A PRELIMINARY PROJECT REPORT

ON

**IPL DATA ANALYSIS AND PREDICTION USING MACHINE LEARNING**

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF

**BACHELOR OF ENGINEERING**

In

**COMPUTER ENGINEERING**

By

**Kaiwalya Sevekar Exam. No: B190814253**

**Yash Padol Exam. No: B190814235**

**Chandan Sinha Exam. No: B190814211**

**Rizwan Shaikh Exam. No: B190814255**

Under the Guidance of

**Dr. Vilas Joshi**

****

Submitted to

**Peoples Empowerment Group**

**ISB&M COLLEGE OF ENGINEERING, NANDE, PUNE**

****

**CERTIFICATE**

This is to certify that the Project Report entitled

**IPL DATA ANALYSIS AND PREDICTION USING MACHINE LEARNING**

Submitted By

**Kaiwalya Sevekar Exam. No: B190814253**

**Yash Padol Exam. No: B190814235**

**Chandan Sinha Exam. No: B190814211**

**Rizwan Shaikh Exam. No: B190814255**

Is a bonafide work carried out by them under the supervision of **Dr. Vilas Joshi** and it is approved for the partial fulfilment of the requirement of **Savitribai Phule Pune University** for the Project in the Final Year of Computer Engineering.

|  |  |  |
| --- | --- | --- |
| **(Dr. Vilas Joshi)**  Project Guide  Department of Computer Engineering | | **(Prof. B. B. Gite)**  Head of Department  Department of Computer Engineering |
|  | | |  |
| Place : Pune  Date : / / 202 | | **(Dr. P. K. Srivastava)**  Principal  ISB&M College of Engineering, Pune | |

**ACKNOWLEDGEMENT**

We would like to take this opportunity to thank all the people who were part of this seminar in numerous ways, people who gave un-ending support right from the initial stage.

In particular we wish to thank **Dr. Vilas Joshi** as internal project guide who gave their co-operation timely and precious guidance without which this project would not have been a success. We thank them for reviewing the entire project with painstaking efforts and more of his, unbanning ability to spot the mistakes.

We would like to thank **our H.O.D Prof. B. B. Gite** for his continuous encouragement, support and guidance at each and every stage of project.

And last but not the least we would like to thank all my friends who were Associated with me and helped me in preparing my project. The project named “**IPL Data Analysis**” would not been possible without the extensive support of people who were directly or indirectly involved in its successful execution.

**Project Group Members:**

**Kaiwalya Sevekar**

**Yash Padol**

**Chandan Sinha**

**Rizwan Shaikh**

**ABSTRACT**

The Indian Premier League (**IPL**) is a professional Indian T20 cricket league conducted by the Board of Control for Cricket in India (**BCCI**), founded in 2008. It is the most attended cricketing event in the world valued at $ 6.7 Billion in 2019, even in the year 2021, the worth of the IPL tournament was $ 4.7 Billion albeit the world was facing Covid-19 catastrophic effects.

It would be an understatement to state that Indians love cricket the game is played in just about every nook and cranny of India, rural or urban, and popular with the young and old alike, connecting billions in India, unlike any other sport. Cricket enjoys lots of media attention there is a significant amount of money from stake over the last several years, technology has literally been a game changer. Audiences are spoilt for choice with streaming media, tournaments and affordable access to mobile-based live cricket watching, and more.

Cricket is a game of numbers – the runs scored by batsmen, the wickets taken by bowlers, the matches won by the cricket team, the number of times a batsman responds in a certain way to a kind of bowling attack, etc. The capability to dig into cricketing numbers for both improving performance and studying the business opportunities, overall market, and economics of cricket via powerful analytics tools, powered by numerical computing software such as NumPy, is a big deal. Cricket analytics provides greater insights into the game and predictive intelligence regarding the game outcomes.

**TABLE OF CONTENTS**

Acknowledgement III

Abstract IV

Table of contents V

List of Abbreviations VI

List of Figures VII

**Chapter 1** **INTRODUCTION**

1.1 Overview 2

1.1.1 Indian Premier League (IPL) 2

1.1.2 Machine Learning 3

1.2 Motivation 11

1.3 Problem Statement 12

1.4 Organization of Thesis 13

**Chapter 2 LITERATURE SURVEY**

2.1 Dr. Sudhamathy’s Research Papers 15

2.2 Other Important Research Papers 15

**Chapter 3 FACTORS AFFECTING MATCH PREDICTION**

3.1 Pitch 17

3.2 Toss 17

3.3 Team Composition 17

3.4 Weather 17

3.5 Home Ground Advantage 18

3.6 Batting and Balling Depth 18

3.7 Team Combination and Balance 18

3.8 Toss Outcome 18

3.9 Strategy and Tactics 18

3.10 Past Records 18

**Chapter 4 METHODOLOGIES**

4.1 Data Collection 20

4.2 EDA – Exploratory Data Analysis 21

4.2.1 Check Shape 21

4.2.2 Check Data Type 21

4.2.3 Check Null Values 22

4.2.4 Check Summary Statistics 22

4.3 Data Cleaning 23

4.3.1 Changing Column Type 24

4.4 Data Visualization 25

4.5 Model Development and Evaluation 25

4.5.1 Random Forest 26

4.5.2 Logistic Regression 26

**Chapter 5 SYSTEM DESIGN**

5.1 Architecture 28

5.2 Proposed Architecture 28

5.3 Work Flow Diagram 29

**Chapter 6** **TECHNOLOGIES USED**

6.1 Hardware and Software Requirement 31

6.2 Libraries Used 31

**Chapter 7 RESULTS**

7.1 Match Prediction Web Application 34

7.2 Web Application Implementation 36

7.3 Power BI Dashboard 37

7.3.1 Summary Page 38

7.3.2 Team Stats Page 39

7.3.3 Batsman Stats Page 40

7.3.4 Bowler Stats Page 41

**Chapter 8 FUTURE WORK**

8.1 Future Work 43

8.2 Applications 44

**Chapter 9 REFERENCES**

9.1 References 46

**LIST OF FIGURES**

**FIGURE ILLUSTRATIONS PAGE NO.**

1.1 Regression 5 1.2 Classification 6 1.3 Clustering 8 1.4 Association 9 4.1 Random Forest 26 4.2 Logistic Regression 27 5.1 System Architecture 28 5.2 Proposed Architecture 28

5.3 Work-Flow Diagram 29

**LIST OF ABBREVATIONS**

**ABBREVIATION ILLUSTRATION**

IPL Indian Premier League

T20 Twenty Twenty

ICC International Cricket Council

SVM Support Vector Machine

ML Machine Learning

DBSCAN Density-Based Spatial Clustering of

Application with Noise

PCA Principle Component Analysis

t-SNE T – Distributed Stochastic Neighbour Embedding

NAN Not A Number

KNN K – Nearest Neighbour

SQL Structured Query Language

**CHAPTER 1**

**INTRODUCTION**

* 1. **OVERVIEW**

**1.1.1 Indian Premier League (IPL)**

Sports have gained much importance at both national and international level. Cricket is one such game, which is marked as the prominent sport in the world. T20 is one among the forms of cricket which is recognized by the International Cricket Council (ICC). Because of the short duration of time and the excitement generated, T20 has become a huge success. The T20 format gave a productive platform to the IPL, which is now pointed as the biggest revolution in the field of cricket. IPL is an annual tournament usually played in the months of April and May. Each team in IPL represents a state or a part of the nation in India. IPL has taken T20 cricket’s popularity to sparkling heights.

It is the most attended cricket league in the world and in the year 2010, IPL became the first sporting event to be broadcasted live. Till date, IPL has successfully completed 13 seasons from the year of its inauguration. Currently, there are 8 teams that compete with each other, organized in a round robin fashion during the stages of the league. After the completion of league stages, the top 4 teams in the points table are eligible to the playoffs. In playoffs, the winner between 1st and 2nd team qualifies for the final and the loser gets another opportunity to qualify for the finals by playing against the winner between 3rd and 4th team. In the end, the 2 qualified teams played against each other for the IPL title. The significance is that IPL employs television timeouts and therefore there is no time constraint in which teams have to complete the innings.

In this paper, we have examined various elements that may affect the outcome of an IPL match in determining the runs for each ball by considering the runs scored by the batsman in the previous ball as the labeled data. The suggested prediction model makes use of SVM and Naïve Bayes Theorem to fulfill the objective of the problem stated. Few works have been carried out in this field of predicting the outcomes in IPL. In our survey, we found that the work carried out so far is based on Data Mining for analyzing and predicting the outcomes of the match.

**1.1.2 Machine Learning**

Machine Learning is the preferred technique of predicting or classifying information to assist folks in creating necessary selections. Machine Learning algorithms are trained over instances or examples through that they learn from past experiences and analyse the historical knowledge. Simply building models isn't enough. You want to conjointly optimize and tune the model appropriately in order that it provides you with 2 correct results. Improvement techniques involve tuning the hyper parameters to succeed in Associate in Nursing optimum results.

As it trains over the examples, more than once it will determine patterns to form selections additionally accurately. Whenever any new input is introduced to the cubic centimeter model, it applies its learned patterns over the new knowledge to form future predictions. Based on the ultimate accuracy, one will optimize their models by exploiting numerous standardized approaches. During this manner, the Machine Learning model learns to adapt to new examples and produce higher results.

Machine learning has numerous applications across different industries and fields. It is used in computer vision for tasks like image classification and object detection, natural language processing for sentiment analysis and language translation, recommendation systems for personalized suggestions, fraud detection in financial transactions, healthcare for diagnosis and treatment recommendations, and many more. The ability of machine learning to analyze large amounts of data and extract valuable insights has revolutionized industries and continues to drive technological advancements.

**Types of Learning:**

Machine Learning Algorithms can be classified into 3 types as follows,

1. Supervised learning

2. Unsupervised Learning

3. Reinforcement Learning

**1.1.2.1 Supervised Learning**

Supervised learning is a fundamental concept in machine learning where the model learns from labelled training data. In this approach, each example in the training dataset is associated with a known output or target value. The goal of supervised learning is to develop a mapping function that can accurately predict the correct output for new, unseen inputs.

The training process involves presenting the model with input features and their corresponding correct outputs. The model then learns the underlying patterns and relationships between the input features and the outputs. It generalizes this learning to make predictions on new, unseen data points. Supervised machine learning includes labelled data which helps machine to learn from the input data and make prediction for the new data. It is called supervised learning as it learns under the supervision of the labelled data.

Supervised learning encompasses various algorithms, each with its own strengths and characteristics. Decision trees are a popular choice, where a tree-like structure is built to make sequential decisions based on feature values. Random forests combine multiple decision trees to improve prediction accuracy. Support Vector Machines (SVM) find an optimal hyper plane to separate different classes in the data space. Neural networks, including deep learning models, consist of interconnected layers of nodes that can capture complex relationships between inputs and outputs.

Supervised learning has a broad range of applications. In classification tasks, the goal is to assign inputs to specific categories or classes. For instance, email spam detection, sentiment analysis, and image classification are common classification problems. In regression tasks, the aim is to predict continuous values. Examples include predicting house prices, stock market trends, or weather forecasts.

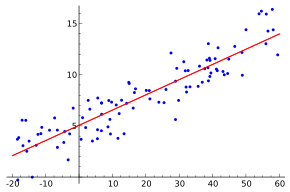
Supervised learning is classified into two types of algorithms which are as follows,

**(i) Regression:**

Regression is a type of supervised learning in machine learning that focuses on predicting continuous numerical values. In regression, the goal is to build a model that can accurately estimate or approximate a target variable based on input features. It is widely used in various fields, including finance, economics, healthcare, and social sciences.

In regression, the relationship between the input features and the target variable is modeled using a mathematical function. The model learns from the training data, which consists of input-output pairs, and tries to find the best-fitting function that can generalize well to unseen data.

There are different regression algorithms available, each with its own characteristics and assumptions. Linear regression is a commonly used algorithm that assumes a linear relationship between the input features and the target variable. It estimates the parameters of a linear equation to make predictions. Polynomial regression extends linear regression by incorporating higher-degree polynomial terms to capture more complex relationships.



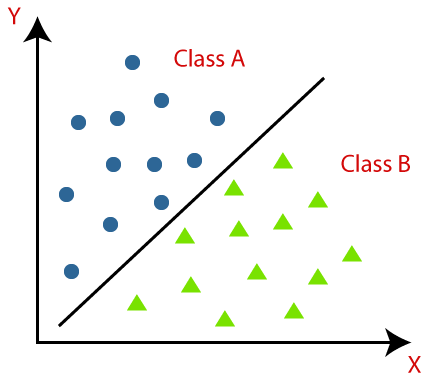
**Fig 1.1: Regression**

**(ii) Classification:**

Classification is a fundamental concept in machine learning that focuses on categorizing or classifying input data into predefined classes or categories. It is a supervised learning approach where the model learns from labeled training data to make predictions about the class labels of unseen instances.

In classification, the input data is typically represented by a set of features or attributes, and each instance is associated with a known class label. The goal is to develop a model that can accurately assign new, unseen data points to the correct classes based on their feature values.

There are various algorithms used in classification, each with its own strengths and characteristics. Decision trees are popular for their interpretability, where a tree-like structure is constructed to make sequential decisions based on feature values. Random forests combine multiple decision trees to improve prediction accuracy and handle complex data relationships. Support Vector Machines (SVM) aim to find an optimal hyper plane that separates different classes in the data space.

****

**Fig 1.2: Classification**

**1.1.2.2 Unsupervised Learning**

Unsupervised learning is a branch of machine learning that deals with unlabeled data, where the goal is to discover patterns, structures, or relationships within the data without any predefined target variables. Unlike supervised learning, unsupervised learning algorithms work with input data that lacks explicit labels or class information.

In unsupervised learning, the model explores the inherent structure of the data to uncover hidden patterns or groupings. Clustering is a commonly used technique in unsupervised learning, where data points are grouped into clusters based on their similarities or proximity in the feature space. The goal is to identify natural clusters or subgroups within the data. Clustering algorithms such as k-means and hierarchical clustering are popular choices for unsupervised clustering tasks.

Another important technique in unsupervised learning is dimensionality reduction. It aims to reduce the number of input features while preserving the essential information or structure of the data. Dimensionality reduction techniques like principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) are commonly used to transform high-dimensional data into lower-dimensional representations, making it easier to visualize and analyze.

Unsupervised learning is valuable in various scenarios. It can be used to discover customer segments or market segments based on purchasing patterns, identify anomalies or outliers in datasets, detect patterns in large datasets for fraud detection, and explore relationships between variables in exploratory data analysis.

One common application of unsupervised learning is anomaly detection. By learning the normal patterns in the data, the model can identify instances that deviate significantly from the norm, indicating potential anomalies or outliers. This is valuable in various domains such as fraud detection, network intrusion detection, and detecting unusual behaviors in sensor data.

Unsupervised learning problems are classified into two categories of algorithms:

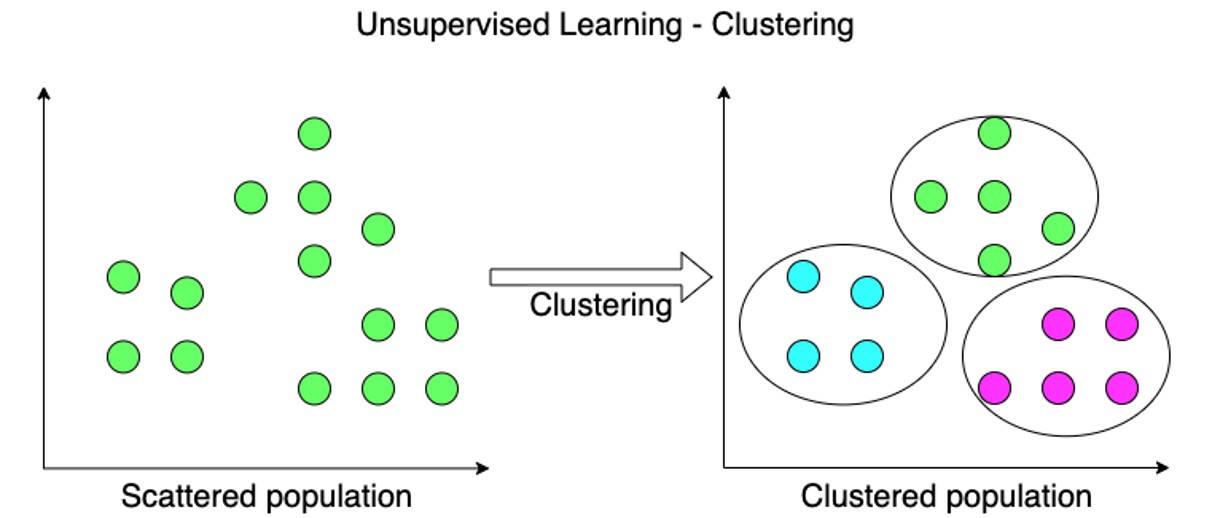
**(i) Clustering:**

Clustering is a technique in unsupervised learning that involves grouping similar data points together based on their characteristics or proximity in the feature space. It aims to discover inherent patterns, structures, or subgroups within the data without any prior knowledge of the class labels or target variables.

K-means Clustering: K-means is a popular centroid-based clustering algorithm. It aims to partition the data into K clusters, where K is a user-defined parameter. It iteratively assigns data points to the nearest centroid and updates the centroid based on the mean of the assigned data points. The process continues until convergence.

Hierarchical Clustering: Hierarchical clustering creates a hierarchical structure of clusters, often represented as a tree-like structure called a dendrogram. It can be agglomerative, starting with each data point as a separate cluster and then merging the closest clusters, or divisive, starting with all data points in one cluster and recursively dividing them into smaller clusters.

Density-based Clustering: Density-based clustering algorithms, such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise), group data points based on their density. It identifies regions of high density as clusters and separates sparse regions as noise.

****

**Fig 1.3: Clustering**

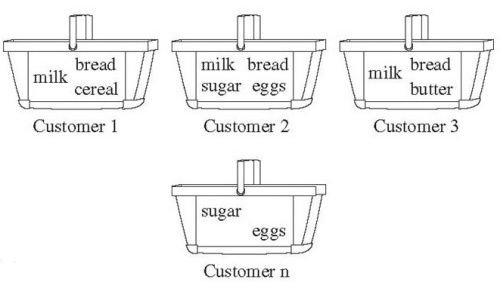
**(ii) Association:**

Association analysis, also known as association rule mining, is a technique used in data mining and unsupervised learning to discover interesting relationships or associations between variables or items within large datasets. It aims to identify frequent patterns or co-occurrences of items and extract valuable insights from the data.

Association analysis is commonly applied to transactional datasets, where each transaction consists of a set of items. The goal is to find associations or correlations between items that occur together more frequently than would be expected by chance.

The most well-known algorithm for association rule mining is the Apriori algorithm. It works by generating candidate item sets and then pruning them based on a minimum support threshold. Support refers to the proportion of transactions in which an item set occurs. The algorithm iteratively explores larger item sets until no more frequent item sets can be found.

Association rules are typically expressed in the form of "if-then" statements, where the antecedent represents the items or item sets that act as the premise, and the consequent represents the items that are likely to follow. For example, a rule could be "if {diapers} then {baby food}," indicating that customers who purchase diapers are also likely to buy baby food.



**Fig 1.4: Association**

**1.1.2.3 Reinforcement Learning**

Reinforcement learning is a branch of machine learning that focuses on how an agent can learn to make decisions or take actions in an environment to maximize a reward signal. It is inspired by the concept of learning through trial and error, similar to how humans learn and adapt based on feedback.

In reinforcement learning, an agent interacts with an environment and learns by receiving feedback in the form of rewards or penalties based on its actions. The agent's goal is to learn an optimal policy—a strategy that determines the actions to take in different situations—to maximize the cumulative rewards over time.

The key components of reinforcement learning are:

Agent: The learner or decision-making entity that interacts with the environment.

Environment: The external system with which the agent interacts. It provides the agent with feedback through rewards or penalties based on its actions.

Actions: The choices available to the agent at each step. The agent selects actions based on its current state.

State: The representation of the environment at a given time. It captures relevant information that the agent can use to make decisions.

Rewards: The feedback provided by the environment to the agent after taking an action. Rewards indicate the desirability of the agent's actions and guide its learning process.

Reinforcement learning algorithms use the concept of value functions and policies to guide the agent's decision-making. A value function estimates the expected cumulative reward from a given state or state-action pair. It helps the agent evaluate the desirability of different actions and states. A policy defines the agent's behaviour by mapping states to actions.

* 1. **MOTIVATION**

IPL data analysis and prediction models have become increasingly popular among cricket enthusiasts and fans. One of the main motivations for undertaking such projects is to enhance fan engagement. By analyzing historical data and developing prediction models, fans can actively participate in discussions and debates surrounding the IPL. They can make informed predictions about match outcomes, player performances, and team strategies, adding an extra layer of excitement to the viewing experience.

Furthermore, IPL data analysis plays a crucial role in fantasy cricket leagues. Fantasy cricket platforms allow fans to create their virtual teams and earn points based on the real-life performances of players. By leveraging data analysis and prediction models, participants can make strategic decisions about team selection, captaincy, and player transfers. This involvement not only increases the entertainment value of watching IPL matches but also encourages fans to delve deeper into the game's intricacies and statistics.

Overall, the motivation behind IPL data analysis and prediction model projects lies in improving team performances, assisting in strategic decision-making, engaging fans, and enhancing the overall experience of following the Indian Premier League. By leveraging data-driven insights, these projects contribute to a deeper understanding of the game and add an element of excitement and competition for fans and stakeholders alike.

**1.3 PROBLEM STATEMENT**

The main objective is to predict the IPL Match Result that would be beneficial for the franchises and authorities who are at a position of decision-making. IPL has a large set of audience. In cricket, particularly IPL is most watched and loved by the people, where no one can guess who will win the match until the last ball of the last over. The main purpose of this research work is to find the best prediction model i.e the best machine learning technique which will predict the match winner out of the two teams.

Developing an accurate and reliable data analysis and prediction model for the Indian Premier League (IPL) to forecast match outcomes, player performances, and team strategies, based on historical data, in order to assist teams, fans, and stakeholders in making informed decisions and enhancing the overall viewing experience."

This problem statement acknowledges the need for an effective model that can leverage historical IPL data to provide valuable insights and predictions. The project aims to address the challenges of accurately forecasting match outcomes, player performances, and team strategies, ultimately benefiting teams, fans, and stakeholders by enabling them to make informed decisions and enhancing their engagement with the IPL.

**1.4 ORGANIZATION OF THESIS**

The chapters of this document describe the following:

**Chapter-1** is about the introduction of our project where we have given clear insights about our project domain and other related concepts.

**Chapter-2** specifies a literature survey where all different existing methods and models are examined.

**Chapter-3** is about the factors which affect the result of an ongoing or upcoming match.

**Chapter-4** specifies the experimental analysis of our system along with performance measures and comparisons between different models. It also specifies about implementation along with sample code.

**Chapter-5** contains the architecture diagram of the project and also explains the workflow of the model

**Chapter-6** is about the technologies that are being used in this project. We have used python libraries like Pandas, NumPy, Matplotlib and Seaborn for data cleaning and visualisation.

**Chapter-7** is the final result of our project where we can see the implementation of the project.

**Chapter -8** is about the future work and applications of the project.

**Chapter-9** contains the referencesto the research papers and studies that helped in making this project.

**CHAPTER 2**

**LITERATURE SURVEY**

**2.1 Dr. SUDHAMATHY’S RESEARCH PAPER**

The research paper of G. Sudhamathy helps to understand the different machine learning algorithms working principles and their implementation. It creates the Model and Training dataset and helps to predict with the help of the model created. The model classifies the data and compares the results and gets accuracy which is the important one.

As in the dataset, there are many parameters present. Out of them which parameters are helpful in the project? The factor affecting concept was taken by Maheshwari in their prediction of live cricket score paper from that we get to know the main factors which required for the prediction of the score and the prediction of the winning team. The role of classification is clear in the paper of Tejinder Singh it gives proper information or use of naive bias and linear regression. They give the proper knowledge of data collection and preparation and also how to train the data and test the data given by them which is more helpful.

The support vector machines brief idea is been taken from Aminul Islam Anik’s paper which is about players’ performance in this paper the idea about the SVM system is given in detail where the player performance prediction is given by collecting the old information or data. From the literature survey, it is concluded that machine learning is needed for prediction.

**2.2 OTHER IMPORTANT RESEARCH PAPERS**

"Predictive Modeling for Indian Premier League Using Machine Learning Techniques" by D. Balakrishnan and M. Santhi. This research paper discusses the application of machine learning techniques such as decision trees, support vector machines, and random forests for predicting IPL match outcomes.

"Data Mining Techniques for IPL Prediction" by R. R. Mahajan and A. V. Kulkarni. This paper explores the use of data mining techniques, including classification algorithms and association rule mining, to predict IPL match results and identify key factors influencing team performance.

"Data Analytics in Cricket: A Study on Predicting IPL Match Outcomes Using Machine Learning" by M. Anandhavalli and A. P. Subashini. The authors analyze IPL match data and employ machine learning algorithms to predict match outcomes. They evaluate the performance of classifiers such as decision trees, random forests, and naïve Bayes.

**CHAPTER 3**

**FACTORS AFFECTING MATCH PREDICTION**

**3.1 FACTORS AFFECTING MATCH PREDICTION**

Cricket winning can be predicted like all other games. We need to find the best attributes or factors that influence the match outcome. The result of a cricket match depends upon more of in-game and more of pre-game attributes. Pre-game attributes like pitch, Toss, Team strength, weather, venue etc. and in-game attributes like run rate, total run, strike rate, wickets in hand etc. influences a match result predominantly. Below are the attributes that decides outcome of the cricket match.

**3.1.1 PITCH**

Unlike other sports, cricket stadiums shape and size are not fixed except the dimensions of the inner circle and pitch which are 30 yards and 22 yards respectively. Outfield variations and pitch can have a substantiate effect on bowling and batting. The spin of the ball, seam movement and the bounce depend upon the nature of the pitch. It depends on how the wet is the pitch. The more wet the pitch, the slower it will play. On the off chance that it is drying out, those balls will change significantly, yet all it will get less difficult those drier it gets. Green pitches tend to get easier to bat on. Wickets can get significantly more dry or wet. They might start to break up if they are soft.

**3.1.2 TOSS**

According to cricket analysts, there is sure measure of advantage for a team if it wins the toss. This might not be deciding factor in a match but it would give the team the opportunity of choosing “what they want”. The outcome of the toss can have an impact, especially in limited-overs matches. The toss-winning captain can choose to bat first or chase based on pitch and weather conditions, which can influence the game's course

**3.1.3 TEAM COMPOSITION**

The composition of the teams, including the strength of the batting and bowling line-ups, the presence of key players, and the team's overall balance, can significantly impact the outcome of a match. The form and fitness of players are crucial considerations.

**3.1.4 WEATHER**

In addition to the pitch, the ground also gets wet in the rain which slows the ball’s travelling speed. Most of the time the pitch is covered but the whole ground is not covered. The ball gets tends to get slippery and it becomes harder to grip the ball for bowlers. Weather conditions, such as rain, humidity, and wind, can affect the outcome of a match. Rain interruptions can result in revised targets or even match abandonment. Overcast conditions may assist swing bowling, while dry and hot conditions might aid spinners.

**3.1.5 HOME GROUND ADVANTAGE**

This is another attribute which determines the winner in the match. If you are playing in the home ground condition everything would be in your hands like climatic factor, pitch nature and the major role is played by the home crowd. Home team gets better motivation.

**3.1.6 BATTING AND BALLING DEPTH**

The depth and quality of a team's batting and bowling line-ups can influence match outcomes. Teams with a deep batting order can recover from early setbacks, while strong bowling attacks can put pressure on the opposition and restrict their scoring.

**3.1.7 TEAM COMBINATION AND BALANCE**

The selection of a well-balanced team, with a combination of different skill sets and playing styles, can contribute to success. A team that strikes the right balance between batting, bowling, and all-round abilities is better equipped to tackle various match scenarios.

**3.1.8 TOSS OUTCOME**

The outcome of the toss can have an impact, especially in limited-overs matches. The toss-winning captain can choose to bat first or chase based on pitch and weather conditions, which can influence the game's course.

**3.1.9 STRATEGY AND TACTICS**

The ability of teams to adapt to match situations, formulate effective strategies, and make tactical decisions can greatly impact the result. Tactical choices such as field placements, bowling changes, and batting order adjustments can shape the outcome of a match.

**3.1.10 PAST RECORDS**

The past team performances can be considered to predict the outcomes of a match. History of games at that venue, how did the team perform, performances at that specific venue, performance against the specific opposition and experience at the specific venue.

**CHAPTER 4**

**METHODOLOGY**

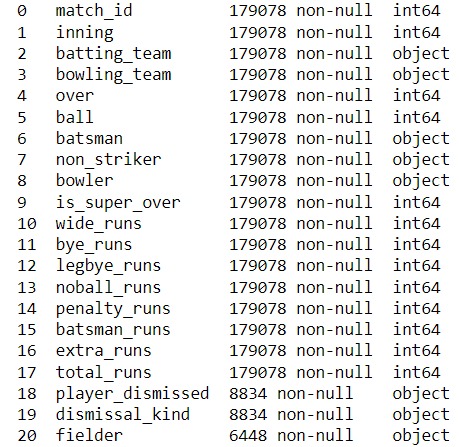
**4.1 DATA COLLECTION**

Before we start our project we need to gather data. This step can be done using the following 3 methods:

* Open Sources – This data is readily available in the form of structured data (rows and columns) and can be downloaded from sites like [**Kaggle**](https://www.kaggle.com/), [**UCI-ML-Repository**](https://archive.ics.uci.edu/ml/index.php), and [**Open Government Data**](https://data.gov.in/).
* Collection by Individuals – Often it happens that in some cases, data is not available so the team gathers data using tools like we scrapper or go out and gather data for themselves.
* Crowd-Sourcing – In this technique, people like ours help in annotating data for eg. Captcha services.

For our use case, we are going to use the IPL Scores Dataset (link in reference) which has 72430 observations and 20 features:

Link of dataset: <https://www.kaggle.com/ramjidoolla/ip>

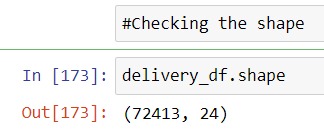


**4.2 EDA – EXPLORATORY DATA ANALYSIS**

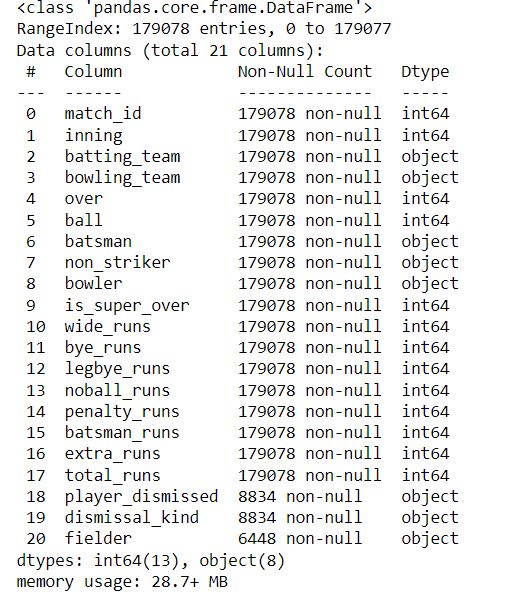
Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

Having looked at the data quickly, let’s dive deeper into the dataset and explore some of the insights. This procedure is very important and will allow us to understand the data and plan our next steps. Luckily pandas provide easy-to-use functions to perform our analysis. So, let’s begin.

**4.2.1 CHECKING SHAPE**

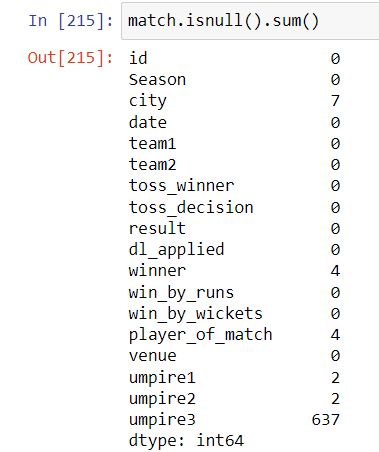
Even though we know the shape of data (observations and features) the info can be wrong. So it’s better to cross-check our data once. So let’s check it.

**4.2.2 CHECK DATA TYPE**

In real-world cases, the data we find are not of the desired data type, so it will be good if we check it too. Pandas provide an info method that returns the data type of all columns present in the dataset.

**4.2.3 CHECKING NULL VALUES**

Having understood the shape and info, we will now check for null values (fields having no data – NAN’s). It is essential for any project as null values can change the whole story depicted by data and can even contribute to making data worse for the use cases. Let’s perform the same

All we have to do is to use the “is null” method of the data-frame and then sum the outputs to get the total count for each column.

**4.2.4 CHECK SUMMARY STATISTICS**

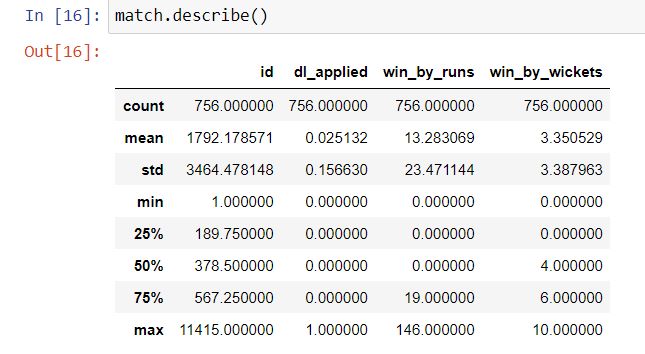
Since there is no Nan value doesn’t mean our data is a good representation. To get the gist, we can plot summary measures which generally include:

1. Measurement of Central Tendency – These measures allows us to understand where most of our data lies and mostly include

* **Mean** – Average of Data
* **Median** – Centre for Data
* **Frequency** – No of occurrence of specific data.
* **Mode** – Highest Observation in Data

2. Measures of Dispersion – These are the measures that allow us to understand how widespread data is and mostly includes:

* **Max & Min** – Highest and Lowest Value in the dataset
* **Range** – Highest-Lowest (captures the reach of data)
* **Variance** – Captures the variation of data (how data is varying) – Usually the sum of deviation of actual data from its mean/no of samples – 1
* **Standard Deviation** – Same as standard deviation but on the same scale as data -sqrt (variance)
* **Percentiles** – Capture the spread of data for a specific value i.e. how much data is above or below it. 50% – Median

Performing each check will be cumbersome so pandas pack all these in a single function called describe. Now let’s check the summary stats of our data.

**4.3 DATA CLEANING**

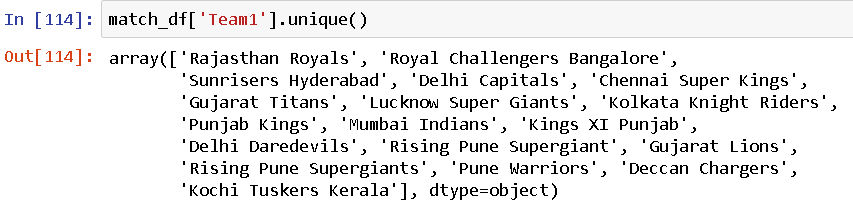
We have collected data information for the past 13 years/seasons there are some teams that were in IPL for very short spam. If we use their data for our prediction model our prediction will slightly differ or will not be that accurate, so to prevent that from happening we have removed those teams:

Rising Pune Super-giants - Played 2 Seasons

Pune Warriors India - Played 2 Seasons

Gujarat Lions - Played 2 Seasons

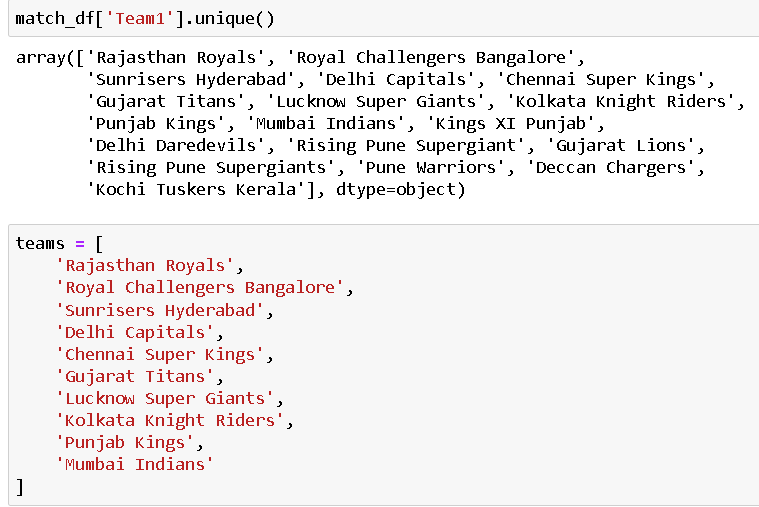
Kochi Tuskers Kerala - Played 1 Season

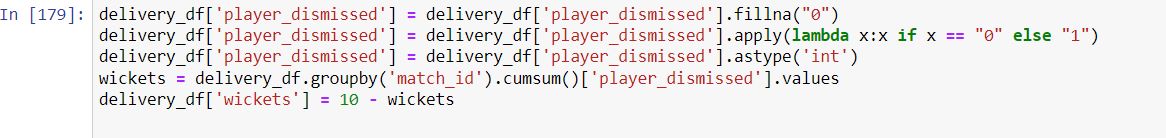
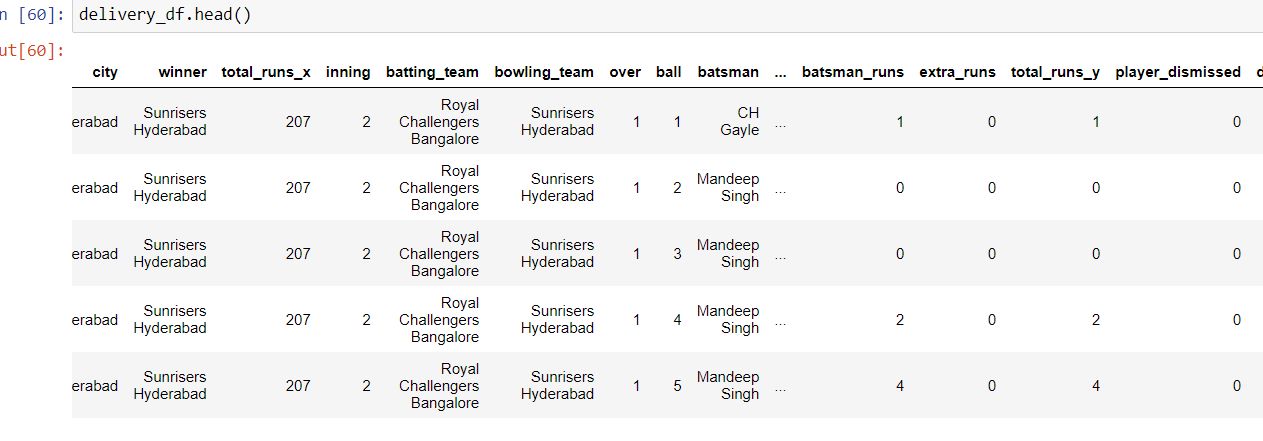


There are some teams which are playing under same franchise but have changed their names, so we have updated their old names to their respective new names.

**4.3.1** **CHANGING COLUMN TYPE**

While examining the info on features column we found that the “player\_dismissed” column is a String Data Type on which no operations can be performed

 So we have to change its data type to int.



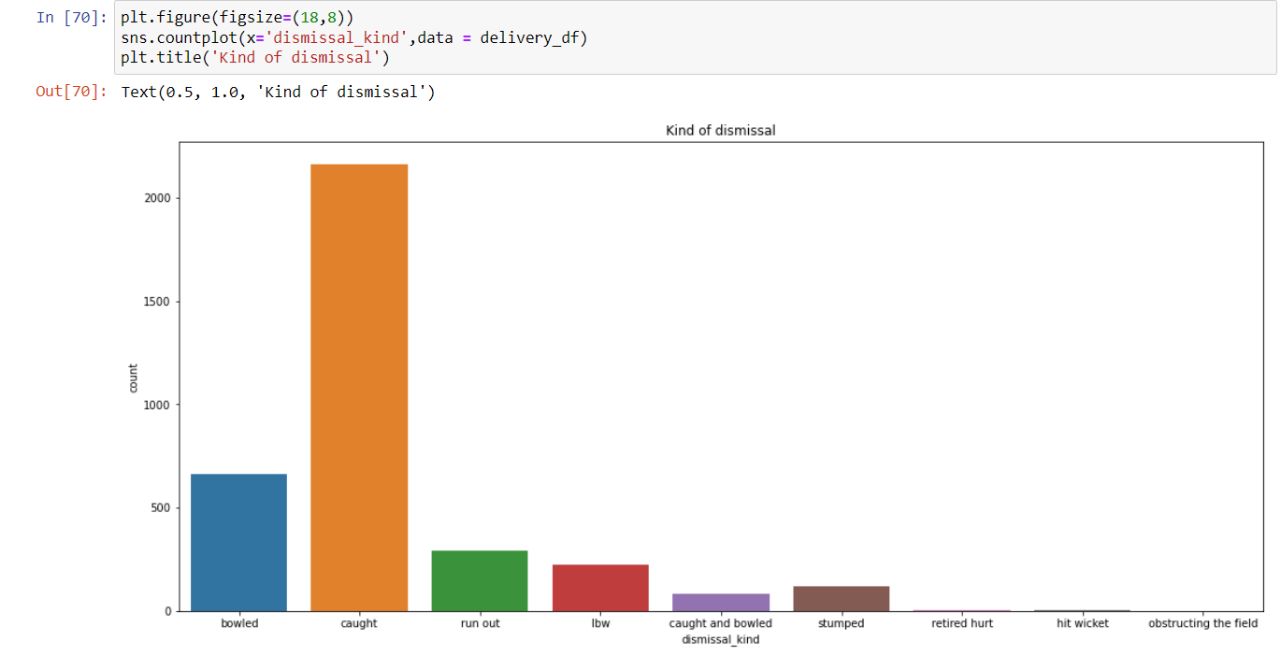
**4.4 DATA VISUALIZATION**

Data visualization is the graphical representation of information and data. By using v[isual elements like charts, graphs, and maps](https://www.tableau.com/data-insights/reference-library/visual-analytics), data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data

In the world of Big Data, data visualization tools and technologies are essential to analyse massive amounts of information and make data-driven decisions.

Matplotlib & Seaborn Library is used here for visualizing the graphs

Some other advantages of data visualization include:

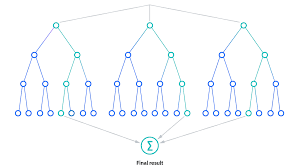
* Easily sharing information.
* Interactively explore opportunities.
* Visualize patterns and relationships.

**4.5 MODEL DEVELOPMENT FOR PREDICTION**

Here, we have developed a generic model and applied all classification methods. The data is split into training data and test data, we train the model using certain features and use it to predict the testing data and then we calculate the performance of the system. The various classification models used are: Logistic Regression, Gaussian Naïve Bayes Classifier, KNN (K Nearest Neighbour) algorithm, Support Vector Machines, Random Forest Classifier. We have tried different algorithms to get most accurate predictions. Some important algorithms were,

**4.5.1 Random Forest**

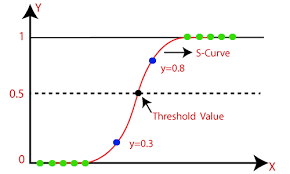
Random forest classifier creates multiple decision trees and find out the individual output. It combines all the results together and gives the results with more accuracy. It can be used as both classification and regression.

****

**Fig 4.1: Random Forest**

**4.5.2 Logistic Regression**

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

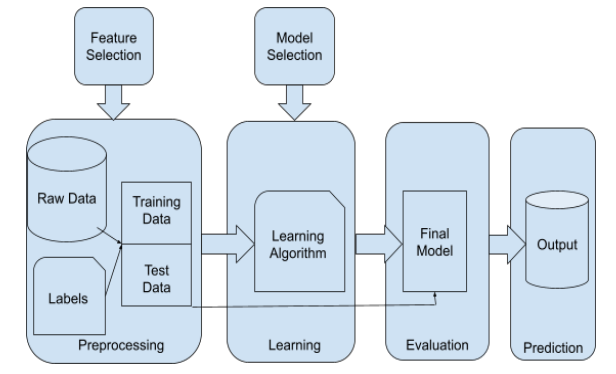


**Fig 4.2: Logistic Regression**

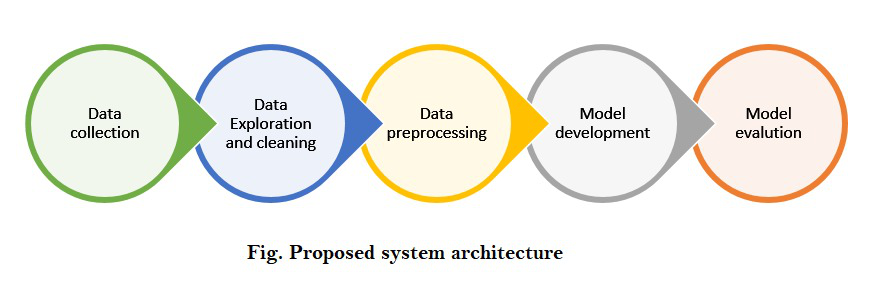
**CHAPTER 5**

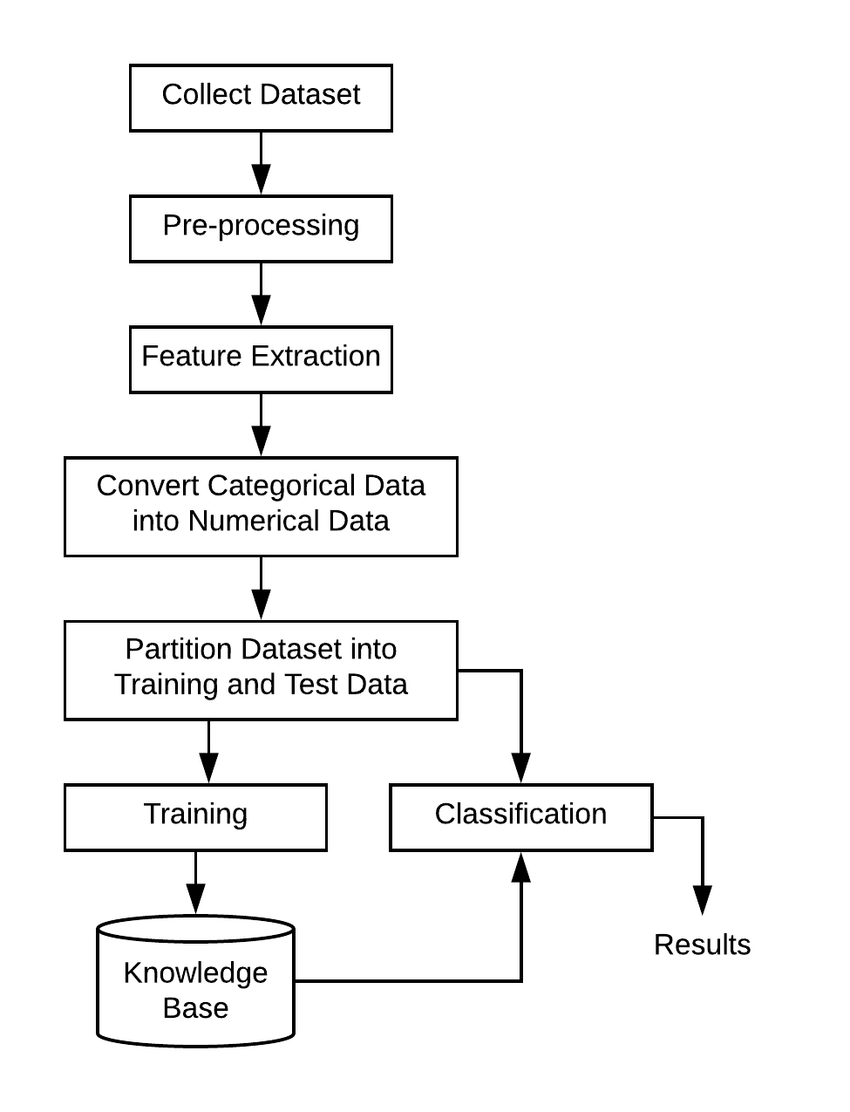
**SYSTEM DESIGN**

**5.1 ARCHITECTURE**



**Fig 5.1: System Architecture**

**5.2 PROPOSED ARCHITECTURE**

**5.3 WORKFLOW DIAGRAM**

**Fig 5.3 : Workflow Diagram**

**CHAPTER 6**

**TECHNOLOGIES USED**

**6.1 HARDWARE AND SOFTWARE REQUIREMENT**

**Programming Language Used**: Python , SQL

**Software Requirements**: Jupyter Notebook, Power BI, PyCharm, Tableau, Streamlit

**Hardware Requirement**:

Processor-Intel(R) Core(TM) i5-6440HQ CPU @ 2.60GHz 2.59 GHz

Installed RAM-8.00 GB (7.88 GB usable)

System type- 64-bit operating system, x64-based processor.

**6.2 LIBRARIES USED**

**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.

**2. Pandas:**

Pandas is an open-source library that is made mainly for working with relational or labelled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library.

Pandas is fast and it has high performance & productivity for users.

**3. Matplotlib:**

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

**4. 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 modelling including classification, regression, clustering and dimensionality reduction via a consistency interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

**5. Seaborn:**

Seaborn is an amazing visualization library for statistical graphics plotting in Python. It provides beautiful default styles and colour palettes to make statistical plots more attractive. It is built on the top of the matplotlib library and also closely integrated to the data structures from pandas. Seaborn aims to make visualization the central part of exploring and understanding data.

**6.2 Power BI**

Microsoft Power BI is an interactive data visualization software product developed by Microsoft with a primary focus on business intelligence.[1] It is part of the Microsoft Power Platform. Power BI is a collection of software services, apps, and connectors that work together to turn unrelated sources of data into coherent, visually immersive, and interactive insights. Data may be input by reading directly from a database, webpage, or structured files such as spreadsheets, CSV, XML, and JSON.

**6.3 PyCharm**

PyCharm is one of the most popular Python IDEs. There is a multitude of reasons for this, including the fact that it is developed by JetBrains, the developer behind the popular IntelliJ IDEA IDE that is one of the big 3 of Java IDEs and the “smartest JavaScript IDE” WebStorm. Having the support for web development by leveraging Django is yet another credible reason.There are a galore of factors that make PyCharm one of the most complete and comprehensive integrated development environments for working with the Python programming language.

**CHAPTER 7**

**RESULT**

**7.1 MATCH PREDICTION WEB APPLICATION**

**Fig 7.1 : Web Application**

Fig. 7.1 Shows the Prediction model web application developed using streamlit. After successfully developing machine learning based prediction model a pickle file was generated. This pickle serves as a database for our web application. Using this pickle file we created a streamlit-based web application which lets user to interact with the model and generate predictions.

Streamlit Code:

import streamlit as st

import pickle

import pandas as pd

teams = ['Rajasthan Royals',

'Royal Challengers Bangalore',

'Sunrisers Hyderabad',

'Delhi Capitals',

'Chennai Super Kings',

'Gujarat Titans',

'Lucknow Super Giants',

'Kolkata Knight Riders',

'Punjab Kings',

'Mumbai Indians']

cities = ['Ahmedabad', 'Kolkata', 'Mumbai', 'Navi Mumbai', 'Pune', 'Dubai',

'Sharjah', 'Abu Dhabi', 'Delhi', 'Chennai', 'Hyderabad',

'Visakhapatnam', 'Chandigarh', 'Bengaluru', 'Jaipur', 'Indore',

'Bangalore', 'Raipur', 'Ranchi', 'Cuttack', 'Dharamsala', 'Nagpur',

'Johannesburg', 'Centurion', 'Durban', 'Bloemfontein',

'Port Elizabeth', 'Kimberley', 'East London', 'Cape Town']

pipe = pickle.load(open('pipe.pkl','rb'))

st.title('IPL Win Predictor')

col1, col2 = st.columns(2)

with col1:

BattingTeam = st.selectbox('Select the batting team',sorted(teams))

with col2:

BowlingTeam = st.selectbox('Select the bowling team',sorted(teams))

selected\_City = st.selectbox('Select host city',sorted(cities))

target = st.number\_input('Target')

col3,col4,col5 = st.columns(3)

with col3:

score = st.number\_input('Score')

with col4:

overs = st.number\_input('Overs completed')

with col5:

wickets = st.number\_input('Wickets out')

if st.button('Predict Probability'):

runs\_left = target - score

balls\_left = 120 - (overs\*6)

wickets = 10 - wickets

crr = score/overs

rrr = (runs\_left\*6)/(balls\_left)

input\_df = pd.DataFrame({'BattingTeam':[BattingTeam],'BowlingTeam':[BowlingTeam],'City':[selected\_City],'runs\_left':[runs\_left],'balls\_left':[balls\_left],'wickets':[wickets],'total\_runs\_x':[target],'crr':[crr],'rrr':[rrr]})

result = pipe.predict\_proba(input\_df)

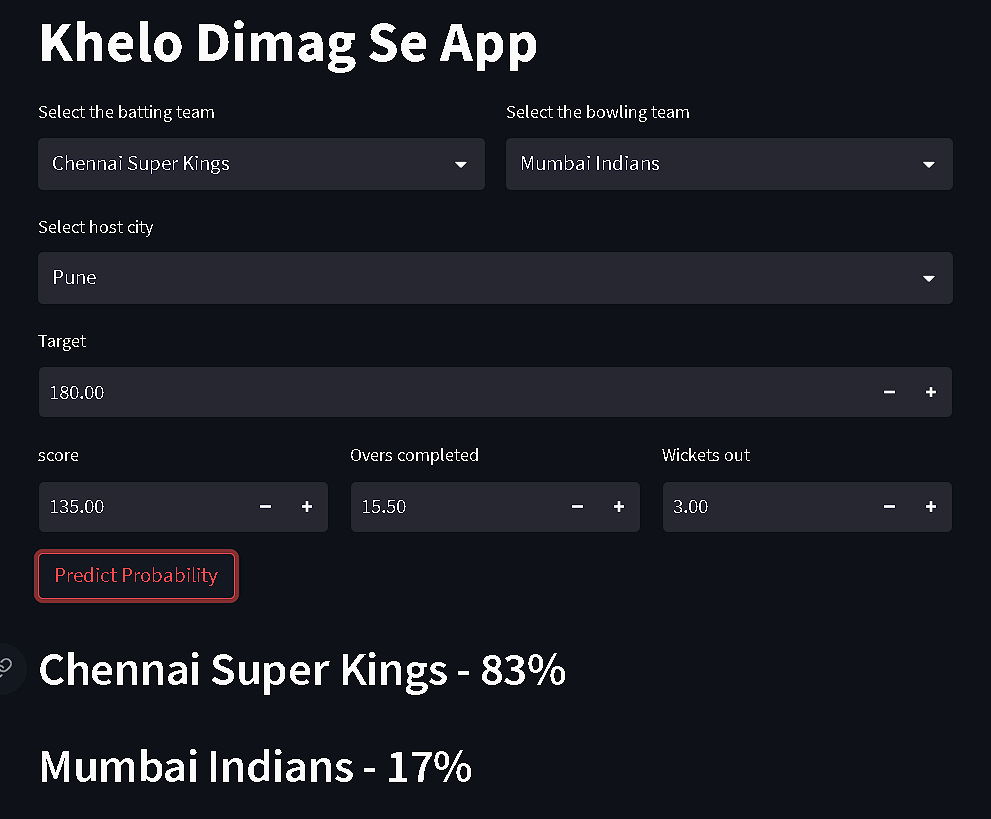
loss = result[0][0]

win = result[0][1]

st.header(BattingTeam + "- " + str(round(win\*100)) + "%")

st.header(BowlingTeam + "- " + str(round(loss\*100)) + "%")

**7.2 WEB APPLICATION IMPLEMENTATION**

 “Khelo Dimag Se App” is an user interface of our model that we have developed using streamlit. This web application lets user to interact with our data base and get accurate predictions for any math ups. The user interface is simple and easy to understand. You just have to enter some data like name of the bowling team, name of the batting team, place where the match is happening, target to be chased, current score, no. of overs down and no. of wickets down then you can hit “Predict Probability” button to get the predictions.

**Fig 7.2 : Web Application Implemantation**

After pressing the “Predict Probability” button you will get the winning chance of both the teams in percentage. The prediction is live that means it changes with every ball. The model is highly accurate so it gives most accurate predictions for every scenario. This prediction is based on past data. To make these predictions the model considers different factors like batting and balling team, place, current run rate, required run rate, weather conditions, player line up etc.

**7.3 POWER BI DASHBOARD**

The second part of our project is IPL data Analysis. We have analysed the data of past 15 seasons of IPL and we have developed a Power BI dashboard to display this data in a more easily graspable format. As the data available on internet is in the form of excel files which is very hard to understand so we have created this dashboard to display the data.

This power BI dashboard has multiple pages where each page focuses on specific part of the game. We have analysed the data of past 15 years and performed pre-processing on this data. After that we have developed relations between different data tables. Then we created multiple pages such as Match Stats, Batsman Stats, Baller Stats and summary.



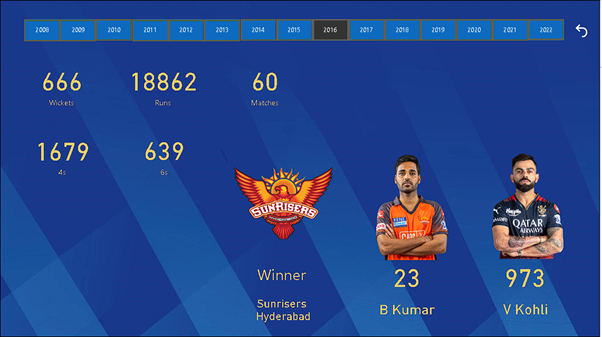
**Fig 7.3 : Landing Page**

Following are the different field specific pages that we have created in this dashboard.

* Summary
* Team Stats
* Batsman Stats
* Baller Stats

**7.3.1 SUMMARY PAGE**

The summary page shows summary of every season of IPL. From the above list you can select any season and you will see the number of matches played in that season, number of runs scored, number of fours and sixes hit, wickets taken winner of the season, orange cap holder and purple cap holder of that season.



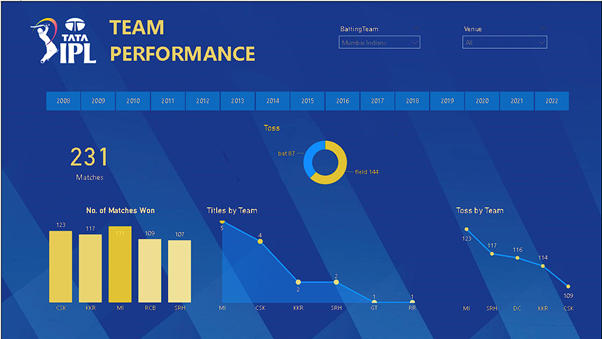
**Fig 7.3.1 : Sumary Page**

As you can see in the above figure when you select the 2016 season of IPL we are getting all the information. In the 2016 season of IPL, 666 wickets were taken, 18862 runs were scored in 60 matches. These runs includes 1697 4s and 629 6s. The winner team of this season was Sun Risers Hyderabad. Bhuvaneshwar Kumar bacame the orange cap holder with 23 wickets while Virat Kohli became the Purple cap holder with a total pf 973 runs.

In this way you can select any other season from the list on the top and get all the data of that season in this same manner.

**7.3.2 TEAM PERFOMANCE PAGE:**

The team performance page highlights the data of specific teams playing in the IPL. Here you can select any team from a list and see all the information related to that team. This information includes no. of matches played by the team in any season, no. of matches won in that season, Toss win rate of the team, titles held by that team and many more.



**Fig 7.3.2 : Team Stats Page**

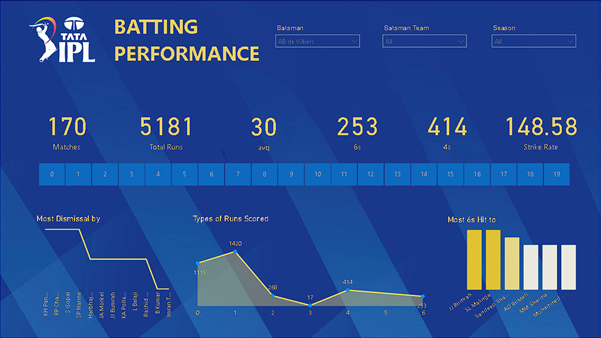
As you can see in the above image when we select Mumbai Indians it shows that MI has played 231 matches in IPL you can select any specific season to see the number of matches played by MI in that season. The page also shows the no. of matches won by other teams. You can also see the no. of times MI chose to bat first or field first after winning the toss.

In the same way you can select any other team and get the information of that team. The page is made using line graph, bar graph, and different charts which makes it simple to understand as well as graspable.

**7.3.3 BATSMAN STATS PAGE**

The batsman stats page is one of the most important parts of our analysis. This page gives you all the information related to any batsman. You can select any batsman from the list which contains all the batsman ever played in IPL. After selecting the batsman you will get the information such as number of matches played, total runs scored in those matches, batting average, total centuries and half centuries scored by that batsman, total number of 4s and 6s scored, Strike rate and many more.

You can also select any season get the information of the batsman related to that specific season. If you don’t select any season then the stats will be combined for all the seasons.



**Fig 7.3.3 : Batsman Stats Page**

As you can see in the above picture when we select AB de Villers it shows that AB has played 170 matches in all the seasons of IPL and has scored a total of 5181 runs with an average of 30. He has scored 253 6s, 414 4s and has a strike rate of 1488.58.

**7.3.4 BOWLING STATS PAGE**

The bowling stats page displays all the information related to specific bowler which is a very important part of cricket. Here also you just have to select the bowler form the list and you the information of that bowler will be displayed on the page. This information includes matches played by that bowler in all the seasons, total wickets taken in those matches, Economic rate, Strike rate and the average of that bowler.

**Fig 7.3.4 : Bowling stats page**

As you can see in the above picture when we select Chahal as bowler from the list the page displays all the information of Chahal. The page shows that Chahal has played 130 matches in 15 seasons and has taken 166 wickets with an economic run of 7.55. His strike rate is 17.71 with an average of 22.28.

You can also select any season to get the information of bowler for that specific season. You can also select opponent team. After selecting the opponent team the information will only be for the matches played against that specific team.

**CHAPTER 8**

**FUTURE WORK**

**7.1 FUTURE WORK**

At present, the data is limited to match and score. It doesn't have details about the players and their stats. There is a great scope for applying this concept to the players and their stats data and can find the batting order and bowling order of a particular match. It will be helpful to franchise people who are at decision making level.

This survey helps to propose a model that helps in the prediction and analysis of the team and player performance. The useful applications are online fantasy games, used by team analyst, which provides stats to cricket lovers and they can also use to access an opponent’s strengths and weakness. The IPL prediction helps people who are willing to play online fantasy games, such as dream11, MPL, and other online platforms. We are planning to build an IPL analysis website which will have all the records and information regarding matches that has been happened in the past seasons.

Improving the accuracy of player performance forecasting is another promising area for future research. By considering historical data, contextual factors, and player-specific models, researchers can strive for more precise predictions regarding player form and performance trajectories.

Additionally, future work can focus on contextual analysis by incorporating match-specific factors like pitch conditions, weather conditions, team strategies, and head-to-head records. Such contextual elements contribute to a more comprehensive understanding of match outcomes and player performances, ultimately enhancing the predictive models' efficacy.

**7.2 APPLICATIONS**

The IPL data analysis and prediction project can have several applications across different stakeholders. Here are some potential applications:

1. Team management and strategy: IPL teams can leverage data analysis and prediction models to gain insights into player performances, identify strengths and weaknesses, and formulate effective strategies. This can help teams make informed decisions regarding team composition, batting order, bowling tactics, fielding positions, and player transfers.

2. Player selection and scouting: Data analysis can aid in player selection and scouting processes. By analysing player statistics, performance trends, and match conditions, teams can identify talented players, assess their potential contributions, and make strategic decisions during player auctions and transfers.

3. Fan engagement and fantasy leagues: IPL data analysis and prediction models can enhance fan engagement by providing accurate and insightful predictions. Fans can use these predictions to participate in fantasy cricket leagues, make informed decisions in fantasy teams, and compete with other fans. This adds an extra layer of excitement and involvement to the IPL viewing experience.

4. Broadcast and commentary: Data analysis findings can be incorporated into IPL broadcasts and commentary to provide deeper insights and enhance the storytelling aspect of the matches. Commentators and analysts can use statistical trends, player comparisons, and predictive analysis to provide informative and engaging commentary during the matches.

5. Betting and gambling: While it is important to note that gambling regulations vary across jurisdictions, some platforms and individuals may use IPL data analysis and prediction models for betting and gambling purposes. These models can provide insights and predictions that inform betting decisions. However, it is essential to follow legal and responsible gambling practices.

6. Sponsorship and marketing: Data analysis can offer valuable insights to sponsors and marketers. By understanding team and player performance trends, marketability, and fan preferences, sponsors can make informed decisions regarding sponsorship deals, endorsements, and marketing campaigns.

7. Research and academic studies: IPL data analysis and prediction projects can contribute to academic research in the field of sports analytics. The findings can be used to develop new methodologies, models, and insights into the dynamics of T20 cricket, player performance analysis, and match outcome prediction.

**CHAPTER 9**

**REFERENCES**

**9.1 REFRENCES**

* G. Sudhamathy and G. Raja Meenakshi, “PREDICTION ON IPL DATA USING MACHINE LEARNING” ICTACT JOURNAL ON SOFT COMPUTING, OCTOBER 2020
* Aminulislamanik, SakifYeaser, A.G.M. Imam Hussain, Amitabha Chakrabarty, “Player’s Performance Prediction in ODI Cricket Using Machine Learning Algorithms,”BRAC university,2019.
* KilariJaswanth, KenchamArun, B Deva Deekshith, KilariAshwik, Pavithra J, “Analyzing and Predicting Outcomes of IPL Cricket Data,”IRJET Volume: 09 Issue: 03, Mar 2022.
* D. Jyothsna, K. Srikanth, “Analyzing and Predicting outcome of IPL Cricket Data”, International Journal of Innovative Research in Science, Engineering and Technology,Vol. 8, Issue 4, April 2019
* SiddharthSinha, PrakashTripathi, AakashVishwakarma, Prof. AmrutaSankhe, “IPL Win Prediction System To Improve Team Performance using SVM.,”International Journal of Future Generation Communication and Networking Vol. 13, No. 1s, (2020), pp. 17- 23
* Ruchitha M and Dr. Preeti Savant, “IPL WINNER PREDICTION USING ML ALGORITHMS”,IJEASTVol. 6, Issue 9, January 2022
* RabindraLamsal and Ayesha Choudhary, “Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning”, arXiv:1809.09813 [stat.AP] (September 2018)
* S. Abhishek, Ketaki V. Patil, P. Yuktha and S. Meghana,“Predictive Analysis of IPL Match Winner using Machine Learning Techniques”,International Journal of Innovative Technology and Exploring Engineering, Vol. 9, No. 1, pp.430-435, 2019.
* Vistro, Daniel, Leo Gertrude David,“The cricket winner prediction with application of machine learning and data analytics,”International Journal of Scientific and Technology Research, Volume 8, Issue 09, 2019.







