions of viewers [8].

Recent work by [21] has almost the same objectives as this study which is to predict the PUBG game results. The authors have used the same dataset which is the PUBG dataset from Kaggle website [22]. They also have made some data engineering process to the dataset which is removing outliers and combined some of the features into one variable. In the research, they have used few machine learning techniques including linear regression, Lasso regression, Elastic net, Ridge regression and Stochastic Gradient Descent (SGD) regression. From all the regression models, the authors have selected a trained model using Ridge regression as the best model due to the lowest mean absolute error (MAE) value obtained from the model as the regression model can handle multicollinearity data very well. Furthermore, the study has used random forest algorithm and LightGBM model to train those data and have got their best MAE

value (MAE = 0.0204) by using the LightGBM technique.

Another study has been made to predict the PUBG game matches outcome as presented in [29]. The authors have used the dataset from Kaggle as well. In this study, the authors used Weka software in the feature selection process to filter out insignificant attributes. On top of that, they have inserted new features to improve model performance. The new features are playerJoined, killsNorm, damageDealtNorm, heals And Boosts, totalDistance, boostsPerWalkDistance, healsPerWalkDistance, killsPerWalkDistance, healsAndBoostsPerWalkDistance and team. With the new features added combined with the existing dataset, they trained the dataset using various regression technique which are LightGBM Regression (Light Gradient Boosting Machine Regression), MultiLayer Perceptron, M5P and Random Forest. The models' performances were measure in the MAE value and found LightGBM regression model as the best model with the lowest MAE value.

C. Supervised Machine Learning: Regression Models

Machine learning is a part of Artificial Intelligence (AI) where it works by training data fed by a human. The training process itself means giving a set of inputs data and also the expected outputs data [30]. From the training, machine learning will produce its model that will be used to map new data in the testing process. Machine learning can be classified into two major categories, which are supervised learning and unsupervised learning. In supervised learning, this method is useful for classification and regression types while unsupervised learning used for clustering data set. This study will be focusing on supervised machine learning to predict the correct response (predicted) in correspond with the input (features) trained to them [31], [32].

The crucial part of the regression technique is feature selection or extraction which is to extract only useful information from the raw data. The feature selection is important to reduce the dimension of the dataset hence increase the model performance [31]. It also benefits in reducing training time and prevent overfitting as fewer data will reduce the model complexity. The performance of this machine learning can be evaluated by the prediction accuracy which equals the percentage of correct prediction divided by the total number of predictions.

D. Decision Tree Regression

One of the famous regression techniques is a decision tree. A decision tree is a hierarchical design that implements the divide-and-conquer approach. A regression tree deals with the prediction of a continuous response variable given the values of the predictors [33], [34]. This model is in the form of a tree structure built top-down from a root node that represents the entire learning set. The root node is partitioned homogeneously into branches and parent nodes based on the impurity measure. A parent node will further split into two child nodes and it will continue to split until a leaf node is constructed where there is no further split that can be made. A node represents a subset of the predictor variable while the leaf node represents the numerical target [31].

In this project, the decision tree will be used to observe the most significant attributes in predicting the player's performance in PUBG matches.

III. PROPOSED SYSTEM

The workflow for this study is based on the following Figure 2 which comprised of five steps explained further in several subsections.

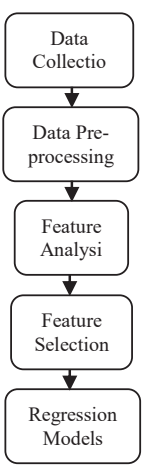

Fig. 2.Research methodology workflow

A. Data Collection

The PUBG dataset used in this study was obtained from Kaggle [22]. It contains a large number of anonymized PUBG game stats, formatted so that each row contains one player's post-game stats. The dataset incorporates 29 variables, including the target class, the winning placement percentile where 1 corresponds to first place and 0 corresponds to the last place in the match. These include all features for solo, duo and squad mode and are aggregated across all regions. The dataset contains three variables with unique columns which are the 'Id', 'groupId' and 'matchId', and one categorical variable which is the 'matchType'. While other 25 variables were characterised as a double type of numerical values. The dataset provided by Kaggle was already divided into the training set and testing set in form of an excel file (.csv file). The only training set was used throughout this study as it only focuses on determining the influential features. The dataset variables and their descriptions are shown in Table I.

B. Data Pre-processing

The training dataset which consists of 4.5 million instances will make the computation expensive and time consuming [35]. Therefore, the training dataset was further divided into five small sets which are randomly selected and saved in a .csv file for five times repetitions. Each set contains 6000 different instances but with the same variables. All five datasets were imported into MATLAB software version R2018b and saved in the workspace. For the data analysis step, Regression Learner App from the machine learning toolbox in MATLAB was used. This built-in app can train regression models to make predictions using supervised machine learning. The first dataset was selected in a new session to start extracting its response and predictors. Predictors such as Id, groupId and matchId have unique columns hence they were removed, as these predictors are not giving any information to the models [36]. Therefore, 25 features were remaining for training the models. For the validation scheme, 25 per cent of holdout validation was selected for faster training as it is recommended for large datasets. This validation method helps in examining the accuracy of the model and protects the model from overfitting [37].

TABLE I. THE DESCRIPTION OF FEATURES IN PUBG DATASET FROM KAGGLE [22]

Variable name	Description
assists	Number of enemy players damaged by this player but were killed by teammates
boosts	Number of boosting items used
damageDealt	Total number of damages dealt with enemy players
DBNOs	Number of enemy players knocked
groupId	Id to identify a group in a match
headshotKills	Number of enemy players killed with headshots
heals	Number of healing items used
Id	Player's id
killPlace	Ranking in match based on the number of enemy players killed
killPoints	Kills-based external ranking of players
killStreaks	Maximum number of enemy players killed in a short amount of time
kills	Number of enemy players killed
longestKill	The longest distance that enemy players killed
matchDuration	Duration played in a match in seconds
matchId	Id to identify the match
matchType	Type of match mode played
maxPlace	Worst placement in the match
numGroups	Number of groups in the match
rankPoints	Number of points gained for ranking
revives	Number of teammates revived by this player
rideDistance	Total distance travelled by riding vehicles in meters
roadKills	Numbers of enemy players killed while riding vehicles
swimDistance	Total distance travelled by swimming in meters
teamKills	Number of teammates killed by this player
vehicleDestroys	Number of vehicles destroyed
walkDistance	Total distance travelled by walking in meters
weaponsAcquired	Number of weapons picked up
winPoints	Win-based external ranking of players
winPlacePerc	Winning placement percentile which is the target of prediction

C. Feature Analysis

Before the process of matchmaking the best predictors, all the predictors were evaluated individually with the response. This phase was to observe the correlation between each predictor and target variable, or the predictor importance which later will assist in finding different combinations of predictors in order to get the list of most significant predictors. This step was performed by training the predictor one by one for all 25 features using the regression tree model. The performances of every predictor were observed and compared in terms of R-squared value. In statistical analysis, R-squared is a measure of the proportion of variation of response explained by the predictor [38], [39]. In a simple meaning, R-squared can

show how well the data fit into the regression model or how strong is their relationships. In conventional statistics, the formula for finding the R-squared value is as follows in Equation 1:

$$R^2 = 1 - \frac{SSE}{SST} \quad (1)$$

where SSE is the sum of squared errors of the actual data in the model, and SST refers to the total sum of squares [38].

SSE and SST can also be defined as Equation 2 and Equation 3 respectively:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (2)$$

$$SST = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2 \quad (3)$$

where $\hat{y}_i$ is the predicted value, $\hat{y}_i$ is the mean of observed value, $\hat{y}_i$ is the observed value, and n is the number of observations [39].

For a regression that uses only one predictor, as what happened during this phase, the R-squared value is equal to the square of the correlation coefficient, r between the predictor and response [38], [39]. During this phase, a correlation features set which are the features that have significant R-squared values, and no-correlation features set which are features with zero or negative R-squared value were obtained.

D. Feature Selection

The focus of this study is to filter out the predictors that are not relevant and insignificant to make predictions in the future outcome and want to keep the remaining predictors who give the most impact to the response. Feature selection was performed on the PUBG dataset by using the feature selection tool in Regression Learner App to choose any predictors to be included in the model. In this feature selection phase, RMSE and R-squared values were observed and compared between different models which have different features combinations. RMSE or root mean square error was beneficial in interpreting the accuracy of the model in prediction [40]. RMSE value measured the errors in the model by measuring the differences between the predicted value and observed value [40]–[42]. Statistically, RMSE is defined as shown in Equation 4:

$$\boldsymbol{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}} \quad (4)$$

where $\hat{y}_i$ is the predicted value, $\hat{y}_i$ is the observed value and n is the number of observations [40]–[42].

In Regression Learner App, all the results value such as R-squared values and RMSE values were computed automatically after training the models and were displayed on the results box as shown in Figure 3. This app also computed other statistical performances such as mean absolute error (MAE), mean squared error (MSE), prediction speed and the training time. However, this study used two statistical values which were RMSE and R-squared values to compare different trained models.

Initially, a regression tree model was trained by using all predictors which were 25 predictors to set a benchmark for the simplified model. The model was trained using all regression tree types, which are fine tree, medium tree and coarse tree. The highest performance of the regression model tree type which has the lowest RMSE value was selected to train models afterwards. In identifying which features are the most significant, the new regression tree model was trained by removing the features one-by-one, from the weakest correlated variable which has the least value of individual R-squared until the strongest correlated variable which has the most value of individual R-squared. A new benchmark was set when the model obtained the smallest value of RMSE after removing one predictor. However, when the model had a higher RMSE value after removing one predictor, the predictor was kept in the next round. These rules were followed for every predictor until it met the stopping criteria which were removing predictor with

the highest R-squared value. The model with the lowest RMSE value and its features set was recorded as the most significant features set. All the steps during data pre-processing, feature analysis and feature selection were performed for all dataset which is dataset 1 until dataset 5.

Results	
RMSE	0.11349
R-Squared	0.87
MSE	0.01288
MAE	0.082644
Prediction speed	~52000 obs/sec
Training time	7.2292 sec

Fig. 3. The results box displayed the computed statistical performance obtained after training the models.

E. Regression Models

All the datasets and their set of significant features were trained using the regression tree model. The best regression tree model with the highest performance was exported to the workspace.

Every dataset may result differently and not precise because of different instances in all datasets. To get a better comparison, all datasets were trained using the set of features from other datasets obtained during the feature selection phase earlier. For example, new regression tree models of dataset 2 until dataset 5 were trained using the significant features set from dataset 1. This method was repeated by using another significant feature set from dataset 2, dataset 3, dataset 4 and dataset 5.

Another comparison made was by comparing the regression tree models with another regression technique which were the linear regression and support vector machine (SVM). Every dataset and its own set of significant features were trained using these two regression models instead of the regression tree model. All linear regression types which are linear, interactions linear, robust linear and stepwise linear were trained and the best-performed model was selected. Similarly, all SVM model types which are linear SVM, quadratic SVM, cubic SVM, fine Gaussian SVM, medium Gaussian SVM and coarse Gaussian SVM were trained and was selected as the best model.

The performances of models were recorded in RMSE and its training time. All these comparison steps were performed by using the same app which is the Regression Learner App in MATLAB as this app provides various regression techniques to be trained. This method also used the same validation scheme which was 25 per cent of holdout validation.

IV. EXPERIMENT AND RESULTS

A. Feature Analysis Results

Figure 4 – Figure 8 shows the R-squared values of the regression tree model trained using individual predictors versus response (winPlacePerc) obtained from dataset 1, dataset 2, dataset 3, dataset 4 and dataset 5. From the figures, it was clearly shown that 'walkDistance' attribute has obtained the highest R-squared value which ranged between 0.7 and 0.8 while the 'killPlace' attribute ranked as the second-highest R-squared value, ranged in between 0.6 and 0.7. It means that 'walkDistance' explained about 70-80% of the variance in the model while 'killPlace explained 60-70%. In most statistical analysis, it was explained that the closest the R-squared value to 1, the better the relationship between the independent variable and the dependent variable (response) [38], [39]. Some researchers use the rule of thumb to interpret the R-squared value. According to the rule, R-squared higher than 0.3 is considered as a strong relationship, R-squared value between 0.1 and 0.2 is moderate, while less than 0.1 is a weak relationship [43]. In other words, a high R-squared value indicates that the predictor provides high information to the response while zero or low R-squared value means that they have no or less information.

There were few inconsistencies in these results where some of the predictors have a weak relationship, meaning they have R-squared value higher than 0 in the certain dataset, but have no relationship or zero R-squared value

in other datasets. For example, 'winPoints' and 'rankPoints' attributes do not correlate dataset 2 but they appeared to have some or least significant to the response in dataset 1. These dissimilarities may cause by the missing values hence effecting inaccurate results [44]. According to these figures, there are also negative R-squared values. The 'matchDuration' attribute recorded negative R-squared value (-0.1 to 0 range) in all datasets. However, most statistics books or previous researches obtained 0 to 1 R-squared value instead of negative values. The negative value may be due to the regression sum squared error greater than the mean value [38].

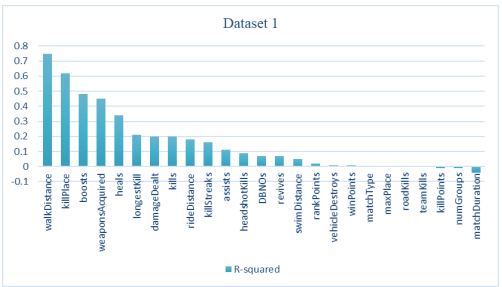

Fig. 4. R-squared value of individual predictors versus WinPlacePerc in dataset 1

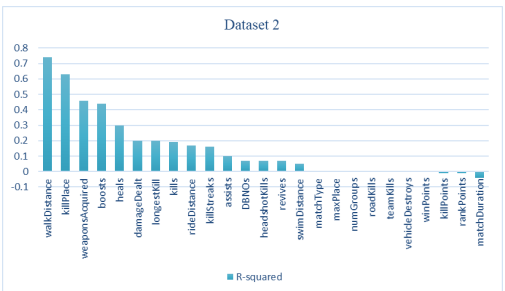

Fig. 5. R-squared value of individual predictors versus WinPlacePerc in dataset 2

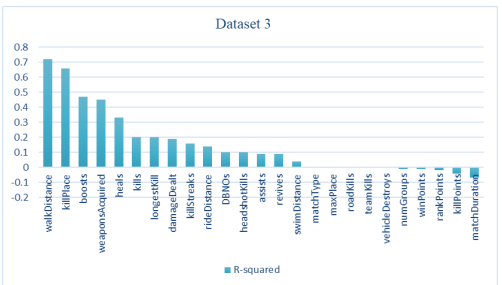

Fig. 6. R-squared value of individual predictors versus WinPlacePerc in dataset 3

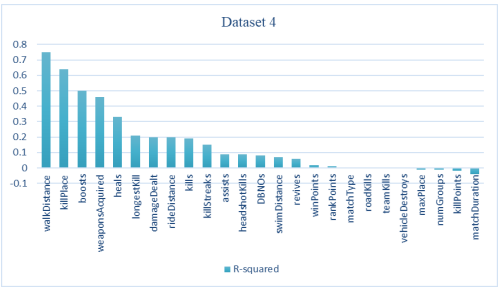

Fig. 7. R-squared value of individual predictors versus WinPlacePerc in dataset 4

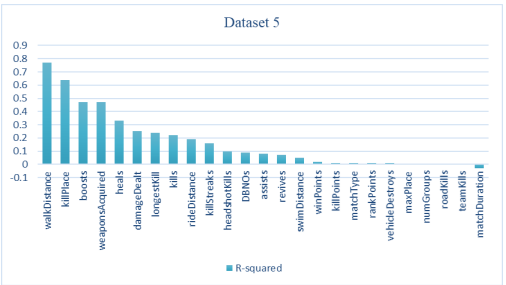

Fig. 8. R-squared value of individual predictors versus WinPlacePerc in dataset 5

B. Feature Selection Results

The set of significant features are shown in Table II below which were trained using the regression tree models. This method managed to reduce the dataset features from 25 to 7 (dataset 1, dataset 2), and from 25 to 8 (dataset 3, dataset 4, dataset 5). The similarities in all dataset are 'DBNOs', 'killPlace', 'kills', 'matchDuration', and 'walkDistance' attributes. The 'rideDistance' and 'boosts' attributes appeared to be significant in three datasets from five, hence it can be assumed to be a significant attribute in winning the PUBG game.

TABLE II. SIGNIFICANT FEATURE SET FROM THE FEATURE SELECTION PROCESS

Dataset	Significant features set
1	DBNOs, killPlace, kills, matchDuration, rideDistance, walkDistance, weaponsAcquired
2	DBNOs, killPlace, kills, matchDuration, rideDistance, walkDistance, weaponsAcquired
3	boosts, DBNOs, killPlace, kills, matchDuration, maxPlace, numGroups, walkDistance
4	boosts, DBNOs, killPlace, kills, longestKill, matchDuration, numGroups, walkDistance
5	boosts, DBNOs, killPlace, kills, matchDuration, rideDistance, walkDistance, weaponsAcquired

Surprisingly, some of the strong correlated features during feature analysis such as 'heals' attribute was not included in the significant features set while some of the weakly

correlated features such as 'DBNOs' has better significant to the model. On top of that, the 'matchDuration' attribute which has a negative R-squared value in all datasets during feature analysis was found to be in the significant features list. The combination of predictors that are statistically significant but have low R-squared value is because the predictors have high variability around the regression line, while at the same time the high variability data may have a significant trend that provides some information about the response.

The performances of every dataset trained using those significant features set are presented in Table III. The results of applying the feature selection method indicate that this method can reduce the datasets dimensionality as well as the models' performance. By comparing the models before applying and after applying feature selection, there are decrements in the RMSE values. According to [41], there is no absolute value or specific range to determine the good or bad RMSE value. However, in statistical books, the lower the RMSE value indicates the better fit or better prediction to the model [40], [41]. Therefore, the RMSE value from the original model (all features) was set as the upper limit. By referring to Table III, the RMSE values for all datasets are lower than the upper limit after applying the feature selection method indicated that the models have been improved with a lesser number of features. This means some of the features are not giving any impact in determining the game results. While the R-squared value did not improve much and cannot be compared and interpreted very well.

TABLE III. PERFORMANCE RESULTS FROM THE FEATURE SELECTION PROCESS

Dataset	All features		After feature selection	
	RMSE	R-squared	RMSE	R-squared
1	0.11659	0.86	0.11323	0.87
2	0.12179	0.85	0.11700	0.86
3	0.11631	0.85	0.11414	0.86
4	0.11628	0.85	0.11534	0.86
5	0.11910	0.85	0.11383	0.86

C. Regression Tree Models Results

From the models in the feature selection method, a decision tree regression diagram was illustrated. A full decision tree diagrams were formulated from dataset 1, dataset 2, dataset 3, dataset 4, and dataset 5. To get a clearer view, the decision tree model was pruned to 110 levels by setting the pruning level in the regression tree viewer in MATLAB. The example of a simplified decision tree from dataset 3 is illustrated in Figure 9. Based on the figure, the 'walkDistance' attribute is on the topmost node which is the root node, indicates the most significant predictor in the dataset [45].

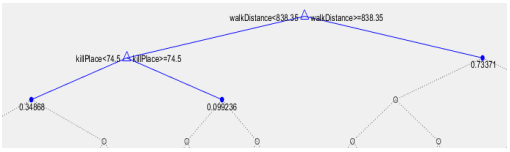

Fig. 9. Simplified decision tree diagram with pruning level of 110 in dataset 1

Another comparison was made to justify the results in the feature selection phase by comparing the model performances when training the significant features set from other datasets. The model performances in terms of RMSE value are presented in the lines graph shown in Figure 10 and the RMSE values are tabulated in Table IV, for all datasets.

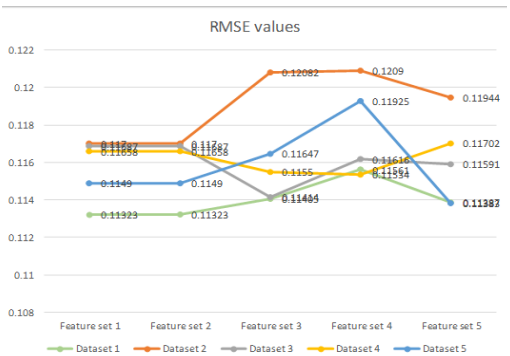

Fig. 10. Line graph of comparison between all datasets using every significant feature set

TABLE IV. RMSE VALUES OF ALL DATASETS USING EVERY SIGNIFICANT FEATURE SET

Dataset	RMSE values				
	Feature set 1	Feature set 2	Feature set 3	Feature set 4	Feature set 5
1	0.11323	0.11323	0.11405	0.11561	0.11387
2	0.11700	0.11700	0.12082	0.1209	0.11944
3	0.11687	0.11687	0.11414	0.11616	0.11591
4	0.11658	0.11658	0.11550	0.11534	0.11702
5	0.11490	0.11490	0.11647	0.11925	0.11383

When all datasets trained using feature set 1 and feature set 2, they will get the same results of RMSE values because of the same variables in those feature set. However, on

average, all datasets obtained higher RMSE value when training with features set from other datasets rather than when training with their own significant feature set. For example, dataset 1 only obtained the best model when training with its own feature set, but the RMSE value became increases when trained with other models. This happens to the other datasets where they only became the most fitted model when trained with their own feature set.

The increments in the RMSE value means that some of the predictors are happened to be significant in a certain dataset only. It may cause by an unbalanced dataset where some of the datasets may contain stats from bots (robots, character controlled by computer), AFK (away from keyboard) players, hackers and cheaters players. These stats will disrupt the model in terms of accuracy. For instances, the cheaters might kill the enemies a lot by headshot which will make the dataset appeared to have 'headshotKills' as the significant attribute in winning the game. However, when this attribute is trained using other datasets that have not cheater stats, it will appear insignificant or less significant to the model. Therefore, it is wise to remove the bad stats that can interrupt the model accuracy.

Another comparison was made between regression models which are the regression tree, linear regression and SVM model using the significant feature set in each dataset, where the models' performances were tabulated in Table V below in RMSE values. Based on the results, SVM models achieved lower RMSE values in all datasets than the regression tree and linear regression model. Some linear regression models have better performances than the regression tree model in dataset 2, dataset 3 and dataset 5.

TABLE V. RMSE VALUES OBTAINED FROM THREE REGRESSION MODELS IN EACH DATASET

Dataset	RMSE values		
	Regression tree	Linear regression	Support vector machines (SVM)
1	0.11323	0.11888	0.11158
2	0.11700	0.11496	0.10431
3	0.11414	0.10625	0.09372
4	0.11534	0.11777	0.10041
5	0.11383	0.10865	0.10052

Table VI shows the comparison of training speed between the three regression models. SVM model has the longest training time among those three models while the decision tree is the fastest prediction model. Therefore, it can be concluded that the SVM model has the most accuracy in predictions between these models, however, in terms of training time, the regression tree model wins in this comparison. A regression tree is more flexible in training large data due to faster prediction speed while training using the SVM model is very time-consuming.

TABLE VI. THE TRAINING TIME OF THREE REGRESSION MODELS IN EACH DATASET

Dataset	Training time (s)		
	Regression tree	Linear regression	Support vector machines (SVM)
1	8.50	150.78	504.53
2	8.33	143.21	420.15
3	9.10	191.20	628.31
4	7.97	155.92	554.20
5	8.94	166.39	541.11

V. CONCLUSION

This research has managed to accomplish the desired objective. The decision tree regression has been successfully modelled during the feature analysis and feature selection process where all the datasets were trained using the regression tree model to find the list of significant features using the Regression Learner App in MATLAB. The tree regression model using the RMSE value was later achieved during the feature selection process that assists in determining the most important attributes of the player's performance in PUBG matches. Several regression models based on the RMSE performance was compared and fully accomplished during the regression model phase where three regression models which are regression tree, linear regression and support vector machine (SVM) were trained.

After investigating the regression tree technique to build the prediction model using the different combination of features from feature selection methods, the important attributes in winning the PUBG games are 'boosts', 'DBNOs', 'killPlace', 'kills', 'matchDuration', 'rideDistance' and 'walkDistance', where 'walkDistance' denoted as the most important

attributes among them. After reducing some of the variables, there was a decrement in the RMSE value, meaning the model is improved. Although more features in the model will result in a higher performance theoretically, this study proves that the reduction in features causes a more effective model and better performances. This research has analysed and discussed the performances of different significant features sets on a dataset.

This study also has trained those significant features sets using three different regression models. The support vector machine (SVM) has the highest performance and better prediction model than the regression model and linear regression model. However, it is faster to train the model using a decision tree model and it is also easier to interpret the model.

REFERENCES

[1] T. Wilde, "PlayerUnknown's Battlegrounds gets its own company, PUBG Corp | PC Gamer," 2017. [Online]. Available: https://www.pcgamer.com/playerunknowns-battlegrounds-gets-its-own-company-pubg-corp/. [Accessed: 09-Oct-2019].

[2] T. H. Jr, "'PUBG' creator went from welfare to making a billion-dollar video game," 2019. [Online]. Available: https://www.cnbc.com/2019/04/26/pubg-creator-went-from-welfare-to-making-a-billion-dollar-video-game.html. [Accessed: 09-Oct-2019].

[3] C. Carter, "Understanding Playerunknown's Battlegrounds," 2017. [Online]. Available: https://web.archive.org/web/20170609175706/https://www.polygon.com/playerunknowns-battlegrounds-guide/2017/6/9/15721366/pubg-how-to-play-blue-wall-white-red-circle-map-weapon-vehicle-inventory-air-drop. [Accessed: 03-Jul-2020].

[4] A. Willingham, "What is eSports? A look at an explosive, billion-dollar industry - CNN," 2018. [Online]. Available: https://edition.cnn.com/2018/08/27/us/esports-what-is-video-game-professional-league-madden-trnd/index.html. [Accessed: 09-Oct-2019].

[5] M. R. Llorens, "eSport Gaming : The Rise of a New Sports Practice eSport Gaming : The Rise of a New Sports Practice," *Sport. Ethics Philos.*, vol. 1321, no. September, pp. 1–13, 2017.

[6] P. Tassi, "2012: The Year of eSports," 2012. [Online]. Available: https://www.forbes.com/sites/insertcoin/2012/12/20/2012-the-year-of-esports/#18d2fda17e11. [Accessed: 03-Jul-2020].

[7] Newzoo, "Global eSports Market Report," 2018.

[8] L. Maximilians, M. Schubert, and T. Mahlmann, "Esports Analytics Through Encounter Detection," pp. 1–18, 2016.

[9] D. Y. Aleksandr Semenov, Peter Romov, Sergey Korolev and and K. Neklyudov, "Performance of Machine Learning Algorithms in Predicting Game Outcome from Drafts in Dota 2," vol. 2, no. December 2018, pp. 305–313, 2017.

[10] L. Tagliaferri, "An Introduction to Machine Learning | DigitalOcean," 2017. [Online]. Available: https://www.digitalocean.com/community/tutorials/an-introduction-to-machine-learning#targetText=Introduction,of artificial intelligence (AI).&targetText=Because of this%2C machine learning,has benefitted from machine learning. [Accessed: 09-Oct-2019].

[11] N. A. Fadi Thabtah, Li Zhang, "NBA Game Result Prediction Using Feature Analysis and Machine Learning," *Ann. Data Sci.*, vol. 6, no. 1, pp. 103–116, 2019.

[12] M. Gupta, "What is Machine Learning?" [Online]. Available: https://www.geeksforgeeks.org/ml-machine-learning/.

[13] S. Ray, "7 Regression Techniques you should know!," 2015. [Online]. Available: https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/.

[14] M. R. Segal, "Machine learning benchmarks and random forest regression. *Center for Bioinformatics and Molecular Biostatistics, UC San Francisco, USA.*," 2004.

[15] A. J. L. George A. F. Seber, *Linear regression analysis*, Second Edi. 2012.

[16] P. Y. Hao, "New support vector algorithms with parametric insensitive/margin model," *Neural Networks*, vol. 23, no. 1, pp. 60–73, 2010.

[17] Y. Yang, T. Qin, and Y.-H. Lei, "Real-time eSports Match Result Prediction," no. Nips, pp. 1–9, 2016.

[18] A. Summerville, M. Cook, and B. Steenhuisen, "Draft-Analysis of the ancients: Predicting draft picks in DotA 2 using machine learning," *AAAI Work. - Tech. Rep.*, vol. WS-16-21-, no. Godec, pp. 100–106, 2016.

[19] A. Katona et al., "Time to die: Death prediction in dota 2 using deep learning," *IEEE Conf. Comput. Intell. Games*, CIG, vol. 2019-Augus, 2019.

[20] M. Manjunath Mamulpet, "Pubg Winner Placement Prediction Using Artificial Neural Network," *Int. J. Eng. Appl. Sci. Technol.*, vol. 3, no. 12, pp. 107–118, 2019.

[21] W. Wei, X. Lu, and Y. Li, "PUBG : A Guide to Free Chicken Dinner," pp. 3–8, 2018.

[22] Kaggle, "PUBG Finish Placement Prediction (Kernels Only) | Kaggle," 2019. [Online]. Available: https://www.kaggle.com/c/pubg-finish-placement-prediction/data. [Accessed: 09-Oct-2019].

[23] G. Kumar, "Machine Learning for Soccer Analytics," *KU Leuven, MSc thesis*, no. SEPTEMBER 2013, pp. 1–2, 2013.

[24] B. C. Alamar, *Sports Analytics: A Guide for Coaches, Managers, and Other Decision Makers.* Columbia University Press, New York, 2013.

[25] L. Steinberg, "CHANGING THE GAME: The Rise of Sports Analytics," Forbes. [Online]. Available: https://www.forbes.com/sites/leighsteinberg/2015/08/18/changing-the-game-the-rise-of-sports-analytics/#1db846eb4c1f. [Accessed: 21-Oct-2019].

[26] R. Baboota and H. Kaur, "Predictive analysis and modelling football results using machine learning approach for English Premier League," *Int. J. Forecast.*, vol. 35, no. 2, pp. 741–755, 2019.

[27] R. Lamsal and A. Choudhary, "Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning," no. September, 2018.

[28] J. Hamari and M. Sjöblom, "What is eSports and why do people watch it?," *Internet Res.*, vol. 27, no. 2, pp. 211–232, 2017.

[29] A. J. Brij Rokad, Omkar Acharya, Tushar Karumudi, "Survival of the Fittest in PlayerUnknown ' s BattleGrounds," 2019.

[30] A. E. Mohamed, "Comparative Study of Four Supervised Machine Learning Techniques for Classification," *Int. J. Appl. Sci. Technol.*, vol. 7, no. 2, pp. 5–18, 2017.

[31] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," pp. 249–268, 2007.

[32] W. H. J. C. R. Gentleman, "Supervised Machine Learning. In: Bioconductor Case Studies. Use R!," in *Bioconductor Case Studies*, Springer, New York, NY, 2008, pp. 121–136.

[33] W.-Y. Loh, "Classification and Regression Tree Methods," *Encycl. Meas. Stat.*, pp. 1–11, 2013.

[34] R. Timofeev, "Classification and Regression Trees (CART) Theory and Applications," 2005.

[35] R. Abielmona, R. Falcon, N. Zincir-Heywood, and H. *Abbass, Recent advances in computational intelligence in defense and security*, vol. 621. 2016.

[36] R. M. F. Tony Fischetti, Eric Mayor, *R: Predictive Analysis.* 2017.

[37] T. A. Reddy, *Applied Data Analysis and Modeling for Energy Engineers and Scientists.* 2011.

[38] J. Miles, "R Squared, Adjusted R Squared," *Wiley StatsRef Stat. Ref. Online*, no. 2, pp. 2–4, 2014.

[39] E. Kasuya, "On the use of r and r squared in correlation and regression," *Ecol. Res.*, vol. 34, no. 1, pp. 235–236, 2019.

[40] V. B. Kamble and S. N. Deshmukh, "Comparision Between Accuracy and MSE,RMSE by Using Proposed Method with Imputation Technique," *Orient. J. Comput. Sci. Technol.*, vol. 10, no. 04, pp. 773–779, 2017.

[41] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature," *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, 2014.

[42] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Clim. Res.*, vol. 30, no. 1, pp. 79–82, 2005.

[43] A. C. Acock, *A Gentle Introduction to Stata*, Second Edi. Stata Press, 2008.

[44] H. Kang, "The prevention and handling of the missing data," *Korean J. Anesthesiol.*, vol. 64, no. 5, pp. 402–406, 2013.

[45] M. Umano et al., "Fuzzy decision trees by fuzzy ID3 algorithm and its application to diagnosis systems," *IEEE Int. Conf. Fuzzy Syst.*, vol. 3, no. 1, pp. 2113–2118, 1994.