Introduction

Cricket, often referred to as the "gentleman's game," is an ancient and popular sport that originated in southeast England in the late 16th century. It became the national sport of England in the 18th century and has since spread worldwide. The International Cricket Council's Cricket World Cup, a one-day international cricket tournament, is the flagship event of the international cricket calendar and takes place every four years with matches played in a 50-over format. It is the biggest cricketing tournament and one of the most widely viewed sporting events in the world. Cricket is a game of uncertainty, with the outcome of a match difficult to predict until the final moments of play. However, probability models can be used to make predictions about the results of matches.

We have taken the data of Virat Kohli across all his formats(ODI, T20 and test matches). We have tried to build the performance predictive model using Machine learning tools. Our aim was to do the analysis of his records across all the formats.

Dataset Used:

The dataset is scrapped from the website https://www.espncricinfo.com/ using python library BeautifulSoup and Selenium. The size of the dataset is small. Given below is the size of the datasets across all the formats

Test match 189*10
One day match: 252*10
T20 match:91*9

Bias Variance Tradeoff

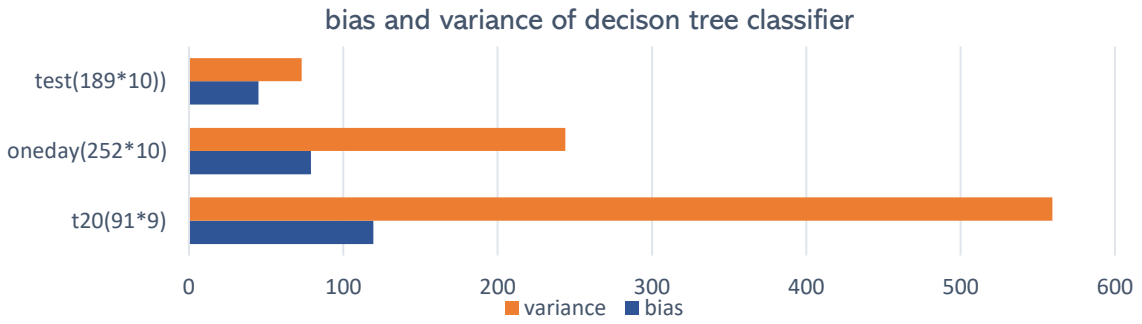
The bias-variance trade-off is a fundamental concept in machine learning and statistical modelling. It refers to the balance between the bias and the variance of a model.

Bias refers to the error introduced by the assumptions made by the model. A model with high bias tends to make predictions that are consistently far from the true values, even on different data sets. On the other hand, variance refers to the variability of a model's predictions. A model with high variance tends to make very different predictions on different data sets, even if they are similar.

In general, a model with low bias and high variance is prone to overfitting, while a model with high bias and low variance is prone to underfitting. The goal in model selection is to find a balance between bias and variance that results in good performance on the test data set. This is known as the bias-variance trade-off.

Since out data set is very small it can lead to a model with higher bias and lower variance. This can be a problem because a model with high bias may not be able to capture the complexity of the data, leading to poor performance, while a model with low variance may not be able to adapt to new data, leading to poor generalization. It is important to balance the bias and variance of a model by selecting an appropriate model complexity. For analysis of the dataset, we have used Decision Tree Classifier for prediction of the out type and for predicting strike rate we have used the linear regression model, since the dataset was the linear.

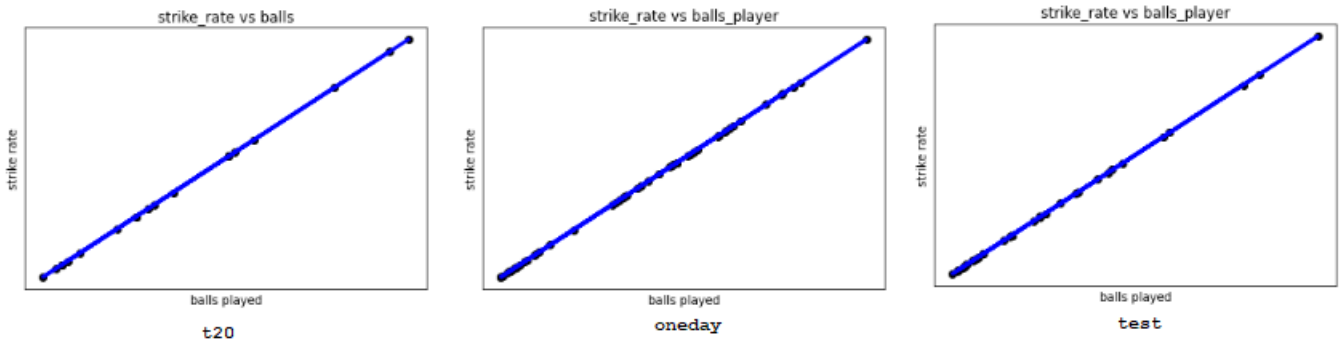
Here is the bar graph showing the bias and the variance of decision tree classifier across all the formats.



Analysis

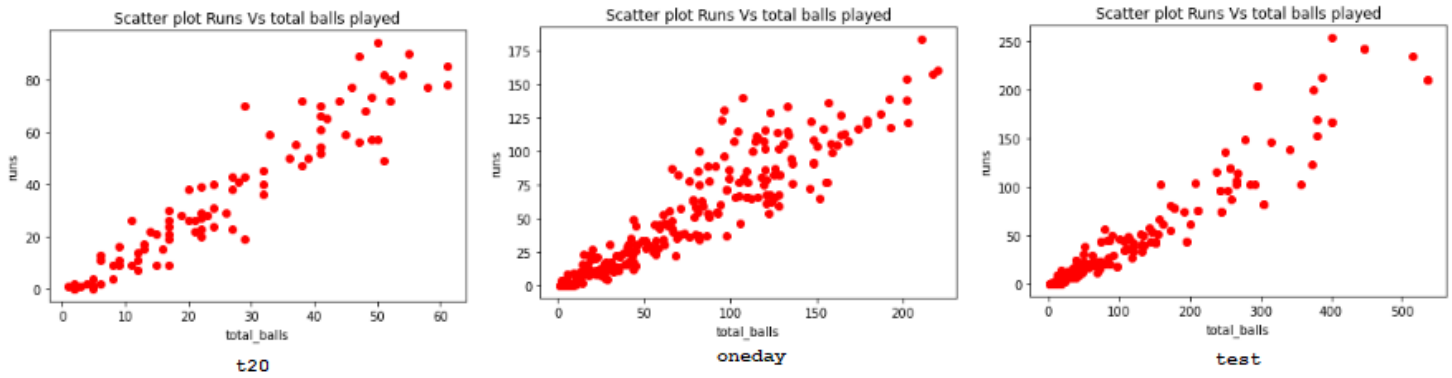
Linear regression Results across all the formats:

Linear regression line strike rate VS balls played across all the formats:



The above figure shows the best fit line of the linear regression across all the formats. The strike rate and the balls played are directly related to each other. Now let's check how the runs and the total balls are related to each other.

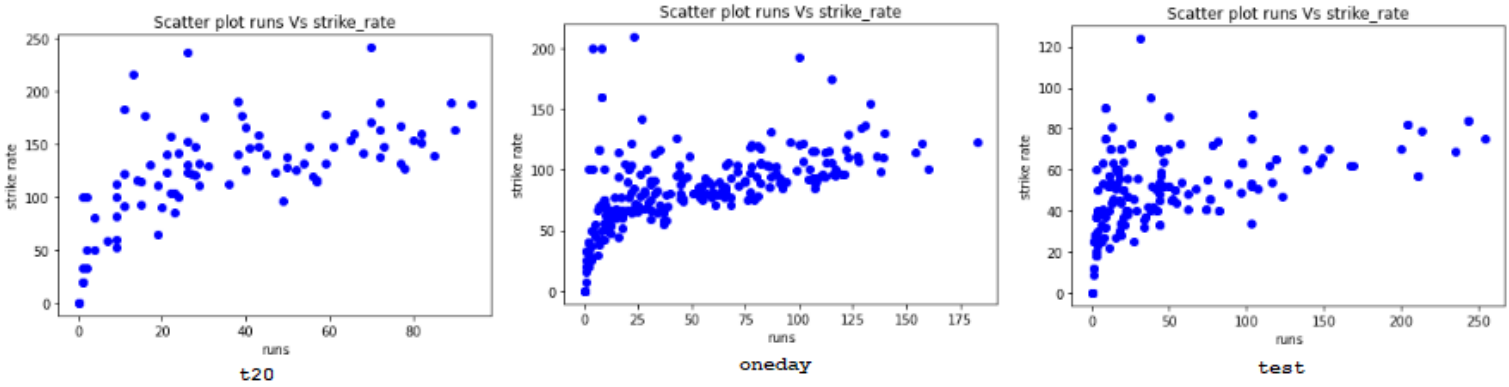
Runs VS balls played across all the formats



We can directly see the difference between the total balls played and the run scored across all the formats. In t20 he played less balls and scored more runs comparatively oneday and test matches.

Now, let's see what the trend for the runs vs strike rate among all the formats. From above figure it should obvious that the strike rate should be more in t20 , then oneday then the test match.

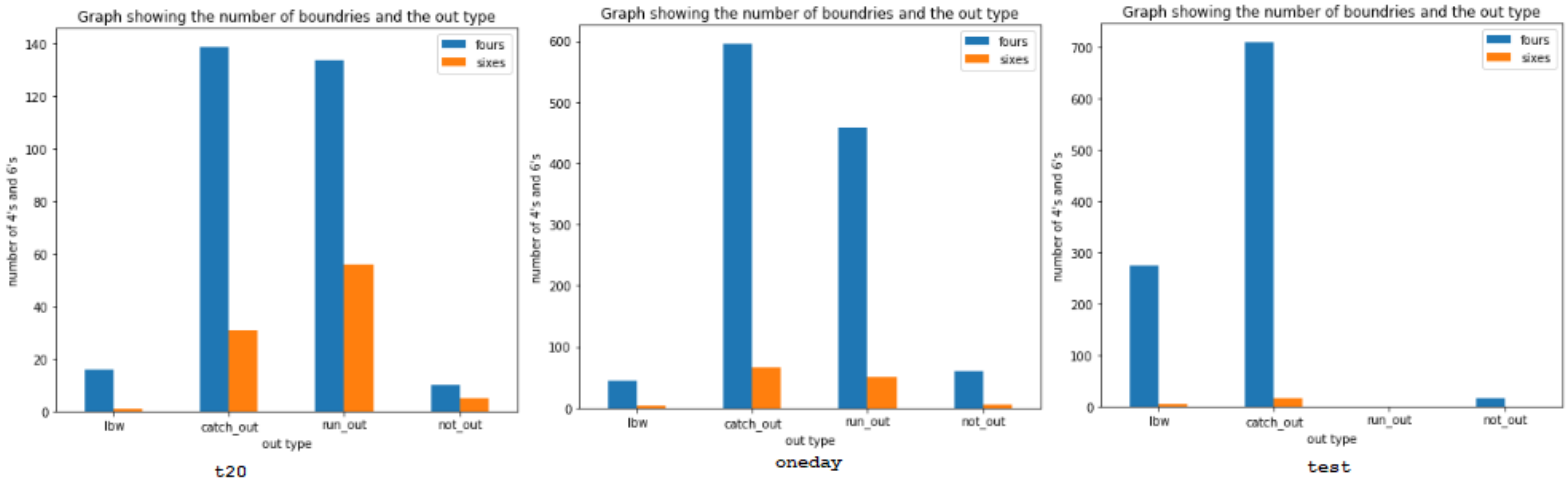
Strike rate VS runs across all the formats



From above figure we can see that most of the strikes rate in t20 formats is around 125 to 150. Where as in oneday it is in between 50 to 100 while in test matches it is in between 40 to 60.

Now let's check that whether scoring more boundaries creates greater risks for catch out. To plot this graph, we have used the bar graphs which is showing the out type and the number of boundaries(sixes, and fours) scored.

Boundaries vs out type across all the formats



From the above figure we found out that the catch type is the most common dismissal across all the formats. And the risk of dismissal is also increasing the probability of hitting boundaries.

Decision Tree Classifier

The decision tree is a supervised learning algorithm that involves dividing data into smaller and smaller subsets until each subset can be classified. It relies on concepts like nodes, edges, and leaf nodes to classify the data. The algorithm starts by calculating the entropy, or uncertainty, of the database. A lower entropy value indicates better classification results. The algorithm then calculates the information gain of each feature, which is a measure of how much uncertainty is reduced by splitting the data on that feature. The data is then split on the feature with the highest information gain, and the process is repeated until all the nodes are cleared.

Here, is the accuracy obtained from across all the formats.

To summarize, this study analyzed the performance of players in three different tournaments. Since we have very small data size therefore, the model may be high biased with higher variance. However, it is also important to note that there may be other factors that can affect the performance of players, such as geographical location, pitch conditions, weather, and lighting conditions in day-night matches. Therefore, it would be worthwhile to also consider these factors when studying the performance of players in order to get a more complete picture.