electronics

MDPI

Article

Cricket Match Analytics Using the Big Data Approach

Mazhar Javed Awan [1,*], Syed Arbaz Haider Gilani [1], Hamza Ramzan [1], Haitham Nobanee [2,3,4,*], Awais Yasin [5], Azlan Mohd Zain [6] and Rabia Javed [7]

[1] Department of Software Engineering, University of Management and Technology, Lahore 54770, Pakistan; shaharbaz5@gmail.com (S.A.H.G.); hamzaramzan7907@gmail.com (H.R.)
[2] College of Business, Abu Dhabi University, Abu Dhabi 59911, United Arab Emirates
[3] Oxford Centre for Islamic Studies, University of Oxford, Marston Road, Headington, Oxford OX3 0EE, UK
[4] Faculty of Humanities & Social Sciences, University of Liverpool, 12 Abercromby Square, Liverpool L69 7WZ, UK
[5] Department of Computer Engineering, National University of Technology, Islamabad 44000, Pakistan; awaisyasin@nutech.edu.pk
[6] UTM Big Data Centre, School of Computing, Universiti Teknologi Malaysia, Skudai 81310, Johor, Malaysia; azlanmz@utm.my
[7] Department of Computer Science, Lahore College for Women University, Lahore 54000, Pakistan; rabia_javeed888@yahoo.com
* Correspondence: mazhar.awan@umt.edu.pk (M.J.A.); nobanee@gmail.com (H.N.)

check for updates

Citation: Awan, M.J.; Gilani, S.A.H.; Ramzan, H.; Nobanee, H.; Yasin, A.; Zain, A.M.; Javed, R. Cricket Match Analytics Using the Big Data Approach. *Electronics* **2021**, *10*, 2350. https://doi.org/10.3390/electronics10192350

Academic Editors: Amin Karami, Fahimeh Jafari and Manel Guerrero-Zapata

Received: 21 August 2021
Accepted: 23 September 2021
Published: 26 September 2021

Abstract: Cricket is one of the most liked, played, encouraged, and exciting sports in today's time that requires a proper advancement with machine learning and artificial intelligence (AI) to attain more accuracy. With the increasing number of matches with time, the data related to cricket matches and the individual player are increasing rapidly. Moreover, the need of using big data analytics and the opportunities of utilizing this big data effectively in many beneficial ways are also increasing, such as the selection process of players in the team, predicting the winner of the match, and many more future predictions using some machine learning models or big data techniques. We applied the machine learning linear regression model to predict the team scores without big data and the big data framework Spark ML. The experimental results are measured through accuracy, the root mean square error (RMSE), mean square error (MSE), and mean absolute error (MAE), respectively 95%, 30.2, 1350.34, and 28.2 after applying linear regression in Spark ML. Furthermore, our approach can be applied to other sports.

Keywords: big data analytics; machine learning; cricket; match prediction; Spark ML; prediction model

1. Introduction

In this era, while many sports are played in different countries, a few have been liked and encouraged a little more than others. Similarly, cricket is one of the most picked and played sports of this modern era. cricket was introduced in England in the sixteenth century [1]. Initially, cricket was introduced and played in a test format only. After some time, and due to some conditions and policies, the international cricket council introduced additional formats such as the T20 and ODI format. There are three official formats in which cricket is played internationally with varying durations and standards. The one-day international (ODI) cricket is one of the most played and liked structures by everyone. In this format, 100 overs of play are designed. Given that there are 100 overs in a one-day game, each team plays 50 overs, with the aim to fight and win. The datasets are available on different electronic databases with the maximum available information [2].

Data extraction is a significant component of information retrieval. A mechanism by which the user gives a query to the system, according to its demanded data, is used. The system works on the question and returns to the user with the desired information and dataset. Many algorithms exist in the information retrieval process, in which strategies work to run the query and produce the valuable and desired outcome in the form of information extraction. Therefore, our data can be available in the form of structured and unstructured data [3,4].

Most of the produced ball-to-ball data from matches were massive data that could not be utilized until or unless big data models and techniques were applied. Therefore, to use these data in our research, we have applied big data analytics techniques using machine learning and tools to gain valuable outcomes [5].

1.1. Big Data

The term big data represents the massive amount of complex data that is impossible to process in a traditional software. For this reason, it is important to apply data science and data analytics techniques [6]. The data are increasing with time and the information in terms of data is becoming more challenging to utilize in a meaningful and valuable way. Therefore, for businesses to use these data to predict and forecast future decisions [7], these processes can be done by applying big data approaches and frameworks that visualize and customize data. Big data has five significant characteristics namely velocity, verity, valence, volume, and veracity [8,9], and together, the five Vs deliver the value, as shown in Figure 1.

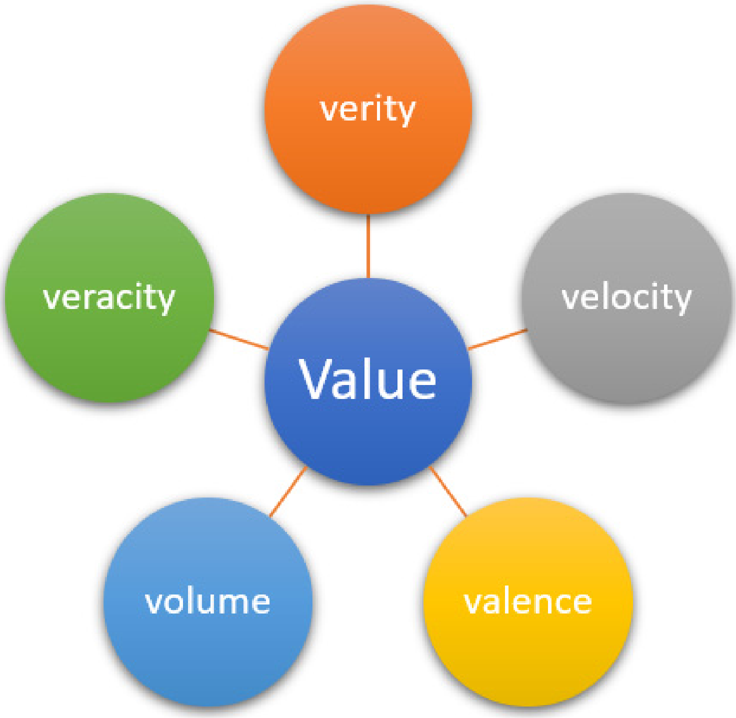

Figure 1. Characteristics of the five Vs in big data.

These are the significant factors that make an overall value in big data, as well as the factors that make the big data.

1.2. Machine Learning

Machine learning is a brand of artificial intelligence (AI) that works on the mechanism of that system, as well as on learning and understanding trends from historical information or data. It can understand patterns of different data types and make decisions with or without minimal human hindrance. Machine learning makes it possible to build and produce automated models that handle complex data and analyze the vast data that traditional applications cannot control. By applying its mechanisms, machine learning enables the models to analyze massive complex data and deliver outcomes faster with possible accurate results [10–13].

Machine learning has two methods that are followed in the process: Unsupervised and supervised machine learning. In supervised machine learning, our designed models are trained based on labels, where the output of a given input is already known. The problems to be solved by the supervised learning methods are categorized into two types: Regression and classification. In the regression problem, we predict the outcomes based on the provided historical datasets, and the results that we expect are absolute values. The category and the effects of the model are in definite forms. In unsupervised machine learning, the user does not need to supervise the respective model, since the user only permits the model to get information, discover some patterns from the datasets, and train automatically by itself. Unsupervised learning deals with the data which is not considered or labeled yet. This learning is categorized into significant types such as clustering, associations, and dimensionality. The primary objective of these machine learning algorithms and models is to get valuable data patterns within complex and massive datasets [14–16].

1.3. Spark ML

The Apache Spark framework is a highly effective and highly recommended framework by most analyst communities and in research. It is an open-source, easy-to-use, and freely accessible framework with a high performance rate. Moreover, the Apache Spark framework is a clustering-based computing system that is freely accessible by everyone. It is used in big data analytics for analysis and to resolve interactions and constants. Spark ML supports a high range of components, performance, and scope using Hadoop [17,18].

In this process, we first did some big data visualization on our dataset to neglect and remove all the unwanted information. Following this step, we used big data frameworks such as Jupiter notebook. Moreover, we checked for the best model with the highest accuracy and performed it in the Spark ML framework, in order to check which framework is better suited for building this prediction model using the Python language. While processing these data for our project outcomes, we have faced many problems in the decision-making process due to the large amount of information available and the number of labels with different classes. We have applied some analytical approaches that visualize data and machine learning algorithms to build the model used in the prediction process.

The study aims to predict the total score through a machine learning model, Scikit learn, and a big data framework, Spark ML. The significant contribution of our research is to achieve good accuracy using big data of the one-day international (ODI) cricket, as per our knowledge.

2. Related Work

With the fast-growing advancement of cricket, it turned into an exceptionally intriguing topic for all sports analysts. However, there are still conflicting and convoluted informational indexes. Despite extensive research, they could not leap forward and precisely anticipate the winner of the match. Techniques such as logistic regression, KNN, Naïve Bayes, and SVM, etc., have been used to predict the winner. Moreover, the collected data of the matches were obtained from websites such as Kaggle, Cricsheet, etc., and the ball-by-ball details along with various rules were applied. The sorted data were split into two parts: Training (80%) and testing data (20%). TensorFlow and Python were used as the main tools. The performance measures were concluded by a confusion matrix [19,20].

In a study by Ahmed [21], the winner of the match was guessed using ICC match evaluations and rating data from the one-day international match outcomes, ICC rankings for batters, bowlers, home factors, ICC rating differentiations, and ground ramifications. The prediction he made showed 80% of precision and accurateness. The authors carried out various measurable methodologies to arrange the datasets and attempted different characterization strategies to anticipate the winner of the one-day cricket match comprising 50 overs. In addition, they executed logistic regression on the data and achieved precision in expecting the outcomes of games by 74.9%. In 81% of the matches, they expected the winner precisely.

Yasir et al. [22] anticipated the cricket match results and the winner expectation techniques. He proposed a method to foresee the team results and the strategy expounded in this work by utilizing properties of the dynamic group for the winner's forecast, such as the player's set of experiences, climate, ground olden times, and winning rate. After applying the same technique to 100 matches, the authors got a prediction of 85%. The Naïve Bayes works on the Bayes probability theorem with a forecast for all class labels to be independent, which can be false. It can be with repetitive feature elimination, as well.

Vistro et al. [23] used SEMMA Modeling namely sample, explore, modify, model, and access. The main aim of their research was to identify a winner of the IPL match using previous data from 2008 to 2017. The Decision Tree Model, which showed 76.9% accuracy, increased to 94%. Similarly, the Random Forest model was applied, which predicted 71% accuracy, but in the XGBoost learning model, an accuracy of 94.23% was seen. In cricket, in order to achieve the convergence of the data science world, a lot more data are required to conclude the winner of the match.

However, no method could reach the level of accuracy and scalability in terms of big data. Recently, studies have shown promising results after applying the big data framework Apache Sharp using Spark ML on various domains [24–26].

Morgulev et al. [27] researched the relationship of big data and sports. They found that different factors affect the relevance of big data with the actual outcomes, such as the fan's use of big data, the athlete's use of big data, challenges, and the benefits of big data. Based on the historical data of the player's performance, the match outcomes were predicted by machine learning algorithms.

Samaria et al. [28] performed a sentimental analysis by a Twitter API during the ICC Cricket World Cup 2015. Different hashtags were used to retrieve the tweets relevant to their research. After noise cancellation, the received data were classified into negative, positive, and neutral. Micro-blogging and the use of the internet have been increased unanimously.

3. Materials and Methods

3.1. Data Description

In this paper, the data are extracted from the Cricsheet. The obtained data are only from the one-day international cricket matches from 2006 to 2017 [29] and consist of the one-day international cricket format. This dataset has 15 labels with additional details of the match. It contains attributes such as mid, date venue, bat team, bowl team, batsman, bowler, runs, wickets, overs, runs_last_5, wickets_last_5, striker, non-striker, and total runs. These are the actual attributes included in our selected dataset and all those attributes that impact our predicted results, will be considered. We have chosen this dataset to produce an effective model that can predict the winner from a considerable amount of historical data and can apply some prediction models for better accuracy and outcomes. The reason for selecting this dataset, which is around 4 years older than this research, was the lack of recent data in any digital databases. Table 1 shows all of the attributes and values in our datasets. As it can be seen, there are 15 columns and 350,899 entries in our dataset of the one-day international cricket format.

In this paper, we have built a prediction model to predict the winner of the match. For this reason, we have made two models using traditional methods and frameworks. Then, we checked the best model with high accuracy—the model with high accuracy was selected for further work. Based on a high accuracy rate, the model decided on will be built again using the Spark machine learning framework, in order to check which one is better suited to make a prediction model. The overall hierarchy of our research that shows the road map of the study is shown in Figure 2.

Table 1. Data structure and counts.

Mid	Date	Venue	Bat_ Team	Bowl_ Team	Batsman	Bowler	Runs	Wickets	Overs	Runs_ Last_5	Wickets_ Last_5	Striker	Non_ Striker	Total
1	13 June 2006	Civil Service Cricket Club, Stormont	England	Ireland	ME Trescothick	DT Johnston	0	0	0.1	0	0	0	0	301
2	13 June 2006	Civil Service Cricket Club, Stormont	England	Ireland	ME Trescothick	DT Johnston	0	0	0.2	0	0	0	0	301
3	13 June 2006	Civil Service Cricket Club, Stormont	England	Ireland	ME Trescothick	DT Johnston	4	0	0.3	4	0	0	0	301
4	13 June 2006	Civil Service Cricket Club, Stormont	England	Ireland	ME Trescothick	DT Johnston	6	0	0.4	6	0	0	0	301
5	13 June 2006	Civil Service Cricket Club, Stormont	England	Ireland	ME Trescothick	DT Johnston	6	0	0.5	6	0	0	0	301

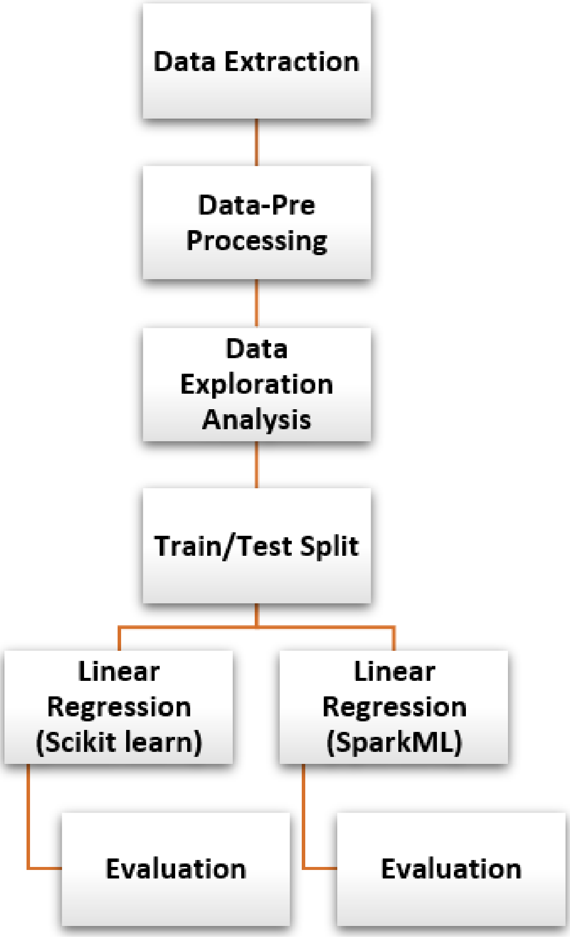
Data Extraction
Data-Pre Processing
Data Exploration Analysis
Train/Test Split
Linear Regression (Scikit learn)
Linear Regression (SparkML)
Evaluation
Evaluation

Figure 2. Hierarchy of our approach.

In Figure 2, all of the steps and the proper hierarchy of the model-building processes are shown. As can be seen, we have taken a dataset and first performed pre-processing on the data, which includes assigning their classes and labels to the attributes. Thereafter, we will evaluate the data, exclude some irrelevant data, and then make a prediction of our model. Finally, we will implement our model on two frameworks to check which framework shows the best accuracy.

3.2. Data Pre-Processing

Data pre-processing in the big data approach is for any kind of prediction or forecasting or, in some cases, for understanding the real meaning of data. By applying some analytical tools, a sorted and well-mannered form of data is achieved. Occasionally, data pre-processing tasks become more complicated and lengthy due to the fact that when the data have transparencies and outliers, they have to be sorted into a good shape.

The following steps are selected for any kind of outlier or noise and result in consistency.

3.2.1. Removing Unwanted Columns

In the first step, we filtered the data and removed all the unwanted columns from the dataset, in order to consider only those columns on which our prediction is based and dependent. The unwanted columns we removed from the dataset are the "first five overs". Moreover, we only considered consistent teams that are valuable for our prediction. Furthermore, we dropped the mid and date columns from the dataset.

3.2.2. Assigning Unique Values

In the second step, we assigned all the unique values from the dataset into our model, to make a prediction based on these particular values. In our prediction model, our values are bat_team and bowl_team.

3.2.3. One-Hot Encoding

In the third step, we converted all the categorical features into one-hot encoding using the Pandas dummies method. The variables are bat_team and bowl_team.

3.2.4. Data Transformation

Lastly, before putting the data into machine learning algorithms, an important step is to transform the features by scaling [0, 1]. We used the min-max Scaler () function to transform the minimum value into 0 and the maximum value into 1. This step is also called standardization.

3.3. Data Exploration Analysis

The data contain attributes such as mid, date venue, bat team, bowl team, batsman, bowler, runs, wickets, overs, runs_last_5, wickets_last_5, striker, non-striker, and total runs. The filtered and primary dependent attributes, as well as the data in which our prediction model predicts the winning team, is dependent. Therefore, after applying the analytical tools, we screened and processed these attributes, wickets, runs, total, and overs, in which we have applied the model and made some predictions through the proposed model.

This paper has processed our data into the analytical framework Spark to make it more precise and valuable for our proposed model, which enhances its accuracy. We have run and modified our dataset through the Databricks community version (an online platform) to suspend all those attributes that are not helpful in our prediction model [30].

The following steps are for data exploration:

- First, log in to the Databricks community and make a new cluster after the login. This cluster is a data frame where you will perform your required task. This cluster assigns you some storage to use a few resources.
- After creating a cluster, you need to upload or make a table from the design tab. In this tab, you can upload your dataset table to do some visualization. This can be done by applying the techniques and tools needed to make it more relevant to your data requirements.
- After finalizing the data table, you will have to create a worksheet that performs all these tasks to predict and customize the data. This is the significant step in which you will build your model and run it on your respective dataset or table.

The above steps are followed to customize your dataset through the Databricks platform. The visualized data structure is shown in Table 2.

Table 2. Data exploration of the selected variables.

Bat_Team	Bowl_Team	Runs	Wickets	Overs	Wickets_Last_5	Total
England	Ireland	0	0	0.1	0	301
England	Ireland	0	0	0.2	0	301
England	Ireland	4	0	0.3	0	301
England	Ireland	6	0	0.4	0	301
England	Ireland	6	0	0.5	0	301

The statistics summary of the mean, as well as the mode of six features are described in Table 3.

Table 3. The statistics description of numeric values.

	Mean	Std	Min	Max
runs	114.801661	77.665959	0.0	444.0
wickets	2.974970	2.298959	0.0	10.0
overs	24.052899	14.235439	0.0	49.6
runs_last_5	23.548303	11.042974	0.0	101.0
wickets_last_5	0.669814	0.833895	0.0	7.0
total	255.355387	62.354412	44.0	444.0

3.4. Linear Regression Model

Linear regression is a predictive analysis algorithm that is used to predict something according to your choice. Moreover, it is a commonly used basic prediction algorithm, which is used by most of the analyst community in the prediction process. The primary purpose of this linear regression is to show the linear relationship between one or more than one independent variable with a single dependent variable. Linear regression follows the principles of supervised machine learning algorithms [31], which is the simplest and most classic linear method for regression. Linear regression finds the parameters w and b that minimize the mean squared error between predictions and the actual regression targets, y, on the training set. In addition, it finds a coefficient parameter w and an offset b to make predictions using a linear combination of features in Equation (1).

$$\mathrm{min}w,b\sum i\,|\,|\,w\mathrm{T}xi+b-yi\,|\,|\,2 \tag{1}$$

The attributes that we have chosen for the model have one independent variable, "total", and all the other features are over-dependent variables. We took the data from these attributes only and applied the respective model to them. The distribution of the independent value of the total score against density is shown in Figure 3.

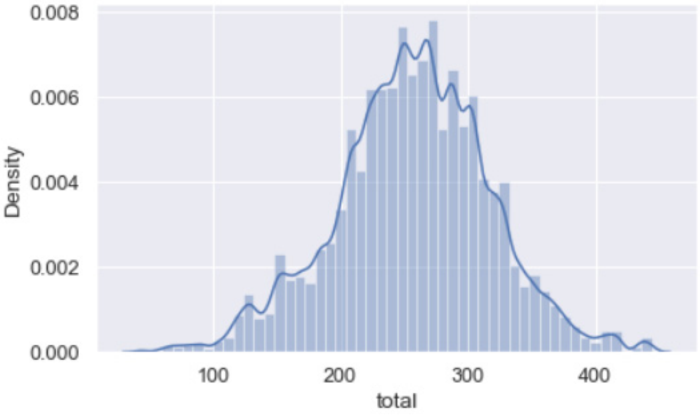

Figure 3. The density of the total score.

Furthermore, linear regression has the problem of multicollinearity. Multicollinearity is a condition of significant dependency between the dependent features and independent features. Multicollinearity reduces the precision of the estimated coefficients [32]. There are many ways to handle this problem, for example, by combining independent variables, performing a designed analysis, and evaluating the covariance between the variables to construct a heatmap. Here, we plotted the heatmap correlation between the features, as shown in Figure 4.

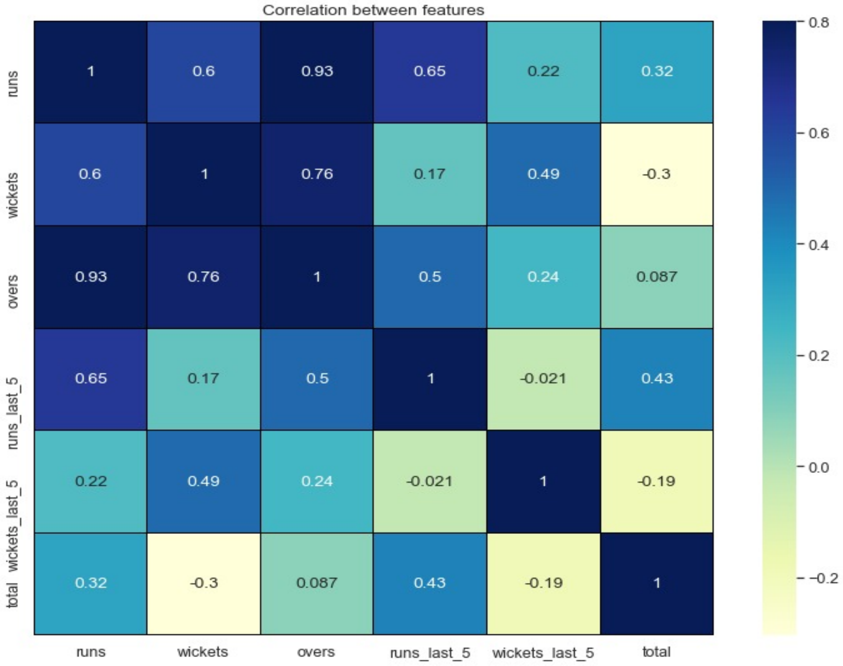

Figure 4. Correlation between all the features.

Figure 4 explains the correlation of all the attributes we have considered in our model to train and predict the winner of a cricket match. Here, we considered five characteristics based on their covariance and correlation to the other attributes. The value of the correlation is between 0 to 1. If the correlation is in a positive value, then the relation is strong among the two attributes. If the value is negative, then the correlation is not strong, as needed. In the threshold value, we selected a correlation coefficient of >0.8. The heatmap matrix for the numeric features of overs and runs indicates a high correlation of 0.93, which is a problem of multicollinearity.

4. Experimental Results

In the proposed models, we applied the data analytics technique and then ran the dataset in our model. Thereafter, we assigned values to the x and y variables in which we have to predict. We used some libraries such as pandas, NumPy, random forest, linear regression, standard scalar, and Matplotlib libraries. Moreover, we applied some of the results from this model on Google Colab [33].

4.1. Train and Test Split Data Sets

The dataset is divided into training and test data. The values above the 2016 date year are related to training the dataset and the rest of the values are related to testing the dataset, in order to check how many predicted values of our model are correct in the test.

Our model's training and testing datasets are divided with the test size of 20% and a random state at zero.

4.2. Converting of String into an Object

This step will convert all the strings present in our dataset into objects to make them understandable for our model. The strings present will be converted into things using the Lambda function and date-time function.

4.3. Building the Model on Scikit Learn and Spark ML

The linear regression model was tested on Spark ML as well as Sklearn. The model is fitted to the encoded train and test samples. Thereafter, we will call a pickle file for the location where the model will work. In this model, building of linear regression, we have chosen linear ridge regression, and in model selection, we have chosen the Grid Search CV model selection. In ridge selection, we will define all those parameters in which our prediction is based. Then, we will apply the ridge repressor model to calculate the mean square error of the model and make the prediction.

4.4. Evaluation Result of Linear Regression Using Scikit Learn

Before making the prediction, we have successfully done the model creation and implementation of the model on the dataset. In this step, we will check how efficient our model is. We will import metrics and NumPy from the Sklearn library to calculate its mean square error, root mean square error, and mean of an absolute error of the model.

4.5. Evaluation Result of Linear Regression Using Spark ML

The primary reason for building this model is to check which framework is better in terms of accuracy. In this model, we have done the same data visualization again and applied the descriptive data analysis on our dataset by checking the covariance of the data attributes that are here considered. Based on these attributes, we have built this model of prediction. In addition, we have used Spark machine learning libraries to apply the model and Spark context libraries to transform our data. The accuracy of the model using this Apache framework is better than the traditional machine learning Scikit learn. In the Spark machine learning framework, the model's accuracy is increased by 96% as training and 94.5% as testing accuracy. Table 4 shows the accuracy, mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE).

The above table shows the accuracy comparison of the same model with the different frameworks. As you can see, the Spark machine learning frameworks offer the best accuracy in all aspects.

Table 4. Comparison of two frameworks in terms of accuracy.

Result of the Model Using a Scikit Learn		**Result of the Model Using Spark ML Framework**	
Mean Absolute Error	30.7	Mean Absolute Error	28.2
Mean Square Error	1576.29	Mean Square Error	1350.34
Rooted Mean Square Error	39.69	Rooted Mean Square Error	30.2
Train Accuracy	90%	Train Accuracy	96%
Test Accuracy	88%	Test Accuracy	95%

5. Discussion

In this paper, firstly, we have researched the cores of our topic related to predicting the winner in a cricket match. We have seen that many other researchers have done some work and proposed their results using different digital databases and libraries. All of these results are available for access by everyone. For the models that show high accuracy in the prediction model, we have taken these two models and implemented them on the new ODI dataset, as well as checked their performances. Then, the best model with high accuracy is again built on the Spark ML framework to check whether the Spark framework is efficient or not for building the prediction model. To make this model, the datasets were obtained from cricsheet.com, an online data-store where you can get any dataset related to any topic. We have built two significant models and compared their performance based on their accuracy and confusion matrixes. The models that have a high accuracy and a high level of response are considered in this paper. Since many other people have done their work on predicting scores, the majority of the research work done until now is compared with our proposed model based on their accuracy results, as shown in Table 5.

Table 5. State-of-art comparison work.

Studies	Dataset	Model/Framework	Accuracy
Agarwal et al. [34]	Team Prediction	Statistical Modeling Approach Hadoop Framework	91%
Aburas et al. [35]	ICC 2019 World Cup	K Nearest Neighbors and MySQL	90%
Aburas et al. [36]	ICC 2019 World Cup	KNN and Business Intelligence	93% and 90%
Vistro, Rasheed, and David [23]	IPL Prediction	Decision tree classifier	94.87%
		Random forest classifier	80.76%
		XGBoost classifier	94.23%
Our Study	**ODI Cricsheet Score Prediction**	**Linear Regression Spark ML**	**Train Accuracy = 96% Test Accuaracy = 95%**

Table 4 explains the difference in the performance between our proposed model and the rest of the existing models in terms of the root mean square error (RMSE), mean square error (MSE), and mean absolute error (MAE).

We have proposed a model that can be applied to any situation and condition in a match, in which you can expect a team's score. Then, our model asks for your input, for example, our model will ask you to input the current runs, wickets, and overs, and then it will apply our proposed model to it according to your given scenario. We used our model, then made some predictions and gave you the output that contains the predicted overall score of teams A and B. If the score of section A is more significant, it will show that team A will be the winner, and if the score is minor, then team B will win. We can use this function in the future for every kind of sport by inserting some input values according to the situation, in which the model will work on historical data and predict the winner based on the provided information. In the future, the deep learning bimodal architectures could be applied to cricket and other sports [37–41].

6. Conclusions

The most significant outcomes of this paper can be summarized as follows: (1) The Spark framework is efficient compared to the traditional one; and (2) the linear regression model is the best model used to predict the total score of the match. In this paper, we have built a model that will indicate the winner of a cricket match with your input conditions of the ongoing game. We have seen that the best accuracy we can get is from the linear regression model. It gives us 96% accuracy using the Spark machine learning framework, in terms of the prediction analysis that tells us how efficient our model is in the prediction process. The other models also show excellent results since they do not have good accuracy in the confusion matrix and R mean squared error. Overall, the performance of both models is outstanding and can be used in any match to predict the winner. It is recommended to apply some analytical tools to your dataset to train your model, since your accuracy will not be as good if your dataset is not according to the accurate data. In the future, we can build and use these models in many other sports. From the marathon race datasets, we can predict the percentage of winning countries and cities and we can predict which city or country has more chances to win. Furthermore, these kinds of predictions can be performed in the future.

Author Contributions: Conceptualization, M.J.A.; Formal analysis, A.Y.; Funding acquisition, H.N.; Investigation, A.M.Z.; Methodology, R.J.; Supervision, M.J.A. and H.N.; Validation, A.M.Z.; Visualization, H.R.; Writing—original draft, S.A.H.G. and H.R.; Writing—review & editing, M.J.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data is publicly available Cricsheet is a collection of projects which collectively provide data for various aspects of cricket on the website Cricsheet ODI Cricket. Availabe online: https://cricsheet.org/ (accessed on 10 July 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Williams, J. *Cricket and England: A Cultural and Social History of the Inter-War Years*; Taylor & Francis: Oxfordshire, UK, 1999.
2. Bailey, M.; Clarke, S.R. Predicting the match outcome in one day international cricket matches, while the game is in progress. *J. Sports Sci. Med.* **2006**, *5*, 480.
3. Rehma, A.A.; Awan, M.J.; Butt, I. Comparison and Evaluation of Information Retrieval Models. *VFAST Trans. Softw. Eng.* **2018**, *6*, 7–14.
4. Alam, T.M.; Awan, M.J. Domain analysis of information extraction techniques. *Int. J. Multidiscip. Sci. Eng.* **2018**, *9*, 1–9.
5. Kaur, A.; Kaur, R.; Jagdev, G. Analyzing and Exploring the Impact of Big Data Analytics in Sports Sector. *SN Comput. Sci.* **2021**, *2*, 1–19. [CrossRef]
6. Ahmed, H.M.; Awan, M.J.; Khan, N.S.; Yasin, A.; Shehzad, H.M.F. Sentiment Analysis of Online Food Reviews using Big Data Analytics. *Elem. Educ. Online* **2021**, *20*, 827–836.
7. Zhou, L.; Pan, S.; Wang, J.; Vasilakos, A.V.J.N. Machine learning on big data: Opportunities and challenges. *Neurocomputing* **2017**, *237*, 350–361. [CrossRef]
8. Aftab, M.O.; Awan, M.J.; Khalid, S.; Javed, R.; Shabir, H. Executing Spark BigDL for Leukemia Detection from Microscopic Images using Transfer Learning. In Proceedings of the 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), Riyadh, Saudi Arabia, 6–7 April 2021; pp. 216–220.
9. Awan, M.J.; Khan, M.A.; Ansari, Z.K.; Yasin, A.; Shehzad, H.M.F. Fake Profile Recognition using Big Data Analytics in Social Media Platforms. *Int. J. Comput. Appl. Technol.* **2021**, in press.
10. Anam, M.; Ponnusamy, V.; Hussain, M.; Waqas Nadeem, M.; Javed, M.; Guan Goh, H.; Qadeer, S. Osteoporosis Prediction for Trabecular Bone using Machine Learning: A Review. *Comput. Mater. Contin.* **2021**, *67*, 89–105. [CrossRef]
11. Ali, Y.; Farooq, A.; Alam, T.M.; Farooq, M.S.; Awan, M.J.; Baig, T.I. Detection of Schistosomiasis Factors Using Association Rule Mining. *IEEE Access* **2019**, *7*, 186108–186114. [CrossRef]
12. Nagi, A.T.; Awan, M.J.; Javed, R.; Ayesha, N. A Comparison of Two-Stage Classifier Algorithm with Ensemble Techniques On Detection of Diabetic Retinopathy. In Proceedings of the 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), Riyadh, Saudi Arabia, 6–7 April 2021; pp. 212–215.
13. Awan, M.J.; Yasin, A.; Nobanee, H.; Ali, A.A.; Shahzad, Z.; Nabeel, M.; Zain, A.M.; Shahzad, H.M.F. Fake News Data Exploration and Analytics. *Electronics* **2021**, *10*, 2326. [CrossRef]
14. Gupta, M.; Jain, R.; Arora, S.; Gupta, A.; Javed Awan, M.; Chaudhary, G.; Nobanee, H. AI-enabled COVID-9 Outbreak Analysis and Prediction: Indian States vs. *Union Territories. Comput. Mater. Contin.* **2021**, *67*, 933–950. [CrossRef]
15. de Souza Junior, A.H.; Corona, F.; Barreto, G.A.; Miche, Y.; Lendasse, A. Minimal learning machine: A novel supervised distance-based approach for regression and classification. *Neurocomputing* **2015**, *164*, 34–44. [CrossRef]
16. Javed, R.; Saba, T.; Humdullah, S.; Jamail, N.S.M.; Awan, M.J. An Efficient Pattern Recognition Based Method for Drug-Drug Interaction Diagnosis. In Proceedings of the 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA), Riyadh, Saudi Arabia, 6–7 April 2021; pp. 221–226.
17. Salloum, S.; Dautov, R.; Chen, X.; Peng, P.X.; Huang, J.Z. Big data analytics on Apache Spark. *Int. J. Data Sci. Anal.* **2016**, *1*, 145–164. [CrossRef]
18. Khalil, A.; Awan, M.J.; Yasin, A.; Singh, V.P.; Shehzad, H.M.F. Flight Web Searches Analytics through Big Data. *Int. J. Comput. Appl. Technol.* **2021**, in press.
19. Singh, T.; Singla, V.; Bhatia, P. Score and winning prediction in cricket through data mining. In Proceedings of the 2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI), Faridabad, India, 8–10 October 2015; pp. 60–66.
20. Kamble, R. Cricket Score Prediction Using Machine Learning. *Turk. J. Comput. Math. Educ.* **2021**, *12*, 23–28.
21. Ahmed, W. A Multivariate Data Mining Approach to Predict Match Outcome in One-Day International Cricket. Master's Thesis, Karachi Institute of Economics and Technology, Karachi, Pakistan, August 2015.
22. Yasir, M.; Chen, L.; Shah, S.A.; Akbar, K.; Sarwar, M.U. Ongoing Match Prediction in T20 International. *Int. J. Comput. Sci. Netw. Secur.* **2017**, *17*, 176–181.

23. Vistro, D.M.; Rasheed, F.; David, L.G. The Cricket Winner Prediction With Application Of Machine Learning And Data Analytics. *Int. J. Sci. Technol. Res.* **2019**, *8*, 985–990.
24. Javed Awan, M.; Shafry Mohd Rahim, M.; Nobanee, H.; Munawar, A.; Yasin, A.; Mohd Zain Azlanmz, A. Social Media and Stock Market Prediction: A Big Data Approach. *Comput. Mater. Contin.* **2021**, *67*, 2569–2583. [CrossRef]
25. Javed Awan, M.; Shafry Mohd Rahim, M.; Nobanee, H.; Yasin, A.; Ibrahim Khalaf, O.; Ishfaq, U. A Big Data Approach to Black Friday Sales. *Intell. Autom. Soft Comput.* **2021**, *27*, 785–797. [CrossRef]
26. Awan, M.J.; Khan, R.A.; Nobanee, H.; Yasin, A.; Anwar, S.M.; Naseem, U.; Singh, V.P. A Recommendation Engine for Predicting Movie Ratings Using a Big Data Approach. *Electronics* **2021**, *10*, 1215. [CrossRef]
27. Morgulev, E.; Azar, O.H.; Lidor, R. Sports analytics and the big-data era. *Int. J. Data Sci. Anal.* **2018**, *5*, 213–222. [CrossRef]
28. Samariya, D.; Matariya, A.; Raval, D.; Babu, L.D.; Raj, E.D.; Vekariya, B. A hybrid approach for big data analysis of cricket fan sentiments in twitter. In Proceedings of the International Conference on ICT for Sustainable Development; Springer: Singapore, 2016; pp. 503–512.
29. Cricsheet ODI Cricket. Available online: https://cricsheet.org/ (accessed on 10 July 2021).
30. Ilijason, R. Getting Started with Databricks. In *Beginning Apache Spark Using Azure Databricks*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 27–38.
31. Goldberger, A.S. Best linear unbiased prediction in the generalized linear regression model. *J. Am. Stat. Assoc.* **1962**, *57*, 369–375. [CrossRef]
32. Wang, G.C. How to handle multicollinearity in regression modeling. *J. Bus. Forecast.* **1996**, *15*, 23.
33. Awan, M.J. Acceleration of Knee MRI Cancellous bone Classification on Google Colaboratory using Convolutional Neural Network. *Int. J. Adv. Trends Comput. Sci. Eng.* **2019**, *8*, 83–88. [CrossRef]
34. Agarwal, S.; Yadav, L.; Mehta, S. Cricket team prediction with hadoop: Statistical modeling approach. *Procedia Comput. Sci.* **2017**, *122*, 525–532. [CrossRef]
35. Aburas, A.A.; Mehtab, M.; Mehtab, Y. Cricket World Cup Predictions Using KNN Intelligent Bigdata Approach. In Proceedings of the 2018 International Conference on Computing and Big Data, Charleston, SC, USA, 8–10 September 2018; pp. 18–22.
36. Aburas, A.A.; Mehtab, M.; Mehtab, Y. ICC World Cup Prediction Based Data Analytics and Business Intelligent (BI) Techniques. In Proceedings of the 2018 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), Zhengzhou, China, 18–20 October 2018; pp. 273–2736.
37. Awan, M.J.; Rahim, M.S.M.; Salim, N.; Mohammed, M.A.; Garcia-Zapirain, B.; Abdulkareem, K.H. Efficient Detection of Knee Anterior Cruciate Ligament from Magnetic Resonance Imaging Using Deep Learning Approach. *Diagnostics* **2021**, *11*, 105. [CrossRef]
38. Awan, M.J.; Raza, A.; Yasin, A.; Shehzad, H.M.F.; Butt, I. The Customized Convolutional Neural Network of Face Emotion Expression Classification. *Ann. Rom. Soc. Cell Biol.* **2021**, *25*, 5296–5304.
39. Mujahid, A.; Awan, M.J.; Yasin, A.; Mohammed, M.A.; Damaševičius, R.; Maskeliūnas, R.; Abdulkareem, K.H. Real-Time Hand Gesture Recognition Based on Deep Learning YOLOv3 Model. *Appl. Sci.* **2021**, *11*, 4164. [CrossRef]
40. Mubashar, R.; Javed Awan, M.; Ahsan, M.; Yasin, A.; Partap Singh, V. Efficient Residential Load Forecasting using Deep Learning Approach. *Int. J. Comput. Appl. Technol.* **2021**, in press.
41. Javed Awan, M.; Shehzad, F.; Muhammad, H.; Ashraf, M. Fake News Classification Bimodal using Convolutional Neural Network and Long Short-Term Memory. *Int. J. Emerg. Technol.* **2020**, *11*, 209–212.