Building a Data Driven Ranking System for Corporate Decision Making

Faculty Name: Sarath S

Course Code: 22AIE213 – Machine Learning

GROUP-4

TEAM MEMBERS	ROLL NO:
BHAVIKA GONDI	AM.EN.U4AIE22013
DVSS SWAPNITH	AM.EN.U4AIE22016
RAHUL JOGI	AM.EN.U4AIE22020
MANNE LEELA NARESH	AM.EN.U4AIE22030

Abstract

The primary goal of this project is to develop a computer program that can rank various business entities, such as products, customers, routes, and projects, to assist management in making informed decisions. The program will learn from historical business data, which includes product performance data, customer review data, data on different orders from a delivery sight and market trends. The model will classify and rank these items according to how they affect predetermined business goals by using clustering algorithms. The quality of the clustering will determine how well the ranking model performs. The model's capacity to raise these ranks over time as it learns more from the data will be what determines how effective it is. With constant data inputs, this method guarantees that the model not only pinpoints the most important entities for company success, but also improves its ranking accuracy over time.

Table of Contents

Introduction

In today's fast-paced world, even small tasks often present multiple options making it challenging to choose the best one. This complexity is magnified in large companies, where each decision can involve numerous choices and significant costs. Decisions may also be time-consuming and constrained by various factors. In such scenarios, a data-driven ranking system becomes an invaluable tool for evaluating and prioritizing options based on relevant data.

The primary advantages of utilizing this model include reliance on data rather than subjective judgment, ensuring fair and unbiased decisions. Additionally, by clearly ranking options, the decision-making process becomes faster, ultimately saving time and resources. Furthermore, evaluating the potential costs and benefits of each option allows companies to make more cost-effective decisions. These benefits underscore the importance and utility of a data-driven ranking system in enhancing decision-making processes in a corporate setting. This project will explore existing approaches and methodologies, aiming to contribute a robust framework for developing an effective ranking system.

Dataset analysis

1. **IPL dataset:**

Initially, the IPL dataset captures the performance of the players from the beginning of the tournament 2017 to 2027, which includes both batting and bowling statistics. The batting dataset shows 239 records that have 7 attributes which are runs scored, balls faced, fours, and sixes hit and the strike rate. There are 6 numerical attributes and 1 categorical attribute in the dataset and 6 columns that contain missing data. The bowling dataset has 214 records with 7 attributes which are overs bowled, maidens, runs given, wickets taken, and dot balls, almost all of these are numerics 1 relational, and there is no missing data in any row. This dataset is made up of separate CSV files covering a 5-year period from 2017 to 2021 with details being listed in each file consisting of players' overall and yearly performances in the game.

2. **BBL dataset:**

The Big Bash League dataset for the 2020-2021 season gives details of both batting and bowling statistics. The batting dataset describes 150 records with 13 attributes that catch information such as runs scored, balls faced, average, strike rate, boundaries, and milestones. It involves 11 numerical attributes and 2 categorical attributes of which 13 columns are with missing data. The bowling dataset includes 102 records with 13 different performance metrics, including average, economy rate, strike rate, runs given, and wickets taken. There are no missing data points; it consists of 10 numerical attributes and 3 categorical attributes. The datasets display the player's performance that has taken part in at least one innings to both leagues respectively.

3. **World Cup ODI dataset:**

The World Cup ODI dataset concentrates on player performance in One Day International World Cup matches to be able to display relevant information to this new generation of players. The batting dataset consists of 146 records with 12 attributes related to runs scored, balls faced, strike rate, and boundaries. It has 11 numeric attributes and 1 categorical attribute with no missing data. The bowling dataset has 102 records with 14 attributes: overs bowled, runs given, wickets taken, and economy rate. It contains 13 numerical attributes and 1 categorical attribute, and for the latter, there is no missing data. Therefore, mentioned data provide in-depth insights into the player's performance during that time.

Literature Review

Clustering, an unsupervised machine learning technique, groups data points based on similarity to maximize intra-cluster cohesion and minimize inter-cluster disparity. Internal validation techniques, such as the S_Dbw index, evaluate clustering quality by assessing monotonicity, noise resilience, and data characteristics handling. External validation compares outcomes against ground truth using metrics like entropy, purity, mutual information, and the Rand statistic. Consensus methods, including K-means-based approaches, enhance stability across various applications, such as corporate information extraction, bike-sharing optimization, and healthcare decision-making. [1]

Machine learning (ML) models like Naive Bayes, K-Nearest Neighbours (KNN), and Random Forest (RF) have revolutionized cricket team selection by predicting player categories and optimizing team composition. Recent studies highlight RF's effectiveness in leveraging performance metrics from tournaments like the Dhaka Premier League and ICC events, including batting averages, bowling accuracy, and match-specific statistics from platforms such as ESPNcricinfo and Cricbuzz. These data-driven approaches enhance objectivity and accuracy in player selection, with ongoing research focused on algorithm refinement and real-time analytics integration. [2]

Advances in sports analytics, particularly in cricket, demonstrate the growing adoption of ML algorithms over traditional statistical methods. Decision Trees, Random Forests, and Support Vector Machines (SVM) excel in categorizing players based on diverse attributes such as batting averages and match conditions. These algorithms improve prediction accuracy and support strategic decision-making in team selection and game planning. Challenges remain in data quality assurance and model interpretability, prompting continuous efforts to refine ML models and integrate advanced analytics for real-time insights in cricket management and other sports domains. [3]

Computational techniques, including advanced clustering methods like Spectral Clustering, enhance cricket team management by recommending substitutes based on player metrics such as bowling economy and batting averages. Similarity metrics such as Euclidean distance and Pearson Correlation Coefficient optimize team compositions, validated against actual team selections to show high accuracy, particularly with the Pearson Correlation Coefficient. These techniques suggest potential applications in state-level player selections and real-time strategy implementations. [4]

Cricket analytics, especially in T-20 formats, has evolved with ML algorithms like K-Means clustering and Gaussian Mixture Models categorizing players into roles such as batsmen, bowlers, and all-rounders. Deep Learning (DL) approaches like auto-encoders extract latent features, improving clustering accuracy and supporting predictive modelling for match outcomes and strategic decision-making. This integration underscores the transformative impact of computational techniques in optimizing team compositions and evaluating performance in modern cricket analytics. [5]

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, strike rate, 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 an 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 an 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 centroids. 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 recognizes clusters using data point density. Every data item is categorized 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 Epsilon 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 clustering.

Data Driven Ranking Model:

After Data cleaning and pre-processing 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, dB scan & Hierarchical clustering's 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 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 ‘*group by*’ 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 batsman always as more no runs, as in the case of bowlers with wickets, but in game of white ball cricket (t20 & ODI) especially T20’s strike rate at which they score is also an important factor in determining his level. So, for batting data we are going to take the average of [‘Runs’] + [‘strike rate’] for each cluster by ‘*group by’* method and cluster having higher combination of average gets ‘A’ and others get corresponding lower grades(i.e.).’

The model also plots the dataset with Cluster defining points like one cluster is represented in yellow color and other with red / purple etc. The model uses PCA (Principal component analysis) 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 an added feature of “GRADE”, which represents the corresponding player’s grade/rank. The model also returns the best clustering method for respective dataset along with the silhouette score for that clustering method.

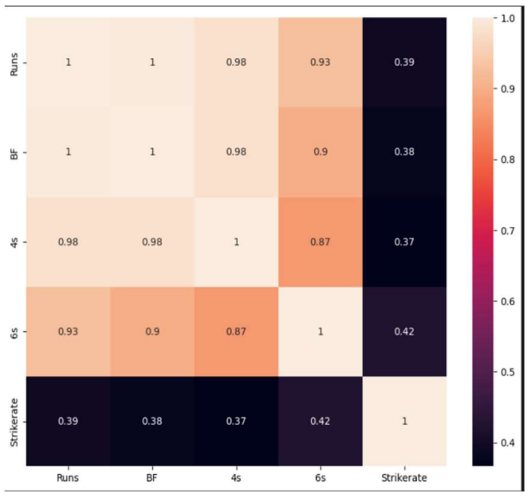
Results

Data Preprocessing Results:

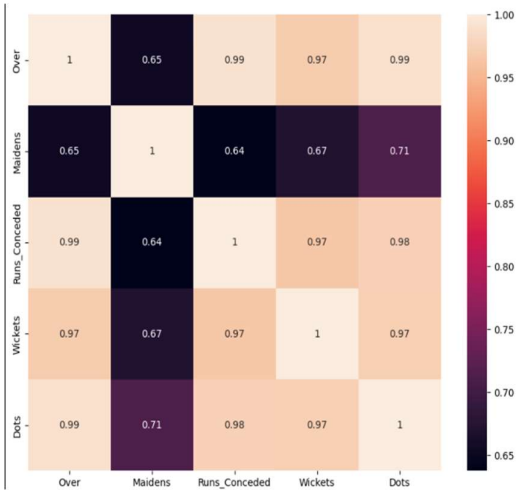
Heatmaps were mainly used in case of analysis of data during preprocessing.

IPL:

Batting:



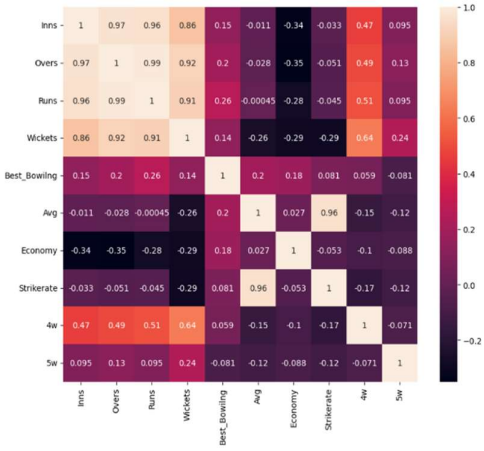
Bowling:



BBL:

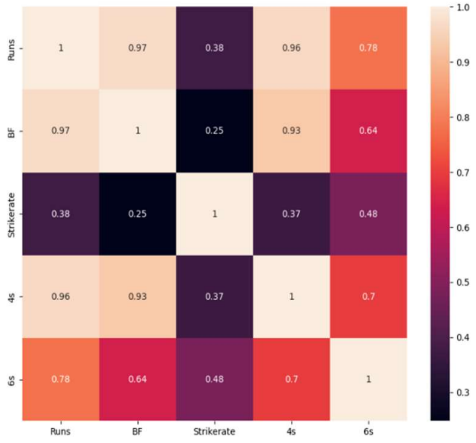
Batting:

Bowling:



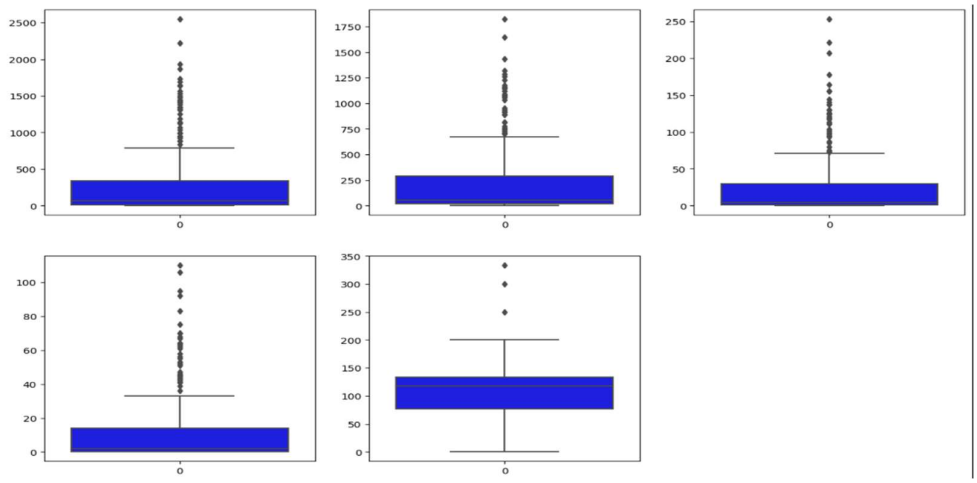
World cup ODI:

Batting:



Bowling:

Outliers in IPL Dataset which are calculated using IQR method, below corresponding boxplots are shown: these are for batting dataset



Clustering Results:

After data preprocessing clustering was done to each dataset and following results were obtained for each dataset. Best cluster is chosen based on cluster performance which was evaluated using silhouette scores.

Table 1: Silhouette Scores for Clustering Methods on the IPL dataset

Clustering Method	Data Type	No. of Clusters	Silhouette Score
K-Means Clustering	Batting	3	0.79
	Bowling	3	0.70
Hierarchical Clustering	Batting	3	0.74
	Bowling	3	0.68
DBSCAN	Batting	2	0.53
	Bowling	2	0.43

Table 2: Silhouette Scores for Clustering Methods on the BBL dataset

Clustering Method	Data Type	No. of Clusters	Silhouette Score
K-Means Clustering	Batting	2	0.644
	Bowling	2	0.644
Hierarchical Clustering	Batting	2	0.67
	Bowling	2	0.68
DBSCAN	Batting	2	0.40
	Bowling	2	0.19

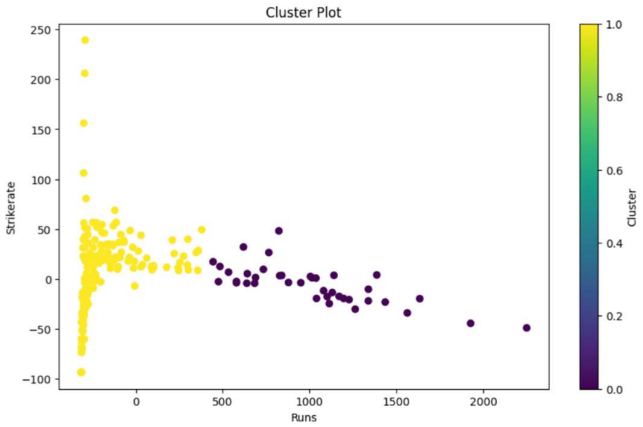
Table 3: Silhouette Scores for Clustering Methods on World Cup ODI dataset

Clustering Method	Data Type	No. of Clusters	Silhouette Score
K-Means Clustering	Batting	2	0.66
	Bowling	2	0.62
Hierarchical Clustering	Batting	2	0.66
	Bowling	2	0.58
DBSCAN	Batting	2	0.48
	Bowling	2	0.36

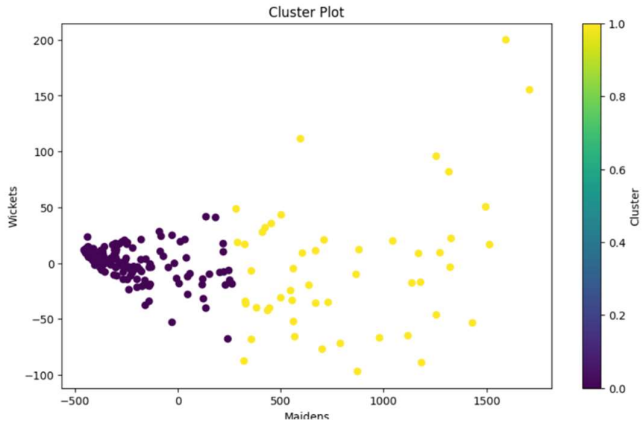
Using the results above, the cluster performance was evaluated, and the following clusters were obtained for each dataset.

IPL DATASET:

BATTING DATA:

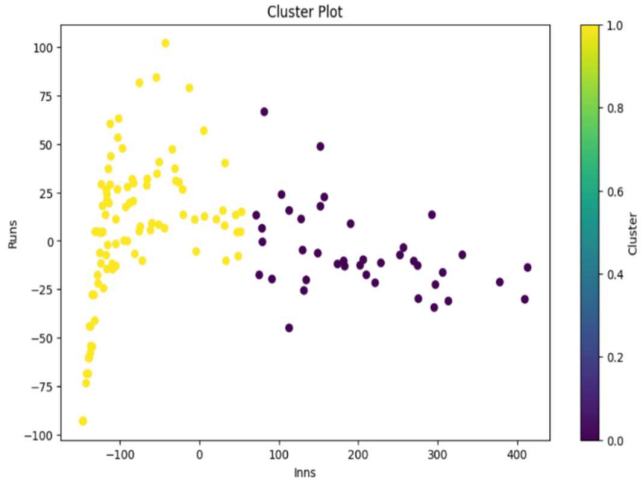


BOWLING DATA:

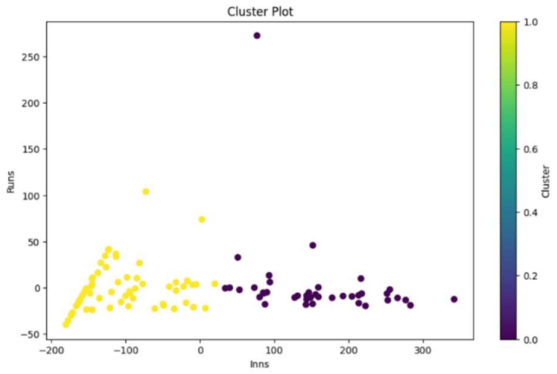


BBL DATASET:

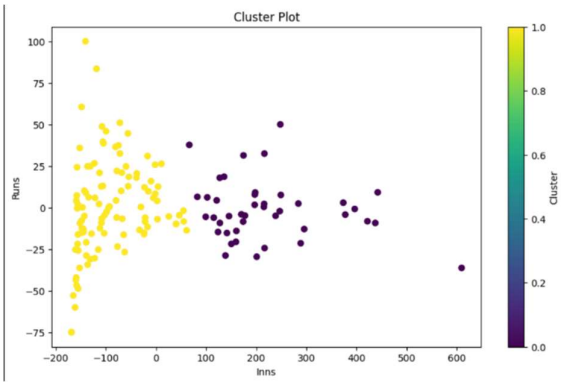
BATTING DATA:



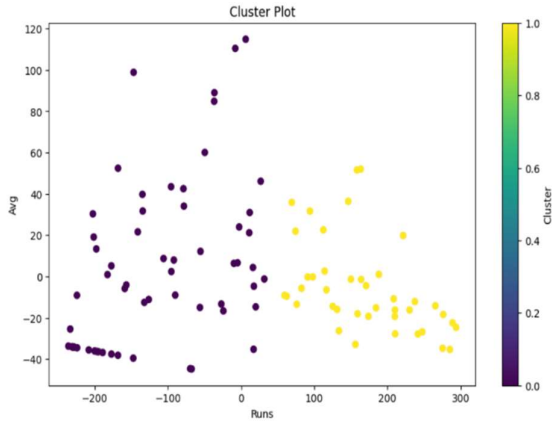
BOWLING DATA:



WORLDCUP ODI DATASET:
BATTING DATA:



BOWLING DATA:



After the clustering analysis, the final step involved assigning grades to players based on the clusters identified. This was done using our data-driven ranking model to present the following ranked results for each pre-processed dataset. The clusters identified were divided into two groups: Grade A and Grade B, with both grading orders being aligned.

Rankings:

IPL DATASET:

BATTING DATA:

GRADE A:

	Player	Runs	Strikerate	4s	6s	Cluster	Grade
0	Ruturaj Gaikwad	839.0	132.125984	80.0	29.0	0	A
1	Faf du Plessis	1640.0	133.659332	155.0	58.0	0	A
2	Robin Uthappa	944.0	124.538259	85.0	43.0	0	A
4	Shubman Gill	1417.0	123.003472	137.0	36.0	0	A
6	Nitish Rana	1383.0	133.623188	130.0	64.0	0	A

GRADE B:

	Player	Runs	Strikerate	4s	6s	Cluster	Grade
3	Moeen Ali	666.0	146.373626	52.0	42.0	1	B
5	Venkatesh Iyer	370.0	128.472222	37.0	14.0	1	B
7	Sunil Narine	683.0	168.226601	70.0	45.0	1	B
8	Eoin Morgan	551.0	124.943311	40.0	30.0	1	B
10	Shakib Al Hasan	295.0	115.686275	29.0	6.0	1	B

BOWLING DATA:

GRADE A:

	Player	Maidens	Wickets	Dots	Runs_Conceded	Cluster	Grade
0	Shakib Al Hasan	0	20	177	769	1	A
1	Shivam Mavi	2	25	218	715	1	A
3	Varun Chakravarthy	0	36	268	839	1	A
4	Sunil Narine	1	58	564	1889	1	A
6	Deepak Chahar	6	59	570	1681	1	A

GRADE B:

	Player	Maidens	Wickets	Dots	Runs_Conceded	Cluster	Grade
2	Lockie Ferguson	1	24	197	648	0	B
5	Venkatesh Iyer	0	3	14	69	0	B
7	Josh Hazlewood	0	12	118	357	0	B
17	George Garton	0	3	32	135	0	B
20	Glenn Maxwell	0	18	133	561	0	B

BBL DATASET:

BATTING DATA:

GRADE A:

	Player	Inns	Runs	Avg	Strikerate	4s	6s	Cluster	Grade
0	Alex Hales	15.0	537.0	38.36	159.82	54	30	0	A
1	James Vince	16.0	536.0	38.29	143.32	59	11	0	A
2	Josh Phillippe	16.0	504.0	31.50	148.24	55	14	0	A
3	Chris Lynn	13.0	456.0	35.08	154.05	39	26	0	A
4	Colin Munro	15.0	443.0	31.64	128.03	32	19	0	A

GRADE B:

	Player	Inns	Runs	Avg	Strikerate	4s	6s	Cluster	Grade
40	Sam Heazlett	9.0	180.0	22.50	135.34	14	9	1	B
41	Joe Burns	11.0	180.0	22.50	125.00	18	5	1	B
42	Aaron Finch	13.0	179.0	13.77	113.29	20	3	1	B
43	Marnus Labuschagne	6.0	176.0	29.33	123.08	13	3	1	B
44	Lewis Gregory	12.0	173.0	19.22	133.08	13	8	1	B

BOWLING DATA:

GRADE A:

	Player	TEAM	Inns	Runs	Avg	Strikerate	Wickets	Best_Bowilng	Economy	4w	5w	Cluster	Grade
0	Jhye Richardson	Perth Scorchers	17	463	15.965517	12.724138	29	20	7.528455	2	0	0	A
1	Ben Dwarshuis	Sydney Sixers	13	403	16.791667	11.375000	24	9	8.857143	1	0	0	A
2	Mark Steketee	Brisbane Heat	16	522	21.750000	14.083250	24	29	9.266327	1	0	0	A
3	Wes Agar	Adelaide Strikers	15	457	20.772727	14.909182	22	23	8.359705	1	0	0	A
4	Tanveer Sangha	Sydney Thunder	15	394	18.761905	13.571429	21	10	8.294737	1	0	0	A

GRADE B:

	Player	TEAM	Inns	Runs	Avg	Strikerate	Wickets	Best_Bowilng	Economy	4w	5w	Cluster	Grade
20	Mujeeb Ur Rahman	Brisbane Heat	8	188	13.428571	12.857143	14	10	6.266667	0	1	1	B
26	Nathan Coulter-Nile	Melbourne Stars	6	172	15.636364	11.908909	11	6	7.877983	1	0	1	B
29	Zak Evans	Melbourne Renegades	5	146	14.600000	9.000000	10	28	9.733333	0	1	1	B
30	Marnus Labuschagne	Brisbane Heat	6	146	14.600000	10.200000	10	10	8.588235	0	0	1	B
34	Brendan Doggett	Sydney Thunder	5	120	13.333333	12.066667	9	18	6.629834	1	0	1	B

WORLD CUP ODI DATASET:

BATTING DATA:

GRADE A:

	Player	Inns	Runs	Avg	Strikerate	4s	6s	Cluster	Grade
1	DJ Malan	9	404	44.89	101.00	50	9	0	A
2	JE Root	9	276	30.67	88.46	21	2	0	A
11	DP Conway	10	372	41.33	101.64	54	4	0	A
13	R Ravindra	10	578	64.22	106.45	55	17	0	A
14	Fakhar Zaman	4	220	73.33	122.91	14	18	0	A

GRADE B:

	Player	Inns	Runs	Avg	Strikerate	4s	6s	Cluster	Grade
0	JM Bairstow	9	215	23.89	88.48	27	3	1	B
3	HC Brook	6	169	28.17	112.67	20	5	1	B
4	MM Ali	6	95	15.83	74.80	8	1	1	B
5	JC Buttler	9	138	15.33	97.18	11	5	1	B
6	LS Livingstone	6	60	10.00	63.83	6	0	1	B

BOWLING DATA:

GRADE A:

	Player	Runs	Avg	Strikerate	Wickets	Best_Bowilng	Economy	4w	6w	Cluster	Grade
0	A Dutt	426	42.60	46.50	10	3/44	5.50	0	0	1	A
1	A Zampa	515	22.39	25.04	23	4/8	5.36	3	0	1	A
7	AU Rashid	413	27.53	31.87	15	3/42	5.18	0	0	1	A
9	BFW de Leede	487	30.44	25.12	16	4/62	7.27	1	0	1	A
12	CAK Rajitha	336	42.00	36.75	8	4/50	6.86	1	0	1	A

GRADE B:

	Player	Runs	Avg	Strikerate	Wickets	Best_Bowilng	Economy	4w	6w	Cluster	Grade
2	AAP Atkinson	146	36.50	36.00	4	2/45	6.08	0	0	0	B
3	AD Mathews	107	17.83	22.17	6	2/14	4.83	0	0	0	B
4	Agha Salman	46	0.00	0.00	0	0/21	9.20	0	0	0	B
5	AK Markram	85	85.00	111.00	1	1/23	4.59	0	0	0	B
6	AL Phehlukwayo	36	36.00	42.00	1	1/36	5.14	0	0	0	B

Interpretations

Data preprocessing results:

Using a heatmap, which is essentially a correlation matrix of attributes, we identified and dropped some columns to select the most relevant attributes for clustering and cluster analysis. This step ensured that the chosen attributes were optimal for producing meaningful clusters.

The boxplot demonstrated in the preprocessing of the data revealed to us we can get the idea that many outlines are present in this dataset. Analogous boxplots for the other datasets let us infer that outliers are a common feature of all the datasets, so the assumption is that the outliers are merely genuine data. In contrast, the true data points are the outliers.

Clustering Results:

As stated, grouping was the method used with the best silhouette score for each record. In the case of IPL match datasets, both the batting and bowling datasets achieved the highest silhouette scores with the K-means algorithm. Hence the last clusters were formed using the K-means algorithm for both IPL batting and bowling data.

For the BBL match datasets, the Hierarchical method scored the highest silhouette for both batting and bowling datasets. Then, of the Hierarchical algorithms, clustering was done for both BBL batting and bowling data.

In the case of the World Cup ODI match datasets, the batting dataset had the highest silhouette scores with both K-means and Hierarchical algorithms. The model could have used one or both algorithms to make the clusters. In the case of the World Cup ODI bowling dataset, the K-means algorithm had the highest silhouette score, thus clustering was done using the K-means algorithm.

The lowest silhouette score was observed for the DBSCAN algorithm in every dataset to other algorithms. The main reason is the impact of the outliers. 10-20% of the data is considered as outliers in each dataset. While K-means and Hierarchical clustering allow these outliers to be in some clusters, DBSCAN categorizes them as noise points. Therefore, when 10-20% noise is labelled in the data, the silhouette score is instantly decreased. Moreover, our datasets have different densities, and in environments with different densities, it is known that DBSCAN fails, so it contributes to the lower silhouette scores.

Cluster analysis Results:

In accordance with the methodology, the data collected from the clusters are subjected to cluster analysis which is simply grading the clusters based on the chosen attributes and, in the clusters that have been graded, the required data has been sorted and ranked.

Conclusion

This project has created a new approach that relies on using historical business data through clustering algorithms. At the end of the system, the work was done on IPL, BBL, and World Cup ODI datasets, accompanied by careful preprocessing that dealt with the missing data, outliers, and attribute selection. For the IPL dataset K-means clustering gave best performance, while Hierarchical clustering was the best with the BBL dataset. In the World Cup ODI, datasets had benefits from both K-means and the hierarchical clustering methods. However, DBSCAN was not quite effective due to its high sensitivity to densities and outliers in the data. The ranking approach effectively divided the companies into two categories, grade A and grade B, based on the financial data received. It is a method for creating a crystal of priorities. The method takes us to make the right decision by using the correct and fair data while still being the most efficient in the corporate setting in the end enhancing the corporate decision-making process.

Future Scope

Developing new methods is the aim for the future of this project which will allow the ranking model to be more flexible and to be used in a great variety of spheres. This involves a wide range of methods to optimize the delivery routes for delivery services and an evaluation of the products based on sales performance for e-commerce businesses. It is highly possible to ameliorate the algorithm performance with an improved technique such as Spectral Clustering and Gaussian Mixture Models. Both the integration of hybrid models that take advantage of multiple clustering algorithms and the development of models through machine learning that can handle diverse datasets more effectively will be considered as well as the confusion of the model's applicability and accuracy across different business scenarios.

References

[1] Sun, L., Chen, G., Xiong, H., & Guo, C. (2017). Cluster analysis in data-driven management and decisions. Journal of Management Science and Engineering, 2(4), 227-251.

[2] Robel, M., Khan, M. A. R., Ahammad, I., Alam, M. M., & Hasan, K. (2024). Cricket players selection for national team and franchise league using machine learning algorithms. Cloud Computing and Data Science, 108-139.

[3] Bharadwaj, F., Saxena, A., Kumar, R., Kumar, R., Kumar, S., & Stević, Ž. (2024). Player Performance Predictive Analysis in Cricket Using Machine Learning. Revue d'Intelligence Artificielle, 38(2).

[4] Das, N. R., Mukherjee, I., Patel, A. D., & Paul, G. (2023). An intelligent clustering framework for substitute recommendation and player selection. The Journal of Supercomputing, 79(15), 16409-16441.

[5] Bose, A., Mitra, S., Ghosh, S., Ghosh, R., Patra, T., & Chakrabarti, S. (2021). Unsupervised learning-based evaluation of player performances. Innovations in Systems and Software Engineering, 17(2), 121-130.