Table of Contents

Abstract

Cricket is a bat and ball game played between two teams in a field of limited area. Both the teams consist of 11 players in it. Cricket in today's time is the most liked, played game that needs a machine learning analysis for more accuracy. As time continues more and more games are played which increases the data for the matches and individual players which requires proper analysis and accuracy. With increasing number of matches there are more players and thus data to analyze and predict the scores and the selection of players in a team, which team will win the game and so on. To do all this we need a machine learning models with high accuracy and analyze the data for it.

KEYWORDS: machine learning, cricket, score prediction

This project involves data analysis and regression model development for predicting scores in cricket matches. I have applied machine learning linear regression model to predict the team scores without big data. The analysis uses historical match data to identify key features influencing match scores. The experimental results are measured through accuracy, the root mean square error (RMSE), mean square error (MSE), and mean absolute error (MAE). The given data are also plotted in graph for better understanding of the problem. The dataset used for this project is available in the README.md file of the GitHub repository. In doing the analysis various libraries were used such as numpy which is employed for numerical operations essential in model computations, pandas for data loading, cleaning, and preprocessing, matplotlib for plotting of graphs for better understanding of the data analysis, scikit-learn is used to train a linear regression model, split the dataset into training and testing sets with train_test_split, and standardize the features using StandardScaler. These tasks ensured efficient model training, evaluation, and data preprocessing.

The flow of the project begins with the importing of various necessary libraries, data loading and preprocessing, handling missing values, encoding categorical variables. The given data set is split between training and testing sets for machine learning model and build the linear regression model. The built model is applied for prediction of test sets and calculate the evaluation metrics and finally combine the actual and predicted scores. Again matplotlib library is imported for plotting of graphs. Using this I have plotted the datasets graph for better understanding of relationships between various data. Histograms, box plots, and rest are drawn.

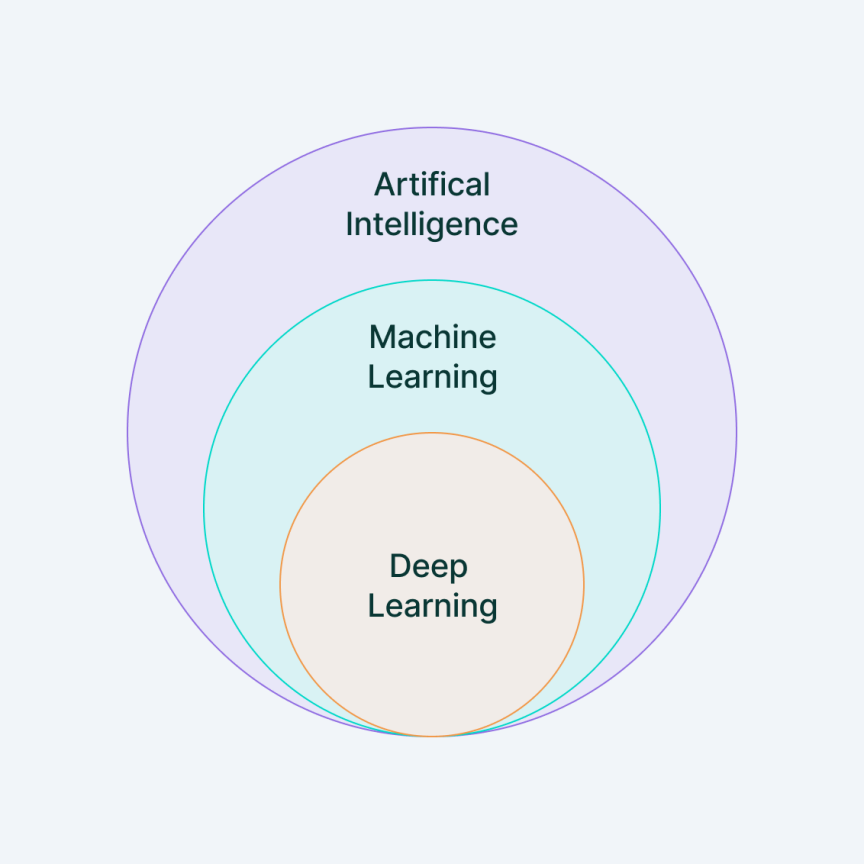
By doing above machine learning regression and data analysis for cricket and score prediction we come to the conclusion that by analyzing various features such as runs, bowl_team, runs_last_t, venue, bat_team, etc that certain features such as wickets and overs had significant influence on score prediction and while others had less impacts.

What is Artificial Intelligence (AI) ?

Artificial intelligence (AI) is a set of technologies that enable computers to perform a variety of advanced functions, including the ability to see, understand and translate spoken and written language, analyze data, make recommendations, and more. On an operational level for business use, AI is a set of technologies that are based primarily on machine learning and deep learning, used for data analytics, predictions and forecasting, object categorization, natural language processing, recommendations, intelligent data retrieval, and more.

What is Machine learning (ML) ?

Machine Learning (ML) is a branch of artificial intelligence (AI) and computer science that focuses on the using data and algorithms to enable AI to imitate the way that humans learn, gradually improving its accuracy. ML allows machines to make predictions and recommendations based on large amounts of data without being explicitly programmed. Machine Learning uses a data-driven approach, It is typically trained on historical data and then used to make predictions on new data. ML can find patterns and insights in large datasets that might be difficult for humans to discover.



BLACKBUCKS
GROUP OF COMPANIES

What is Deep Learning ?

The definition of Deep learning is that it is the branch of machine learning that is based on artificial neural network architecture. An artificial neural network or ANN uses layers of interconnected nodes called neurons that work together to process and learn from the input data. In a fully connected Deep neural network, there is an input layer and one or more hidden layers connected one after the other. Each neuron receives input from the previous layer neurons or the input layer. The output of one neuron becomes the input to other neurons in the next layer of the network, and this process continues until the final layer produces the output of the network. The layers of the neural network transform the input data through a series of nonlinear transformations, allowing the network to learn complex representations of the input data.

How machine learning algorithms work

Machine Learning works in the following manner.
A machine learning algorithm works by learning patterns and relationships from data to make predictions or decisions without being explicitly programmed for each task. Here's a simplified overview of how a typical machine learning algorithm works:

1. Data Collection:
First, relevant data is collected or curated. This data could include examples, features, or attributes that are important for the task at hand, such as images, text, numerical data, etc.

2. Data Preprocessing:
Before feeding the data into the algorithm, it often needs to be preprocessed. This step may involve cleaning the data (handling missing values, outliers), transforming the data (normalization, scaling), and splitting it into training and test sets.

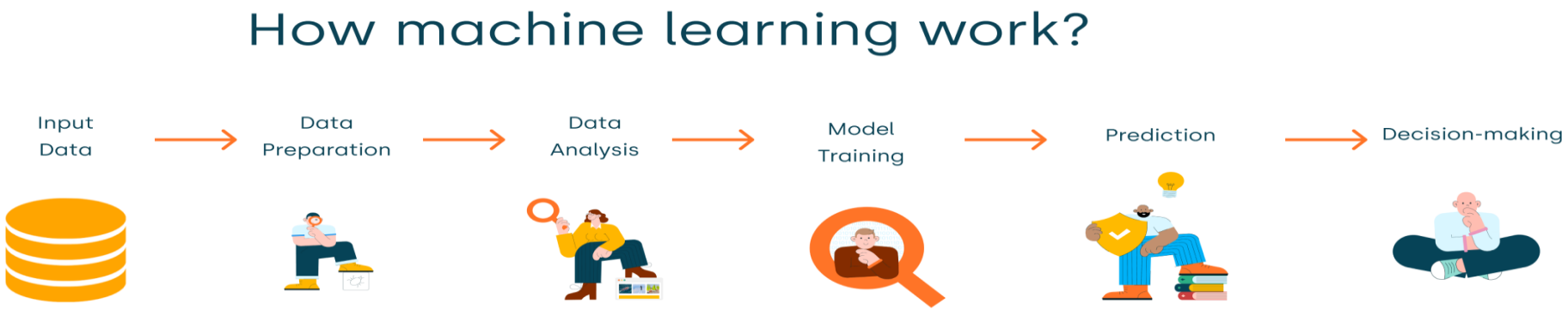
3. Choosing a Model:
Depending on the task (e.g., classification, regression, clustering), a suitable machine learning model is chosen. Examples include decision trees, neural networks, support vector machines, and more advanced models like deep learning architectures.

4. Training the Model:
The selected model is trained using the training data. During training, the algorithm learns patterns and relationships in the data. This involves adjusting model parameters iteratively to minimize the difference between predicted outputs and actual outputs (labels or targets) in the training data.

5. Evaluating the Model:
Once trained, the model is evaluated using the test data to assess its performance. Metrics such as accuracy, precision, recall, or mean squared error are used to evaluate how well the model generalizes to new, unseen data.

6. Fine-tuning:
Models may be fine-tuned by adjusting hyperparameters (parameters that are not directly learned during training, like learning rate or number of hidden layers in a neural network) to improve performance.

7. Prediction or Inference:
Finally, the trained model is used to make predictions or decisions on new data. This process involves applying the learned patterns to new inputs to generate outputs, such as class labels in classification tasks or numerical values in regression tasks.



BLACKBUCKS
GROUP OF COMPANIES

Machine Learning lifecycle:

The lifecycle of a machine learning project involves a series of steps that include:

1. Study the Problems:

The first step is to study the problem. This step involves understanding the business problem and defining the objectives of the model.

2. Data Collection:

When the problem is well-defined, we can collect the relevant data required for the model. The data could come from various sources such as databases, APIs, or web scraping.

3. Data Preparation:

When our problem-related data is collected. then it is a good idea to check the data properly and make it in the desired format so that it can be used by the model to find the hidden patterns. This can be done in the following steps:

- Data cleaning
- Data Transformation
- Explanatory Data Analysis and Feature Engineering
- Split the dataset for training and testing.

4. Model Selection:

The next step is to select the appropriate machine learning algorithm that is suitable for our problem. This step requires knowledge of the strengths and weaknesses of different algorithms. Sometimes we use multiple models and compare their results and select the best model as per our requirements.

5. Model building and Training:

- After selecting the algorithm, we have to build the model.
- In the case of traditional machine learning building mode is easy it is just a few hyperparameter tunings.
- In the case of deep learning, we have to define layer-wise architecture along with input and output size, number of nodes in each layer, loss function, gradient descent optimizer, etc.
- After that model is trained using the preprocessed dataset.

6. Model Evaluation:

Once the model is trained, it can be evaluated on the test dataset to determine its accuracy and performance using different techniques. like classification report, F1 score, precision, recall, ROC Curve, Mean Square error, absolute error, etc.

7. Model Tuning:

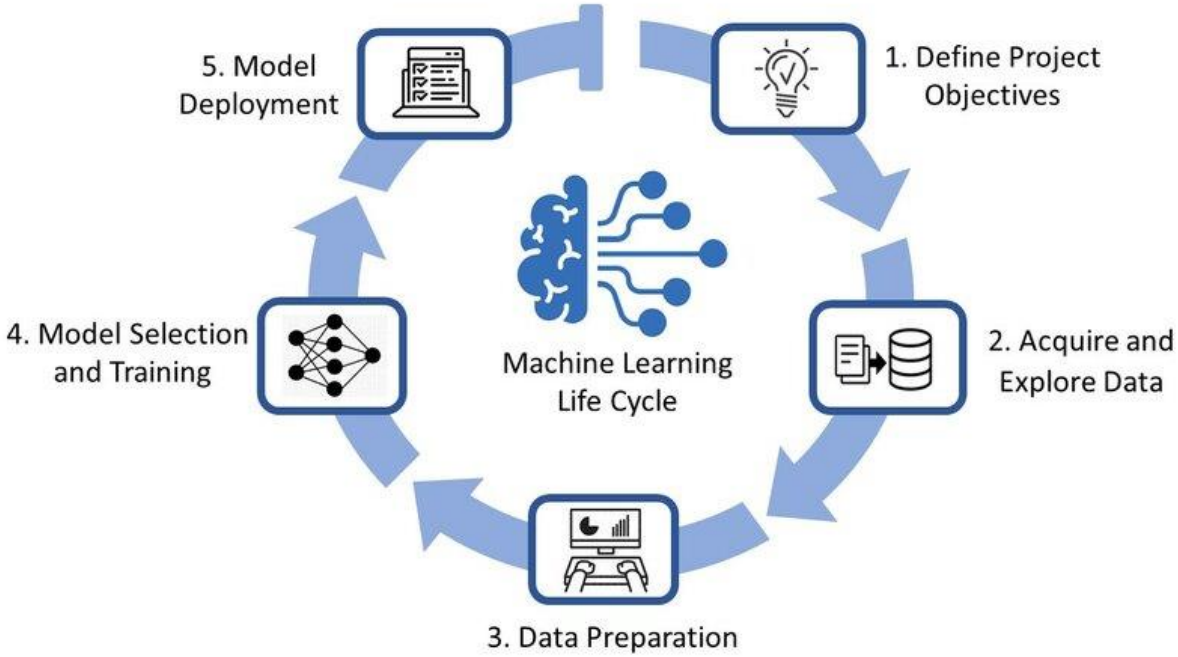
Based on the evaluation results, the model may need to be tuned or optimized to improve its performance. This involves tweaking the hyperparameters of the model.

8. Deployment:

Once the model is trained and tuned, it can be deployed in a production environment to make predictions on new data. This step requires integrating the model into an existing software system or creating a new system for the model.

9. Monitoring and Maintenance:

Finally, it is essential to monitor the model's performance in the production environment and perform maintenance tasks as required. This involves monitoring for data drift, retraining the model as needed, and updating the model as new data becomes available.



Machine learning methods

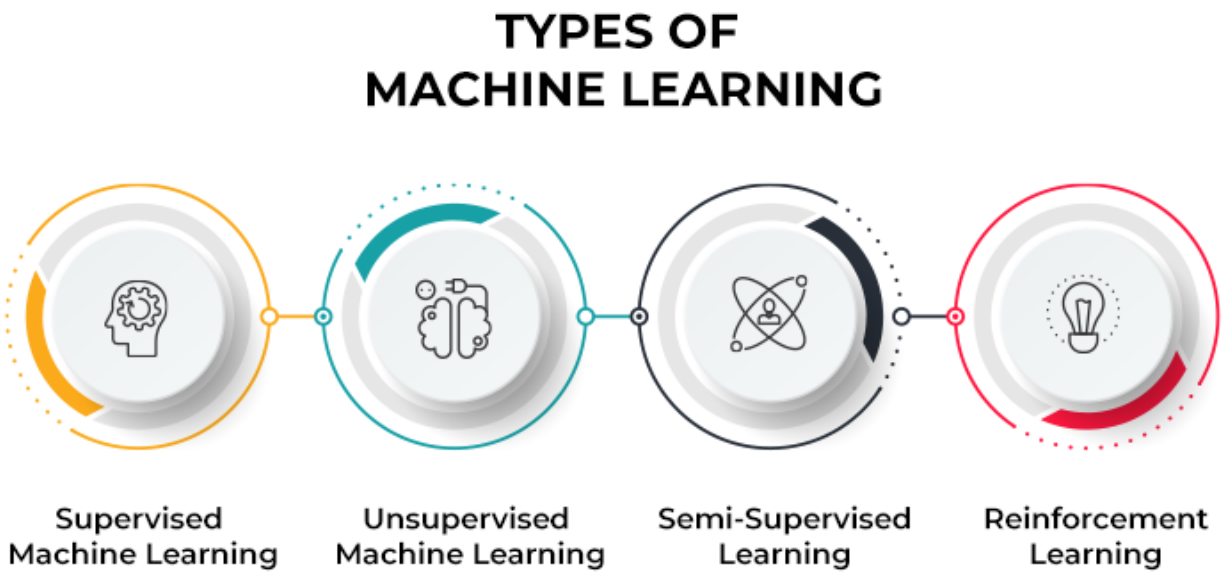
Machine learning models fall into these primary categories.

Supervised machine learning:

Supervised Learning, also known as supervised machine learning, is defined by its use of labeled datasets to train algorithms to classify data or predict outcomes accurately. As input data is fed into the model, the model adjusts its weights until it has been fitted appropriately. This occurs as part of the cross validation process to ensure that the model avoids overfitting or underfitting. Supervised learning helps organizations solve a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox. Some methods used in supervised learning include neural networks, naïve bayes, linear regression, logistic regression, random forest, and support vector machine (SVM).

Unsupervised machine learning:

Unsupervised Learning, also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets (subsets called clusters). These algorithms discover hidden patterns or data groupings without the need for human intervention. This method's ability to discover similarities and differences in information make it ideal for exploratory data analysis, cross-selling strategies, customer segmentation, and image and pattern recognition. It's also used to reduce the number of features in a model through the process of dimensionality reduction. Principal component analysis (PCA) and singular value decomposition (SVD) are two common approaches for this. Other algorithms used in unsupervised learning include neural networks, k-means clustering, and probabilistic clustering methods.



BLACKBUCKS
GROUP OF COMPANIES

Semi-supervised learning:

Semi-supervised learning offers a happy medium between supervised and unsupervised learning. During training, it uses a smaller labeled data set to guide classification and feature extraction from a larger, unlabeled data set. Semi-supervised learning can solve the problem of not having enough labeled data for a supervised learning algorithm. It also helps if it's too costly to label enough data.

Reinforcement Learning:

Reinforcement learning is an area of Machine Learning. It is about taking suitable action to maximize reward in a particular situation. It is employed by various software and machines to find the best possible behavior or path it should take in a specific situation. Reinforcement learning differs from supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of a training dataset, it is bound to learn from its experience.

Difference Between Machine Learning vs Deep Learning vs Artificial Intelligence

Machine Learning

- ML stands for Machine Learning, and is the study that uses statistical methods enabling machines to improve with experience
- ML is the subset of AI.
- ML is an AI algorithm which allows system to learn from data.
- If you have a clear idea about the logic(math) involved in behind and you can visualize the complex functionalities like K-Mean, Support Vector Machines, etc., then it defines the ML aspect.
- The aim is to increase accuracy not caring much about the success ratio.
- Three broad categories/types Of ML are: Supervised Learning, Unsupervised Learning and Reinforcement Learning
- Less efficient than DL as it can't work for longer dimensions or higher amount of data.

Deep Learning

- DL stands for Deep Learning, and is the study that makes use of Neural Networks(similar to neurons present in human brain) to imitate functionality just like a human brain.
- DL is the subset of ML.
- DL is a ML algorithm that uses deep(more than one layer) neural networks to analyze data and provide output accordingly.
- If you are clear about the math involved in it but don't have idea about the features, so you break the complex functionalities into linear/lower dimension features by adding more layers, then it defines the DL aspect.
- It attains the highest rank in terms of accuracy when it is trained with large amount of data.
- DL can be considered as neural networks with a large number of parameters layers lying in one of the four fundamental network architectures: Unsupervised Pre-trained Networks, Convolutional Neural Networks, Recurrent Neural Networks and Recursive Neural Networks
- More powerful than ML as it can easily work for larger sets of data.

Artificial Intelligence

- AI stands for Artificial Intelligence, and is basically the study/process which enables machines to mimic human behavior through particular algorithm.
- AI is the broader family consisting of ML and DL as its components.
- AI is a computer algorithm which exhibits intelligence through decision making.
- Search Trees and much complex math is involved in AI.
- The aim is to basically increase chances of success and not accuracy.
- Three broad categories/types Of AI are: Artificial Narrow Intelligence (ANI), Artificial General Intelligence (AGI) and Artificial Super Intelligence (ASI)
- The efficiency Of AI is basically the efficiency provided by ML and DL respectively.

BLACKBUCKS
GROUP OF COMPANIES

About Project

The use of Machine Learning for Cricket data analysis has become more necessary and important as time passes due to the sheer amount of data generated as the time is passing. Day by day lots and lots of cricket matches are conducted and every time lots of data are created using this we can train and create a model for data analysis for more accurate score prediction and further more tasks to be performed. So, I have created this project which uses following python libraries and a Linear algorithm model for data analysis and score prediction. The project contains following libraries:

1. Numpy:

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more. At the core of the NumPy package, is the ndarray object. This encapsulates *n*-dimensional arrays of homogeneous data types, with many operations being performed in compiled code for performance.

2. pandas:

The name of Pandas is gotten from the word Board Information, and that implies an Econometrics from Multi-faceted information. It was created in 2008 by Wes McKinney and is used for data analysis in Python. Processing, such as restructuring, cleaning, merging, etc., is necessary for data analysis. Numpy, Scipy, Cython, and Panda are just a few of the fast data processing tools available. Yet, we incline toward Pandas since working with Pandas is quick, basic and more expressive than different apparatuses. Pandas is built on top of the Numpy bundle, it is expected that Numpy will work with Pandas.

3. sklearn (scikit-learn):

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon Numpy, Scipy, and Matplotlib.

- train_test_split: For splitting the dataset into training and test sets.
- LinearRegression: For building a linear regression model.
- LabelEncoder: For encoding categorical features.
- mean_absolute_error and mean_squared_error: For evaluating the model performance.

4. Matplotlib:

Matplotlib is a python library used to create 2D graphs and plots by using python scripts. It has a module named pyplot which makes things easy for plotting by providing feature to control line styles, font properties, formatting axes etc. It supports a very wide variety of graphs and plots namely - histogram, bar charts, power spectra, error charts etc. It is used along with NumPy to provide an environment that is an effective open source alternative for MatLab. It can also be used with graphics toolkits like PyQt and wxPython.

5. seaborn:

Seaborn is an amazing visualization library for statistical graphics plotting in Python. It provides beautiful default styles and color palettes to make statistical plots more attractive. It is built on top matplotlib library and is also closely integrated with the data structures from pandas. Seaborn aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs so that we can switch between different visual representations for the same variables for a better understanding of the dataset.

DataSet Columns:

Each dataset consists of following columns(features):

- mid: Each match is given a unique number
- date: When the match happened
- venue: Stadium where match is being played
- bat_team: Batting team name
- bowl_team: Bowling team name
- batsman: Batsman name who faced that ball
- bowler: Bowler who bowled that ball
- runs: Total runs scored by team at that instance
- wickets: Total wickets fallen at that instance
- overs: Total overs bowled at that instance
- runs_last_5: Total runs scored in last 5 overs
- wickets_last_5: Total wickets that fell in last 5 overs
- striker: max(runs scored by striker, runs scored by non-striker)
- non-striker: min(runs scored by striker, runs scored by non-striker)
- total: Total runs scored by batting team after first innings

Data Dictionary

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	mid	date	venue	bat_team	bowl_team	batsman	bowler	runs	wickets	overs	runs_last_5	wickets_last_5	striker
2	1	2006-06-13	Civil Service Cri	England	Ireland	ME Trescothick	DT Johnston	0	0	0.1	0	0	0
3	1	2006-06-13	Civil Service Cri	England	Ireland	ME Trescothick	DT Johnston	0	0	0.2	0	0	0
4	1	2006-06-13	Civil Service Cri	England	Ireland	ME Trescothick	DT Johnston	4	0	0.3	4	0	0
5	1	2006-06-13	Civil Service Cri	England	Ireland	ME Trescothick	DT Johnston	6	0	0.4	6	0	0
6	1	2006-06-13	Civil Service Cri	England	Ireland	ME Trescothick	DT Johnston	6	0	0.5	6	0	0
7	1	2006-06-13	Civil Service Cri	England	Ireland	ME Trescothick	DT Johnston	6	0	0.6	6	0	0
8	1	2006-06-13	Civil Service Cri	England	Ireland	EC Joyce	D Langford-Smit	6	0	1.1	6	0	0
9	1	2006-06-13	Civil Service Cri	England	Ireland	EC Joyce	D Langford-Smit	6	0	1.2	6	0	0
10	1	2006-06-13	Civil Service Cri	England	Ireland	EC Joyce	D Langford-Smit	6	0	1.3	6	0	0
11	1	2006-06-13	Civil Service Cri	England	Ireland	EC Joyce	D Langford-Smit	7	0	1.3	7	0	0
12	1	2006-06-13	Civil Service Cri	England	Ireland	EC Joyce	D Langford-Smit	8	0	1.4	8	0	1
13	1	2006-06-13	Civil Service Cri	England	Ireland	ME Trescothick	D Langford-Smit	8	0	1.5	8	0	1
14	1	2006-06-13	Civil Service Cri	England	Ireland	ME Trescothick	D Langford-Smit	9	0	1.6	9	0	1
15	1	2006-06-13	Civil Service Cri	England	Ireland	ME Trescothick	DT Johnston	10	0	2	10	0	1
16	1	2006-06-13	Civil Service Cri	England	Ireland	ME Trescothick	DT Johnston	10	0	2.1	10	0	1
17	1	2006-06-13	Civil Service Cri	England	Ireland	ME Trescothick	DT Johnston	11	0	2.2	11	0	1
18	1	2006-06-13	Civil Service Cri	England	Ireland	EC Joyce	DT Johnston	11	0	2.3	11	0	1
19	1	2006-06-13	Civil Service Cri	England	Ireland	EC Joyce	DT Johnston	12	0	2.4	12	0	2
20	1	2006-06-13	Civil Service Cri	England	Ireland	ME Trescothick	DT Johnston	12	0	2.5	12	0	2
21	1	2006-06-13	Civil Service Cri	England	Ireland	ME Trescothick	DT Johnston	16	0	2.6	16	0	5
22	1	2006-06-13	Civil Service Cri	England	Ireland	EC Joyce	D Langford-Smit	17	0	3	17	0	5
23	1	2006-06-13	Civil Service Cri	England	Ireland	EC Joyce	D Langford-Smit	17	0	3.1	17	0	5
24	1	2006-06-13	Civil Service Cri	England	Ireland	EC Joyce	D Langford-Smit	21	0	3.2	21	0	6
25	1	2006-06-13	Civil Service Cri	England	Ireland	EC Joyce	D Langford-Smit	21	0	3.3	21	0	6
26	1	2006-06-13	Civil Service Cri	England	Ireland	EC Joyce	D Langford-Smit	21	0	3.4	21	0	6

Cricket Match Score Prediction:
Using Linear Regression Model

Algorithm Used and It's Introduction:

This project includes the Cricket data analysis and focuses
more on score prediction using linear regression model. Linear regression is a data analysis technique that predicts the value of unknown data by using another related and known data value.
It uses a single independent variable to predict a dependent variable by fitting a linear equation to observed data. The value you want to predict is the dependent variable and the value you use to predict the other variable's value is an independent variable. Further details like the features used are given in the README file in github.

The linear regression model was trained using the training datasets. The model's performance was evaluated using metrics such as mean absolute error(MAE) and mean squared error(MSE).
The various steps in this project are:
Data Processing:
1. Data loading and preprocessing
2. Splitting data into training and testing sets
Model Evaluation:
1. Correlation Matrix: To visualize the relationships between variables.
2. Feature Importance: To identify the most significant features influencing the target variables.

Why Did I Use This Algorithm:

1. Linear Regressions are relatively easy to understand and
Apply.
2. They can be used to predict the future values based on the existing data.
3. Linear Regressions are highly interpretable.
4. Linear Regression comes with assumptions that can be diagnosed and verified.
5. Linear Regression is computationally efficient.