IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023

DOI: 10.17148/IJARCCE.2023.12686

Exploring Data Mining and Machine Learning Techniques to Enhance the Prediction of Marathon Running Times

Brijal M. Panwala[1], Dr. Sanjay H. Buch[2]

Research Scholar, College Of Computer Applications, Bhagwan Mahavir University, Surat, India[1]

Dean, College Of Computer Applications, Bhagwan Mahavir University, Surat, India[2]

Abstract: Machine learning is a powerful modelling approach that has applications in multiple industries. Advanced data analysis techniques, such as data mining, which place an emphasis on exploration and the creation of new insights, are becoming more and more effective tools for analysing the performance data of elite athletes and assisting in the crucial decision-making that is necessary to succeed. The aim is to enhance the precision of marathon running time prediction by leveraging data mining domains and machine learning. We introduce a model that utilizes performance analysis through data mining methods and employs regression techniques such as Linear Regression, Random Forest, K-nearest Neighbor, Support Vector Regression, and Decision Tree.

Keywords: Performance prediction, running, athletes, Smart watches, supervised machine learning algorithms.

I. INTRODUCTION

As far as we are aware, the phrase "exercise snacks" was introduced by Dr. Howard Hartley in a 2007 article published in a weekly news magazine (https://www.newsweek.com/exercise-snack-plan-96095) [1].

Exercise is physical activity that is done in order to become healthier and stronger. It can take many forms, including aerobic activities like running or cycling, strength training with weights, or flexibility exercises like yoga or Pilates. Exercise is important for maintaining good health, as it can help to reduce the risk of numerous health problems, including heart disease, obesity, type 2 diabetes, and certain types of cancer. It can also improve mood, increase energy levels, and improve sleep quality. It is recommended that adults get at least 150 minutes of moderate-intensity exercise or 75 minutes of vigorous-intensity exercise each week, along with muscle-strengthening activities on at least two days per week. One of the most well-liked forms of exercise worldwide is running. The marathon is a highly recognized long-distance race that has seen a rise in participation across all ages and genders, making it one of the most symbolically significant events globally [2, 3]. For example, the TCS New York City Marathon, which is sponsored by Tata Consultancy Services and organized by New York Road Runners (NYRR), is the leading event in the world and the largest marathon in existence. Since its inception in 1970, over 1.2 million individuals have completed the race [3].

Machine learning is a potent modelling technique that can be used in a variety of fields such as education, sports, health, finance, transportation, marketing, computer, security, robotics etc [4].

Healthcare: Predictive analysis of patient data, medical imaging analysis, drug discoveryFinance: Fraud detection, stock market predictions, loan approval predictions

Retail: Customer behaviour prediction, product recommendationTransportation: Traffic prediction, autonomous vehicles

Marketing: Predictive analysis of customer buying behaviour, target advertising

Natural Language Processing: Sentiment analysis, language translation, chatbotsComputer Vision: Image and video analysis, object recognition

Robotics: Manipulation and control of robots, autonomous navigation

Education: Adaptive learning systems, student performance predictionsSecurity: Network intrusion detection, facial recognition.

These are just a few examples; machine learning is rapidly being adopted by many industries for various purposes.

For performance outcomes in sports, strategic decisions are essential [5]. For that Data mining, a subfield of computer science and artificial intelligence is one such cutting-edge technology. Sports data analysis techniques can be used to examine sports performance data like running performance [6].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023

DOI: 10.17148/IJARCCE.2023.12686

Machine learning can be used to predict the performance of runners in a variety of ways. For example:

1. Training performance prediction: Machine learning algorithms can analyse a runner's training data, such as pace,distance, and heart rate, to predict future performance and make recommendations for improvement.
2. Race performance prediction: Machine learning can also be used to predict a runner's performance in a specificrace based on factors such as their past race times, weather conditions, and course difficulty.
3. Injury prediction: Machine learning can be used to analyse a runner's training data and identify factors that increasethe risk of injury, allowing for proactive measures to be taken to prevent injury.

However, it's important to note that machine learning predictions are only as accurate as the data they are trained on. Therefore, the quality and quantity of data used to train the algorithms are critical for accurate performance predictions. This article focuses on the analysis of marathon-running performance data using data mining techniques.We discuss various regression methods to improve running time accuracy. To accomplish our objectives, we thoroughly reviewed data mining techniques and sports performance analysis in our earlier papers. We will specifically explain what data mining is, what data mining techniques are available, and how these methods can be effectively used in running performance time. In our previous publication, we delved into the basics of data analysis; the primary focus of our previous paper was on uncovering valuable insights from historical data, including definitions and explanations of pre-processing, dataset division, dimensionality reduction, and feature significance.

II. METHODOLOGY

Here are some of the commonly used regression algorithms in machine learning for predicting the performance of runners:

Table 1: Machine learning regression algorithm

Regression Algorithm	Description
Simple Linear Regression	A basic regression algorithm that models the relationship between one dependent and one independent variable.
Multiple Linear Regression	Extension of simple linear regression, which models the relationship between a dependent variable and multiple independent variables.
Polynomial Regression	A regression algorithm that models the relationship between a dependent variable and an independent variable by fitting a polynomial equation to the data.
Support Vector Regression (SVR)	A regression algorithm that uses support vectors to model the relationshipbetween a dependent variable and independent variables.
Decision Tree Regression	A regression algorithm that builds a tree-based model to predict the outcome based on the input variables.
Random Forest Regression	An ensemble algorithm that uses multiple decision trees to make predictions.
Gradient Boosting Regression	An ensemble algorithm that builds multiple simple models to make predictions.

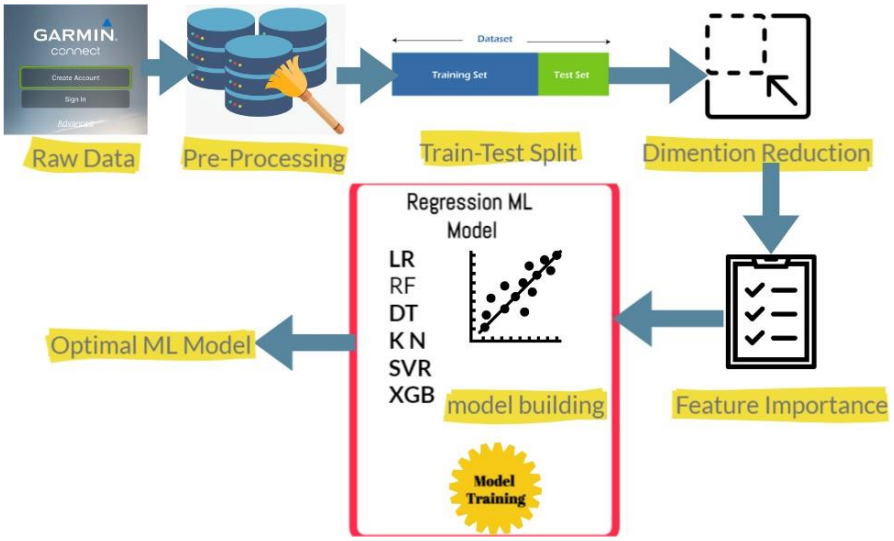


Fig. 1 Working model

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023

DOI: 10.17148/IJARCCE.2023.12686

The flowchart in Figure 1 illustrates the research methodology employed in this study. The primary goal of the machine learning models is to develop accurate models that can provide precise results for athlete running performance. Data cleaning is a step in the data processing process that removes invalid and irrelevant data. Collecting and getting ready the data for ML algorithms is the first step. Data reduction, cleansing, normalisation, and feature engineering are examples of discrete tasks that fall under the category of data preparation. In this study, data reduction refers to using a subset of the dataset instead of the entire dataset. Sampling and filtering are two of the most popular methods for reducing data. Data cleaning is a step in the data processing process that removes invalid and irrelevant data. Additionally, we standardised or normalised the variety of data features. The feature engineering step comes after the pre-processing step, where we take the raw features and turn them into some new features that we think might be usefulfor the predictive power. In addition, one-hot encoding is used in this step to encode categorical data so that it can befed into machine learning algorithms. The data is then fed to various ML models. Decision Tree, Random Forest, Logistic Regression, Support Vector Regression, and K-Nearest Neighbor are some of the algorithms used. Since these models function as a "black box," no additional information about how to construct them is needed. Below define algorithm for performance prediction of athletes.

Algorithm 1 Implementation of prediction Model using ML

Input: *ListML* of models (DT, RF, NB, LR, SVR, KNN) & raw dataset Output: Best model along performance and optimal architecture

1: Call PreProcessing
2: for every model in the *ListML* do
3: Call ListML model one by onea Calculate Accuracy(Acc), Model
4: if ListML$_{(i)}$.Acc larger than ListML$_{(i+1)}$.Acc then
5: select best ListML$_{(i)}$.Acc model,
6: Calculate Runners_perofmance
7: end if
8: end for
9: Return FinalModel, Runners_perofmance, Final.Acc

In this algorithm study, we compare the various ML models to find the model with the best performance. Pick the ultim ate best model, and then forecast the runner's performance. To check the accuracy of model first understands Performance metrics.

Performance metrics (error measurements) are critical components of several evaluation frameworks. A performance metric is a logical and mathematical construct that measures how close actual outcomes are to what was expected or forecasted. Root Mean Squared Error (RMSE) is a popular alternative to Mean Absolute Error (MAE) [7].

Performance metrics are used in machine learning regression studies to compare trained model predictions to actual (observed) data from the testing data set. The outcomes of these comparisons can have a direct impact on the decision-making process for picking the types of machine learning algorithms to use.
In the following sections, we describe and explore numerous metrics that may be used to calculate the prediction error of such a model [8].

Mean Absolute Error(MAE) is one of the simplest metrics, which measures the absolute difference between actual and predicted values, where absolute means taking a number as Positive. To understand MAE, let's take an example of Linear Regression, where the model draws a best fit line between dependent and independent variables. To measure theMAE or error in prediction, we need to calculate the difference between actual values and predicted values. But in order to find the absolute error for the complete dataset, we need to find the mean absolute of the complete dataset. The below formula is used to calculate

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |Y - Y'|$$

Here, Y is the Actual outcome, Y' is the predicted outcome, and N is the total number of data points [9].

The Mean Squared Error (MSE) is a prominent regression-related statistic that measures the average squared difference

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023

DOI: 10.17148/IJARCCE.2023.12686

between predicted and actual values. It accepts positive or negative values and is denoted by

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y - Y')^2$$

Here, Y is the Actual outcome, Y' is the predicted outcome, and N is the total number of data points [9].

R squared error is also known as the Coefficient of Determination, which is another useful statistic for evaluating Regression models. The R-squared measure allows us to compare our model's performance to that of a constant baseline. To get the constant baseline, we must take the mean of the data and draw a line through it. The R squared score will always be less than or equal to 1 without concerning if the values are too large or small.

$$R^2 = 1 - \frac{MSE(model)}{MSE(baseline)}$$

To avoid the problem of R square, adjusted R squared is employed, which always returns a lower number than R2. This is due to the fact that it adjusts the values of growing predictors and only indicates progress when there is a genuine improvement. The adjusted R squared may be calculated as follows

$$R^2_a = 1 - \left[\frac{n-1}{(n-k-1)} * (1 - R^2)\right]$$

Here, n is the number of observations, k denotes the number of independent variables and Ra2 denotes the adjusted R2 [9].

The Root Mean Squared Error (RMSE) is another often used measure of the discrepancies between expected values (sample or population values) and actual values. It is equal to the square root of MSE. RMSE, as contrast to MSE, gives an error measure in the same unit as the target variable. It is supplied by and accepts values in the range [0, +∞] [9].

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y - Y')^2}$$

The range of values for MAE, MSE, RMSE, R2 depends on the scale and range of the target variable. If the target variable is on a small range (e.g., 0 to 1), an RMSE value of 0 to 0.1 might be considered great, 0.1 to 0.2 as good, and values over 0.2 may imply a bigger prediction error relative to the variable's scale. For a target variable on a broader scale (e.g., 0 to 1000), an RMSE number less than 10 may be deemed excellent, 10 to50 as good, and values greater than 50 may suggest a significant prediction error relative to the variable's scale [10].We are currently focused on basic machine learning algorithms, ensemble approaches, and hyper parameter tuning strategies for our research study after understanding the above performance assessment.

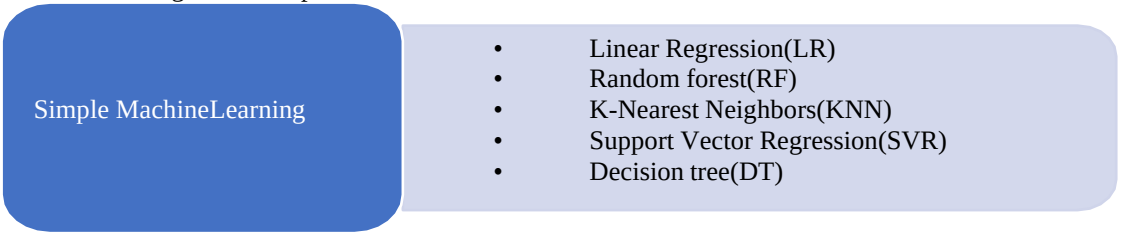


Fig. 2 Machine learning algorithm approach.

Linear Regression Algorithm

Regression is a supervised learning strategy. It may be used to model and predict continuous variables.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023

DOI: 10.17148/IJARCCE.2023.12686

This method has frequently been applied to numerous correlational researchers. Finding the best-fitted line between the independent variables (the cause or features) and the dependent variables, sometimes referred to as the "least squares regression line," in linear regression models enables us to determine the target variable (the effect or target). This method's ultimate objective is to fit a straight line to the given data set.

Linear regression has the benefit of being simple to grasp and straightforward to avoid overfitting through regularisation. Linear regression has the disadvantage of being unsuitable for dealing with non-linear connections. It is difficult to manage complicated patterns [11].

Accuracy

```
Train Score : 97.05% and Test Score : 97.14% using Linear Regression.
```

Actual and Predicted value

	Actual	Predicted
2677	0.474892	0.442741
1073	0.061693	0.061749
2231	0.050215	0.054022
3564	0.163558	0.162760
2747	0.034433	0.031133
...	...	...
2845	0.060258	0.055641
1680	0.170732	0.190436
780	0.153515	0.152331
2829	0.050215	0.038959
2921	0.093257	0.080011

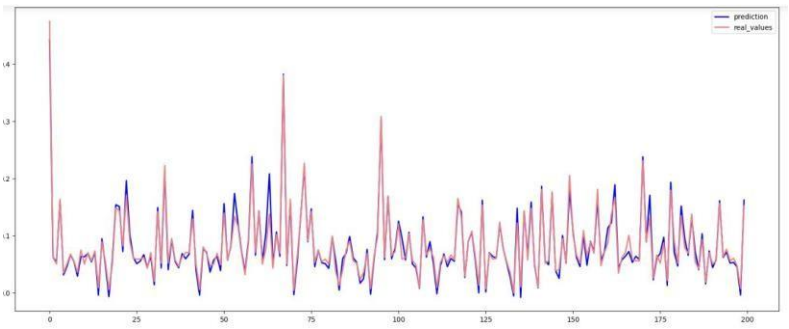


Fig. 3 Outcome of Linear regression.

Model Performance Evaluation

```
MAE 0.007705391112581709
MSE 0.0001500494463935615
RMSE 0.01224946718814992
R2 0.9714498299659755
R2 Score =  0.9714327348400328
```

Support Vector Regression (SVR)

The goal of SVR is to create a model that can accurately predict continuous output values based on a set of input features. It does this by finding a hyperplane in a high-dimensional space that maximally separates the output values from the input data points. This hyperplane is then used to make predictions for new input data.

In SVR, the choice of the kernel function is very important because it determines the shape of the hyperplane that is used to make predictions. Some common kernel functions used in SVR include linear, polynomial, radial basis function(RBF), and sigmoid.

SVR has several advantages over other regression algorithms. It works well with high-dimensional data and outliers, and it is less prone to overfitting. However, it can be computationally expensive for large datasets and requires careful selection of hyperparameters [12].

Accuracy

```
Train Score : 57.10% and Test Score : 51.05% using SVR Regression.
```

Actual and Predicted value

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023

DOI: 10.17148/IJARCCE.2023.12686

	Actual	Predicted
2677	0.474892	0.473692
1073	0.061693	0.102500
2231	0.050215	0.148954
3564	0.163558	0.170874
2747	0.034433	0.054052
...	...	...
2845	0.060258	0.084758
1680	0.170732	0.206272
780	0.153515	0.173154
2829	0.050215	0.108413
2921	0.093257	0.154975

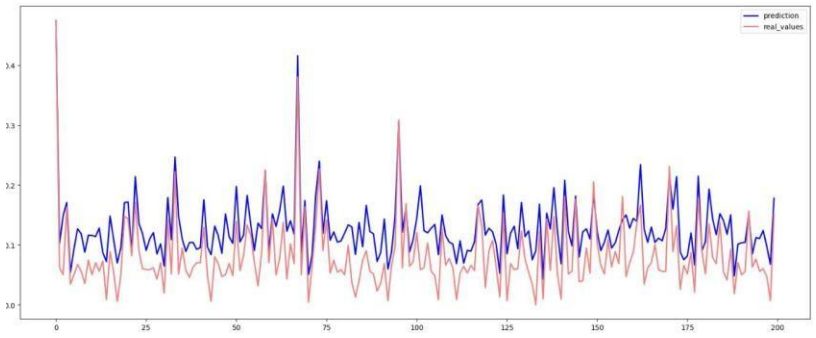


Fig. 4 Outcome of Support vector regression.

Model Performance Evaluation

```
MAE 0.04450171875041671
MSE 0.0025713117281861942
RMSE 0.050708103180716536
R2 0.8588198494754931
```

Random Forest Regression

For group learning, Random Forest employs a network of decision trees. This method frequently uses the bootstrap technique to create the randomly generated data sets that may later be utilised to train the ensemble of decision trees.

It does this by constructing a large number of decision trees and then combining their predictions to create a more accurate and robust model. In Random Forest regression, each decision tree is constructed using a random subset of the input features and arandom subset of the training data. This helps to reduce overfitting and increase the generalization ability of the model. The final prediction is then made by averaging the predictions of all the decision trees.

Random Forest regression has several advantages over other regression algorithms. It is effective in handling data with high dimensionality and outliers, and it can capture non-linear relationships between the input features and output values. It is also less prone to overfitting than other regression algorithms, and it can provide estimates of the importance of each input feature. However, for big datasets, Random Forest regression may be computationally costly, and it necessitates carefulselection of hyperparameters such as the number of decision trees and the maximum depth of each tree [13].

Accuracy

```
Train Score : 99.78% and Test Score : 99.34% using Random Forest Regression.
```

Actual and Predicted value

	Actual	Predicted
2677	0.474892	0.495237
1073	0.061693	0.061693
2231	0.050215	0.052869
3564	0.163558	0.163931
2747	0.034433	0.034433
...	...	...
2845	0.060258	0.060258
1680	0.170732	0.170775
780	0.153515	0.154720
2829	0.050215	0.050215
2921	0.093257	0.094032

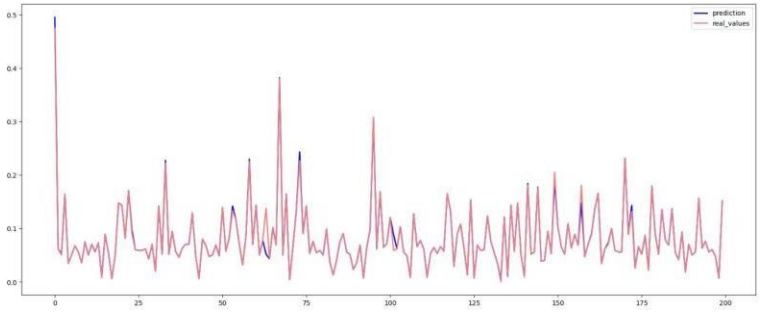


Fig. 5 Outcome of Random forest regression.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023

DOI: 10.17148/IJARCCE.2023.12686

Model Performance Evaluation

```
MAE 0.00128518516295184
MSE 3.467310144929478e-05
RMSE 0.005888386998940778
R2 0.9934140077618095
R2 Score =  0.9933987381704665
```

k-Nearest Neighbor Regression

The K-nearest neighbor technique is a basic yet effective algorithm. When the datasets are vast, high dimensional, or unpredictable, typical query processing techniques fail miserably to extract the needed data in the allotted time. In these cases, closest neighbor approaches are critical. The value of K is significant since it dictates the algorithm's accuracy and efficacy. The closest neighbor (NN) approach is extremely simple, highly efficient, and useful in pattern recognition, text classification, object identification, Stock Market Forecasting, healthcare and other fields. It has numerous advantages, such as simplicity, robustness to noisy training data, improved query time and memory needs, and so on, but it also has drawbacks, such as computational complexity, memory limitation, and high cost in algorithm execution. In k-NN, the model uses the entire training dataset as its knowledge base. When a new input is given, it compares it with all the training data points and selects the k-nearest data points based on a distance metric. The distance metric canbe Euclidean, Manhattan, or other distance measures. The k-nearest data points are used to determine the output value by taking the average or median value of the k-nearest neighbors. One advantage of k-NN is that it can handle complex decision boundaries and nonlinear relationships between input features and output values. Another advantage is that it can easily incorporate new training data points without retraining the model.

However, k-NN can be sensitive to the choice of distance metric and the value of k. A small value of k can lead to over-fitting, while a large value of k can lead to under-fitting. Also, the algorithm can be computationally expensive forlarge datasets. However, because they are distance-based, they perform better with smaller number of input variables, need feature scaling or normalisation, and are sensitive to outliers in the data set[78].

Accuracy

```
Train Score : 77.27% and Test Score : 65.90% using KNN Regression.
```

Actual and Predicted value

	Actual	Predicted
2677	0.474892	0.253945
1073	0.061693	0.091822
2231	0.050215	0.121377
3564	0.163558	0.096700
2747	0.034433	0.051937
...	...	...
2845	0.060258	0.059684
1680	0.170732	0.070875
780	0.153515	0.101865
2829	0.050215	0.084075
2921	0.093257	0.066858

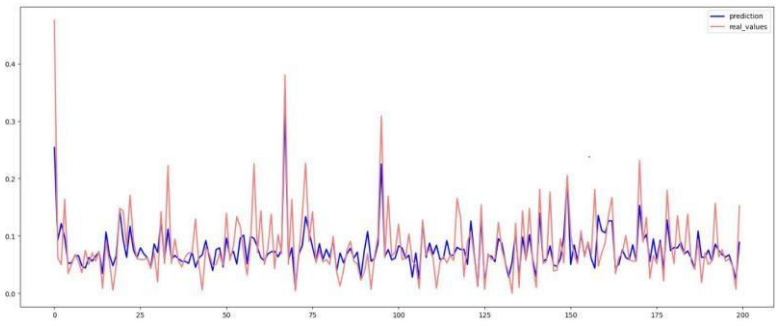


Fig. 6 Outcome of K-nearest neighbor regression.

Model Performance Evaluation:

```
MAE 0.02626838673013853
MSE 0.0017911235342580997
RMSE 0.0423216674323933
R2 0.6732881705664582
R2 Score =  0.658995736624029
```

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023

DOI: 10.17148/IJARCCE.2023.12686

Decision Tree Regression

Decision Tree is a Supervised Machine Learning solution to classification and regression issues that solves them by continually separating data depending on a certain parameter. The purpose of Decision Tree regression is to develop a model that can predict continuous output values reliably based on a collection of input characteristics. It accomplishes this by recursively splitting the input space into smaller sections based on the input feature values. The partitioning is done in such a way that the variance of the output values within each zone is minimised.

The method finds the input feature that best separates the output values at each phase of the partitioning procedure. This is accomplished by estimating the variance reduction that would follow from separating the data depending on the values of that characteristic. The partitioning procedure is repeated until the data is entirely partitioned or a stopping requirement is reached. The decision tree is traversed from the root node to a leaf node that corresponds to the area holding the new input to make the final prediction for a new input. The expected output is then the output value for that location.

Decision Tree regression has several advantages over other regression algorithms. It is easy to interpret and visualize, and it can handle both numerical and categorical input features. It can also capture non-linear relationships between the input features and output values.

However, Decision Tree regression can be sensitive to small changes in the input data and prone to over-fitting. It can also be biased towards input features with many levels or categories. Therefore, care must be taken when selecting hyper-parameters such as the maximum depth of the tree or the minimum number of data points required to split a node[14].

Accuracy:

```
Train Score : 100.00% and Test Score : 98.07% using Decission Tree Regression.
```

Actual and Predicted value:

	Actual	Predicted
2677	0.474892	0.494978
1073	0.061693	0.061693
2231	0.050215	0.050215
3564	0.163558	0.163558
2747	0.034433	0.034433
...	...	...
2845	0.060258	0.060258
1680	0.170732	0.170732
780	0.153515	0.157819
2829	0.050215	0.050215
2921	0.093257	0.093257

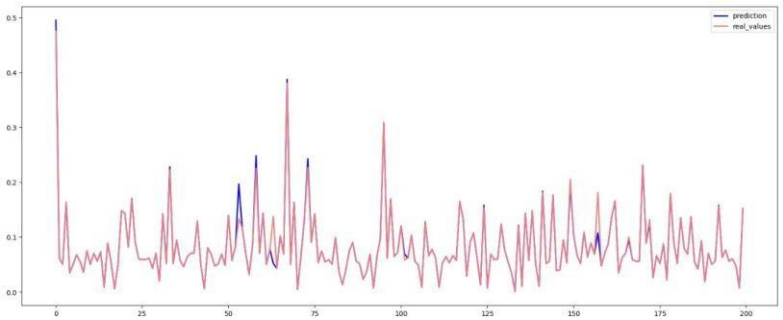


Fig. 7 Outcome of Decision tree regression.

Model Performance Evaluation:

```
MAE 0.0017748909952796457
MSE 0.00010125312462092704
RMSE 0.010062461161213346
R2 0.9807393282170167
```

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023

DOI: 10.17148/IJARCCE.2023.12686

Comparison of simple ML Algorithms

Prediction results of simple ML algorithms

Train Score : 97.05% and Test Score : 97.14% using Linear Regression.
Train Score : 99.78% and Test Score : 99.34% using Random Forest Regression.
Train Score : 77.27% and Test Score : 65.90% using KNN Regression.
Train Score : 57.10% and Test Score : 51.05% using SVR Regression.
Train Score : 100.00% and Test Score : 98.07% using Decission Tree Regression.

	model	Train_Score	Test_Score
0	lr_acc	97.050960	97.143273
1	rf_acc	99.783711	99.339874
2	knn_acc	77.265660	65.899574
3	svr_acc	57.098575	51.045908
4	DT_acc	100.000000	98.072286

Fig. 8 Result of simple ML algorithm.

Prediction and Performance Assessment results of simple ML algorithms

```
Linear Regression
Accuracy =  97.14327348400327
Mean Absolute Error =  0.007705391112581709
Mean Squared Error =  0.0001500494463935615
Root Mean Squared Error =  0.01224946718814992
R-Squared =  0.9714327348400328

Support Vector Regression
Accuracy =  51.04590805662554
Mean Absolute Error =  0.04450171875041671
Mean Squared Error =  0.0025713117281861942
Root Mean Squared Error =  0.050708103180716536
R-Squared =  0.5104590805662554

Random Forest Regression
Accuracy =  99.33987381704665
Mean Absolute Error =  0.00128518516295184
Mean Squared Error =  3.467310144929478e-05
Root Mean Squared Error =  0.005888386998940778
R-Squared =  0.9933987381704665

K-Nearest Neighbors Regression
Accuracy =  65.8995736624029
Mean Absolute Error =  0.02626838673013853
Mean Squared Error =  0.0017911235342580997
Root Mean Squared Error =  0.0423216674323933
R-Squared =  0.658995736624029

Decission Tree Regression
Accuracy =  98.07228555063476
Mean Absolute Error =  0.0017748909952796457
Mean Squared Error =  0.00010125312462092704
Root Mean Squared Error =  0.010062461161213346
R-Squared =  0.9807228555063476
```

Fig. 9 Performance evaluation outcome of ML Algorithm.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023

DOI: 10.17148/IJARCCE.2023.12686

Table 2: Experimental compression of ML algorithms

Models	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	R-Squared(R2)	Accuracy
Linear Regression	0.0077	0.00015	0.0122	0.9714	97.1432
Support VectorRegression	0.0445	0.0025	0.0507	0.5104	51.0459
Random Forest Regression	0.0012	3.4673	0.0058	0.9933	99.3398
K-Nearest Neighbors Regression	0.0262	0.0017	0.0423	0.6589	65.8995
Decission Tree Regression	0.0017	0.0001	0.0100	0.9807	98.0722

Table 3: Train and Test experimental Comparisons of Different ML Algorithms

ML Algorithm	Train Accuracy score	Test Accuracy score	Training time Root Mean Squared Error (RMSE)	Test time Root Mean Squared Error (RMSE)
Linear Regression(LR)	97.0509	97.1432	1.3067	0.0122
Random Forest Regression(RF)	99.7837	99.3398	3.5389	0.00588
K-Nearest Neighbors Regression(KNN)	77.2656	65.8995	3.6282	0.04232
Support Vector Regression(SVR)	57.0985	51.0459	4.9841	0.0507
Decission Tree Regression(DT)	100.0000	98.0722	2.3195	0.0100

Table 2 and 3 shows that Random Forest has the greatest accuracy when comparing the aforementioned simple machine learning models. The train score is 99.783711, the test score is 99.33, and the RMSE value is 0.0058.

Performance Assessment of simple ML algorithms

RMSE for ML model performance evaluation

	Error	rmse_list
0	lr_rmse	0.012249
1	rf_rmse	0.005888
2	knn_rmse	0.042322
3	svr_rmse	0.050708
4	DT_rmse	0.010062

Accuracy of ML model

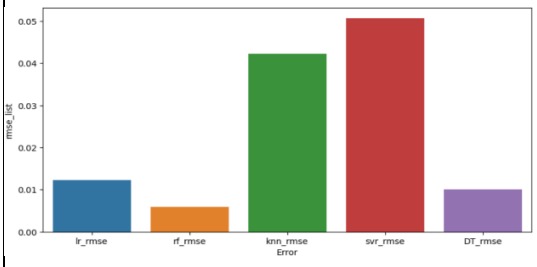
	model	accuracy
0	lr_acc	97.143273
1	rf_acc	99.339874
2	knn_acc	65.899574
3	svr_acc	51.045908
4	DT_acc	98.072286

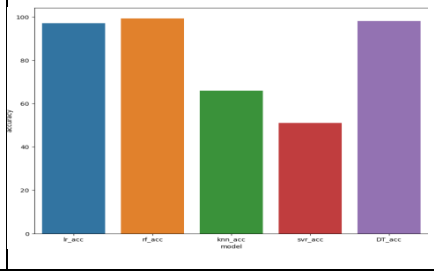
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 6, June 2023

DOI: 10.17148/IJARCCE.2023.12686





The accuracy result and graph depicts the highest level of Random forest accuracy, whilst the RMSE graph depicts the lowest error rate. We compute MAE, MSE, and R2 in addition to RMSE. When compared to other models, the overall error level is minimal; Random Forest is best suited when employing simple ML methodologies.

Comparison between accuracy and performance evaluation of machine learning algorithm

	model	Train_acc_Score	Test_acc_Score	Train_rmse_list	Test_rmse_list
0	Linear Regression(LR)	97.050960	97.143273	1.306770e-02	0.012249
1	Random forest(RF)	99.783711	99.339874	3.538965e-03	0.005888
2	K-Nearest Neighbors(KNN)	77.265660	65.899574	3.628273e-02	0.042322
3	Support Vector Regression(SVR)	57.098575	51.045908	4.984191e-02	0.050708
4	Decision tree(DT)	100.000000	98.072286	2.319525e-17	0.010062

Fig. 10 Outcome of simple ML algorithm.

III. RESULT AND DISCUSSION

The entire model building table is defined here to show you the performance of each model in relation to our study direction. It can include the practical outcomes of basic machine learning algorithms.

Table 4: Model Comparison table

Model	**Model**	**Train_acc_Score**	**Test_acc_Score**	**Train_rmse_list**	**Test_rmse_list**
Machine Learning Model	Linear Regression(LR)	97.050960	97.143273	1.306770e-02	0.012249
	Random forest(RF)	99.783711	99.339874	3.538965e-03	0.005888
	K-Nearest Neighbors(KNN)	77.265660	65.899574	3.628273e-02	0.042322
	Support Vector Regression(SVR)	57.098575	51.045908	4.984191e-02	0.050708
	Decision tree(DT)	100.000000	98.072286	2.319525e-17	0.010062

Our study focused on utilizing different machine learning (ML) algorithms to predict the performance of runners. We employed fundamental ML techniques such as Decision Tree, Random Forest, Logistic Regression, Support Vector Regression, and K-Nearest Neighbors. To evaluate their performance, we employed metrics including MAE, MSE, RMSE, and R-squared. Among the algorithms examined, Random Forest demonstrated the highest accuracy, achieving a remarkable accuracy rate of 99.33%. The training RMSE was calculated as 3.5389, while the testing RMSE stood at 0.0058, indicating a strong fit to the data.

IV. CONCLUSION

This article focuses on a practical analysis of machine learning regression techniques to improve the accuracy of a sport athlete's running times. In conclusion, our study focused on employing various machine learning (ML) algorithms to predict the performance ofrunners. By evaluating algorithms such as Decision Tree, Random Forest, Logistic Regression, Support Vector Regression, and K-Nearest Neighbors, we aimed to identify the most accurate model. Among the

DOI: 10.17148/IJARCCE.2023.12686

algorithms tested, Random Forest emerged as the top performer, achieving an impressive accuracy rate of 99.33%. This indicates that Random Forest was highly effective in predicting runner performance based on the given data. Additionally, the calculated metrics of training RMSE (3.5389) and testing RMSE (0.0058) demonstrated a strong fit of the model to the dataset.

These findings highlight the potential of machine learning algorithms, particularly Random Forest, in accurately predicting the performance of runners. By leveraging these algorithms, trainers, athletes, and sports analysts can make informed decisions and optimize training strategies to enhance performance outcomes.

Further research can explore the application of additional ML algorithms and the inclusion of more features to improve prediction accuracy. Additionally, conducting experiments on larger and diverse datasets can help validate the generalizability of the findings. Overall, our study contributes to the growing body of knowledge in the field of ML-based performance prediction for runners and provides valuable insights for improving training and performance analysis in the realm of athletics.

ACKNOWLEDGEMENTS

We would like to express our sincere appreciation to the Marathon runners who generously contributed their valuable data for this study. Their participation and ongoing support have been instrumental in enabling us to conduct this research and derive meaningful insights. Their dedication to their sport and willingness to contribute to scientific endeavors is truly commendable. We extend our heartfelt gratitude to each of them for their invaluable contribution.

REFERENCES

[1] Islam, H., Gibala, M. J., & Little, J. P. (2022). Exercise snacks: A novel strategy to improve cardiometabolic health. Exercise and sport sciences reviews, 50(1), 31-37.
[2] Reusser, M.; Sousa, C.V.; Villiger, E.; Alvero Cruz, J.R.; Hill, L.; Rosemann, T.; Nikolaidis, P.T.; Knechtle, B. Increased Participation and Decreased Performance in Recreational Master Athletes in “Berlin Marathon” 1974–2019. Front. Physiol. 2021, 12, 631237. [CrossRef]
[3] Vitti, A.; Nikolaidis, P.T.; Villiger, E.; Onywera, V.; Knechtle, B. The “New York City Marathon”: Participation and performance trends of 1.2M runners during half-century. Res. Sport. Med. 2020, 28, 121–137. [CrossRef]
[4] Zahedi, L., Mohammadi, F. G., Rezapour, S., Ohland, M. W., &Amini, M. H. (2021). Search algorithms for automated hyper-parameter tuning. arXiv preprint arXiv:2104.14677.
[5] Ofoghi, B., Zeleznikow, J., MacMahon, C., &Raab, M. (2013). Data mining in elite sports: a review and a framework. Measurement in Physical Education and Exercise Science, 17(3), 171-186.
[6] Schumaker, R. P., Solieman, O. K., & Chen, H. (2010). Sports data mining.Integrated series in information systems (Vol. 26). New York, NY: Springer.
[7] A. Botchkarev, “Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology,” pp. 1–37, 2018, [Online]. Available: http://arxiv.org/abs/1809.03006.
[8] A. Botchkarev, “A new typology design of performance metrics to measure errors in machine learning regression algorithms,” Interdiscip. J. Information, Knowledge, Manag., vol. 14, pp. 45–76, 2019, doi: 10.28945/4184.
[9] V. Plevris, G. Solorzano, N. P. Bakas, and M. E. A. Ben Seghier, “Investigation of Performance Metrics in Regression Analysis and Machine Learning-Based Prediction Models,” World Congr. Comput. Mech. ECCOMAS Congr., pp. 0–25, 2022, doi: 10.23967/eccomas.2022.155.
[10] D. Chicco, M. J. Warrens, and G. Jurman, “The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation,” PeerJ Comput. Sci., vol. 7, pp. 1–24, 2021,doi: 10.7717/PEERJ-CS.623.
[11] S. Ray, “A Quick Review of Machine Learning Algorithms,” Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019, pp. 35–39, 2019.
[12] K. Functions et al., “Support Vector Machines for Classification and Regression - SVM.pdf,” 2016.
[13] M. R. Segal, “Machine Learning Benchmarks and Random Forest Regression Publication Date Machine Learning Benchmarks and Random Forest Regression,” Cent. Bioinforma. Mol. Biostat., p. 15, 2004, [Online]. Available: https://escholarship.org/uc/item/35x3v9t4.
[14] A. Kashvi Taunk, Sanjukta De, Srishti Verma, “A Brief Review of Nearest Neighbor Algorithm for Learning and Classification Kashvi,” 2019 Int. Conf. Intell. Comput. Control Syst. ICCS 2019, no. Iciccs, pp. 1255–1260, 2019.

This paper exemplifies the practical implementation of the research work.