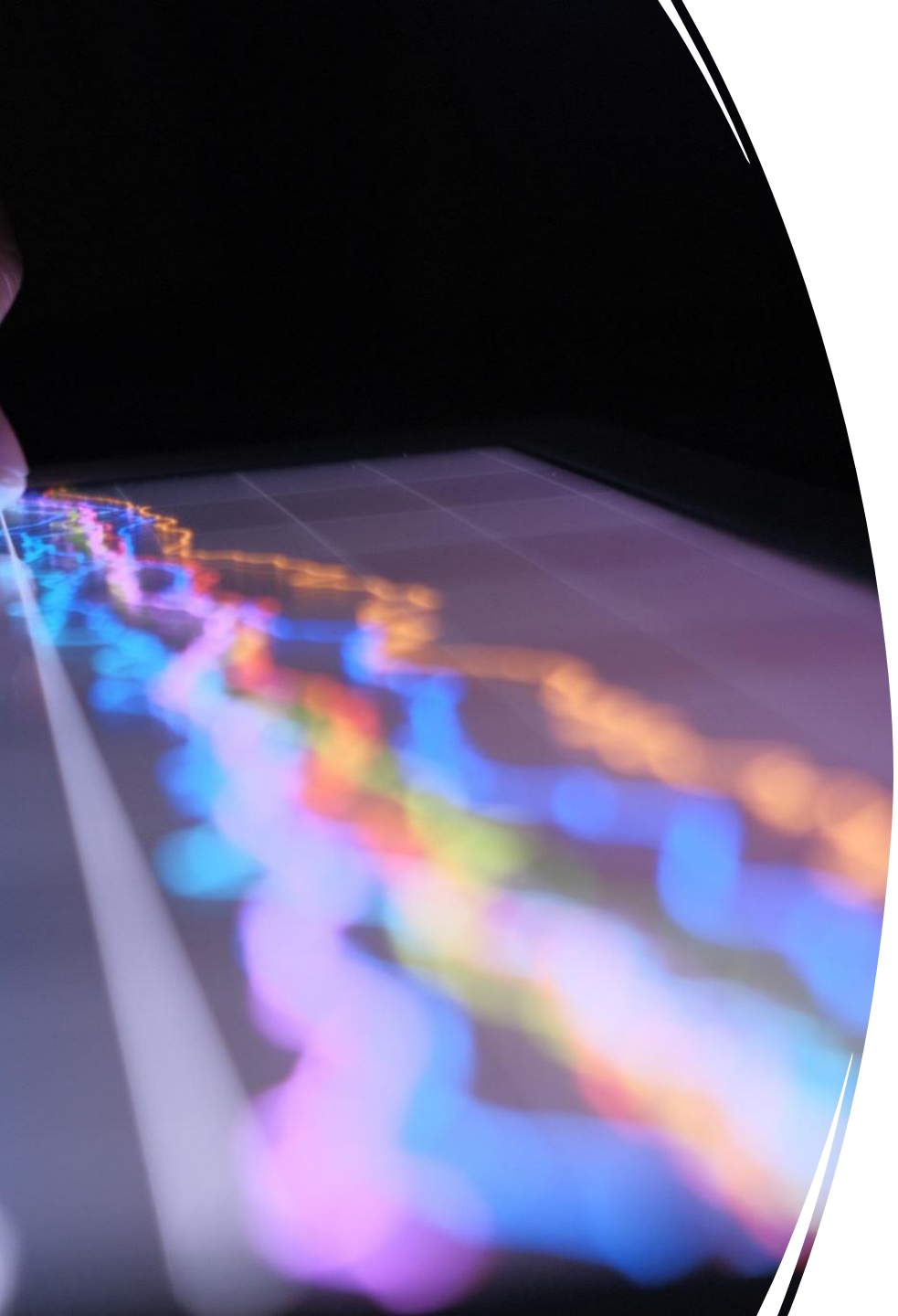
Data Analysis on Cricket dataset Big Data-CSCI 527

Sandeep Ramakrishnan

Mentored By: Milton King

Introduction

- Cricket is a bat-and-ball game played between two teams of eleven players on a field at the center of which is a 22-yard pitch.
- Cricket is a sport where it is difficult to anticipate the performance of a player.
- Our model can predict the performance of a cricketer in terms of his batting performance which is an efficient as it help recruiters in a selection process.



Dataset

- The dataset has the batting statistics of different player from the world.
- The dataset has 2500 rows of unique values.
- The dataset has information of retired player and also Active player.

Dataset

Player	Span	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	0
SR Tendulkar (INDIA)	1989-2012	463	452	41	18426	200*	44.83	21367	86.23	49	96	20
KC Sangakkara (Asia/ICC/SL)	2000-2015	404	380	41	14234	169	41.98	18048	78.86	25	93	15
RT Ponting (AUS/ICC)	1995-2012	375	365	39	13704	164	42.03	17046	80.39	30	82	20
ST Jayasuriya (Asia/SL)	1989-2011	445	433	18	13430	189	32.36	14725	91.2	28	68	34
DPMD Jayawardene (Asia/SL)	1998-2015	448	418	39	12650	144	33.37	16020	78.96	19	77	28
Inzamam-ul-Haq (Asia/PAK)	1991-2007	378	350	53	11739	137*	39.52	15812	74.24	10	83	20
V Kohli (INDIA)	2008-2019	242	233	39	11609	183	59.84	12445	93.28	43	55	13
JH Kallis (Afr/ICC/SA)	1996-2014	328	314	53	11579	139	44.36	15885	72.89	17	86	17
SC Ganguly (Asia/INDIA)	1992-2007	311	300	23	11363	183	41.02	15416	73.7	22	72	16
R Dravid (Asia/ICC/INDIA)	1996-2011	344	318	40	10889	153	39.16	15284	71.24	12	83	13
MS Dhoni (Asia/INDIA)	2004-2019	350	297	84	10773	183*	50.57	12303	87.56	10	73	10
CH Gayle (ICC/WI)	1999-2019	301	294	17	10480	215	37.83	12019	87.19	25	54	25
BC Lara (ICC/WI)	1990-2007	299	289	32	10405	169	40.48	13086	79.51	19	63	16
TM Dilshan (SL)	1999-2016	330	303	41	10290	161*	39.27	11933	86.23	22	47	11
Mohammad Yousuf (Asia/PAK)	1998-2010	288	273	40	9720	141*	41.71	12942	75.1	15	64	15
AC Gilchrist (AUS/ICC)	1996-2008	287	279	11	9619	172	35.89	9922	96.94	16	55	19

Interesting Findings

- The numbers of hundred as a positive correlation to the total runs scored.
- The number of matches played has negative correlation with innings played
- The dataset can be viewed as regression task
- The dataset had very little presence of outliers.

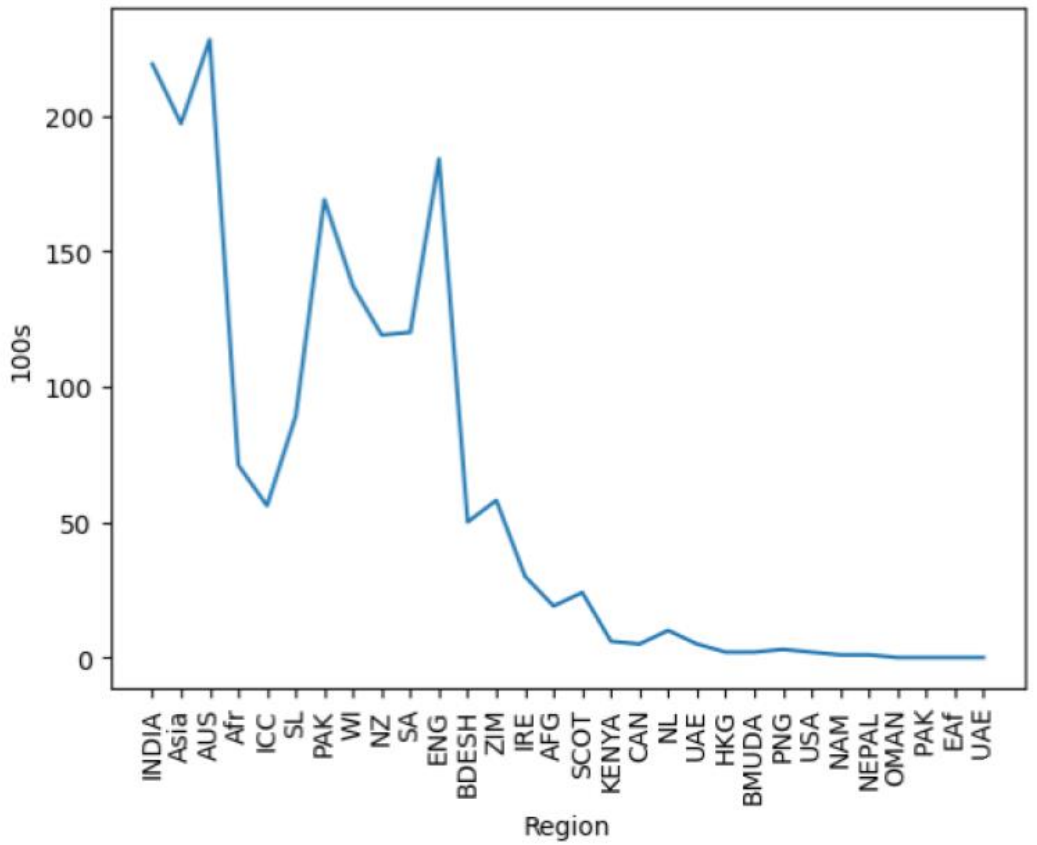
Interesting Findings

	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100
Range	[463.0, 1.0]	[452.0, 0.0]	[84.0, 0.0]	[18426.0, 0.0]	[264.0, 0.0]	[145.0, 0.0]	[21367.0, 0.0]	[328.57, 0.0]	[49.0, 0.0]
Mean	37.1616	29.4872	5.4748	673.5296	47.716	17.330384	901.6104	63.09454	0.7228
Mode	[2.0]	[1.0]	[0.0]	[0.0]	[0.0]	[0.0]	[11.0]	[0.0]	[0.0]
Median	13	9	2	100	34	15.33	163.5	63.925	0
Standard Deviation	58.88507507	51.31165745	9.349841323	1614.175019	44.04416078	13.11480049	2059.036146	27.27368219	2.932405115

- On calculating the mean, median and mode values for the attributes in the dataset, following findings are discovered:

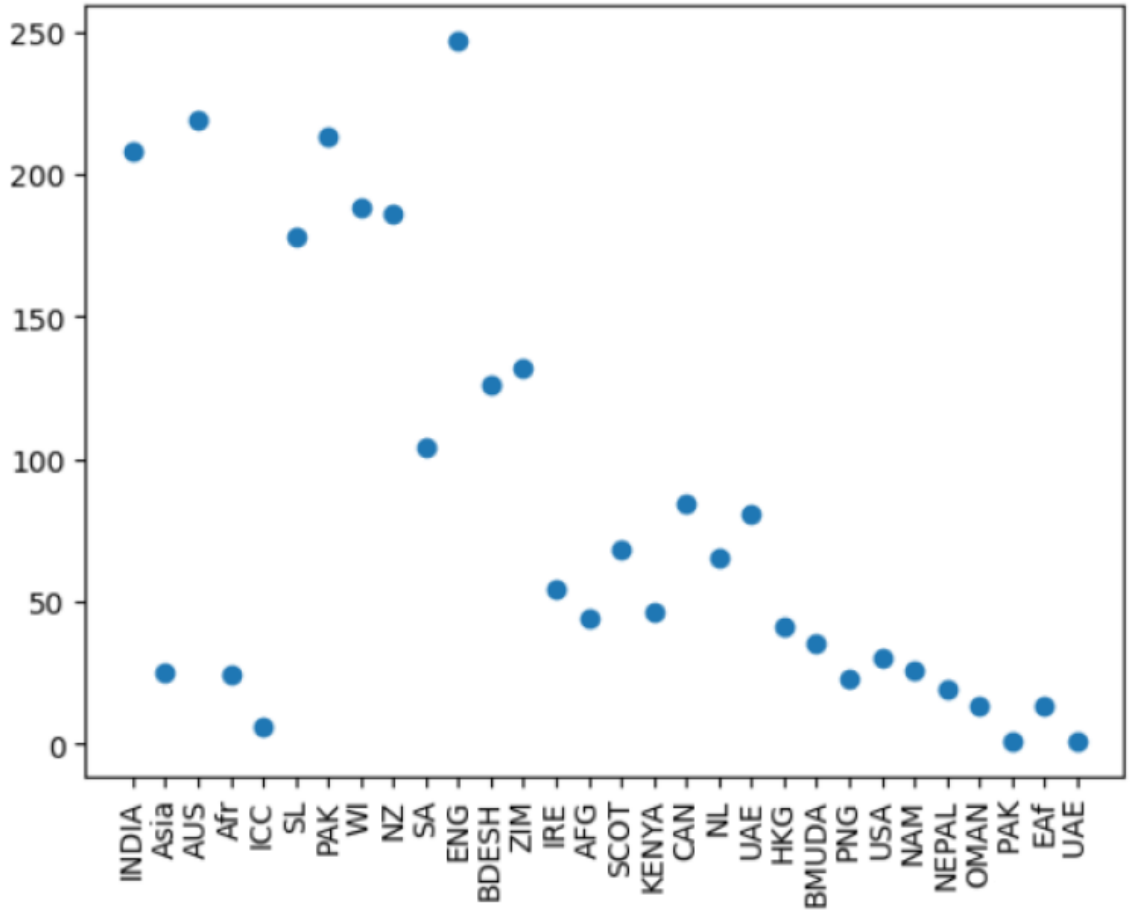
Visualization 1

The given line chart below represents the visualization between no of hundreds (100s) scored by a player to a Region(Country). The chart shows the maximum number of 100s has been scored by the teams.



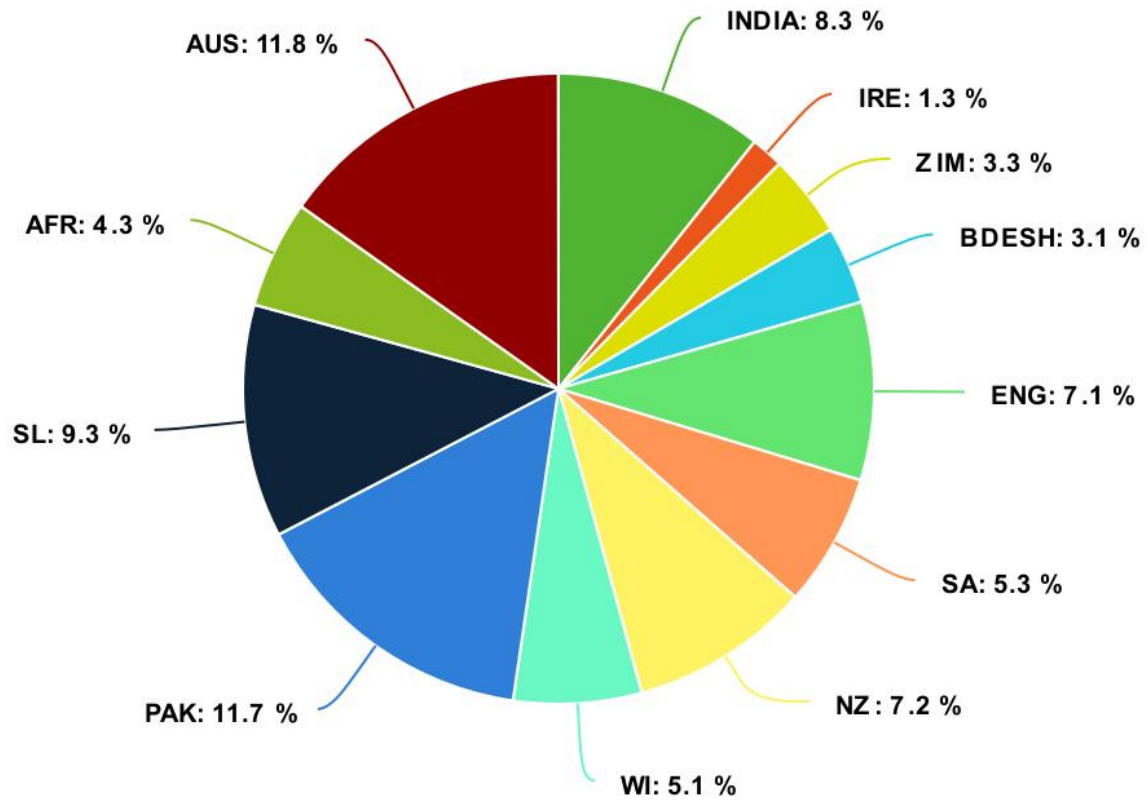
Visualization 2

In the following Scatter Plot, the visualization is between region being the country and player's count. This visualization helps to find the maximum number of players a team has.



Visualization 3

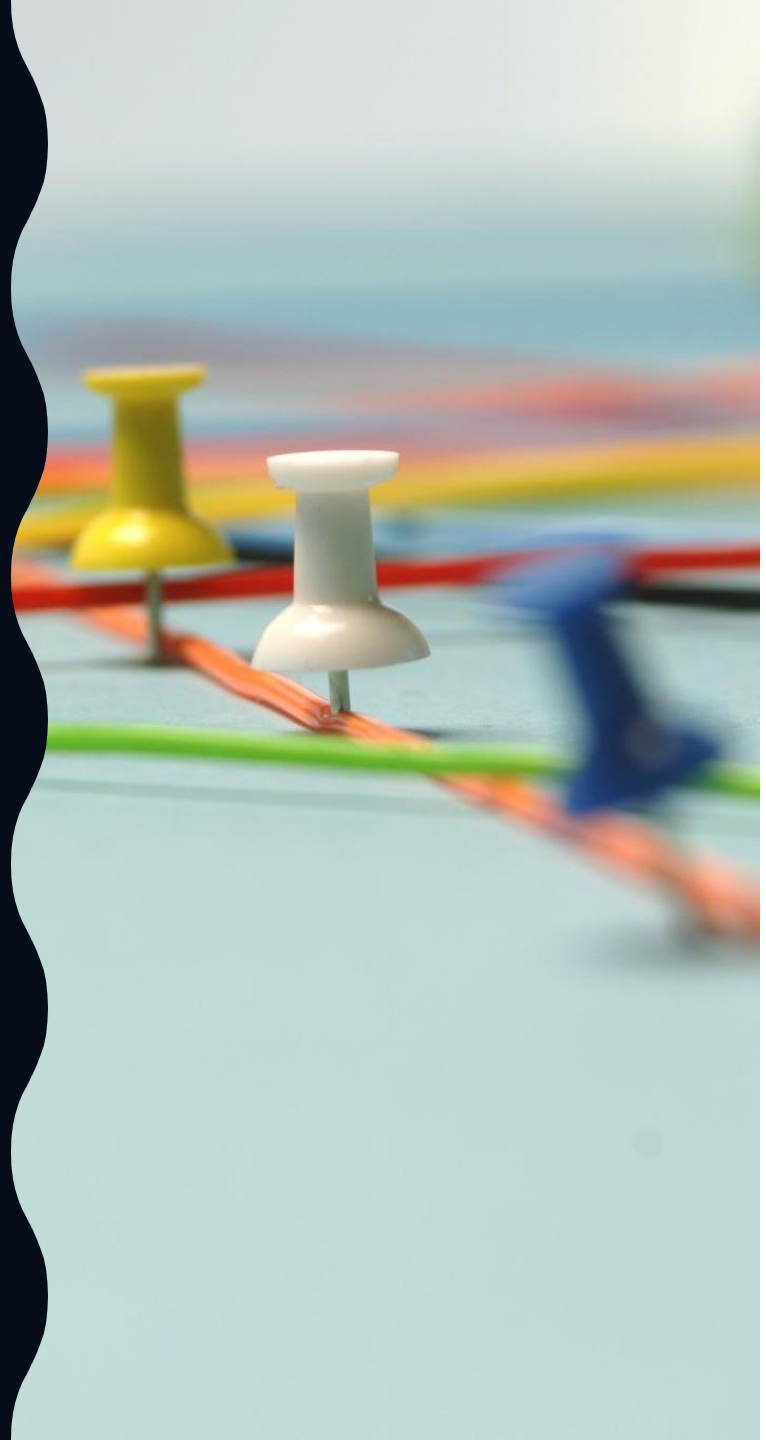
The pie chart depicts the players percentage of having ducks outs (0s) in a region. Duck out means that they've been bowled out, or dismissed, before getting any runs.



Task

1. Description:

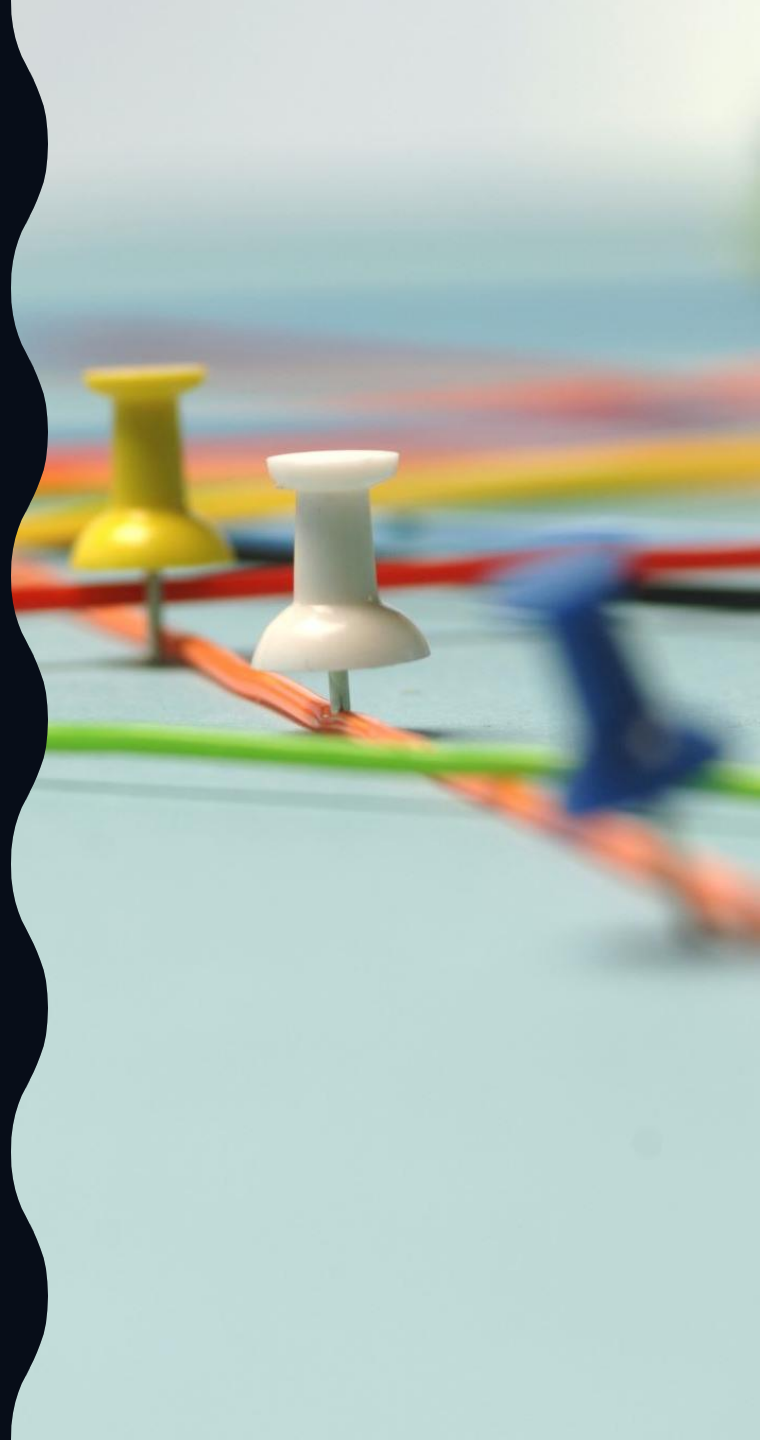
- Target Attribute: 100 (no. Of 100 scored by a player)
- Output: Predicted regression
- Task Type: Classification and Regression



Task

2. Preprocessing:

- After properly importing the dataset, preprocessing was done
- Involves cleaning inconsistent and irrelevant data
- Managed blank and null values
- Some of the values are modified and replaced with appropriate one
- After the preprocessing, model training was done.



Task

3. Modelling:

- In this stage, we divided the dataset into two parts
- One part goes for Training and other for Testing.
- 75% of the data is used for training and 25% for testing.
- After training, we tested the model for accuracy.

Logistic Regression

- Logistic regression is commonly used for **prediction and classification problems**
- Logistic regression is **a statistical analysis method to predict the regressive outcome of a target variable.**
- A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.

SVM

- Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems.
- SVC works by mapping data points to a high-dimensional space and then finding the optimal hyperplane that divides the data into two classes.
- The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes.

Models vs Accuracy

Model	Accuracy	F1 Score	Precision	Recall
Logistic Regression	0.8816	0.8702	0.8604	0.8816
SVM	0.8752	0.8468	0.8226	0.8752

Applications

THESE RESULTS CAN BE USED BY CRICKET ANALYST TO AUCTION PLAYER FOR A LEAGUE.

TEAM COACH CAN VIEW THESE ANALYSIS AND PREPARES A STRATEGY FOR A MATCH.

RESULTS CAN BE USED BY CAPTAIN FOR MAKING THE RIGHT DECISION ON AND OFF THE FIELD.


THANK YOU