Methodology

**Data Collection :**

When using any Machine Learning Algorithm that consists of clustering, the initial most important thing to do is to collect data that is of a better quality, because in clustering we are finding the Patterns within the dataset or we are doing some kind of segregation of groups with similar traits and combining them into clusters. For clustering of cricket dataset , data should cover various aspects such as player statistics, such as his runs, strikerate, average and maidens, dots, wickets if it is bowlers data. Datasets are sourced from reliable databases, like official IPL, BBL, ICC records, and verified cricket analytics platforms (Cricbuzz, Espncricinfo, cricmetrics etc..,) to ensure accuracy.

**Data Preprocessing:**

Before conducting clustering analysis, the data needs to go through a process called data preprocessing which aims at preparing it adequately. This contains a variety of mini processes aimed at cleansing, changing and getting it ready to enable the clustering algorithms effective functioning. In batting datasets, we have Highest Score metric, so here if the player is not out then a Asterik(ex:132\*) will be present along with score. So, we have to use replace function and convert into a integer. In bowling datasets, Best bowling figures will be presented as a/b (ex:3/14) format. So, again here we have to convert it into integer.we are going to use a spilt function and convert in to a integer by replacing value with (b-a).Now, coming to null values we are going to replace them with mean or runs or wickets based on the feature.

**Clustering:**

We are going to use three different types of clustering methods for our dataset. They are K–means clustering, Hierarchical Clustering & DBSCAN.

1. **K–means clustering:**

In This Unsupervised learning algorithm we work on the principle of partitioning data into k clusters which do not intersect each other. This algorithm seeks to reduce the sum of squared differences among elements within a given cluster. At the outset, all individual observations are grouped into k clusters depending on some notions of proximity to their respective centriods. Thereafter, averaging on these clusters gives new values of centroids. The centroids should not change again following this process or be stopped after reaching a certain number of iterations. The Elbow method is used to determine the most appropriate cluster number ‘K’. More specifically, this involves how WCSS changes in relation to cluster numbers, with an observation made here that helps pinpoint where WCSS starts slowly reducing after increasing substantially before this point giving an estimate about optimal clusters. Additionally, the silhouette score is used to determine the cluster quality. It shows how much each data point resembles its own cluster. K-means algorithm is both simple and fast, and operates well with round and similar-sized clusters.

**2. Hierarchical clustering:**

Hierarchical clustering groups similar objects into groups called clusters. Clusters are linked together in a tree-like structure called a dendrogram. Each data point starts as its own cluster. Hierarchical clustering involves no calculation of the number of clusters to be predefined. It involves merging of the nearest clusters successively with respect to linkage distances — be it minimum, maximum, average, or any other algorithm such as Ward's method etc .The process goes on until the data falls within one cluster or a certain criterion (typically distance) is achieved. The best number of groups can be established through the increase in the linkage distances along the dendrogram that represents different groups. Thus, the best number of clusters is defined by the jump in linkage distances in the dendrogram at a particular jump cut point.

**3. DBSCAN(Density Based Spatial Clustering of Applications with Noise):**

This clustering algorithm is based on density , it recognises clusters using data point density. Every data item is categorised into one of these three types: core point; border point; and noise depending on epsilon and at the least number of neighbors. One problem is that if they are mixed up then we cannot distinguish between them. DBSCAN will fail or will find difficulty in clustering, when the data is not densely separated , due its definition of density reachability. The problem arises if datapoints are differently dense and non – spherical. It highly depends on initial parameters of Episilon and minimum no of points. If they are taken incorrectly or biased then the resulting cluster formation will be very poor as DBSCAN is sensitive to initial parameters. It will work superbly for datasets with outliers as it does not consider outliers and ignore them. But, In Sports data, a top-class can outplay everyone , then this player will be considered as outlier by DBSCAN as no will be nearing the top player. So, use of DBCAN can be little awkward for Sports bases datasets clusterings.

**Data Driven Ranking Model:**

After Data cleaning and pre-proccesing of datasets we are passing the data to our “Data Driven Ranking Model” , which will take a dataset and no of clusters ‘K’ value as input and outputs a dataset with “Grading” system to players .(i.e.,grade A | grade B) and also gives the best clustering method, which is determined by silhouette scores.

Initially we are scaling and transforming the datasets and then go for clustering process.But,if datasets are going to perform well on raw data then we may skip the scaling process for data.

Now, In the Model we are having all the clustering techniques of K-means, Dbscan & Hierarchical clusterings as different functions, which are going to return the respective silhouette scores and labels of clustered data. The model has a special function named “**best\_clusters**” , which returns the best clustering method and best cluster tables based on silhouette scores of each method.

After clustering, the bigger task we have is to to rank the clusters, in order to know which cluster or player is of higher rank and which is of lower rank (like 0 to A,1 to B etc..,). And, we have to use different ranking or grading system for both batting and bowling datasets. So, In the model a function “**grade\_clusters”** will be called after clustering . Initially we check for type of dataset that is being passed for grading, we can do this by checking if [‘**Wickets**’] column is present in dataset, which automatically tells us that it is a bowling dataset. so we will rank the clusters based on the average no of wickets for each cluster by using ‘*groupby*’ method. We rank cluster having higher average of wickets as ‘A’ and remains cluster as ‘B’.(as, better bowler has more wickets).

Now, if the [‘**Wickets**’] column is not present in dataset, we can identify it as a batting dataset. so ,we are going to use another metric for ranking batting data, as we cannot use the same metric used for bowling dataset. In cricket a better batsmen always as more no runs , as in the case of bowlers with wickets, but in game of whiteball cricket (t20 & ODI) especially T20’s strike rate at which they score is also a important factor in determining his level.So, for batting data we are going to take the average of [‘Runs’] + [‘strikerate’] for each clusters by ‘*groupby’* method and cluster having higher combination of average gets ‘A’ and others get corresponding lower grades(i.e..,’B).’

The model also plots the dataset with Cluster defining points like one cluster is represented in yellow colour and other with red / purple etc. The model uses PCA (Principal component analysis) in order to find top two features (like runs, Inns from batting and maidens, wickets from bowling data) from the dataset and plotting them for defining clusters. The no of components passed for PCA will be same as ‘K’ value given as input for no of clusters.

One thing to be noted, grades are given from all alphabets(A-Z), the no of grades to be considered will be taken from “K’ value using ‘ :len() ’ function.

After grading the given datasets the results are stored in new files with a added feature of “GRADE”, which represents the corresponding players grade/rank.The model also returns the best clustering method for respective dataset along with the silhouette score for that clustering method.