EFFICIENT AUCTION PRICE PREDICTION USING MACHINE LEARNING FOR IPL TOURNAMENT

DISSERTATION

Submitted to the University of Kerala in partial fulfillment of the requirement for the completion of MSc Computer Science with specialization in Artificial Intelligence Degree
University of Kerala

By
SHAHAL AHAMMED K

85721607017

Department of Computer Science

University of Kerala

Thiruvananthapuram

August 2023

DEPARTMENT OF COMPUTER SCIENCE

UNIVERSITY OF KERALA

THIRUVANANTHAPURAM,

KERALA-695581

CERTIFICATE

This is to certify that the project report entitled **'EFFICIENT AUCTION PRICE PREDICTION USING MACHINE LEARNING FOR IPL TOURNAMENT'** is a bonafide record of the project done by **SHAHAL AHAMMED K (85721607017)** in partial fulfillment of the requirements for the completion of MSc Computer Science with specialization in Artificial Intelligence Degree of University of Kerala.

Supervisor
Dr. Aswathy A.L
Assistant Professor
Department of Computer Science
University of Kerala
Thiruvananthapuram

Head of the Department
Dr.D.Muhammad Noorul Mubarak
Associate Professor
Department of Computer Science
University of Kerala
Thiruvananthapuram

Date of viva-voice:

External Examiner:

ACKNOWLEDGEMENT

First and foremost, I extend my sincere thanks to God for His divine guidance and inspiration throughout this project, which has allowed me to accomplish my goals and deliver this work successfully.

I would like to express my sincere thanks to **Dr. D. Muhammad Noorul Mubarak**, Head of the Department, Department of Computer Science, University of Kerala, for his valuable suggestion and vital encouragement.

I convey my heartfelt gratitude to my supervisor **Dr. Aswathy A.L**, Assistant Professor, Department of Computer Science, University of Kerala, for encouraging me to investigate my interests and support me throughout the work. I am extremely humbled by her guidance during the time of research and the writing of this thesis.

I am greatly obliged to **Dr. Vinod Chandra S.S,** Professor, Department of Computer Science, University of Kerala, for his support and valuable comments rendered to refine my work.

I am greatly obliged to **Dr. Aji S**, Associate Professor, Department of Computer Science, University of Kerala, for his support and valuable comments rendered to refine my work.

I take immense pleasure in thanking **Dr. Philomina Simon**, Assistant Professor, Department of Computer Science, University of Kerala, for motivating me throughout the course and for encouraging me to always work on challenging areas.

I am grateful to **Ms. Rhythu N Raj, Ms. Shyja Rafeek S, Ms. Neethu M S, Ms. Krishna S S, Ms. Hazeena A J** Assistant Professors, Department of Computer Science, University of Kerala, for the unwavering and kind support extended during the work.

Last but not least, I would like to extend my sincere love and gratitude to my family members and friends for being there and for the unconditional support and prayers, without which the completion of this thesis would be a far-fetched dream.

SHAHAL AHAMMED K

DECLARATION

I here by declare that the work presented in the Dissertation titled **"EFFICIENT AUCTION PRICE PREDICTION USING MACHINE LEARNING FOR IPL TOURNAMENT"** is completed by me under the guidance of Dr. Aswathy A L, Assistant Professor, Department of Computer Science, University of Kerala, Kariavattom Campus, Thiruvananthapuram and it has not been included in any other thesis submitted previously for the award of any degree.

Place: Kariavattom SHAHAL AHAMMED K

Date:

ABSTRACT

Cricket is one of the major worldwide sports played in different countries across the world. It is a game where two teams each consisting of eleven players play to win the match. The Indian Premier League (IPL) is one most entertaining and popular franchise-based cricket tournament that is held in India. IPL has immensely grown its popularity from the 2008 inaugural season to now. The tournament is conducted by the Board of Control for Cricket in India (BCCI). For each edition franchisees select their players through auction. The players were categorized into batters, bowlers, all-rounders, and wicket-keepers in the auction. This study provides finding optimal prices for each player according to their performance. In this approach, machine learning models evaluate players' performance. The best evaluation is selected for the prediction. According to that evaluation and predict the auction price for each individual player. The predicted auction price is compared to the real price that the player was sold. The predictions help franchises to select the best players at a particular position within their budget allocated for the player.

CONTENTS

1. INTRODUCTION

A tournament participated by ten franchises containing thrilling 74 matches across three months in a year. The Indian Premier League become the most popular cricket tournament in the world. It is a franchise-based tournament that was initially started on April 2008 by the Board of Control for Cricket in India (BCCI), IPL has attracted fans not only in India it has fans across worldwide. The viewership of IPL has grown in each season also in the case of growing fans in sports. Cricket is a bat and ball game that is played in different formats, such as Test, ODI, and T20. Nowadays many tournaments like IPL are played around the world such as the South African T20 league, Sri Lankan T20, MLC Cricket America, etc. where execute auctions. It is a team game where each player put their best performance for their particular roles. The game has different roles for players such as Batters, Bowlers, All-rounders, and wicket-keepers. Each player performs according to their role and position in the game, for example, batters need to run between the wickets and make runs and bowlers need to convert each a wicket. Fans eagerly wait for each season to see how well players from different countries perform in IPL. The ten participating teams have been split into two groups of five through a random draw. This draw not only determined the composition of the groups but also established the match schedule across the groups, with each team playing a total of 14 matches during the group stage. Within their group, each team competes against the other four teams twice – once at their home venue and once at the opponent's venue. Additionally, they play against four teams from the opposite group once and face the remaining team twice. Points are awarded as follows: a victory grants 2 points, no points are given for a loss, and in the event of a tie or no

result, both teams receive 1 point. The competition advances to a four-game playoff stage comprising Qualifier 1, Qualifier 2, Eliminator, and Final matches.

According to BCCI, each franchise is mandated to have at least 18 players with a maximum squad depth of 25 players. Each team can have a maximum of 8 players as foreigners in their squad. Only 4 foreign players are allowed to appear in a match. Given the massive pool system players were represented by their potential and roles such as Marquee, Batters, Bowlers, All-rounders, and Wicket keepers as pool wise. Each player was given a base price according to their potential. The franchise has a budget for the auction, the franchise must spend 75% of their budget and buy players without exceeding the budget amount. The bidding process of IPL auction is dynamic in nature franchises frequently change their strategy of buying players by player economic value, potential of the player, and budget of the franchise. The team selection and player performance have many subjective involving in auction team rivalry, auction atmosphere, etc. The actual money spent on the player and the performance of the player on the price sometimes is enough, for example, Kyle Jamieson was bought by a franchise for 15 crores in 2021, and the performance to the money they spent was not economic value, the same year Ruturaj Gaikwad bought by franchise costing 20 Lakh and he was top runs scorer that season. After analyzing the example, In the study can say that the selection of players in the auction is not done by player's optimal value.

1.1 PREDICTION

Prediction is the act of using available information and data to make an informed guess or estimation about a future event or result. It entails examining historical patterns, trends, and present circumstances to anticipate potential outcomes. By analyzing these factors, individuals can forecast what might occur in the future.

1.2 TYPES OF PREDICTION

Point prediction: This kind of prediction involves offering a precise forecast of a future outcome. For instance, it may entail predicting that a particular stock will achieve a specific price level at a designated point in time.

Interval prediction: This type of prediction offers a spectrum of potential results. For instance, it anticipates that there is a 90% likelihood of a hurricane reaching land at some location within a specific region in the upcoming week.

Categorical prediction: This type of prediction involves predicting the likelihood of an event occurring in a specific category. As an example, estimating the probability of an individual acquiring a specific illness or the chances of a sports team securing a victory in a match.

Long-term prediction: This type of prediction involves forecasting events or trends that are expected to occur over a longer period, such as predicting climate change or population growth.

Short-term prediction: This form of prediction pertains to anticipating events or trends that are projected to happen over a brief period, like forecasting the weather or predicting the stock market's performance for the following day.

Qualitative prediction: This form of prediction entails forming subjective assessments or expert viewpoints rooted in unquantifiable data, like foreseeing

the societal repercussions of a novel technology.

Quantitative prediction: This form of prediction entails employing mathematical models and statistical techniques to anticipate upcoming events or trends, like projecting consumer demand for a novel product.

Probabilistic prediction: This form of prediction entails assessing the chance or probability of a forthcoming event or result taking place. For instance, foreseeing the likelihood of an individual's survival following a medical procedure.

Deterministic prediction: This form of prediction entails offering an absolute or guaranteed result relying on information that is already known. For instance, foreseeing the result of a coin toss or the solution to a mathematical equation.

Black box prediction: This type of prediction involves using machine learning algorithms or other complex models to make predictions without necessarily understanding how the model arrived at its conclusion. This type of prediction is often used in applications such as fraud detection or image recognition.

1.3 METHODS FOR PREDICTION

Statistical method: These methods involve analyzing historical data and using statistical models to identify patterns and trends that can be used to make predictions about future events or outcomes.

Machine learning method: These methods involve training algorithms to learn patterns and relationships in data and using these models to make predictions about new data.

Expert judgment: This method involves relying on the knowledge and expertise of individuals who have specialized knowledge in a particular area to

make predictions.

Stimulation methods: These methods involve creating computer models that simulate real-world situations to predict outcomes. For example, simulating the spread of a virus in a population to predict the impact of different intervention strategies.

Rule-based method: These methods involve using a set of rules or decision trees to make predictions based on specific criteria. For example, using a set of rules to predict the likelihood of a loan being approved based on a person's credit history and income.

Time-series forecasting: This method involves analyzing historical data to identify patterns and trends over time and using these patterns to make predictions about future values in a series, such as predicting stock prices or demand for a product.

Neural networks: These are a type of machine learning method that involve building networks of interconnected nodes that can learn to make predictions based on input data.

1.4 MOTIVATION AND CHALLENGES

This approach can be used to evaluate players' past performance and get an optimal price for each individual player. The evaluation helps franchises in the Indian premier league to buy players with what they actually worth according to their performance. The study provides select best players to the franchise based on roles and budget allocated for the player.

Existing models could be improved with more sensible and semantic data. Existing model has mathematical evaluation and used supervised machine learning model for prediction. The model has less

number of facts for the perfect prediction. Current model does not consider features like popularity, current form, T20 domestic based performance of the player in data, etc. Due to lack of data prediction of the young players were made error in the prediction. The model had high errors on prediction.

OVERVIEW OF THE WORK

In this study, Regression and Clustering models are used for Machine Learning. Many industries have begun to employ Machine Learning and Prediction techniques and applications. To get a better result and analyze players' price prediction, in study attempted to evaluate and predict using Machine learning models on a dataset.

1.5 CONTRIBUTION OF THE WORK

In the study data sets were created from www.IPLT20.com,www.espncricinfo .com, www.icc cricket.com, www.kaggle.com, and www.cricbuzz.com. Two datasets were created, the first dataset containing more than 40 features in auction data including last performance of the player before the IPL season. The second dataset contained more than 20 features in that season player performance which was sold at the auction. In the study evaluated player performance by building a mathematical model on the basis of role players (Batter, Bowler, All-rounder, Wicket-keeper). The evaluated model is used as a feature in datasets as rating. In the study use K-means clustering to categorize players according to their performances. Standardization of clusters were done using ratings in the dataset. Decision tree were used to predict data, achieved accuracy 95% and 96% for the data sets. The price for each individual player were assigned according to their standard of performance. In the study compare how players performed to their price in the IPL season. Finally, In the study calculate the mean difference

between actual price the player sold and predicted price in the model. The study is compared to the previous study which achieved less errors than previous study.

1.6 ORGANIZATION OF THESIS

The report is divided into 5 chapters. Chapter 1 gives a formal introduction and overview about this project work. Chapter 2 gives review of literature. Chapter 3 explains about the dataset, neural networks, block diagram. Chapter 4 explains the implementation and results. Chapter 5 summarizes the conclusions drawn from the work and future work.

2. LITERATURE REVIEW

Several techniques were implemented for predicting the auction price of players using several Machine Learning prediction models such as Linear Regression, ANN, Decision Tree Regressor, KNN, etc.

Gaurav Malhotra et.al. proposed a hedonic pricing model to predict price of players in **A Comprehensive Approach To Predict Auction Prices And Economic Value Creation Of Cricketers In The Indian Premier League (2022**). The paper aims to produce more accurate results using the hedonic pricing method. They created a Hedonic model by linear regression using several statistics to estimate auction price. The goal of the project was to find the difference between the price predicted by the model and actual price they got in the auction. Various computation was performed to consistency and analyzability. Python tools were used to create visual representation to investigate and represent the correlation between the independent factors and the dependent variable of the Machine Learning model. Various Machine Learning models were attempted for getting desired Machine Learning model. Machine Learning Models like Linear regression, Bayesian ridge regression and Random Forest were attempted. Upon evaluating different aspects including simplicity, accuracy, and reliability, the decision was made to formulate a straightforward equation. This equation should be easily usable by inputting relevant values, aligning with the preference for an explicit approach. Linear regression was used as Machine Learning Model. In result computed the auction price for the players and found the Mean difference between the predicted price and actual price they got in the auction.

Dr. Jhansi Rani et.al. proposed a Multi models' comparison in predicting cost of players sold in **Prediction of Player Price In IPL Auction Using Machine Learning Regression Algorithms (2020).** They Applied Machine Learning algorithm to predict the cost players being sold in IPL Auction. The work estimated the player's selling price using their past performance parameters like runs, balls, innings, wickets and matches played, etc. In the paper they applied algorithms like K-Nearest Neighbors (KNN), Support Vector Regressor (SVR), Decision Tree Regression, Linear Regression, Stochastic Logistic Regression, Random Forest Regressor predicting which one gives best result**.** Concluded that SVR is the best model for batsmen and Linear regression is the best model for bowlers. The result where the model has been able to capture all parameters results and predict the prices. Most of the results were accurate where some results have fluctuation were there not considered popularity and experience.

AHP-Neural Network Based Player Price Estimation in IPL proposed by Dey et.al. in 2014. A hybrid model combining the Analytical Hierarchy Process (AHP) and Artificial Neural Network (ANN) has been developed to assess player valuations in the Indian Premier League (IPL). The AHP method is utilized to determine the relative significance of various attributes, while the latter stage involves training an Artificial Neural Network (ANN-BP) to generate accurate player price estimations. This approach incorporates expert opinions to derive optimal player prices by assigning appropriate weights. The estimation process considers essential factors such as Player's Performance Appraisal, Player's Experience Contribution, and Player's Recent Form. The outcomes indicate instances where certain players received prices higher than their warranted value, while others did not receive the deserved estimation based on the model's output. Table 2.1 shows the overall literature review.

Author	Paper	Method	Year
Gaurav Malhotra et. al.	A Comprehensive Approach To Predict Auction Prices And Economic Value Creation Of Cricketers In The Indian Premier League	Linear Regression used for Creating Hedonic Model	2022
Jhansi Rani et.al.	Prediction Of Player Price In IPL Auction Using Machine Learning Regression Algorithms	KNN, SVR, Decision Tree Regressor, Linear Regression, Stochastic Logistic Regression, Random Forest Regressor for Models comparison	2020
Pabitra Kr. Dey et. al.	AHP-Neural Network Based Player Price Estimation	Create model based on Analytical hierarchy Process (AHP) and Artificial Neural Network (ANN)	2014

Table 2.1:Literature Review

3. EFFICIENT AUCTION PRICE PREDICTION USING MACHINE LEARNING FOR IPL TOURNAMENT

Cricket is the one of the sports that is popular around world. Now days franchise-based tournaments grow around the world. The tournaments were played by franchises across the states in the country. In Indian Premier League it is 10 franchise-based tournaments. Tournament played around the states host the franchises has fans all over the world which make IPL the most profitable T20 franchise-based tournament in the world make a profit of around 48,000 crores. It makes one major contribution to Indian economy as sports event. In IPL players were selected to franchises by bidding in the auction. The franchises need to acquire the service of at least 18 players in the squad with maximum 25 players in the squad. Each team has limited budget to bring players to their squad. The actual money spent on the player and the performance of the player on the price sometime is enough. Selecting the players with what they worth is the one most solid thing in the auction. Spending money on a player was inaccurate then that decision is affecting the change in the whole squad for the franchise. It makes affect to the performance of the franchise in the tournament. The inaccurate evaluation is the reason behind spending more money to players where spending more than they worth. This project aims to evaluate and predict optimal price which actually each individual player is worth according to their performances.

This study aims to evaluate players' performances and predict optimal price players by Machine Learning models. In this aims to create data sets that contain the necessary features to evaluate player performance. The main demerits of previous models had fewer features to evaluate and predict the price of players which made the model inaccurate. The study aims to include maximum features to evaluate and predict the performance

of the player accurately. This study proposes a mathematical model to evaluate the player performance by roles of players and their ability according to their roles such as batting ability and bowling ability. The evaluated result is taken by rating a feature in the dataset. The study uses K-means clustering for categorizing data by the performance of the player. The study proposes to Decision tree regressor to evaluate clustering and predict price. This study standardizes clusters using a rating that is evaluated by the mathematical model. The price for each player was assigned according to their standard of performance. The study compares how players performed to their prices in the IPL season. The comparison on evaluating mean difference by actual price player was sold at and predicted price according to the model. The study aims to perform a comparison of this models with the previous model on the auction price prediction in IPL.

3.1 MODEL BUILDING

Two well-known Machine Learning models are used for predicting the price of the player are the K-means clustering and Decision Tree Regression this study, In the study created a mathematical model for standardization of clusters. The model was accurate in the player price prediction.

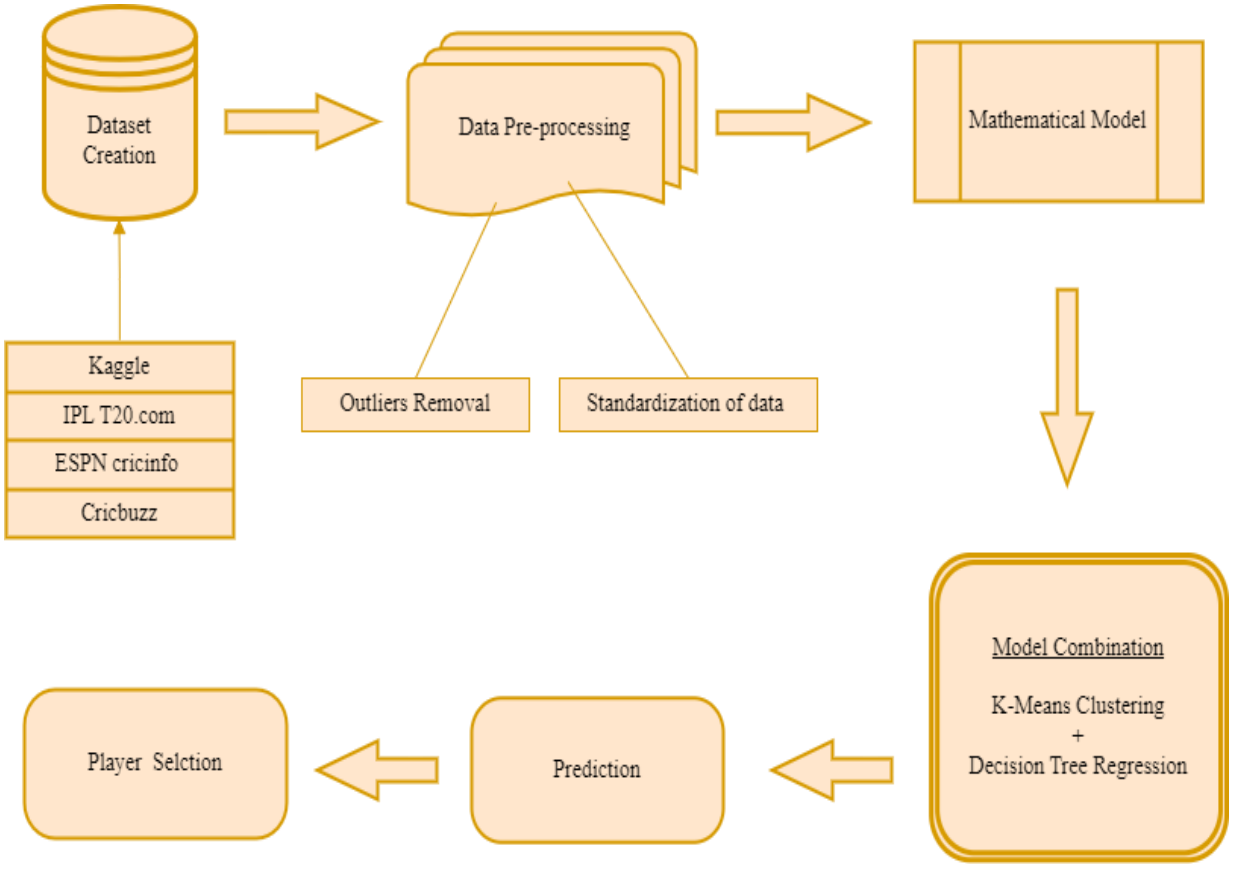


Fig 3.1: Methodology

3.2 DATASET

In the study creates two data sets , Auction dataset for whole players in the auction which is used evaluate the player performance and predict price according to their performance, Stats dataset have the players who were sold at that auction which used to compare actual price players sold at and predicted price for each individual player which is evaluated by this model. The data set were created from www.IPLT20.com,www.espncricinfo.com,www.icccricket.com,www.kaggle.com, and www.cricbuzz.com.In the study data sets were created for 2022 IPL season.

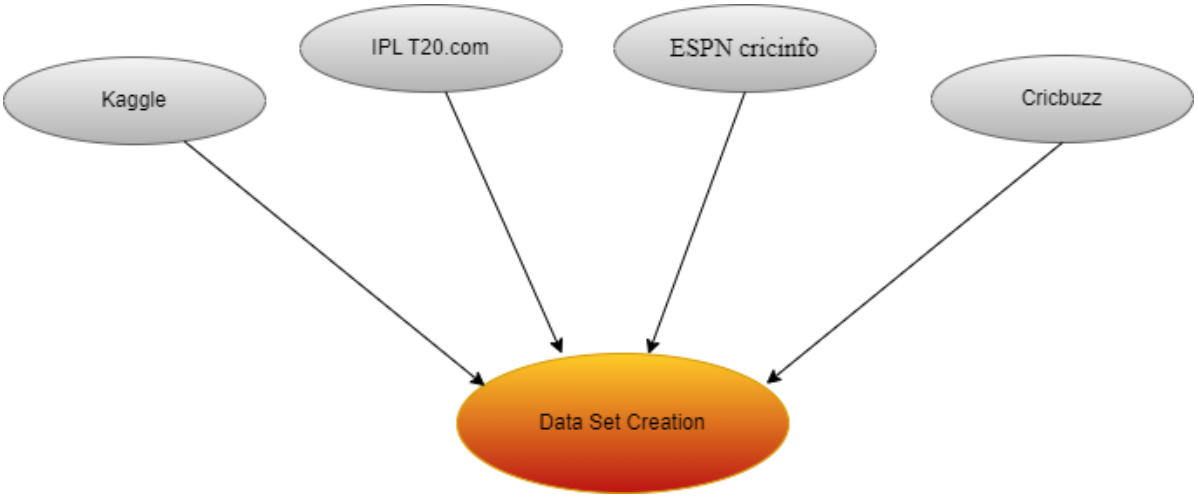


Fig 3.2: Dataset Creation

The datasets contain features which are essential for evaluating players performance and predict price for their performance. The dataset contains features like player current form, popularity, T20 domestic data, etc., which assist predict auction price of young players too.

Auction Dataset For 2022 IPL	
Type of Feature	Features
Player Personal characteristics	Name
	Age
	SLOT (Indian or Foreigner)
Player Role	ROLE (Batter, Bowler, All-rounder or Wicket-keeper)
Player Playing style	Batting (Batting style)
	Bowling (Bowling style)
Caps	U/C/A (international capped or uncapped)
	Test caps (Number of matches played in international test cricket)
	ODI caps (Number of matches played in international One day cricket)
	T20 caps (Number of matches played in international T20 cricket)
	IPL (Number of matches played in IPL)
T20 International performance	T20I Runs (Total runs scored in T20 international)
	T20I HS (High score in T20 international)
	T20I Bat avg (Batting average in T20 international)

	T20I Bat strike rate (Batting strike rate in T20 international)
	T20I Wickets (Wickets taken in T20 international)
	T20I BBI (Best Bowling in T20 international)
	T20I Bowl avg (Bowling average in T20 international)
	T20I Bowl strike rate (Bowling strike rate in T20 international)
	T20I Bowl Economy (Bowl economy in T20 international)
IPL performance	IPL Runs (Total runs scored in IPL)
	IPL HS (High score in IPL)
	IPL Bat avg (Batting average in IPL)
	IPL Bat strike rate (Batting strike rate in IPL)
	IPL Wickets (Wickets taken in IPL)
	IPL BBI (Best Bowling in IPL)
	IPL Bowl avg (Bowling average in IPL)
	IPL Bowl strike rate (Bowling strike rate in IPL)
	IPL Bowl Economy (Bowl economy in IPL)
Last IPL performance	Last IPL Runs (Total runs scored in Last IPL)
	Last IPL HS (High score in Last IPL)
	Last IPL Bat avg (Batting average in Last

	IPL)
	Last IPL Bat strike rate (Batting strike rate in Last IPL)
	Last IPL Wickets (Wickets taken in Last IPL)
	Last IPL BBI (Best Bowling in Last IPL)
	Last IPL Bowl avg (Bowling average in Last IPL)
	Last IPL Bowl strike rate (Bowling strike rate in Last IPL)
	Last IPL Bowl Economy (Bowl economy in Last IPL)
T20 Domestic performance	T20 Runs (Total runs scored in T20 Domestic)
	T20 HS (High score in T20 Domestic)
	T20 Bat avg (Batting average in T20 Domestic)
	T20 Bat strike rate (Batting strike rate in T20 Domestic)
	T20 Wickets (Wickets taken in T20 Domestic)
	T20 BBI (Best Bowling in T20 Domestic)
	T20 Bowl avg (Bowling average in T20 Domestic)
	T20 Bowl strike rate (Bowling strike rate in T20 Domestic)
	T20 Bowl Economy (Bowl economy in T20 Domestic)

Players Additional Information	Current Form (Player form at the auction time)
	Popularity (popularity of the player across the country)
	Captain (Captaining ability of the player)
Base Price	Base Price (The starting bid price evaluated by the player)

Table 3.1: Auction Dataset For 2022 IPL

The auction data set contains 590 data with more than 40 features for evaluating and predicting player price for their performance. Table 3.1 shows the type of feature and features auction dataset.

Stats Dataset for 2022 IPL	
Type of Feature	Features
Player Personal characteristics	Name
	Age
	SLOT (Indian or Foreigner)
Player Role	ROLE (Batter, Bowler, All-rounder or Wicket-keeper)
Season Performance	Matches (Number of matches played)
	Batting innings (Number of batting innings)
	Not out (Number of not outs in the season)

	Runs scored (total runs scored in the season)
	High score (high score in the season)
	Batting avg (batting avg in the season)
	Balls Faced (total number of balls faced in the season)
	Batting strike rate (batting strike rate in the season)
	100 scored (Number of 100's scored in the season)
	50 scored (Number of 50's scored in the season)
	4s (Number of 4's scored in the season)
	6s (number of 6s scored in the season)
	Bowling innings (Total number of bowling innings in the season)
	Overs bowled (overs bowled in the season)
	Runs conceded (Total runs conceded in the season)
	Wickets taken (total wickets taken in the season)
	Bowling average (Bowling average in the season)
	Bowling economy (Bowling economy in the season)
	Bowling strike rate (Bowling strike rate in the season)

Sold Price	Sold Price (The price of a player sold in the auction)

(Table 3.2: Stats Dataset for 2022 IPL)

The Stats data set contains 135 data with more than 20 features to evaluate how the player performed in the season, comparing the actual and predicted price. Table 3.2 shows the type of feature and features stats dataset.

3.3 DATA PREPROCESSING

In the study created the dataset for the prediction. As data pre-processing select essential features for the prediction by abolishing features like player personal characteristic features. The null values in the set in the datasets were replaced by '0'. As datasets stats of the players, In the study can't replace it mean value of the features.

The labeled data were converted into categorical data as it is necessary for clustering and prediction. The categorical conversion done by first replace labeled data based on the feature standard and convert feature into categorical.

The dataset was splits into four datasets by the roles of each individual player. Splits datasets were normalized into the values 0 to 1 for evaluating mathematical model and K-means clustering.

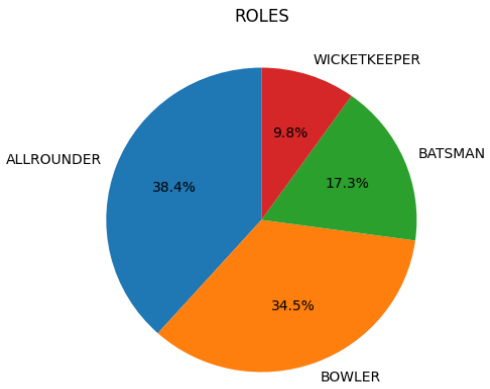


Fig 3.3: Auction dataset Roles %

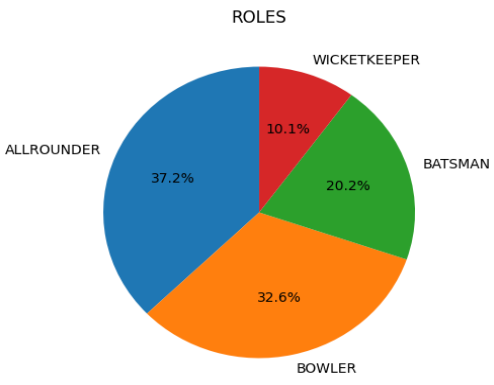


Fig 3.4: Stats dataset Roles %

3.4 MATHEMATICAL MODEL

Ability Type	Features
Bowling Ability	T20I wickets, T20I Bowl avg, T20I Economy, T20I Bowl Strike rate, IPL wickets, IPL Bowl avg, IPL Bowl Economy, IPL Bowl Strike rate, Last IPL wickets, Last IPL Bowl avg, Last IPL Economy, Last IPL Bowl Strike rate, T20 WICKETS, 'T20 BOWL AVG, T20 ECONOMY, T20 BOWL STRIKE RATE
Batting Ability	T20I Runs, T20I HS, T20I Bat avg, T20I Bat Strike rate, IPL Runs, IPL HS, IPL Bat avg, IPL Strike rate, Last IPL Runs, Last IPL HS, Last IPL Bat avg, Last IPL Strike rate', 'T20 MATCHES, T20 RUNS, T20 HS, T20 BAT AVG, T20 BAT STRIKE RATE
Common Ability	C/U/A, Test caps, ODI caps, IPL, T20 cap, captain, Current Form, T20 MATCHES, popularity, Base Price (Rs Lakh)

Table 3.3: Abilities in case of Auction dataset

The mathematical model was used to evaluate the player performance by the mathematical evaluation. The model was used as feature (Rating) in the dataset. It is used to standardize the cluster by the model. The model is assessed by the ability of the player according to their roles. There is main two type of abilities are in cricket, bowling ability and batting ability. Bowling ability is how capable a player in the bowling part of his performance. Batting ability is how capable a player in the batting part of his performance. The player ability is

considered by their roles as Batter and Wicket-keepers consider batting ability, For Bowlers only consider Bowling ability where in case of All-rounders considers both batting and bowling ability. The mathematical model evaluated by the standardized data which were evaluated from the standardization of whole data into the range of 0 to 1.

Ability Type	Features
Bowling Ability	Bowling Innings, Overs Bowled, 'Runs Conceded, Wickets taken, Bowling Average, Bowling Economy, Bowling Strike rate
Batting Ability	Batting Innings, Not out, Runs Scored, High score, Batting avg, Balls faced, Batting Strike rate,50 scored, 4s, 6s
Common Ability	Matches

Table 3.4: Abilities in case of Stats dataset

Mathematical Equation

The mathematical model was created by our own mathematical equation were evaluated by the ability of the player according to their roles. The equation was made to create a rating for the player performance used as a feature.

$$\begin{aligned}\textbf{Rating} = &\left(\left(\frac{\sum \textbf{IPL Performance}}{\textbf{No.of fields in IPL Performance}} \times \frac{25}{100}\right) + \left(\frac{\sum \textbf{Last IPL Performance}}{\textbf{No.of fields in Last IPL Performance}} \times \frac{25}{100}\right)\right.\\ &+ \left(\frac{\sum \textbf{Caps}}{\textbf{No.of fields in Caps}} \times \frac{10}{100}\right) + \left(\frac{\sum \textbf{T 20 Performance}}{\textbf{No.of fields in T20I Performance}} \times \frac{10}{100}\right)\\ &+ \left(\frac{\sum \textbf{T 20 Domestic Performance}}{\textbf{No.of fields in T20 Domestic Performance}} \times \frac{10}{100}\right) + \left(\textbf{Current form} \times \frac{10}{100}\right)\\ &\left. + \left(\textbf{Popularity} \times \frac{5}{100}\right) + \left(\textbf{Captaincy} \times \frac{5}{100}\right)\right) \times \textbf{100}\end{aligned}$$

Equation 3.1: Rating Equation for Auction Dataset

$$\text{Rating} = \left(\left(\frac{\sum \text{Season Performance}}{\text{No of fields in Season Performance}}\right)\right) \times 100$$

Equation 3.2: Rating Equation for Stats Dataset

Rating was used to evaluate player performance and standardize the cluster according to the rating. The standardized clusters were used to set prices for the players and compare player performance in the season.

MODEL COMBINATION

In the study, a hybrid model was employed to group players according to their performance, resulting in 10 distinct player evaluation clusters. Initially, clustering models were utilized to create these clusters, which were subsequently validated using a regression model. The primary goal of this methodology was to assess how accurately the clusters grouped players based on their performance. After experimenting with multiple machine learning approaches, the combination of K-means clustering and Decision Tree regression emerged as the most effective strategy.

Model Combination		
Clustering Model	Regression Model	Accuracy
K-Means Clustering	KNN	55%
	Decision Tree Regression	95%
	Linear Regression	85%
Fuzzy-C-Means	KNN	19%
	Decision Tree Regression	17%
	Linear Regression	8%
Agglomerative Clustering	KNN	63%
	Decision Tree Regression	84%
	Linear Regression	79%

Table 3.5: Model Combination

3.5 K-MEANS CLUSTERING

K-means clustering serves as an unsupervised machine learning approach utilized to address clustering conundrums. Clustering, in this context, pertains to the procedure of aggregating akin data points based on their resemblances. The central objective of the K-means algorithm involves constructing clusters by allotting data points to cluster centers in a manner that minimizes the total squared distance between each data point and its corresponding cluster center. This algorithm functions on untagged data, signifying its independence from predefined categories or classes. Instead, K-means autonomously identifies patterns and correlations within the data to forge distinct clusters. This is accomplished by iteratively updating cluster centers until convergence is reached.

K-means represents an unsupervised machine learning technique that ingests an untagged dataset and segregates it into a pre-established number of clusters (k). It progressively enhances the clusters by computing centroids and reallocating data points to the nearest centroid until convergence is attained. The algorithm's objective is to downsize intra-cluster variance while amplifying inter-cluster variance, ultimately yielding well-defined clusters. The selection of k must be specified before initiating the algorithm. K-means enjoys extensive employment for clustering and uncovering patterns across an array of applications. The algorithm chiefly undertakes two functions: identifying the optimal values for centroid or center points through an iterative process, and assigning each data point to its nearest k-center. As a result, data points in proximity to a specific k-center coalesce into a cluster, with each cluster distinct from the others.

WORKING

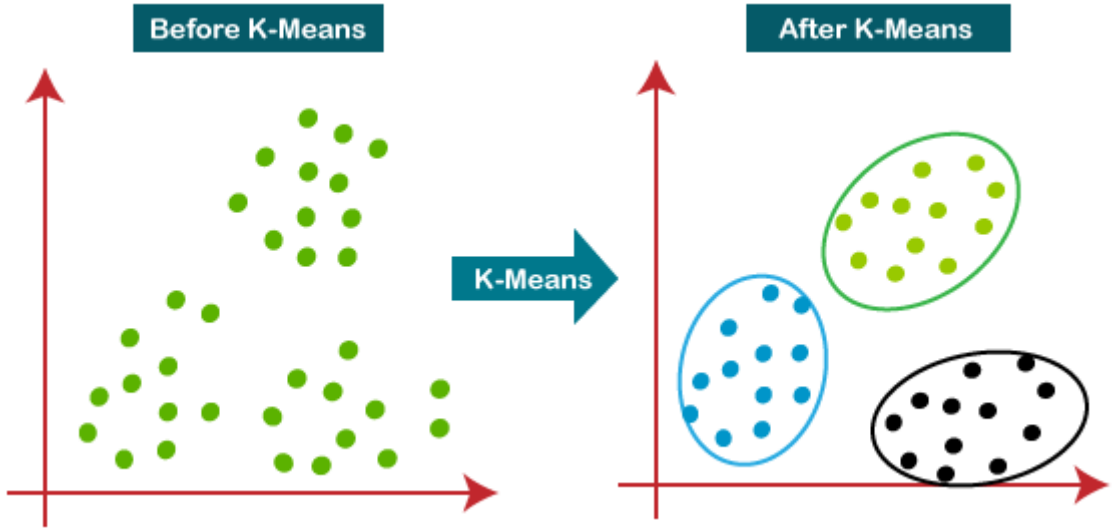


Fig 3.4: K-means clustering

Steps:

1.Initialization: Choose K initial cluster centroids randomly from the data points. These centroids represent the center of each cluster.

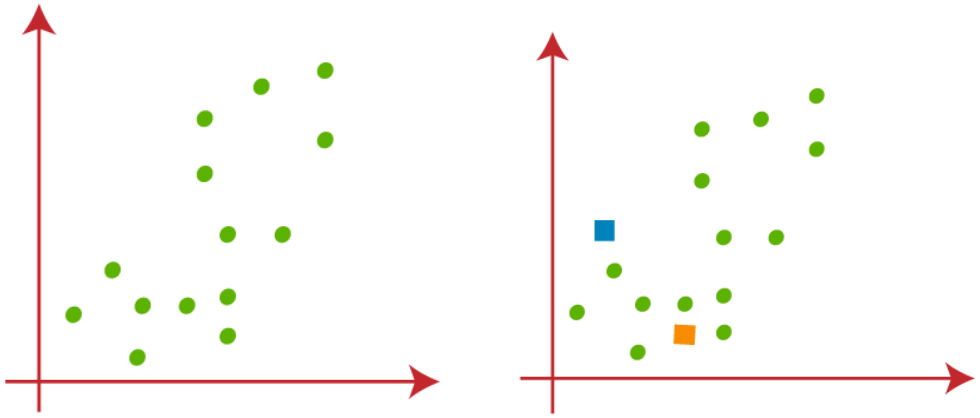


Fig 3.5: K-means clustering from

2.Assignment: Each data point was assigned to its nearest clusters by the calculation of Euclidean distance. This step makes clusters.

$$d = \sqrt{[\ (x2 - x1)^2 + (y2 - y1)^2]}$$

In equation (x1, y1) are the coordinates of one point, (x2, y1) are coordinates another point and d is the distance between (x1, y1) and (x2, y2).

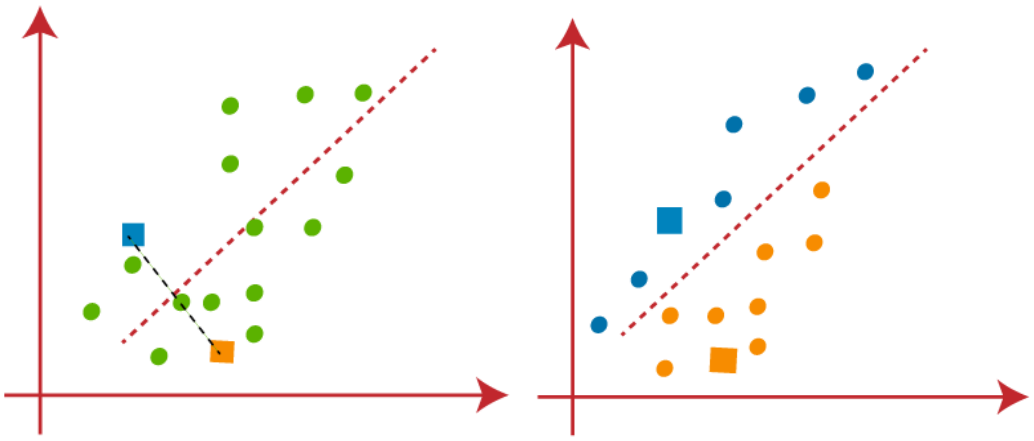


Fig 3.6: K-means clustering

3.Update: Calculate the new cluster centroids by taking the mean of all data points assigned to each cluster.

4.Iteration: Iterations are performed until the centroids no longer change significantly or until a maximum number of iterations is reached.

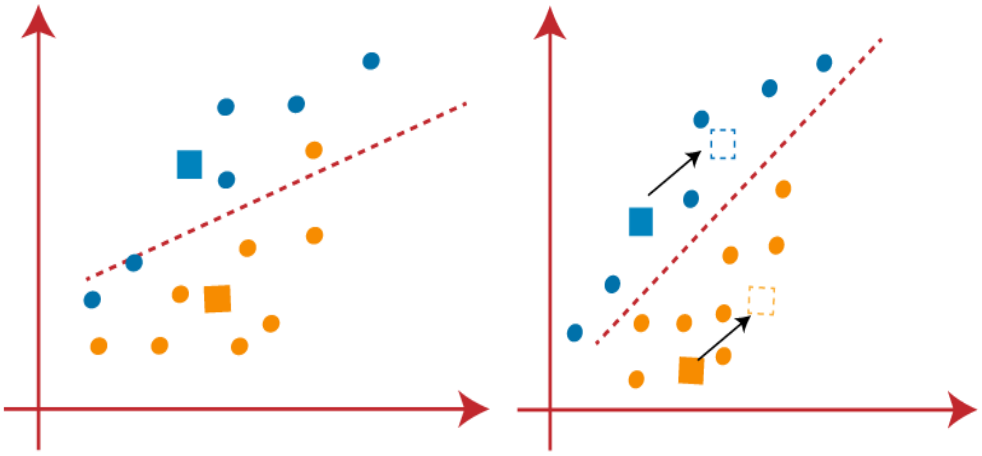


Fig 3.7: K-means clustering

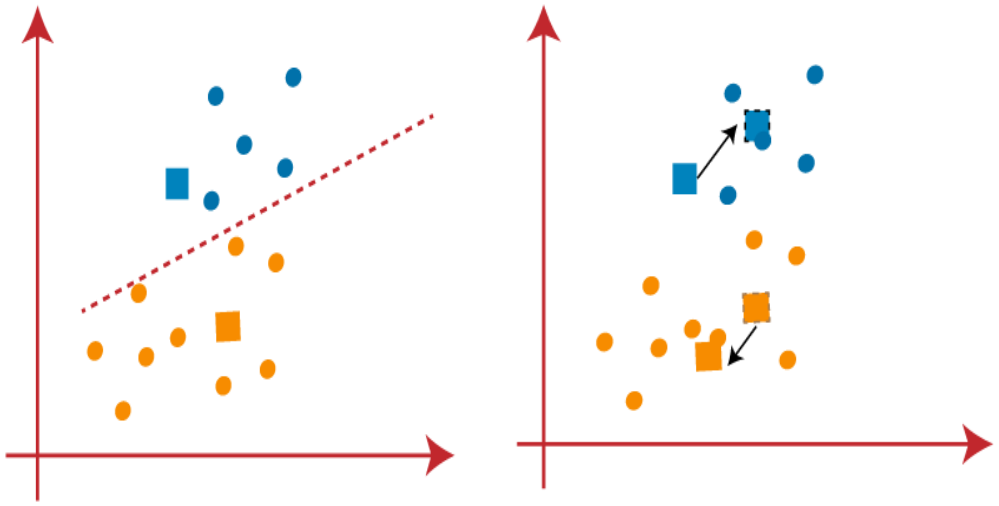


Fig 3.8: K-means clustering

5.Convergence: The algorithm converges when the cluster centroids stabilize, and data points are no longer switching clusters.

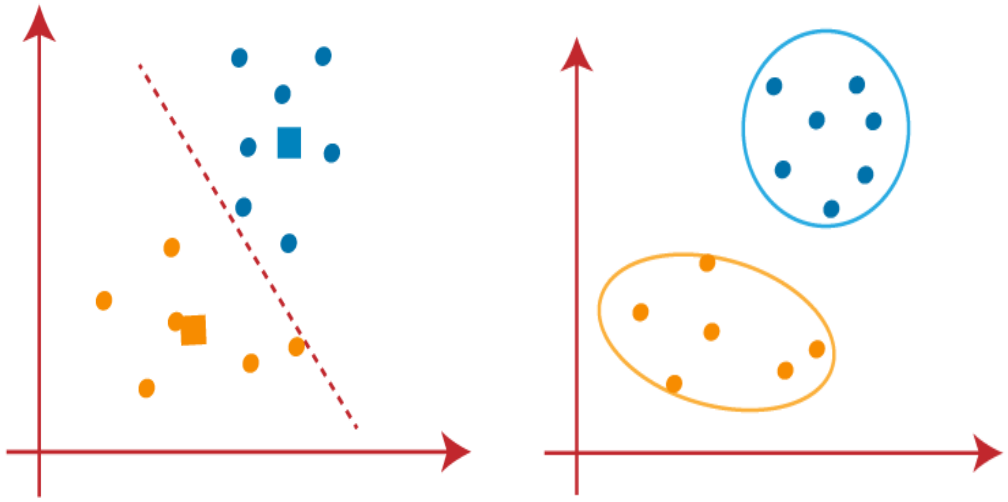


Fig 3.9: K-means clustering

The fig 3.4 to fig 3.9 to shows the working of K-means clustering from unlabeled data points to dividing making them to clusters.

3.6 DECISION TREE REGRESSION

Decision tree regression is a type of supervised machine learning algorithm employed for prediction and classification purposes. This technique involves systematically partitioning the dataset into progressively smaller subsets while simultaneously constructing a decision tree. The end result is a tree structure comprising a root node, intermediate decision nodes, and terminal leaf nodes. Each decision node bifurcates into two or more branches, signifying potential attribute values, while the leaf nodes encapsulate decisions related to numerical targets. The primary decision node at the tree's pinnacle, known as the root node, corresponds to the most influential predictor. Notably, decision trees are adaptable to both categorical and numerical data, rendering them highly effective for resolving decision-oriented challenges.

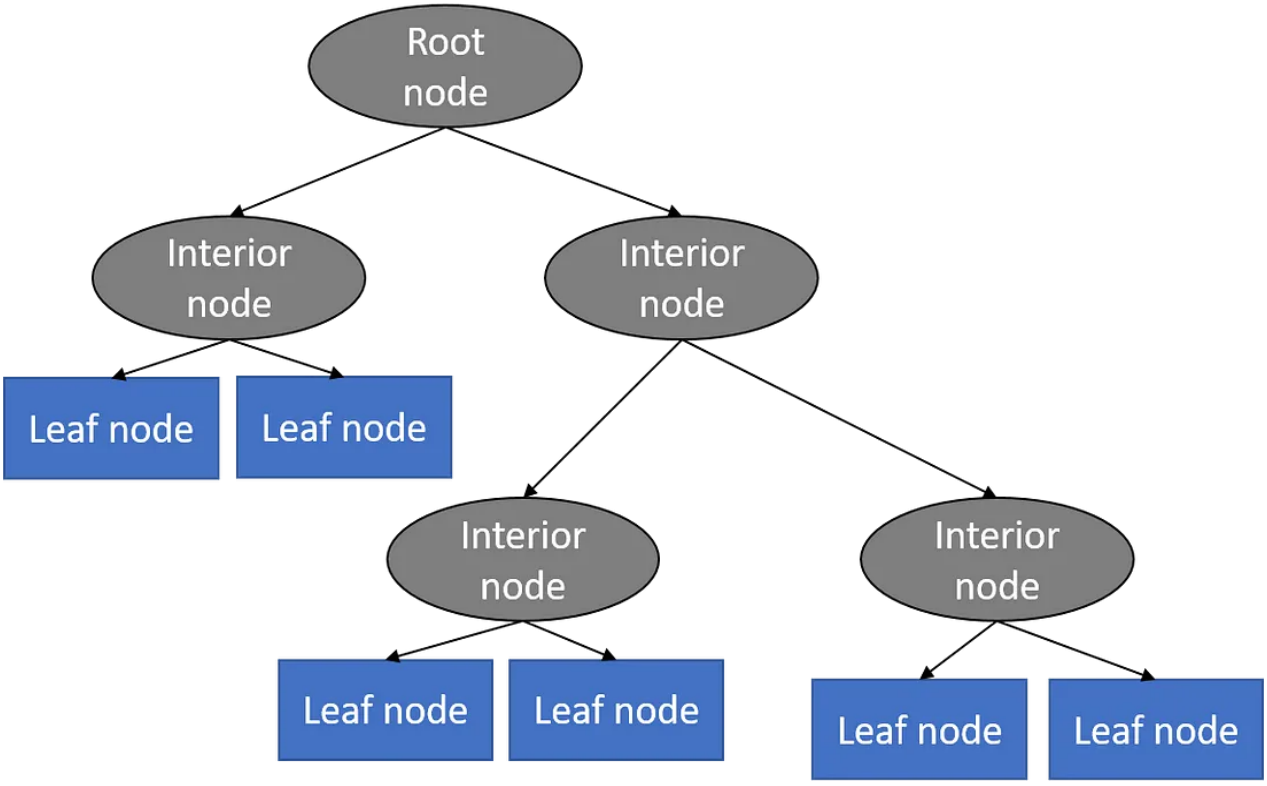


Fig 3.9: Decision Tree Regression

The core algorithm for building a decision tree called ID3 algorithm which is proposed by J. R. Quinlan which employs a top-down, greedy search through the space of possible branches with no backtracking. The Decision tree is built in top-down approach from a root node and involves partitioning the data into subsets that contain instances with similar values. In the study use standard deviation to calculate the similarity of numerical data. Numerical data where completely homogeneous then its standard deviation is zero.

WORKING

In a decision tree used to predict the class of a given dataset, the process begins at the root node. Here, the algorithm evaluates the root attribute's values against the corresponding attribute values in the actual dataset. Depending on this comparison, the algorithm traverses down a specific branch to the next node. This process is repeated for subsequent nodes, as the algorithm continues to compare values within different sub nodes, progressing until it reaches a leaf node. Fig 3.9 shows working architecture of Decision tree regressor.

Steps:

1. Begin the tree with the root node.

2. Find the best attribute from dataset based on Attribute Selection Measures

3. Divide the root node into possible subset which contain possible best attributes

4. Generate decision tree with having best attributes

5. Recursively make new decision trees using the subsets of the dataset created and continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

Attribute Selection Measures (ASM)

Attribute selection measures are used in decision tree algorithm to split data to

get best attributes. The goal is to find the attribute that provides the most information gain. It leading to better separation of classes or values in the resulting subsets. There are two popular techniques for ASM:

- Information Gain
- Gini Index

Information Gain:

The concept of information gain is based on the notion of entropy, which is a measure of the level of impurity or randomness in a dataset. Entropy is calculated for each node in the decision tree and is a way of representing the uncertainty about the class labels at that node.

Information Gain= Entropy(S)- [(Weighted Avg) *Entropy (each feature)

- **Entropy-**: Entropy is a measure of the level of impurity or randomness in a dataset

  Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no)

S= Total number of samples, P(yes)= probability of yes, P(no)= probability of no

Gini Index:

The Gini index measures the level of impurity or randomness in a dataset. For a given node in the decision tree, the Gini index is calculated by summing the probabilities of each class multiplied by the probability of a misclassification for that class. Gini Index calculated by formula below.

$$\textbf{Gini Index= 1- } \sum_j \mathbf{P_j^2}$$

P_j -probability of choosing a sample of class j.

4.RESULT ANALYSIS AND DISCUSSIONS

The clustering was done by the K-means clustering and predicted the target label by Decision tree regression. The accuracy of the model was calculated. The model was compared with Previous study and found to be the model more accurate than previous study.

4.1 Cluster Validation

The clustering was done by the K-means clustering using standardized data from datasets. The clusters must evaluate that it is valid. The validation done by predicting using Decision tree regression. The two datasets were clustered into 10 clusters. Prediction done by decision tree regression and mean squared calculated.

```
from sklearn.tree import DecisionTreeRegressor
model = DecisionTreeRegressor()
model.fit(x_train, y_train)

from sklearn.metrics import mean_squared_error
y_pred = model.predict(x_test)
print(mean_squared_error(y_test, y_pred))
```

```
0.22972972972972974
```

Fig 4.1: Mean squared error for Auction dataset

The fig 4.1 shows the mean squared error which is calculated from auction dataset from Decision Tree Regressor.

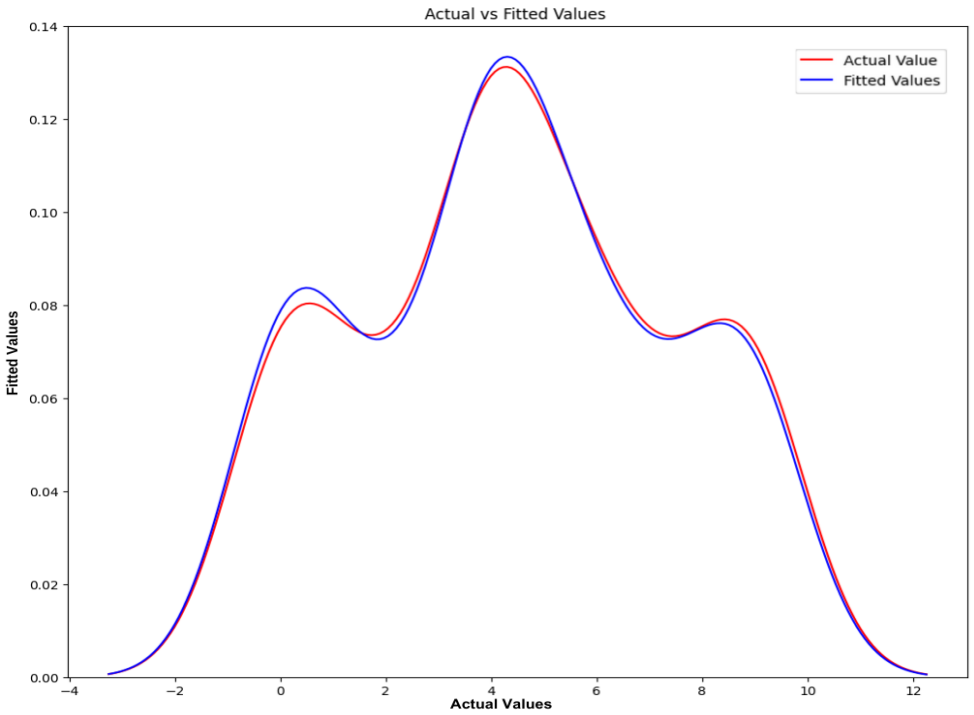


Fig 4.2: Auction dataset: Actual VS Predicted Plotting (1)

The fig 4.2 shows how the actual values and predicted values are plotted in auction dataset

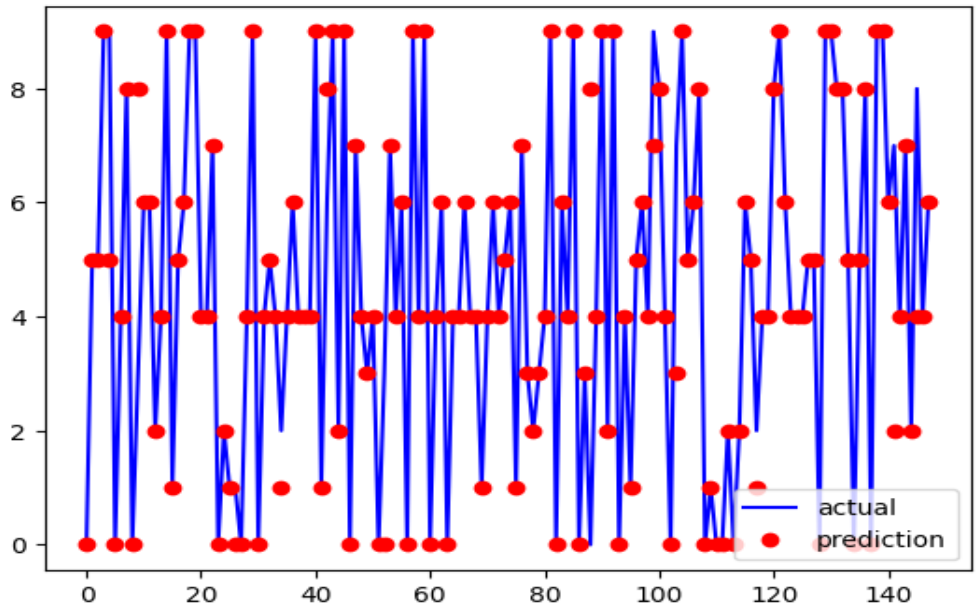


Fig 4.3: Auction dataset: Actual VS Predicted Plotting (2)

The fig 4.3 shows how the actual values and predicted values are plotted in auction dataset

The mean squared error and plotting evaluated by actual and predicted in the Auction dataset.

```
from sklearn.tree import DecisionTreeRegressor
model = DecisionTreeRegressor()
model.fit(x_train, y_train)
from sklearn.metrics import mean_squared_error
y_pred = model.predict(x_test)
print(mean_squared_error(y_test, y_pred))
```

```
0.2727272727272727
```

Fig 4.4: Mean squared error for Season Stats dataset

The fig 4.1 shows the mean squared error which is calculated from auction dataset from Decision Tree Regressor.

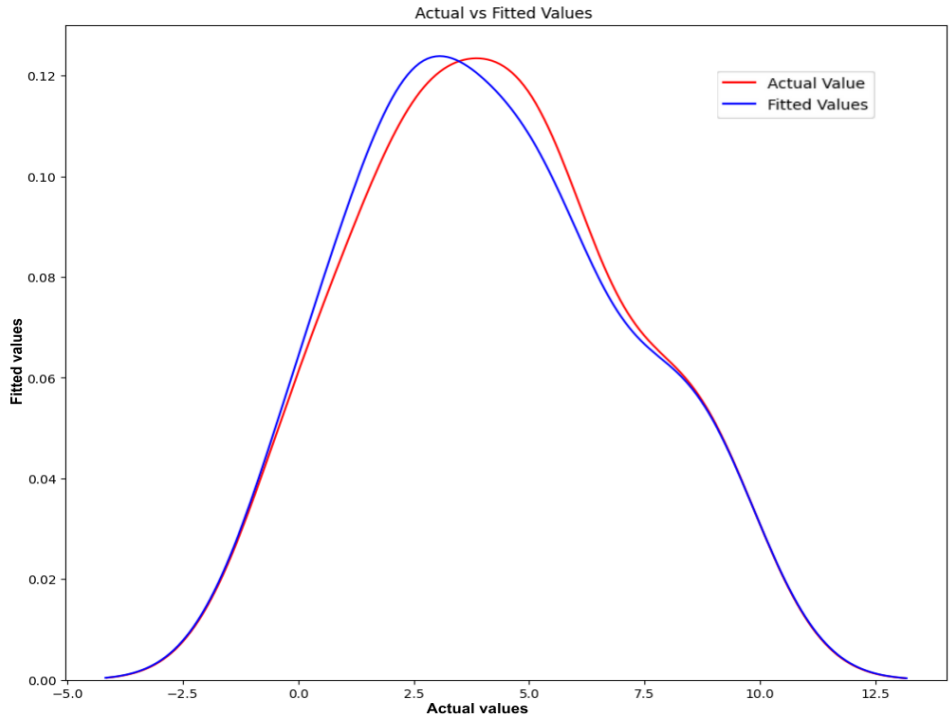


Fig 4.5: Season Stats dataset: Actual VS Predicted Plotting (1)

The fig 4.5 shows how the actual values and predicted values are plotted in stats dataset

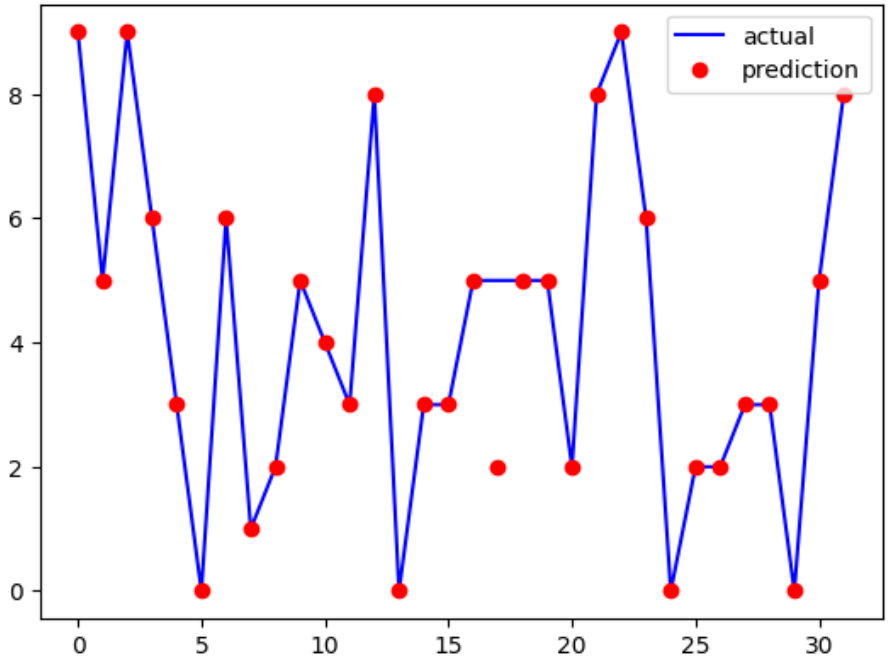


Fig 4.6: Season stats dataset: Actual VS Predicted Plotting (2)

The fig 4.2 shows how the actual values and predicted values are plotted in auction dataset. The mean squared error and plotting evaluated by actual and predicted in the Season stats dataset.

4.2 Confusion Matrix

The confusion matrix is a matrix that is used to assess the performance of classification models for a given set of test data. It can be determined only if the true values of the test data are known. The matrix itself is simple to grasp, but the associated terminologies can be perplexing. Because it displays mistakes in model performance as a matrix, it is also known as an error matrix. In the study have 10 labels so, the matrix in the 10×10 form.

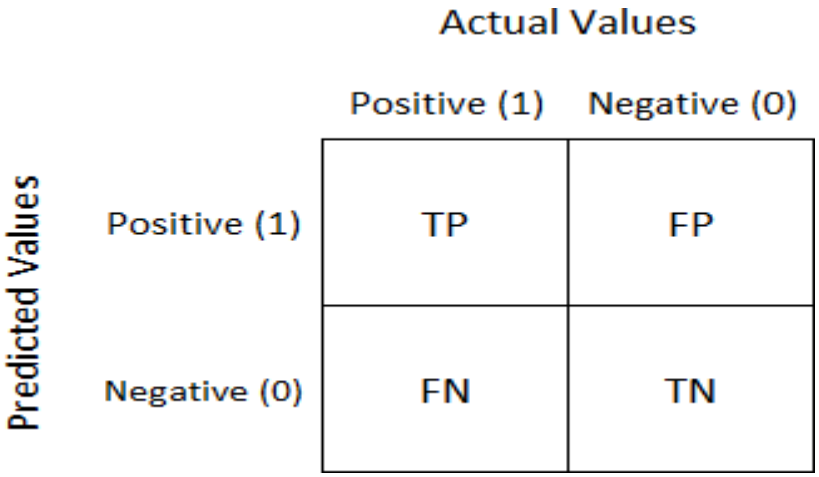


Fig 4.7: Confusion Matrix

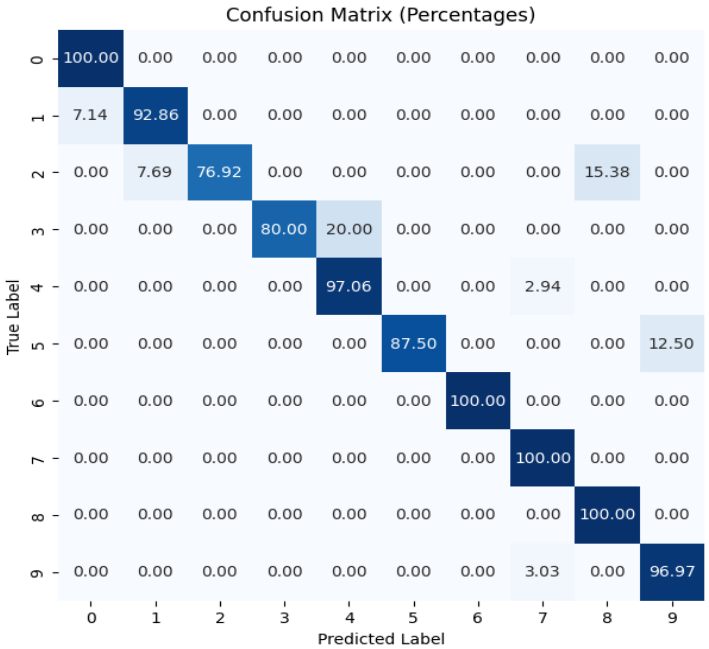


Fig 4.8: Confusion Matrix for Auction dataset

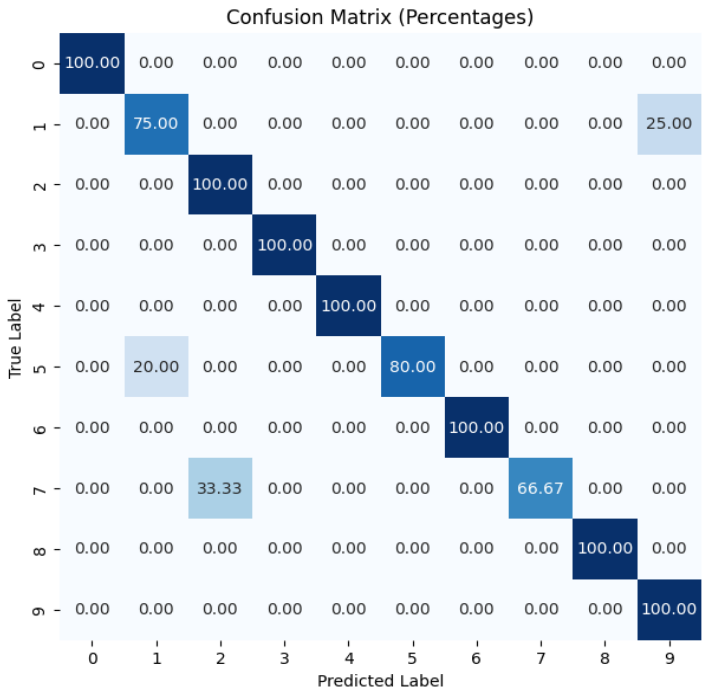


Fig 4.9: Confusion Matrix for Season stats dataset

The fig 4.8 and 4.9 show confusion matrix how the true labels are predicted as true.

4.3 Model accuracy

```
accuracy = accuracy_score(y_test, y_pred)
print(f"{accuracy * 100:.2f}%")
```

```
95.27%
```

Fig 4.10: Accuracy for cluster validation in Auction dataset

```
accuracy = accuracy_score(y_test, y_pred)
print(f"{accuracy * 100:.2f}%")
```

```
96.97%
```

Fig 4.11: Accuracy for cluster validation Season stats dataset

The Decision tree model obtained 95% accuracy for the auction-based dataset and obtained 96% accuracy for Season stats-based dataset.

4.4 Standardization of Clusters

The standardization of clusters was evaluated on the basis of Rating which was evaluated from the mathematical model. The 10 clusters were standardized into 10 standards. Mean of Rating is taken as primary consideration and Mean of Base price is considered as secondary. On the basis of standard set price for each player are done by mapping in Auction dataset.

Cluster Standardization for Auction Dataset		
Cluster	**Mean Rating**	**Mean Base Price**
0	14.608712	28.421053
1	22.963482	74.907407
2	37.202775	88.617021
3	28.795325	72.222222
4	14.604351	21.176471
5	10.779360	20.000000
6	51.168042	142.209302
7	58.330401	155.833333
8	31.687965	32.692308
9	17.185609	21.264368

Table 4.1: Cluster Standardization for Auction Dataset

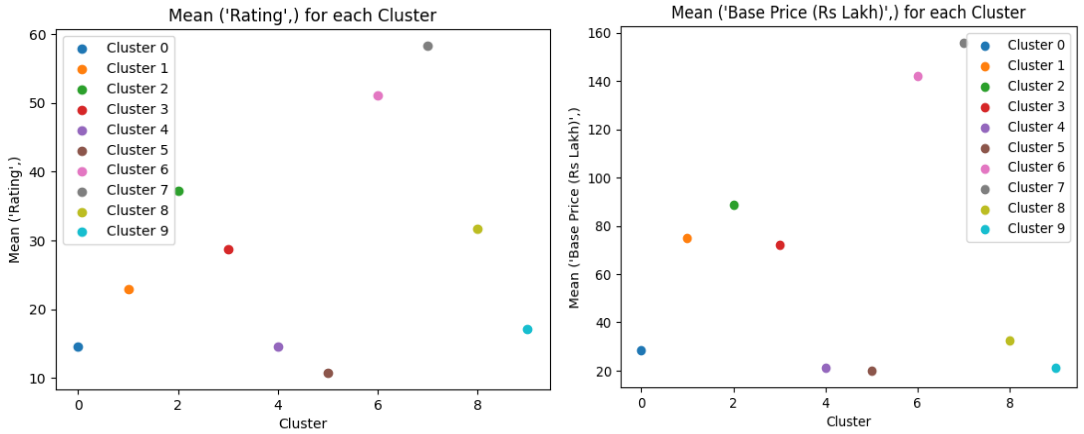


Fig 4.12: Auction dataset Standardization in case of Mean Rating and Mean Base Price

Cluster Standardization for Stats Dataset		
Cluster	**Mean Rating**	**Mean Base Price**
0	7.283501	46.333333
1	39.237750	77.105263
2	67.331450	141.666667
3	37.039387	41.250000
4	48.235682	65.000000
5	62.075786	144.333333
6	25.388785	85.714286
7	52.585088	111.071429
8	62.101384	150.000000
9	24.486389	70.833333

Table 4.2: Cluster Standardization for Stats Dataset

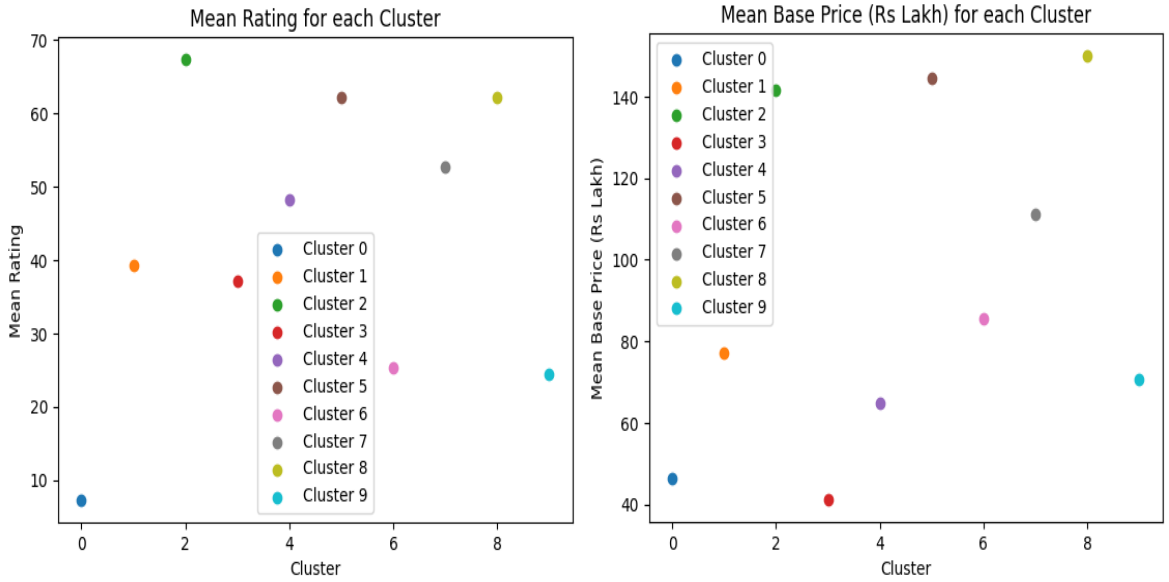


Fig 4.8: Stats dataset Standardization in case of Mean Rating and Mean Base Price.

Price Mapping According to standard		
Standard	Starting Price	Maximum Price
A	1200	1500
B	1000	1200
C	800	1000
D	650	800
E	500	650
F	350	500
G	200	350
H	100	200
I	50	100
J	20	50

Table 4.3: Price Mapping According to Standard

The price mapping is done in the range for each player. For each player there is starting price and the maximum price for the player.

4.5 Season Performance VS Performance Standard by the model

```
Equal_Performance = (df3['price range'] == df3['cluster']).sum()

print("Equally Performed Players:", Equal_Performance)
```

```
Equally Performed Players: 29
```

Fig 4.11: Equal Performed Players according to standard

The comparison of player season performance Actual performance by our model is calculated by equality performance. In the study got equally performed 30 players according standard. As stands player performance of the player depends many things such as player selection in the playing eleven, Price of the player, Match pitch variation, team mentalities.

4.6 Comparison of Current Model and Previous Model

The current model was pitted against the prior study, revealing its superior accuracy compared to the earlier research.

Mean Difference in Actual price and Predicted price Previous study			
	Batters	Bowlers	All-rounders
Average	14.4%	21.3%	38.4%

Table 4.4: Mean Difference in Actual price and Predicted price Previous study

Mean Difference in Actual price and Predicted price Current study			
	Total Sold Price (Lakh)	Difference From Sold Price	Mean Difference (%)
All-rounders	15675	2130	13.588517%
Batters	8205	80	0.975015%
Bowlers	10385	420	4.044295%
Wicket-Keepers	5680	25	0.440141%

Table 4.5: Mean Difference in Actual price and Predicted price Current study

Mean Difference is calculated by taking the difference between starting price and Maximum price considered as zero as it is beneficial for the franchise to select players between the range. The difference calculated by sold price is more than price predicted from the model.

4.7 Players Selection

The model assists franchises to select the best player for their squad by their role to select players within their budget. The player selected by budget allotted to the player, role of the player, batting style and bowling style. The model selects best 5 players for the requirement of franchises.

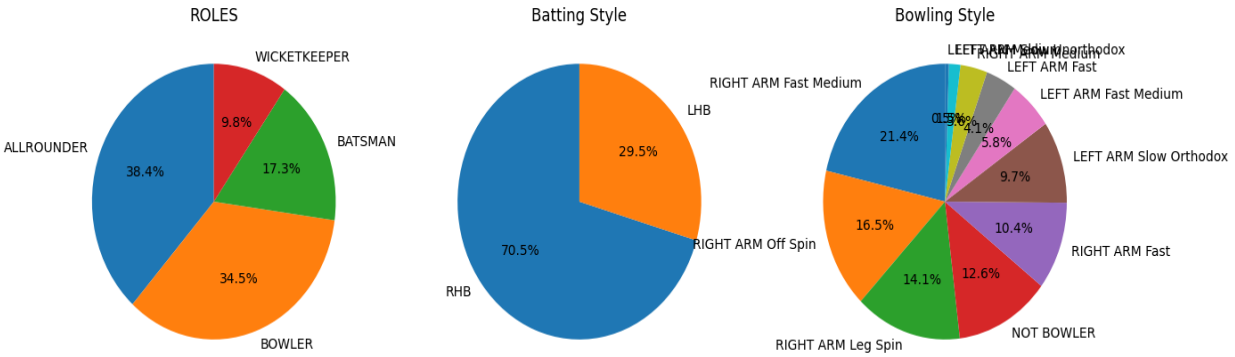


Fig 4.12: Players Selection Criteria

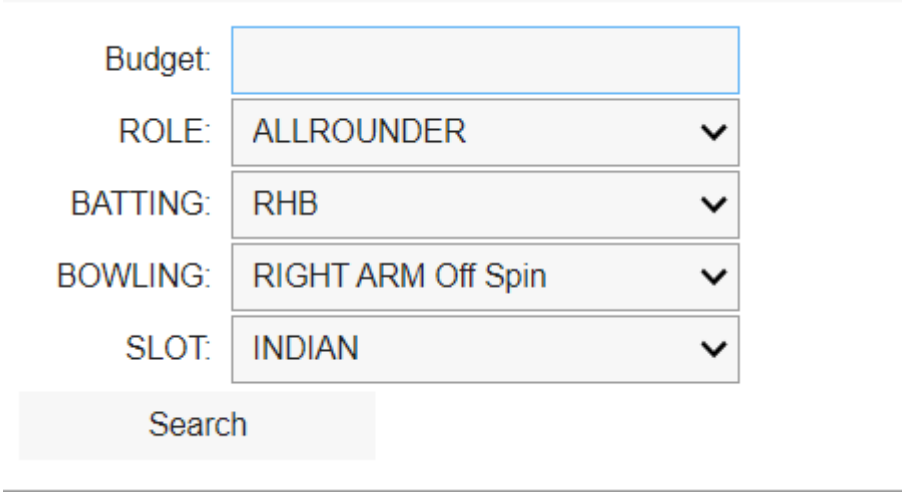
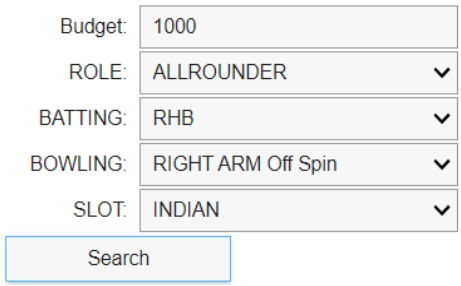


Fig 4.13: Players Selection Widget



Name	SLOT	Age	ROLE	BATTING	BOWLING	Base Price (Rs Lakh)	Maximum price
Lalit Yadav	INDIAN	25.0	ALLROUNDER	RHB	RIGHT ARM Off Spin	20.0	800
Riyan Parag	INDIAN	20.0	ALLROUNDER	RHB	RIGHT ARM Off Spin	30.0	800
K. Gowtham	INDIAN	33.0	ALLROUNDER	RHB	RIGHT ARM Off Spin	50.0	1000
Jayant Yadav	INDIAN	32.0	ALLROUNDER	RHB	RIGHT ARM Off Spin	100.0	800
Parvez Rasool	INDIAN	33.0	ALLROUNDER	RHB	RIGHT ARM Off Spin	50.0	1000

Fig 4.14: Players Selected for certain selections

Fig 4.13 shows the user interface as widget with selection criteria's and fig 4.14 shows the player selected for certain criteria in the interface.

5. CONCLUSION AND FUTURE WORK

The IPL auction price prediction can be used for several T20 franchise-based tournaments. In study model combinations are attempted found that K-means and Decision Tree Regressor as best. This study used K-means clustering and Decision tree regression for price prediction from created datasets. As result, the proposed model predicted the price for the player more accurately than traditional models.

5.1 Future Enhancement

The model can be used to predict the price of the players in other franchise-based tournaments. Nowadays franchise-based tournaments grow around the world. The players selection can be connected to websites which is more precise for the selection.

References

- Gaurav Malhotra,15 September 2022, A comprehensive approach to predict auction prices and economic value creation of cricketers in the Indian Premier League (IPL), Journal of Sports Analytics 8 (2022) 149–170.

- Dr. Jhansi Rani Apurva Kulkarni, Aditya Vidyadhar Kamath, November 04,2020, Prediction Of Player Price In IPL Auction Using Machine Learning Regression Algorithms.

- Pabitra Kr. Dey, Abhijit Banerjee, Dipendra Nath Ghosh, Abhoy Chand Mondal,2014, AHP-Neural Network Based Player Price Estimation in IPL, International Journal of Hybrid Information Technology, Vol.7, No.3 (2014), pp.15-24.

- S. K. Rastogi and Satish Y. Deodhar," Player Pricing and Valuation of Cricketing Attributes: Exploring the IPL Twenty-Twenty Vision", in W.P. No. 2009-01-02, Indian Institute of Management, Ahmedabad, pp. 1-20. International Journal of Hybrid Information Technology Vol.7, No.3 (2014).

- Ahmed,F,Deb,K.,&Jindal,A.,2013.Multi-objective optimization and decision making approaches to cricket team selection.

- Prakash, C. D, & Verma ,S ,2022.Anewin-form and role-based deep player performance index for player evaluation in T20 cricket. Lytics Journal .Retrieved from https://www.sciencedirect.com/science/article/pii/S2772

662222000029 [Accessed 30th April 2022]

- Kristina P. Sinaga and Miin-Shen Yang, May 13 2020, Unsupervised K-Means Clustering Algorithm, IEEE Access, Ministry of Science and Technology, Taiwan, under Grant MOST 107-2118-M-033-002-MY2.

- Harsh H. Patel , Purvi Prajapati, October 31 2018, Study and Analysis of Decision Tree Based Classification Algorithms, International Journal of Computer Sciences and Engineering, Vol.-6, Issue-10, Oct. 2018.

PLAGIARISM REPORT

Ouriginal
by Turnitin

Document Information

Analyzed document	Project document OG First copy.pdf (D172723987)
Submitted	2023-08-11 03:18:00
Submitted by	Vinod Chandra S S
Submitter email	vinod@keralauniversity.ac.in
Similarity	1%
Analysis address	vinod.kerala@analysis.urkund.com