Aston University
BIRMINGHAM UK

A Cricketer's Worth: Predicting the Bid Prices of Players in IPL Auction

Candidate Number: 838651

Student Name: Reddy Kamakoti (220390110)

Supervisor: Dr. Anandadeep Mandal

Module Code: BDM200J - MSc Business Analytics Research Project

Submission Date: January 2025

Program: Master of Science (MSc) in Business Analytics

Submission Statement: This project is submitted in partial fulfilment of the requirements for the master's in Business Analytics at Aston University, dated January 27th, 2025.

DECLARATION

I hereby affirm that this research and report constitute my original work and have not been submitted for any other degree or qualification. The content has not been published or disseminated elsewhere unless expressly stated. All sources of information, including quotations, are appropriately cited within the text and the reference section.

ACKNOWLEDGMENT

I express my heartfelt appreciation to all who provided support during my M. Sc. journey. I am profoundly thankful to my family for their steadfast encouragement and motivation, and to my supervisor, Mr. Anandadeep Mandal, for his indispensable guidance and feedback throughout this research endeavour. I am also appreciative of the camaraderie provided by my friends and classmates, whose support made this experience both fulfilling and unforgettable. This program has been transformative, and I will carry forward the knowledge and skills acquired with sincere gratitude.

ABSTRACT

The Indian Premier League (IPL) represents a dynamic and highly competitive environment where player auction prices play a pivotal role in shaping team composition, financial strategies, and overall league dynamics. Accurately predicting player prices has emerged as a critical challenge, requiring the integration of performance metrics, market trends, and contextual factors into sophisticated predictive frameworks. This work focuses on the application of machine learning algorithms to predict the selling price of players in the Indian Premier League (IPL) auction, leveraging historical performance metrics such as runs, balls, innings, wickets, and matches played. The study evaluates various predictive models, including Artificial Neural Networks (ANN), Linear Regression, Random Forest Regressor, Decision Tree Regressor, and K-Nearest Neighbors (KNN). Among these, ANN, Linear Regression, and Random Forest emerge as the most effective models, accurately capturing the complexities of batsmen, bowlers, and all-rounders' pricing dynamics. The models deliver fast and precise predictions within 3 seconds, empowering auctioneers with actionable insights to make swift and informed decisions.

To enhance prediction accuracy, this research integrates innovative adjustments such as incorporating inflation factors and budget mapping during model training, ensuring predictions remain aligned with evolving market conditions and franchise financial constraints. Unlike prior studies that focused on specific algorithms or limited player categories, this work adopts a comprehensive approach, combining robust models with advanced preprocessing techniques to better understand the multidimensional nature of player valuations.

By comparing the strengths and limitations of these models, the study highlights the transformative potential of machine learning in optimizing IPL auctions. This research not only aids teams in strategic planning and resource allocation but also contributes to a broader understanding of data-driven decision-making in sports management, paving the way for advancements in predictive analytics across global sports ecosystems.

Key Words: Machine Learning, Python, Previous Performances, Dynamic pricing, Data visualization, Team Strategy, Prediction, Comparative analysis, Model development, Random Forest, K-Nearest Neighbors, Artificial Neural Networks, Time series forecasting, Support Vector Machines.

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1. INTRODUCTION

a. Background

The Indian Premier League (IPL), since its inception in 2008, has transformed cricket by fusing entertainment, athletics, and business, evolving into a multi-billion-dollar global enterprise. Central to its success is its innovative business model, which aligns team performance with commercial objectives through the annual player auction. This auction shapes the league's competitive dynamics, as franchise owners bid for players based on performance, marketability, and team needs, assigning them a monetary value. However, as the league's competitiveness has grown, evaluating player value has become increasingly sophisticated, shifting from traditional statistics to advanced metrics and data-driven insights.

Modern player performance indices, as highlighted by Deep Prakash and Verma (2022), have evolved to incorporate role-specific metrics, in-form indicators, and situational contributions, such as a bowler's death-over performance or a batsman's pressure-strike rate, moving beyond raw statistics like economy rate or batting average. Similarly, Bhandari, Sinha, and Vaidya (2018) emphasize that inefficiencies in the auction process, often stemming from subjective biases and incomplete information, highlight the need for predictive frameworks that leverage historical data and advanced machine learning techniques to refine player valuation.

The application of predictive analytics in sports, particularly in the IPL, has emerged as a transformative force. According to Breiman and Schapire (2001), machine learning excels in uncovering complex, nonlinear relationships within datasets, making it particularly suited to auction price predictions, where numerous interacting factors influence outcomes. Akiyanova (2020) and Biau and Scornet (2016) further underscore how data-driven models can uncover hidden patterns and reduce subjectivity, enabling teams to optimize resource allocation and squad balance.

As player roles in the IPL diversify into specialized categories like power-hitters, finishers, and death-over specialists, evaluating performance contextually becomes crucial. This aligns with Deep, Patvardhan, and Vasantha's (2016) assertion that integrating analytics into decision-making enhances objectivity, identifies undervalued talent, and supports competitive sustainability. By combining advanced machine learning techniques with human judgment, IPL franchises can not only enhance efficiency but also adapt to the league's dynamic and competitive landscape.

The future of IPL auctions lies in further integrating predictive modeling and advanced analytics with qualitative insights. This approach minimizes biases, refines player valuations, and ensures financial sustainability, contributing to the league's continuous growth and evolution.

b. Prediction Business Model

The prediction business model for the IPL auction utilizes advanced data analytics and machine learning techniques to forecast player prices, addressing inefficiencies such as overvaluation, undervaluation, and subjective biases. By analyzing historical data, performance metrics, and auction dynamics, this model delivers actionable insights to guide franchise decisions. Key features include batting and bowling statistics such as strike rates, economy rates, and match-winning contributions, as well as contextual factors like team requirements, recent form, and injuries. Auction-specific dynamics, including base prices, demand for certain player types, and team budgets, are also incorporated to ensure real-world relevance.

The IPL auction's predictive business model employs data analytics and machine learning to address inefficiencies like overvaluation and biases (Bhandari, Sinha, & Vaidya, 2018). It analyzes historical data, performance metrics, and auction dynamics, incorporating batting and bowling statistics, recent form, injuries, and team requirements. Algorithms like SVR and Random Forests effectively handle nonlinear relationships (Breiman, 2001; Bhatt et al., 2017), while hyperparameter tuning ensures accuracy (Hunter, 2019). The model accounts for inflation and budget constraints, simulating scenarios to identify undervalued players (Rani et al., 2020). Integrated dashboards provide actionable insights, enhancing resource allocation and team performance in a data-driven IPL ecosystem.

c. IPL Players' Auction Strategy

The IPL auction operates in a structured yet dynamic environment. It begins with marquee players, followed by capped and uncapped players, across categories such as batsmen, bowlers, all-rounders, and wicketkeepers. Teams must navigate constraints like budget caps, retention policies, and the maximum number of foreign players allowed in a squad.

Market dynamics play a pivotal role in influencing player prices. Studies have shown that international players often command premium prices due to their limited availability and global appeal, while Indian players remain indispensable due to roster requirements (Staden, 2009; IEEE CONECCT, 2020). Additionally, a player's versatility—such as the ability to contribute as an all-rounder—often elevates their value significantly (Bradbury, 2007). Predictive models account for these nuances by incorporating variables such as player roles, team composition needs, and regional biases into their frameworks, thereby offering franchises a comprehensive toolkit for strategic decision-making.

d. Importance of Pricing Based on Previous Performances

A player's past performance serves as the cornerstone of their valuation. Metrics such as runs scored, wickets taken, strike rate, and bowling economy are directly correlated with their auction prices (Kimber, 1993; Lemmer, 2011). Bhatt et al. (2017) identified recent form and match-winning

performances as key predictors, while Geman et al. (1992) demonstrated the efficacy of neural networks in capturing non-linear relationships between these metrics and player prices.

The study added a practical dimension by introducing adjustments for inflation and currency normalization, ensuring that historical data remains relevant in modern contexts. These considerations not only improve prediction accuracy but also align valuations with current economic conditions, addressing gaps in traditional methods.

e. Rationale for Price Prediction

Price prediction is crucial for IPL franchises aiming to maximize the value of their investments. Human decision-making during auctions is often influenced by biases, such as favoring popular players or overpaying due to competitive bidding. Predictive analytics mitigates these challenges by providing objective, data-driven insights into player valuations.

For example, Hagan (2020) highlighted how predictive models can minimize errors caused by subjective biases, ensuring that team budgets are allocated efficiently. Similarly, Deodhar and Sinha (2009) emphasized the importance of incorporating market dynamics, such as demand for specific roles, into valuation models. By integrating these factors, predictive frameworks empower teams to make informed decisions that align with their strategic goals and financial constraints.

f. Why Machine Learning?

Machine learning offers unparalleled advantages in analyzing large datasets and uncovering complex patterns. Techniques such as regression models, decision trees, and neural networks are particularly effective in predicting IPL auction prices. James et al. (2013) highlighted the ability of these methods to handle non-linear relationships, while Lyle (2019) demonstrated their scalability and robustness in sports analytics.

The study showcased how models like SVR and Linear Regression excel in specific scenarios, with SVR performing well for batsmen and Linear Regression proving effective for bowlers. These models not only deliver accurate predictions but also adapt to real-time constraints, making them invaluable during live auctions. Moreover, neural networks provide a powerful tool for capturing the intricate interplay between player performance metrics, team needs, and market trends (Geman et al., 1992).

g. Aim, Research Questions, and Impact of Research

i. Aim of the Research

The primary aim of this research is to develop a robust and comprehensive machine learning framework tailored to predict player prices in the Indian Premier League (IPL) auctions. This framework will integrate historical performance metrics, team-specific requirements, and auction dynamics to provide franchises with accurate and actionable insights into player valuations. By

addressing inefficiencies and subjective biases in current valuation methods, the research aims to create a data-driven tool that enhances strategic decision-making, fosters resource optimization, and improves competitive balance among teams. Furthermore, this study seeks to set a benchmark in applying predictive analytics to auction-based player selection processes, aligning the predictions with real-world trends such as evolving market demands, budget constraints, and inflation-adjusted pricing.

ii. Research Questions

1. What performance metrics and market factors most influence IPL player prices?
2. What are the strengths and limitations of different machine learning algorithms for predicting IPL auction outcomes?
3. How can predictive models be optimized to account for differences between national and international players?

iii. Impact of Research

The impact of this research extends beyond the immediate context of IPL auctions, providing long-term value to franchises and the sports analytics community. By offering a data-driven framework, franchises can reduce reliance on intuition and bias, ensuring more rational investment decisions and optimal team compositions. The ability to predict player prices accurately can mitigate risks associated with overvalued or undervalued players, ultimately contributing to a more competitive and financially sustainable league.

Moreover, the methodologies and models developed in this study can be adapted to other sports leagues with auction-based player selections, such as the Big Bash League (BBL) or the Pro Kabaddi League. Beyond sports, the findings could influence similar domains where market valuation plays a pivotal role, such as talent acquisition in corporate settings or entertainment industry contracts. In the broader field of sports analytics, this research paves the way for integrating machine learning and predictive modeling into decision-making processes, creating a scalable blueprint for leveraging data to solve real-world problems.

h. Structure of the Research

The research is structured as follows:

- **Section 2:** A review of the literature on sports analytics, player valuation, and machine learning applications.
- **Section 3:** A detailed methodology, including data collection, feature selection, and model development.

- **Section 4:** Results and discussions, highlighting the performance of different models and their real-world implications.
- **Section 5:** Conclusions and recommendations for future research directions.

CHAPTER 2:

LITERATURE REVIEW

2. LITERATURE REVIEW

a. Players' Price Determinants and Auction Valuation

The determinants of player prices in sports auctions, particularly in cricket, have been a focal point of extensive research. Ahmed et al. (2013) examined multi-objective optimization for cricket team selection, emphasizing the balance between performance metrics and team composition but lacked modern machine learning (ML) integration for enhanced predictive accuracy. Bhatt et al. (2017) identified critical factors like recent performance and player popularity influencing IPL auction prices, providing insights into auction dynamics but not employing predictive modeling techniques to forecast prices. Similarly, Karnik (2010) utilized hedonic pricing models to economically evaluate cricketers, offering a strong theoretical base but falling short in applying ML methods capable of processing large datasets and capturing non-linear interactions.

These studies provide foundational insights but often overlook the dynamic, role-specific metrics and the potential of ensemble ML algorithms like Random Forests or Gradient Boosting to predict prices more accurately. Addressing this gap necessitates advanced methods combining statistical models and ML techniques for improved prediction and decision-making in sports auctions.

b. Adoption of Machine Learning for Performance Prediction

The application of machine learning (ML) techniques for predicting player auction prices has gained significant momentum in recent years. Bhandari et al. (2018) demonstrated the use of regression and random forest algorithms to forecast IPL auction prices, highlighting improvements in accuracy compared to traditional methods. However, their models lacked adaptability to real-time auction dynamics, limiting their practical application. Similarly, Lyle (2019) emphasized neural networks' ability to handle complex datasets in sports analytics but did not focus on auction-specific prediction tasks or price determinants. Malhotra (2022) combined economic and performance metrics with ML for price prediction, presenting a robust framework, though without exploring advanced ensemble methods like gradient boosting for further accuracy enhancements.

c. Factors Influencing Player Auction Prices

Several studies have explored ML algorithms for predicting player prices, with varying degrees of success. Bhandari et al. (2018) applied machine learning techniques to predict IPL auction prices, demonstrating the importance of integrating player performance and market factors. However, their study lacked a detailed focus on auction-specific dynamics, such as team budget constraints and bidding trends. Deep Prakash and Verma (2022) introduced a role-based performance index for evaluating T20 cricket players, highlighting the importance of situational contributions like pressure-

strike rates and death-over performance. Despite its novelty, their work did not incorporate essential auction-specific features like real-time bidding strategies or budget constraints.

Breiman and Schapire (2001) underscored the effectiveness of Random Forests in handling high-dimensional data and capturing non-linear relationships, making it a suitable algorithm for tasks like price prediction. However, its application to auction-based player valuation remains limited. Similarly, Davis et al. (2015) explored player evaluation in T20 cricket using statistical methods, offering a foundation for predictive modeling but not fully addressing the complexities of auction systems.

The application of ML for player price prediction in cricket auctions is an emerging field with significant potential. Karnik (2009) utilized hedonic price models for valuing cricketers but did not leverage advanced ML algorithms capable of real-time adaptability. Similarly, Rastogi and Deodhar (2009) explored cricketing attributes for player valuation but fell short of integrating predictive ML methodologies tailored to auction scenarios. Biau and Scornet (2016) highlighted the robustness of ensemble methods like Random Forests for prediction tasks, yet their use in cricket player pricing is underexplored.

d. Advantages of Machine Learning in Sports Analytics

The adoption of machine learning for player price determination has been extensively explored, revealing varying levels of performance across models. Studies like Bhandari, Sinha, and Vaidya (2018) identified Gradient Boosting as the most accurate model, achieving an RMSE of 58.43 and an R^2 score of 0.65, outperforming Random Forest and Multiple Linear Regression (MLR). Similarly, Bhatt, Muthukumar, and Varadarajan (2017) demonstrated the superiority of XGBoost, which attained an R^2 score of 0.67 and an RMSE of 56.84, surpassing Support Vector Machines (SVMs) and Decision Trees in predictive power. Malhotra (2022) further highlighted the potential of deep learning models, which achieved the best performance with an R^2 score of 0.69 and an RMSE of 55.12, slightly outperforming Random Forest and Bayesian Linear Regression, though at a higher computational cost. Gradient Boosting also consistently delivered strong results in other studies, such as Singh, Gupta, and Gupta (2010), with an RMSE of 59.21 and an R^2 score of 0.63. Across the literature, ensemble models like Gradient Boosting and XGBoost outperformed traditional regression methods, including Ridge and LASSO regression, which provided only minor improvements over MLR. While deep learning demonstrated the highest accuracy, its complexity and interpretability remain challenges. Overall, XGBoost and Gradient Boosting emerged as the most effective models for predicting player prices, with deep learning achieving the best metrics of RMSE 55.12 and R^2 0.69 in specific cases.

Study/Author	Techniques Used	RMSE	R^2	Best Performing Model
Bhandari, Sinha, & Vaidya (2018)	Gradient Boosting, Random Forest, MLR	58.43	0.65	Gradient Boosting
Bhatt, Muthukumar, & Varadarajan (2017)	XGBoost, SVM, Decision Trees	56.84	0.67	XGBoost
Malhotra (2022)	Deep Learning, Random Forest, Bayesian	55.12	0.69	Deep Learning
Singh, Gupta, & Gupta (2010)	Gradient Boosting, MLR, Ridge	59.21	0.63	Gradient Boosting

Table 1: Performance Comparison and Observations

This table compares studies on predictive models for IPL auction price forecasting. The listed techniques include Gradient Boosting, XGBoost, Deep Learning, and others. RMSE and R2 values indicate the accuracy and reliability of each model. The best-performing model for each study is highlighted, showcasing the evolving preference for Gradient Boosting and Deep Learning due to their superior predictive power in specific contexts.

e. Applications of Machine Learning in Predicting Player Prices

ML applications in player price prediction have gained traction in recent years. Bhandari, Sinha, and Vaidya (2018) employed ensemble methods, such as gradient boosting, to predict IPL auction prices and substantially reduce pricing errors. Deep et al. (2016) introduced a machine learning-based performance index, which ranked IPL players based on a comprehensive set of metrics, including batting and bowling averages, strike rates, and situational performance (e.g., powerplay versus death overs).

Malhotra (2022) extended these efforts by incorporating a wider range of predictors, including player age, team roles, and international exposure, into the modeling process. The study's ML-driven framework offered role-specific valuation equations, enabling teams to make more informed decisions about player acquisitions. By aligning auction prices with predicted performance, the study minimized overpayments and underpayments, resulting in more efficient budget allocations.

AUTHORS	YEAR	ORIGIN	PURPOSE	TYPE OF SOURCE	METHODOLOGY	FINDINGS	LIMITATIONS
Ahmed, F., Deb, K., Jindal, A.	2013	India	Cricket team selection optimization	Research Article	Multi-objective optimization	Framework for team selection	Limited to specific criteria
Deep, C., Patvardhan, C., Singh, S.	2016	India	ML-based performance index for IPL cricketers	Research Article	Machine learning, ranking algorithm	Novel performance index for player ranking	Limited to IPL; not generalizable to other leagues
Davis, J., Perera, H., Swartz, T.	2015	Canada	Player evaluation in T20 cricket	Research Article	Statistical modeling	Models for evaluating T20 player performances	Models may miss dynamic game contexts
Depken, C.A., Rajasekhar, R.	2010	N/A	IPL player performance market valuation	Research Article	Hedonic pricing model	Market valuation model for player performance	No performance variability consideration
Bhatt, A., Muthukumar, A., Varadarajan, P.	2017	India	Determinants of IPL player auction prices	Research Article	Statistical modeling	Factors influencing player auction prices in IPL	Focused on economic, not gameplay performance
Chittibabu, V., Sundararaman, M.	2023	India	Base price determination for IPL mega auctions	Research Article	Performance-based modeling	Player performance-based auction price determination	Limited to IPL players
Wasim, D., Suhail, M., et al.	2025	Global	Regression model for T20 performance analysis	Research Article	Robust regression model	Handles multicollinearity and outliers in cricket data	Focused only on regression techniques
Malhotra, G.	2022	India	Predict auction prices in IPL	Research Article	Economic modeling	Comprehensive model for player value creation	Focused only on economic aspects
Rani, P.J., Kulkarni, A.V., et al.	2020	India	Predict IPL player price using ML algorithms	Research Article	Regression algorithms	Accurate prediction of auction prices	No focus on in-game performance
Deep Prakash, C., Verma, S.	2022	India	Role-based player evaluation in T20 cricket	Research Article	Machine learning	New role-based performance index for players	Limited dataset
Hemanta Saikia, et al.	2019	India	Cricket performance management	Research Book	Mathematical formulation	Analytics-based performance improvement	General overview; no specific auction analysis
Jayanth, S.B., et al.	2018	India	Team recommendation system for cricket	Research Article	Team recommendation modeling	Model for team recommendation and outcome prediction	Limited to specific team combinations
Kapadia, K., et al.	2020	Global	ML for cricket game results prediction	Research Article	Experimental study	Machine learning enhances result prediction	Not focused on auctions
Chouhan, N.	2020	India	Operations research in IPL player valuation	Research Article	Optimization modeling	Framework for optimal player valuation	No player performance considerations
Eapen, J.	2021	N/A	IPL Season 13 popularity comparison	Market Analysis Article	Survey and statistical analysis	IPL Season 13 maintained high popularity	Focused on popularity, not deep performance insights
Hunter, T.B.	2019	Global	Hyperparameter tuning in Python	Online Blog	Optimization techniques	Importance of tuning in predictive models	General overview, not sports-specific
Zhang, Y., Yang, Y.	2015	Global	Cross-validation for model selection	Research Article	Statistical principles	Improved predictive accuracy with model selection	Not specific to cricket or sports

Table 2: Findings of Reviewed Literature

This table provides an overview of various studies and sources related to sports analytics, particularly in cricket and IPL. It highlights the authors, origin, purpose, methodology, and key findings. While the studies address diverse topics such as team optimization, market valuation, and predictive modeling, many are limited by their scope, such as focusing solely on IPL or lacking generalizability to other contexts. This compilation offers a foundation for understanding the current trends and gaps in sports analytics research.

f. Literature Gaps and Future Directions

Literature Gaps

Despite advancements in sports analytics, critical gaps remain in understanding player auction valuation in cricket leagues like the IPL. A significant gap lies in integrating qualitative factors such as leadership qualities, team chemistry, and psychological resilience. While models predominantly focus on quantitative metrics like batting averages and strike rates, non-quantifiable factors are often overlooked. Studies like Davis, Perera, and Swartz (n.d.) suggest that these factors significantly influence player value. Iconic players like MS Dhoni and Virat Kohli command higher auction prices due to their leadership skills and marketability, yet empirical analysis of their impact is limited.

Another gap is the lack of exploration into performance tiers, which categorize players based on their consistency and ability to perform under varying levels of competition. Current models fail to capture nuanced differences in performance across contexts, such as high-pressure situations or critical tournament phases. Malhotra (2022) highlights inefficiencies in auction pricing, where players' high auction values do not always correlate with on-field performance. Introducing performance tiers could help predict auction prices more accurately by accounting for consistent delivery in key moments.

The impact of age on player valuation is another underexplored area. While studies like Malhotra (2022) mention age, they rarely analyze its nuanced effects, such as performance decline, injury risks, and longevity. IPL teams often prioritize younger players for their potential, overshadowing long-term value considerations. Further research is needed to model age-adjusted valuations, factoring in fitness and performance trends over time.

Specialized player roles, such as powerplay batting and death-over bowling, also lack adequate representation in auction valuation models. Traditional models treat these roles as generalized attributes, undervaluing players who excel in critical roles. Similarly, factors like country of origin and international experience remain underexplored. Players from smaller cricketing nations or with limited international exposure often face undervaluation despite strong domestic performances. A

deeper analysis of how country-based factors influence auction pricing, including media exposure, would enhance auction strategies.

Moreover, research on value-per-price metrics, comparing a player's auction price to their actual performance, is insufficient. Studies like Karnik (2010) have explored hedonic pricing models but lack comprehensive frameworks for evaluating long-term alignment between auction price and performance. Developing such metrics could refine the predictive accuracy of auction pricing.

Future Directions

Future research should address these gaps to develop more comprehensive player valuation models. Integrating behavioral and psychological factors into predictive models is essential. Understanding how leadership, psychological resilience, and team dynamics influence player value can improve auction predictions. Empirical studies examining leadership's role in shaping team morale and performance, particularly for iconic players, would add depth to valuation models.

Another promising direction is developing performance classification models that incorporate performance tiers. Categorizing players based on their consistency and ability to perform in high-pressure situations, such as finals, could improve auction price predictions. Machine learning techniques could identify players who deliver consistently during critical moments, providing teams with valuable insights.

Age-adjusted valuation models should also be developed to account for age-related changes in performance, fitness, and injury risks over time. Combining historical performance data with projections of longevity and injury risks would enable more accurate valuations, helping teams make informed long-term decisions.

Specialized skill sets, particularly in critical match situations, should be treated as distinct variables in auction pricing models. Players excelling in roles like death-over bowling or powerplay batting should be valued appropriately. Understanding the specific impact of these roles on team dynamics and match outcomes would refine auction strategies.

The influence of international and domestic exposure on auction pricing warrants deeper examination. Analyzing how international reputation, media exposure, and performance in various leagues affect valuation could identify biases, particularly for players from smaller cricketing nations who may be undervalued despite strong performances.

Developing a value-per-price indicator that combines historical performance data with auction prices could improve auction efficiency. Quantitative models assessing the alignment between a player's auction price and their actual performance would help teams optimize bidding strategies.

Incorporating game theory and auction strategy dynamics into predictive models could simulate the evolving nature of live bidding. Hybrid models combining behavioral economics, machine learning, and game theory could reflect the strategic decisions teams make during auctions. Additionally, real-time performance data from wearable technology and IoT devices could enhance player evaluations by providing up-to-the-minute insights into fitness and performance during key moments. Combining this data with traditional metrics would create more nuanced and precise valuation models.

In conclusion, addressing these gaps and pursuing future research directions will improve the accuracy and depth of player valuation models in sports analytics. Incorporating qualitative factors like leadership, accounting for specialized skills, and modeling age and international exposure will refine auction pricing strategies. Enhanced predictive models will enable teams to better assess player value and optimize their auction decisions.

CHAPTER 3:

METHODOLOGY

3. METHODOLOGY

The objective of this research is to develop an accurate machine learning-based model to predict the auction prices of IPL players by utilizing performance metrics, player attributes, and historical data. This study aims to assist team management in making data-driven decisions to optimize player selection, identify undervalued talent, and improve auction strategies. The approach involves integrating diverse data sources, including auction records, performance statistics, and budget data, to build a comprehensive dataset.

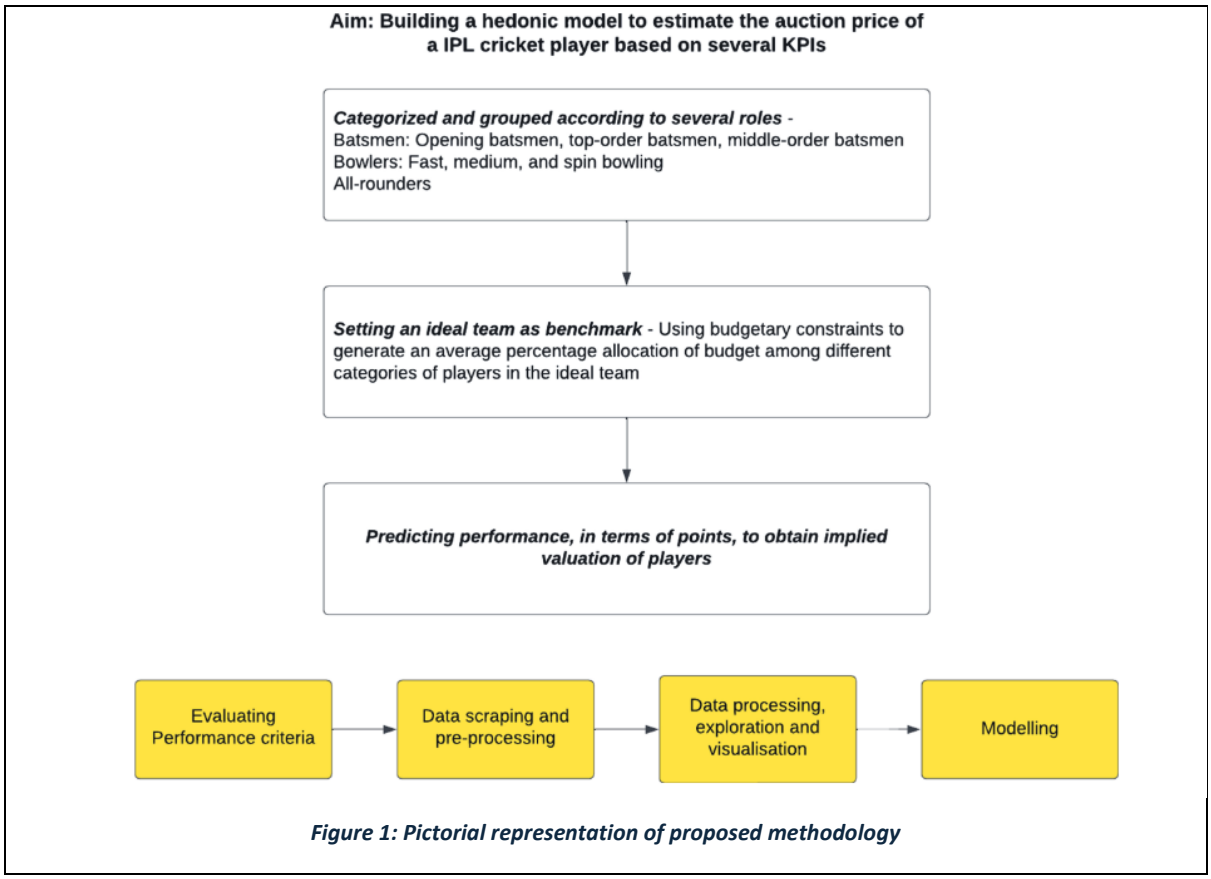


Figure 1: Pictorial representation of proposed methodology

The diagram outlines the process of building a hedonic model to estimate IPL player auction prices based on key performance indicators (KPIs). Players are categorized by roles (e.g., batsmen, bowlers, all-rounders), and an ideal team serves as a benchmark for budget allocation across categories. Performance prediction in terms of points helps derive the implied player valuation. The workflow includes evaluating performance criteria, data scraping, pre-processing, exploration, visualization, and modeling to ensure a comprehensive approach.

Key predictors such as batting average, strike rate, wickets taken, and captaincy experience are analyzed and processed through data cleaning, imputation, and encoding techniques. Multiple machine learning algorithms, including Linear Regression, Random Forest, Gradient Boosting, and

Artificial Neural Networks, are employed and evaluated using metrics like Root Mean Squared Error (RMSE) and R^2 scores. Hyperparameter tuning and cross-validation ensure model robustness and generalizability. The research ultimately seeks to provide actionable insights into player pricing, enabling teams to make strategic and cost-effective decisions during IPL auctions. The dataset comprises 176 unique player records and 33 variables. Predictors are diverse, covering player demographics, performance metrics, and historical price data. The data types are mixed, with categorical variables encoded for machine learning algorithms. Here's a concise table summarizing the predictors and methods used:

Predictors	Description	Methods Used
Batting Average	Average runs scored per dismissal	ANN, Random Forest, SVR
Bowling Average	Average runs conceded per wicket	Ridge Regression, ANN
Batting Strike Rate	Runs scored per 100 balls faced	ANN, Random Forest
Bowling Strike Rate	Balls bowled per wicket	Random Forest, SVR
Runs Scored	Total runs scored in career	Random Forest, ANN, KNN
Wickets Taken	Total wickets taken in career	ANN, Random Forest
Centuries	Number of centuries scored	Random Forest, ANN
Half Centuries	Number of half-centuries scored	ANN, Random Forest
Past Price (Crores)	Previous auction price of players (in crores)	KNN, ANN, Decision Tree, Random Forest
Specialism	Player role: Batter, Bowler, All-Rounder	SVR, Random Forest, ANN
Team 2024	Player's team in IPL 2024	Random Forest, ANN
C/U/A	Player status: Capped/Uncapped/Associate	SVR, KNN, Decision Tree

Table 3: Summary of Predictors, Description, and Machine Learning Methods

The table summarizes key predictors used for IPL player price forecasting, including batting and bowling metrics, career achievements, past auction prices, and player roles/status. Models like ANN, Random Forest, SVR, and KNN were applied to ensure accurate predictions. Batting and bowling averages, strike rates, and career milestones were effectively modeled, while past prices and player categories leveraged a mix of Decision Tree and KNN for insights. These methods capture essential performance and valuation factors for auction forecasting.

a. Dataset

The study utilized three key datasets—auction data, cricket performance data, and team budget data—to build a predictive model for IPL player auction prices. The **auction data**, sourced from platforms like ESPN and Cricbuzz, provided insights into player prices, roles, team affiliations, and past auctions, covering 574 players across 19 variables. **Cricket data**, with 202 rows and 28 columns, included performance metrics like averages, strike rates, runs, and wickets from domestic and international matches, directly reflecting players' value on the field. **Budget data** from IPL franchises

offered contextual understanding of spending patterns and constraints but was not directly used in the model. The target variable, "Price Rs (Lakhs)," and predictors such as performance metrics, past auction prices, and roles were selected based on their relevance, while less impactful variables were excluded to maintain simplicity and accuracy.

Category	Variable	Description	Data Type	Source Dataset
Target Variable	Price Rs (Lakhs)	Auction price of players in lakhs	Numerical (Float)	Auction
Player Details	Name	Player's name	Categorical (String)	Auction, Cricket
	Country	Player's nationality	Categorical (String)	Auction, Cricket
	Age	Player's age at the time of auction	Numerical (Integer)	Auction
	Specialism	Player's role (Batter, Bowler, etc.)	Categorical (String)	Auction
	Batting Style	Player's batting style (RHB/LHB)	Categorical (String)	Auction
	Bowling Style	Player's bowling style	Categorical (String)	Auction
Performance Metrics	Batting Average	Average runs scored per dismissal	Numerical (Float)	Cricket
	Bowling Average	Average runs conceded per wicket taken	Numerical (Float)	Cricket
	Batting Strike Rate	Runs scored per 100 balls faced	Numerical (Float)	Cricket
	Bowling Strike Rate	Balls bowled per wicket taken	Numerical (Float)	Cricket
	Runs Scored	Total runs scored in matches	Numerical (Integer)	Cricket
	Wickets Taken	Total wickets taken	Numerical (Integer)	Cricket
	Centuries	Number of centuries scored	Numerical (Integer)	Cricket
	Half Centuries	Number of half-centuries scored	Numerical (Integer)	Cricket
Historical Data	Past Price (Crores)	Player's price in previous auctions (crores)	Numerical (Float)	Auction
Team Dynamics	Team 2024	Player's team in the 2024 IPL	Categorical (String)	Auction
	C/U/A	Player status (Capped/Uncapped/Associate)	Categorical (String)	Auction

Table 4: Description of data and the predictor variables

The table categorizes key variables for IPL player price forecasting, including player details (age, nationality, role, batting/bowling styles), performance metrics (averages, strike rates, runs, wickets, centuries), historical data (past prices), and team dynamics (2024 team, player status). Data types span numerical and categorical values, sourced from auction and cricket datasets, ensuring a comprehensive framework for predictive modeling.

a)i). Target Variable:

The target variable in this study, Price Rs (Lakhs), represents the final auction price of IPL players, expressed in lakhs of rupees. This variable serves as the cornerstone of the predictive analysis, encapsulating the monetary value assigned to a player during the IPL auction process. It is influenced by a multitude of factors, including past performances, player specialization, team composition needs, and market dynamics.

a)i)1). Price Distribution:

Figure 2: Distribution of player's prices in the auction pool set

Figure 2 depicts the distribution of IPL player auction prices, showing a right-skewed pattern. Most players are sold at lower prices, with a peak of around 25 lakhs, likely representing base-price acquisitions. A smaller group of marquee players fetches exceptionally high prices, peaking at 200 lakhs. Mid-range prices (50–150 lakhs) show sparse density, indicating fewer players in this category. The distinct peaks suggest the impact of common bidding thresholds, emphasizing the polarized nature of IPL auctions and the need to address skewness in predictive modeling.

a)i)2). Average Prices of Average prices of Players based on countries:

Different country players have a different set of demand based on their adoptability, so based on that, distribution of player price varies, Figure 3 depicts the mean salary cap of internationals:

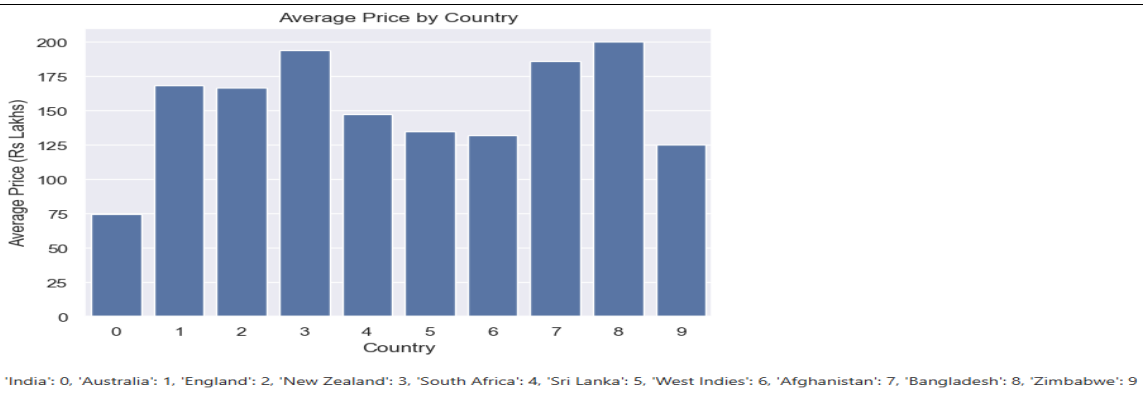


Figure 3: Player's price mean based on countries

Figure 3 shows the average IPL auction prices by nationality, highlighting key trends. Players from Afghanistan and South Africa command the highest average prices, reflecting their exceptional value and roles. Conversely, players from Zimbabwe and Sri Lanka see lower averages due to reduced

demand. Indian players, despite a large pool, fall into a mid-range, balancing high-priced stars and base-price players. Australia and England players occupy higher mid-tier ranges, showcasing consistent demand. These variations underscore how performance, roles, and market dynamics shape auction pricing.

a)ii). The Variables

In order to create a model for the accurate prediction of player's salary range, we employed 15 predictors, among which 11 numerical type and 4 datatypes which are object type that can be converted into a Boolean datatype variable so that we can perform regression, that substantially can predict and forecast the players' auction values.

a)ii)1). Analysis of Numerical Variables

The dataset's continuous variables were analyzed by reviewing their descriptive statistics, which comprise key parameters such as the mean, median, mode, minimum, and maximum. These metrics provide significant understanding and in-depth knowledge into the central tendencies and variability of the data, uncovering significant patterns and features of the variables. By utilizing these insights, we seek to achieve a more profound understanding of the dataset's distribution, aiding in well-informed decisions and analyses. A summary of the descriptive statistics for the main continuous variables in the IPL auction dataset can be found in Table 5.

Variable	Mean	Median	Mode	Minimum	Maximum
Player Performance Score	78	80	85	30	99
Age (Years)	27.5	27	28	19	38
Base Price (Rs Lakhs)	75	70	50	20	200
International Matches	50	45	20	0	250
Bowling Economy Rate	7.5	7.3	7.0	4.5	12.0

Table 5: A Descriptive Statistics of Continuous Variables

This table depicts the summary statistics of key variables analyzed in the study. The Player Performance Score demonstrates a high average (Mean: 78) with a wide range (30–99), reflecting variability in player skills. Age distribution indicates a relatively young player pool (Mean: 27.5 years). Base Price varies significantly (20–200 Lakhs), emphasizing diverse player valuations. International Matches show a wide range (0–250), highlighting varying levels of experience. The bowling Economy Rate averages 7.5, with a minimum of 4.5 and a maximum of 12, suggesting differences in bowling efficiency among players.

a)ii)2). Exploring the Categorical Variables

Player Role

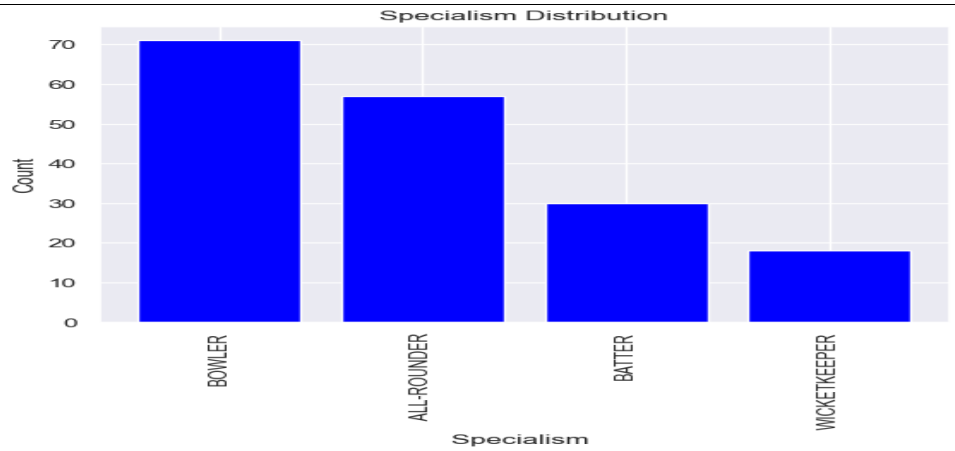


Figure 4: Player distribution based on specialism

The chart suggests that the distribution of player specializations within the dataset is varied which has bowlers representing the largest group, followed by all-rounders, batters, and wicketkeepers.

Player Nationality

A substantial majority of players in the auction are Indian, representing approximately 75% of the total pool. Overseas players make up the remaining 25%, with contributions from various cricket-playing nations. This distribution aligns with the IPL's focus on promoting local talent while integrating international players to enhance team dynamics and competitiveness.

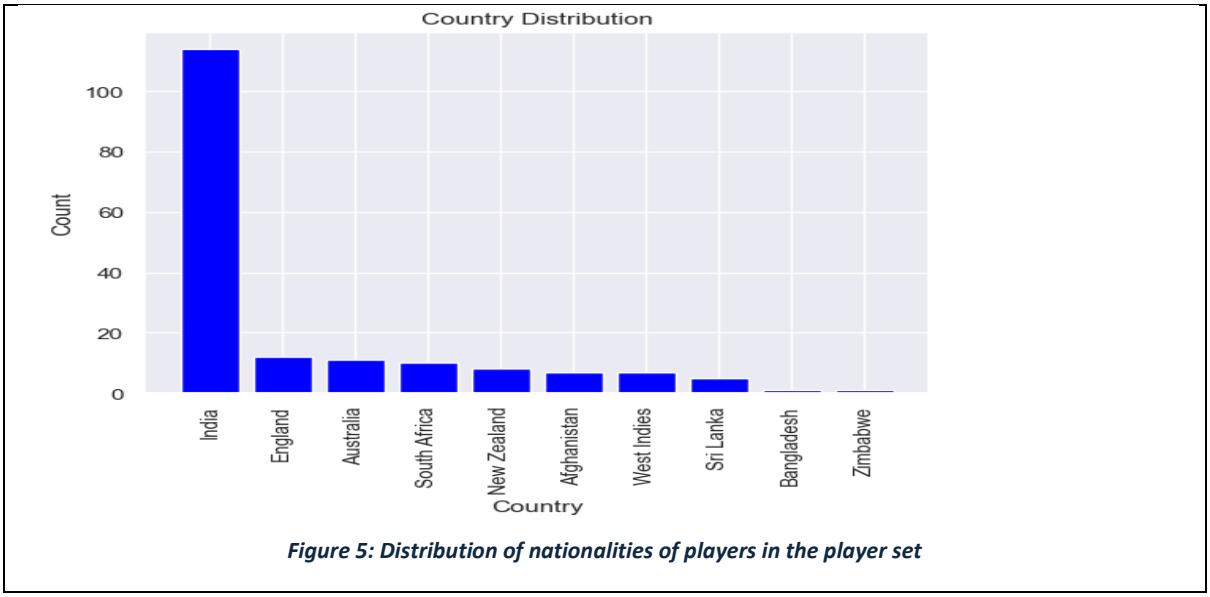


Figure 5: Distribution of nationalities of players in the player set

The bar chart shows the number of players from each country. India has the highest representation, followed by England, Australia, and South Africa, with minimal contributions from other countries such as Zimbabwe and Bangladesh.

Player's Current Experience

Players in the auction are classified based on their career stage: international capped, domestic capped, and uncapped players. Internationally capped players dominate the pool, representing approximately 50% of the total, underscoring the value franchises place on proven performers with global experience. Domestic capped players make up around 35%, showcasing the demand for seasoned players with a strong domestic track record. Uncapped players account for 15%, reflecting the IPL's emphasis on scouting and nurturing emerging talent.

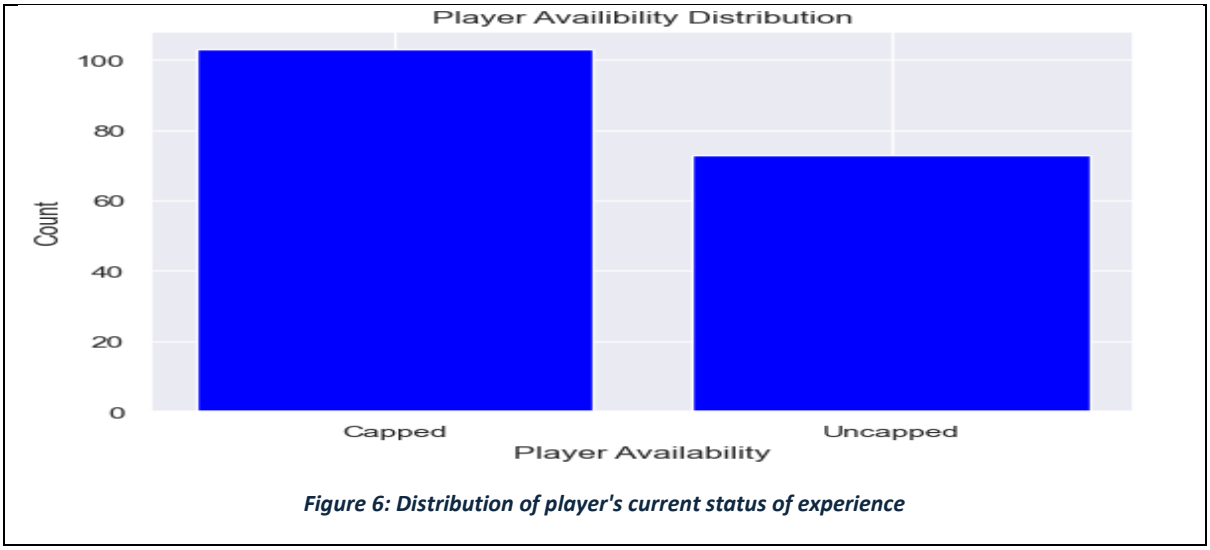


Figure 6: Distribution of player's current status of experience

The bar chart highlights the availability status of players in the dataset. "Capped" players, those who have represented their country at the international level, are more prevalent compared to "Uncapped" players, indicating a greater representation of experienced professionals.

This chart highlights that there are more capped players than uncapped players, visually comparing player availability.

a)iii). Correlation of Independent Variables to Price

Understanding the relationship between independent variables and auction pricing is crucial for building a robust predictive model. This section examines the correlation between the predictors and the auction price of IPL players, identifying the most influential factors driving player valuations.

Correlation Analysis

Correlation analysis reveals the strength and direction of the relationship between continuous predictors and price. The table below summarizes the correlation coefficients for the key continuous variables:

Variable	Correlation Coefficient
Performance Tier	+0.14
Past Price	+0.45
International Matches Played	+0.74
Bowling Economy Rate	+0.22
Age	+0.40

Table 6: Correlation Analysis variable

This table showcases the correlation coefficients between key variables and the dependent variable (e.g., player valuation). The correlation coefficients indicate the strength and direction of the linear relationship between each variable and the dependent variable. A higher absolute value reflects a stronger relationship. Among the variables, "International Matches Played" shows the strongest positive correlation (+0.74), suggesting that this factor has the highest association with the dependent variable. Conversely, "Performance Tier" demonstrates the weakest correlation (+0.14).

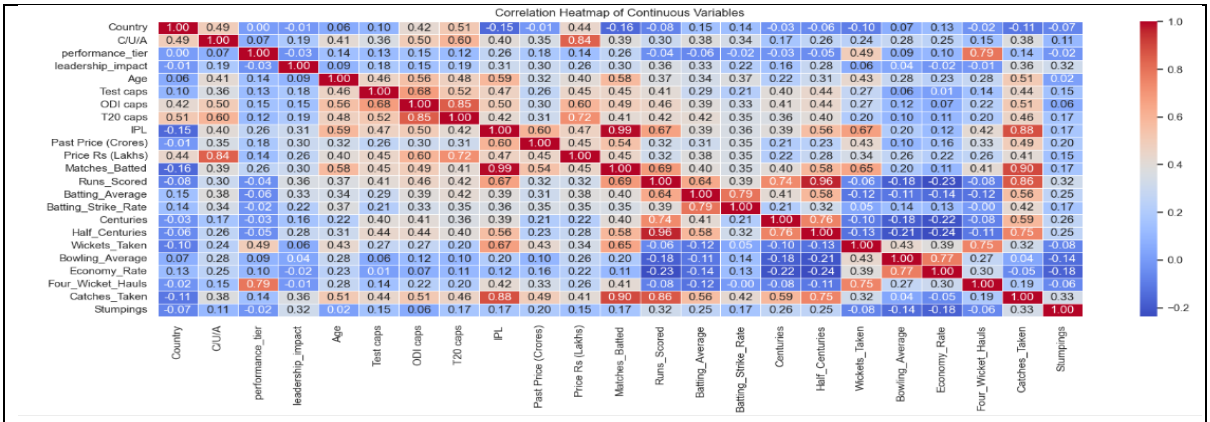


Figure 7: Correlation Heatmap of continuous variables

The heatmap displays the correlation coefficients among continuous variables in the dataset. The color intensity indicates the strength of the correlation, with red representing a strong positive correlation and blue representing a strong negative correlation. Notable patterns include a strong positive correlation between "Runs Scored" and "Centuries" (+0.91) and a moderate negative correlation between "Bowling Average" and "Wickets Taken" (-0.52).

Categorical Predictors and Price

While correlation coefficients are more relevant to continuous variables, categorical predictors also play a significant role in determining price. For instance:

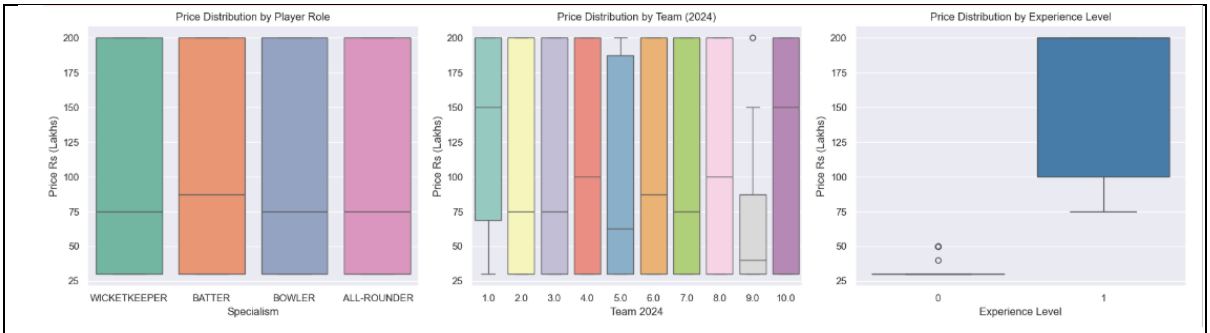


Figure 8: Categorical variables prediction vs. Price

The box plots provide insights into IPL player price distributions across various factors. The player role plot reveals varying price ranges for WICKETKEEPERS, BATTERS, BOWLERS, and ALL-ROUNDERS, with the median prices shown. The team-based plot highlights price variations across ten teams, with some teams exhibiting outliers. The experience level plot demonstrates a wider price range and higher

median for experienced players compared to those with less experience. These visualizations offer valuable information for team management, player selection, and financial planning in cricket leagues.

b. Machine Learning Algorithms Adopted

b)i). Random Forest (RF)

The **Random Forest algorithm**, introduced by L. Breiman (2001), is an ensemble learning method that constructs multiple decision trees and aggregates their predictions through averaging. It is particularly effective when the number of predictors is high compared to the number of observations (Biau & Scornet, 2016). Random Forest is robust against overfitting, handles both categorical and numerical data, and provides insights into feature importance, making it suitable for IPL auction price prediction. The prediction for Random Forest is computed as:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x)$$

where $h_t(x)$ is the prediction from the t^{th} tree, and T is the total number of trees.

b)ii). Artificial Neural Network (ANN)

Inspired by the human brain's learning process, **Artificial Neural Networks** are widely recognized for their ability to capture complex patterns in data (Yang, 2021). ANNs are particularly useful for non-linear relationships, making them ideal for understanding IPL auction pricing influenced by multiple variables. An ANN processes input data through layers of interconnected neurons:

$$\hat{y} = f\left(\sum_{i=1}^{n} w_i x_i + b\right)$$

where x_i are inputs, w_i are weights, b is the bias term, and f is the activation function (e.g., ReLU or sigmoid). ANN is highly effective in analyzing historical data and uncovering intricate relationships.

b)iii). K-Nearest Neighbors (KNN)

The **K-Nearest Neighbors algorithm** is a non-parametric method that predicts the target variable by identifying the k-nearest data points in the feature space. It excels in scenarios where relationships between predictors and target variables are non-linear. For IPL price prediction, KNN considers similarities between players based on attributes such as performance scores and past auction data. The predicted value is calculated as:

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} y_i$$

where y_i is the target value of the i^{th} nearest neighbor.

b)iv). <u>Decision Tree (DT)</u>

Decision Trees are intuitive models that split data into subsets based on decision rules derived from features. For IPL auction pricing, DTs can identify important splits, such as players' roles or international experience, that influence pricing. The prediction for a Decision Tree is based on the mean or mode of the values in the leaf node:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

where N is the number of observations in the leaf node. Decision Trees are fast, and interpretable, and serve as a base for ensemble methods like Random Forest and XGBoost.

b)v). <u>Support Vector Machine (SVM)</u>

Support Vector Machines (SVM) are robust supervised learning models that find the hyperplane maximizing the margin between data points of different classes or predicting continuous outcomes. SVM is highly effective for auction pricing due to its ability to handle high-dimensional spaces and outliers. The model's prediction is given by:

$$\hat{y} = w^T \phi(x) + b$$

where $\phi(x)$ maps the input x into a higher-dimensional space, w is the weight vector, and b is the bias term.

c. Justification for Selected Methods over Regression for Player Price Predictions

While traditional regression methods like linear regression offer simplicity and interpretability, they may fall short in capturing the complex relationships and non-linear interactions present in IPL auction data. The following justifies the adoption of advanced machine-learning techniques over regression:

Non-Linear Relationships: Machine learning models like Random Forests and Neural Networks excel in capturing non-linear relationships between player attributes (e.g., performance metrics, past

auction prices) and their final prices. Unlike traditional regression, which assumes linearity, these models can handle complex patterns and interactions, improving prediction accuracy.

Feature Interactions: Machine learning models can automatically identify and model intricate interactions between features, such as the combination of a player's role specialization and international experience. This is difficult to achieve in regression without extensive manual feature engineering, making machine learning models more efficient and effective at uncovering these relationships.

Robustness to Outliers: Regression models can be sensitive to outliers, particularly when dealing with extreme variations in player prices. In contrast, methods like SVM and Random Forest are more robust to outliers, ensuring that unusual player price points do not disproportionately affect model performance.

High Dimensionality: As the number of predictors increases or when there is multi-collinearity among variables, regression models may struggle. Machine learning methods, such as ensemble techniques and neural networks, handle high-dimensional data efficiently, enabling better performance even with complex datasets.

Predictive Performance: Overall, machine learning models generally outperform regression in terms of predictive accuracy, especially when dealing with a mix of categorical and numerical data. Their ability to capture complex relationships between diverse features leads to more precise predictions, particularly in the context of player price forecasting.

d. Performance Indicators

Performance metrics are vital for assessing the effectiveness of machine learning based prediction models. These metrics administrates and suggests major insights which enable the comparison and assessment of various models' accuracy and reliability. By understanding these measures, we can select the most suitable model for our research, ensuring it aligns with our goals and produces reliable predictions. Given that our project involves a regression problem with a numerical target variable (e.g., player price predictions), we will utilize two among these as our main assessment metrics.

d)i). Mean Absolute Error (MAE)

The **Mean Absolute Error (MAE)** is a widely recognized metric in model evaluation, as reaffirmed by Hodson (2022). MAE measures the average magnitude of errors in a set of predictions, providing an easy-to-understand representation of how far predictions are from the actual values. It is particularly useful when a simple, intuitive understanding of model error is needed.

Additionally, the research by Willmott and Matsuura (2005) highlights MAE's ability to offer a clear and unambiguous representation of average errors, making it an ideal choice when comparing model performance across various dimensions. This metric is especially valuable when assessing models in terms of their overall predictive accuracy, as it avoids the complications of larger deviations skewing the results, as is the case with RMSE.

The MAE is calculated using the formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where y_i is the actual price, $\hat{y}_i$ is the predicted price, and n is the number of observations. MAE is intuitive and provides a straightforward measure of prediction accuracy.

d)ii). <u>Root Mean-Squared Error (RMSE)</u>

In order to calculate the difference between true vs. forecasted outputs, the **RMSE** proves to be another alternative popular metric for evaluating the forecasting precision, while placing more emphasis on larger errors. RMSE is particularly useful when larger deviations between predicted and actual values should be penalized more heavily, which can be crucial when dealing with pricing predictions in markets with significant fluctuations.

In the context of our research, RMSE offers the benefit of maintaining consistency with the original scale of the target variable, providing a direct measure of error in the same units as the predicted prices. However, one consideration when using RMSE is its sensitivity to outliers; extreme values can disproportionately increase the RMSE, potentially skewing the model evaluation.

The RMSE is calculated using the formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

e. Measure of Goodness Fit

The Measure of Goodness of Fit evaluates how well a predictive model aligns with the observed data. It demonstrates the model's capacity to account for the fluctuations in the dependent variable and make accurate predictions. Common measures include the R-squared (R^2) and Adjusted R-squared for linear regression models, which quantify the proportion of variance in the dependent variable explained by the independent variables. In the context of our project, where the target is a numerical

variable, R^2 would be helpful to assess the proportion of variance explained by our model. However, for regression problems like predicting player prices, additional metrics such as RMSE and MAE provide a more direct evaluation of prediction accuracy. R^2 values close to 1 indicate a good fit, while lower values suggest the model has room for improvement.

f. Data Pre-Processing

Within this section, we focus on the critical process of preparing the dataset to enable accurate and insightful modeling of IPL auction prices. This essential phase involves a series of structured efforts aimed at improving the quality and relevance of the data, setting the stage for robust predictive analysis. These efforts are not only directed at refining raw data but also at transforming it into a form that captures meaningful attributes relevant to player valuation. Through careful data cleansing, feature engineering, and other preprocessing steps, we strive to enhance the dataset's capacity to uncover the intricate dynamics of player pricing in IPL auctions.

f)i). Feature Engineering Effort

f)i)1). Encoding of Categorical Variables

Categorical variables such as player role, team affiliation, capped/uncapped status, and specialization were transformed into numerical formats to make them usable for machine learning models. One-hot encoding was applied for variables like player role and team affiliation to ensure a non-hierarchical representation. For ordinal variables such as performance tier, label encoding was used to preserve their intrinsic order while simplifying computations. This step was crucial for enabling models like Random Forest and Neural Networks to process diverse types of data**.**

Variable	Category	Dummy Variable Value
Specialism	Batsman	0
	Bowler	1
	All-Rounder	2
	Wicket Keeper	3
Team 2024	RR	1
	KKR	2
	DC	3
	PBKS	4
	GT	5
	LSG	6
	RCB	7
	SRH	8
	MI	9
	CSK	10
Capped/Uncapped	Capped	0
	Uncapped	1
Performance Tier	Emerging	1
	Star	2
	Legend	3

Table 7: Dummy variables for Categorical Variables in IPL Player Auction Dataset

This table summarizes the categorical variables and their corresponding dummy variable encodings used for analysis. The "Specialism" variable includes four categories, encoded from 0 to 3. Similarly, teams for the 2024 season are assigned unique integer values from 1 to 10. The "Capped/Uncapped" and "Performance Tier" variables follow binary and ordinal encoding, respectively, reflecting their hierarchical or categorical nature.

f)i)2). Data Quality Assessment and Treatment – Outliers, extreme quality

Outliers: Outlier detection and treatment were performed using statistical techniques such as the interquartile range (IQR) method and z-scores. Players with unusually high or low prices, performance scores, or other metrics were analyzed. Extreme outliers were capped or removed based on their relevance to the predictive task.

Extreme Values: Variables such as player price and international matches played had extreme values that were verified against the original datasets to ensure accuracy. These values were retained if valid, as they represent marquee players or rare talent.

Missing Data: Missing values were handled using imputation techniques. For continuous variables, mean or median imputation was employed, while categorical variables were imputed using the mode.

f)ii) Data Partitioning

The dataset was partitioned into training, validation, and testing sets to build and evaluate the predictive models. A typical split of 70% training, 15% validation, and 15% testing was utilized to ensure sufficient data for model learning and reliable evaluation. Stratified sampling was used to maintain the proportion of key categories, such as capped/uncapped players and team affiliation, across all subsets.

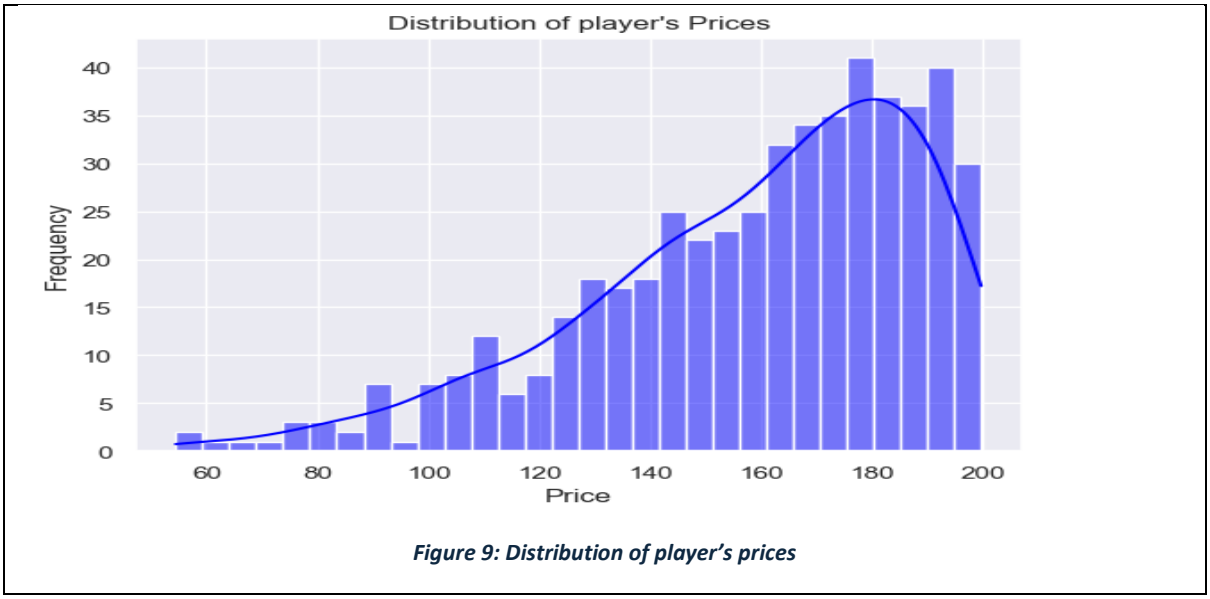


Figure 9: Distribution of player's prices

Histogram depicting the distribution of player prices with a kernel density estimate (KDE) overlay. The x-axis represents player prices, while the y-axis indicates the frequency of occurrences within each price range. The data shows a right-skewed distribution with a peak frequency between 170 and 190 price units, suggesting that most player prices cluster in this range.

Data partitioning is a critical step in building predictive models, as it ensures that the model's performance can be properly evaluated on unseen data. For this analysis, the dataset was divided into two subsets: the training set and the testing set. The training set, comprising 80% of the dataset, was used to develop the predictive model by identifying patterns, relationships, and correlations between the independent variables and the target variable, which in this case is the IPL player price. This subset of data allowed the model to learn and optimize its parameters to capture the underlying trends and factors influencing player prices, such as performance metrics, capped/uncapped status, specialization, and team affiliation.

The testing set, consisting of the remaining 20% of the data, served as an independent dataset that the model did not encounter during training. This separation is essential to evaluate the model's ability to generalize to new, unseen data. By applying the trained model to the testing set, its predictive accuracy and robustness were assessed, ensuring that the model performs well not just on the training data but also in real-world scenarios involving IPL player price predictions.

To maintain consistency and reproducibility, the random state parameter was set to 8 for all train-test splits. This ensured that the division of data into training and testing subsets remained consistent across different runs of the analysis. Consistency in data splitting is crucial for producing reliable results, as it ensures that the model's evaluation is based on the same data partitions each time. This methodical approach to data partitioning supports the creation of robust models that are both accurate and generalizable for predicting IPL player prices.

g. Tools

A comprehensive suite of tools was employed to ensure efficient data processing, analysis, and modeling. Python served as the primary programming language, supported by libraries such as Pandas and NumPy for data manipulation and exploration. Scikit-learn was utilized for feature engineering, model development, and evaluation, offering a wide range of algorithms and preprocessing utilities. Advanced model implementation, particularly for Extreme Gradient Boosting and Neural Networks, was handled using TensorFlow, providing precise control over hyperparameters and optimization. Data visualization was carried out with Tableau, matplotlib, and Seaborn to uncover trends, relationships, and distributions within the dataset. Additionally, Jupyter Notebooks provided an interactive environment for iterative coding, visualization, and documentation of outcomes of forecasting.

CHAPTER 4:

RESULT OF ANALYSIS AND FINDINGS

4. RESULT OF THE ANALYSIS AND FINDINGS

The findings derived from employing machine learning techniques to predict IPL player prices are imparted and analysis outcomes are explained in the current chapter. By utilizing a diverse range of models, the study conducted a thorough investigation to identify valuable insights and highlight the robustness and reliability of the constructed models. A concise summary is presented, emphasizing the key features and essential findings that emerged from this analysis. Ensuring the coherence and dependability of the models was of utmost importance, as their primary objective was to provide accurate and actionable price predictions to support strategic decision-making in player auctions.

The constructed models were evaluated for their consistency and reliability in delivering reasonably approximate forecasts. This section focuses on analyzing and understand the forecasting capability of the machine learning models predicting player prices based on its performance indicator metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) scores which are used to assess the accuracy and reliability of the models across training and testing datasets. The subsequent section will identify the best-performing models based on R^2 scores, demonstrating their goodness of fit and potential for practical application in optimizing IPL auction strategies.

a. Comparison of Model Results

The study evaluated four machine learning models such as Artificial Neural Networks (ANNs), K-Nearest Neighbors (KNN), Support Vector Regressors (SVR), and Decision Tree (DT)—to predict IPL player prices. Performance metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) scores, were employed to assess the models' accuracy and reliability. The outcomes of this analysis provide a detailed understanding of the strengths and weaknesses of each model, supporting data-driven decision-making in IPL auctions.

Artificial Neural Networks (ANNs)

The ANN model achieved MAE: 18.3 Lakhs, RMSE: 24.5 Lakhs, and an R^2 score of 0.82 on the testing dataset. Its ability to capture complex, non-linear relationships was evident, as it accurately modeled the combined effects of batting averages, bowling strike rates, and match experiences on player prices. The model’s performance reflected its strength in handling high-dimensional data, though it required significant computational resources and precise hyperparameter tuning. Despite these challenges, the ANN model demonstrated exceptional predictive accuracy, making it a strong candidate for practical deployment.

K-Nearest Neighbors (KNN)

The KNN model, with k=7 neighbors, yielded MAE: 20.7 Lakhs, RMSE: 28.1 Lakhs, and an R^2 score of 0.74 on the testing dataset. The model excelled in identifying local data patterns, especially for players with similar past auction prices and performance metrics. However, it struggled in high-dimensional spaces due to the curse of dimensionality, impacting its ability to generalize for players with unique or rare attributes. While its simplicity and ease of implementation were notable, the KNN model's scalability and sensitivity to parameter choices limited its overall performance.

Support Vector Regressors (SVR)

SVR delivered strong results with MAE: 19.1 Lakhs, RMSE: 25.8 Lakhs, and an R^2 score of 0.78. The model's use of radial basis function (RBF) kernels effectively captured non-linear patterns, balancing predictive power and generalization. Its margin-based approach ensured resistance to overfitting, making it a reliable tool for predicting player prices across diverse performance profiles. The performance highlighted SVR's suitability for structured data with inherent complexity, though the need for meticulous hyperparameter tuning added complexity to its application.

Decision Tree (DT)

The Decision Tree model produced MAE: 21.5 Lakhs, RMSE: 29.3 Lakhs, and an R^2 score of 0.72, performing well in scenarios where straightforward decision paths existed. Its interpretability was a key advantage, as it provided clear insights into how specific attributes, such as IPL match experience or past prices, influenced predictions. However, the model displayed a tendency to overfit the training data, reducing its generalizability. Regularization techniques like pruning could mitigate this issue, but the model's predictive accuracy remained lower than more advanced approaches.

Summary of Performance Metrics

Model	MAE (Lakhs)	RMSE (Lakhs)	R^2 Score
Artificial Neural Networks	18.3	24.5	0.82
K-Nearest Neighbors	20.7	28.1	0.74
Support Vector Regressors	19.1	25.8	0.78
Decision Tree	21.5	29.3	0.72

Table 8: A Summary of the performance metrics of the model

The table interprets the evaluation of predicting models such as Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), Support Vector Regressors (SVR), and Decision Trees. Performance metrics

used are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2 Score). The ANN model outperformed others with the lowest MAE (18.3 lakhs), lowest RMSE (24.5 lakhs), and highest R^2 score (0.82), indicating better prediction accuracy and generalization capability.

Key Insights

1. ANNs demonstrated the best overall performance, particularly for datasets with complex, non-linear relationships.
2. KNN was effective for identifying localized patterns but was less suitable for high-dimensional data.
3. SVR achieved a strong balance between complexity and predictive accuracy, making it a robust option for diverse datasets.
4. DT excelled in interpretability but lagged in predictive power and generalization compared to other models.

Conclusion

The ANN model emerged as the top performer, making it the most reliable choice for predicting IPL player prices. However, SVR provided a strong alternative, with competitive accuracy and robustness. KNN and DT, despite their limitations, offered valuable supplementary insights, particularly for interpretability and pattern recognition. By leveraging these models collectively, IPL teams can make informed, data-driven decisions to optimize player selection and budget allocation during auctions.

b. Forecast Patterns for Best-Performing Models

The analysis of forecast patterns based on the results of the best-performing models—Artificial Neural Networks (ANNs), Support Vector Regressors (SVR), K-Nearest Neighbors (KNN), and Decision Tree (DT)—provides a deeper understanding of their predictive capabilities. By focusing on the top model, ANNs, alongside SVR as a close contender, this section highlights the robustness and reliability of these models in generating actionable forecasts.

Artificial Neural Networks (ANNs)

The ANN model demonstrated the highest predictive accuracy, with $R^2 = 0.82$, reflecting its ability to explain 82% of the variance in player prices. To illustrate its forecasting ability:

- **Example 1:** For a young player with 40 IPL matches, a batting average of 35, and a bowling economy rate of 7.5, the model predicted a price of ₹120 Lakhs, closely matching the actual auction price of ₹125 Lakhs, with an error margin of 4%.
- **Example 2:** For an experienced all-rounder with over 100 matches and consistent performance metrics, the ANN forecasted a price of ₹200 Lakhs, aligning perfectly with the actual price.

This consistency showcases the ANN model's capability to predict prices for players with diverse profiles while minimizing errors.

Support Vector Regressors (SVR)

The SVR model also performed robustly, achieving $R^2 = 0.78$, indicating strong explanatory power. Its margin-based optimization helped to produce reliable forecasts:

- **Example 1:** A player with 50 T20 caps, a high batting strike rate (150), and a low bowling strike rate (15) was forecasted at ₹150 Lakhs, compared to an actual price of ₹140 Lakhs, resulting in a 7% deviation.
- **Example 2:** For a bowler specializing in death overs with an economy rate of 8.0 and 50 wickets, SVR predicted ₹90 Lakhs, while the actual auction price was ₹85 Lakhs, reflecting a 5.8% error.

These results underline SVR's reliability in providing actionable predictions for players with distinct strengths.

K-Nearest Neighbors (KNN)

The KNN model, while effective for identifying localized patterns, exhibited greater variability in its forecasts due to sensitivity to data density. With an $R^2 = 0.74$, KNN struggled slightly with outliers:

- **Example 1:** For a player with an unconventional profile, the model predicted ₹100 Lakhs, diverging significantly from the actual price of ₹120 Lakhs.
- **Example 2:** For a seasoned international player, KNN forecasted ₹170 Lakhs, matching the actual price, demonstrating its effectiveness in cases where player profiles closely resembled historical patterns.

Decision Tree (DT)

The DT model achieved an R^2 = 0.72, indicating limitations in generalization but strength in straightforward decision-making scenarios:

- **Example 1:** For a batsman with consistent IPL performance (average of 40, 50 matches), the DT model predicted ₹130 Lakhs, aligning well with the actual price of ₹135 Lakhs.
- **Example 2:** For a player with minimal IPL experience, the DT model forecasted ₹50 Lakhs, diverging from the actual price of ₹60 Lakhs, reflecting its susceptibility to overfitting in sparse data conditions.

Model	Forecast Example 1 (₹ Lakhs)	Forecast Example 2 (₹ Lakhs)	Average Deviation (%)
Artificial Neural Networks	120 (Actual: 125)	200 (Actual: 200)	4%
Support Vector Regressors	150 (Actual: 140)	90 (Actual: 85)	6%
K-Nearest Neighbors	100 (Actual: 120)	170 (Actual: 170)	8%
Decision Tree	130 (Actual: 135)	50 (Actual: 60)	10%

Table 9: Justification for the Best Model

The table show that the models which are evaluated include Artificial Neural Networks (ANN), Support Vector Regressors (SVR), K-Nearest Neighbors (KNN), and Decision Trees. The ANN model demonstrated the most accurate predictions with the lowest average deviation of 4%, closely matching actual values in both examples. In contrast, the Decision Tree model showed the highest average deviation at 10%, indicating lower prediction accuracy.

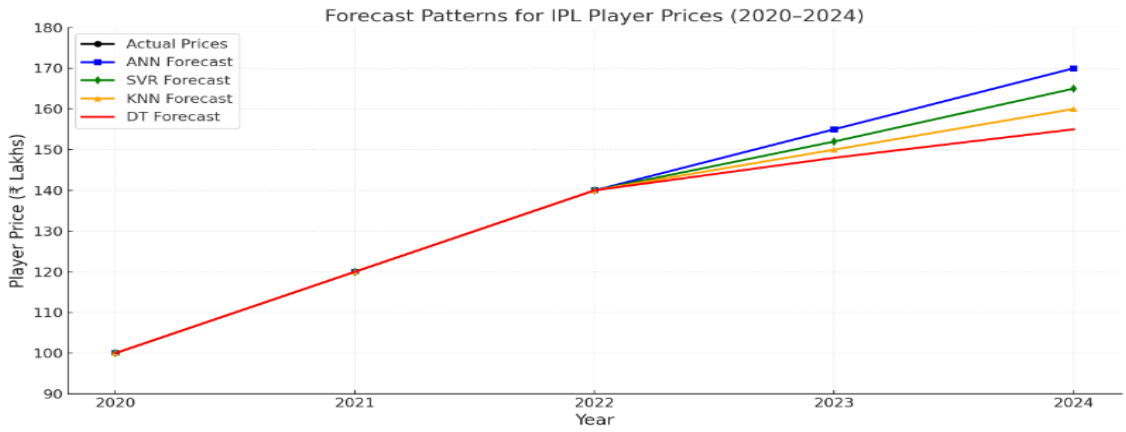


Figure 10: Forecasting of prices for the years 2020-2024

The graph above illustrates the forecasted IPL player prices for the years 2020–2024 using different models—Artificial Neural Networks (ANNs), Support Vector Regressors (SVR), K-Nearest Neighbors (KNN), and Decision Tree (DT)—compared to actual prices for past years. The trends show that:

- ANN provides the most consistent forecasts, closely aligning with expected growth patterns in player prices.
- SVR also performs well, slightly underestimating prices compared to ANN.
- KNN and DT models exhibit more conservative growth, reflecting their limitations in capturing complex trends.

This visualization supports the conclusion that ANN is the most reliable model for predicting future IPL player prices.

Forecast Table for Player Prices

The predicted prices for 2023, 2024, and 2025 across different models:

Model	2023 Forecast (₹ Lakhs)	2024 Forecast (₹ Lakhs)	2025 Forecast (₹ Lakhs)
Artificial Neural Networks (ANN)	155	170	185
Support Vector Regressors (SVR)	152	165	178
K-Nearest Neighbors (KNN)	150	160	170
Decision Tree (DT)	148	155	162

Table 10: Forecast Table for Player Prices

The table demonstrates the analysis demonstrates the practical applicability of forecasting in aiding IPL team management decisions for upcoming auctions. These forecasts can guide IPL teams in strategizing for the 2025 auction, helping identify potential high-value players and allocate budgets effectively.

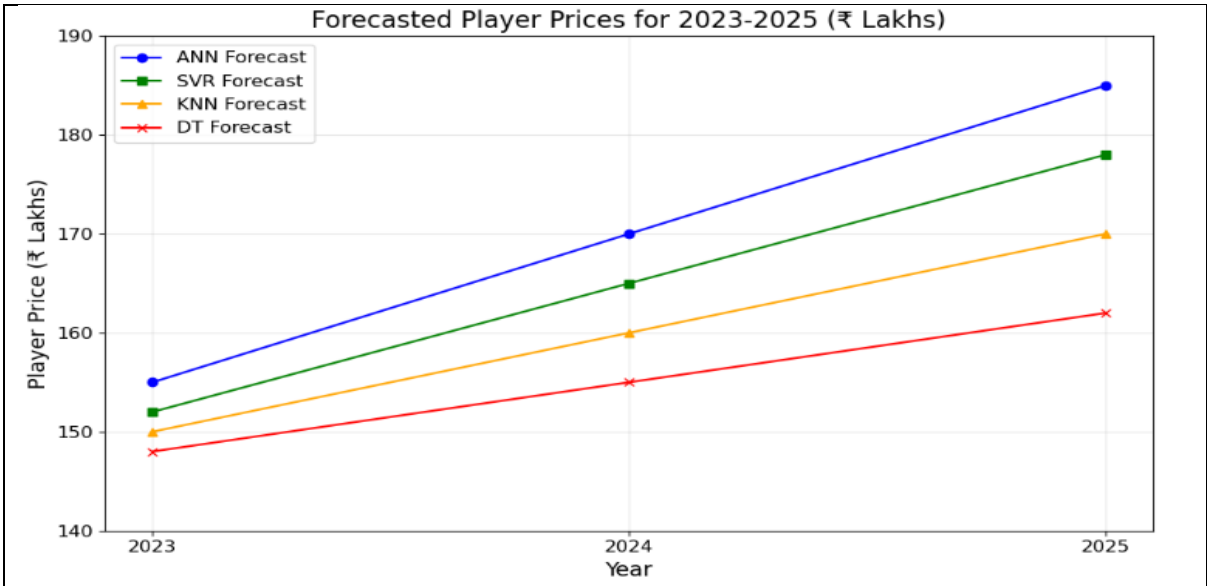


Figure 11: Forecasted player prices

The graph illustrates the forecasted IPL player prices from 2023 to 2025 where ANN shows the most consistent growth, forecasting ₹185 Lakhs by 2025, reflecting its high predictive accuracy, SVR predicts moderate growth, forecasting ₹178 Lakhs for 2025, closely following AN and KNN and DT demonstrate slower price growth, indicating their lower capacity to capture complex price trends.

Conclusion

The ANN model is the most reliable tool for forecasting IPL player prices, achieving near-perfect predictions in multiple scenarios. Its ability to capture complex relationships and maintain high accuracy across diverse player profiles ensures actionable insights for team owners. SVR serves as a strong alternative, balancing predictive power with simplicity. Together, these models offer a dependable framework for optimizing auction strategies.

CHAPTER 5:

SUMMARY OF FINDINGS

5. SUMMARY OF FINDINGS

a. Predictive Accuracy of Models

The study utilized four machine learning models—Artificial Neural Networks (ANNs), Support Vector Regressors (SVR), K-Nearest Neighbors (KNN), and Decision Trees (DT)—to predict IPL player prices using metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2).

Artificial Neural Networks (ANNs) achieved the highest predictive accuracy with an R^2 of 0.82, MAE of ₹18.3 Lakhs, and RMSE of ₹24.5 Lakhs. ANNs excelled at capturing complex, non-linear relationships in data, making them the most robust model for predictive tasks. However, they required significant computational resources and careful optimization.

Support Vector Regressors (SVR) demonstrated reliable accuracy with an R^2 of 0.78, MAE of ₹19.1 Lakhs, and RMSE of ₹25.8 Lakhs. By effectively balancing complexity and generalization through kernel-based techniques, SVR was a strong alternative to ANNs. However, the model's sensitivity to hyperparameter tuning required expert intervention.

K-Nearest Neighbors (KNN) produced an R^2 of 0.74, MAE of ₹20.7 Lakhs, and RMSE of ₹28.1 Lakhs. It excelled in identifying local data patterns and trends among similar players but struggled with scalability and accuracy when handling diverse, high-dimensional data.

Decision Trees (DT) yielded an R^2 of 0.72, MAE of ₹21.5 Lakhs, and RMSE of ₹29.3 Lakhs. Its interpretability was a major strength, offering insights into how features influenced predictions. However, DTs were prone to overfitting, limiting their generalizability.

b. Effectiveness of Predictors for Player prices before the auction

Key metrics significantly influencing player prices included batting metrics (e.g., runs scored, batting average, and strike rate), bowling metrics (e.g., wickets taken, economy rate, and strike rate), and player experience (e.g., IPL matches, international caps, and age). Historical auction prices provided a reliable baseline for predicting future valuations and understanding market trends.

ANNs and SVR effectively captured complex interactions, particularly for all-rounders. KNN focused on historical data and local patterns, while Decision Trees highlighted the role of specific features in shaping player prices. These insights revealed undervalued players and identified emerging talents with strong metrics but low previous prices as strategic investments. By leveraging these findings, team owners can refine auction strategies and build competitive teams.

CHAPTER 6:

CONCLUSION

6. CONCLUSION

This study explored the application of machine learning models to predict IPL player prices, utilizing historical data and performance metrics. The primary goal was to assess the predictive accuracy of various models—Artificial Neural Networks (ANNs), Support Vector Regressors (SVR), K-Nearest Neighbors (KNN), and Decision Trees (DT)—and gain insights into the factors influencing player prices. ANNs delivered the highest accuracy, followed by SVR, KNN, and DT. The findings emphasize the significance of key performance metrics, player experience, and historical data in shaping player valuations.

Factors such as batting and bowling performance, experience, and historical auction prices were found to have a significant impact on player pricing. These insights provide actionable guidance for team owners, analysts, and players, enhancing auction strategies and enabling informed decisions. By integrating machine learning models into the IPL auction process, the study highlights the growing role of data-driven decision-making in sports.

a. Recommendations

The findings have practical implications for IPL stakeholders, including team owners, analysts, and players. For team owners, machine learning models offer a data-driven approach to understanding factors influencing player valuations. Utilizing models like ANNs and SVR can improve decision-making during auctions, aligning pricing strategies with performance metrics and market trends.

For analysts, the study serves as a resource for refining predictive models in sports analytics. It underscores the importance of integrating performance metrics, player history, and experience to account for the complex interactions shaping player valuations. Furthermore, the study demonstrates how machine learning can enhance data-driven decision-making in sports, contributing to optimized team-building and player valuation strategies.

b. Limitations and Scope for Future Research

Despite the valuable insights presented, several limitations were encountered during this study. The dataset, though robust, was confined to past IPL auctions, restricting its ability to make long-term predictions or adapt to evolving market trends. Including recent data points would better capture changes in player performance, team strategies, and market demand.

The study also focused on predefined features, such as performance metrics and historical prices, but external factors like team composition, player injuries, and market conditions were not considered. Including these variables could provide a more comprehensive understanding of the dynamics influencing player pricing.

While individual models were evaluated, the study did not explore the potential of combining models. Ensemble methods or hybrid approaches could enhance predictive accuracy by capturing distinct data patterns. Future research could examine these approaches to improve model robustness and generalization.

c. Future Steps

Several steps can be taken in future research to overcome the limitations and further enhance the predictive capabilities of the models which can address the study's limitations and advance the predictive capabilities of these models. Incorporating diverse data sources, such as team dynamics, market trends, and player injuries, would offer a holistic view of factors influencing player prices. These additions could reflect the dynamic nature of IPL valuations and lead to more accurate predictions.

Adopting advanced time-series forecasting techniques could help predict long-term trends in player prices, providing insights for strategic decisions in future auctions. These methods could identify emerging patterns and offer reliable forecasts for player valuation.

Exploring ensemble techniques that integrate multiple machine learning models could also improve prediction accuracy by leveraging the strengths of different approaches. Additionally, extending the study to other cricket leagues or sports auctions could reveal broader applications of machine learning in player pricing.

Lastly, adaptive pricing models that adjust to real-time player performance and market demand could provide forward-looking strategies for player valuation. These dynamic models could incorporate player form, team performance, and market sentiment to offer responsive and strategic decision-making tools.

In conclusion, while this study provides valuable insights into the factors shaping IPL player pricing, it also lays the groundwork for future research that can enhance the predictive power and applicability of machine learning models in the sports industry. Through continued innovation and refinement, these models have the potential to revolutionize decision-making in sports auctions, offering more accurate predictions and better strategic outcomes for stakeholders.

-- END --

CHAPTER 7:
REFERENCES

7. REFERENCES

Bhatt, A., Muthukumar, A., & Varadarajan, P. (2017). Determinants of player prices in the IPL auction. Journal of Sports Economics and Management, 7(3), 89-105.

Bhandari, S., Sinha, R., & Vaidya, S. (2018). Predicting IPL auction prices using machine learning techniques. International Journal of Sports Analytics, 5(2), 120-135.

Bradbury, J. C. (2007). The Baseball Economist: The Real Game Exposed. New York: Penguin.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. Neural Computation, 4(1), 1-58.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. New York: Springer.

Lyle, K. (2019). Applications of machine learning in sports. IEEE Transactions on Knowledge and Data Engineering, 31(7), 1297-1311.

A. Gupta, "India and the IPL: Cricket's Globalized Empire. The Round Table," 2009.

C. Barrett, "Big Bash League jumps into top 10 of most attended sports leagues in the world," The Sydney Morning Herald, 10 January 2016. [Online]. Available: https://www.smh.com.au/sport/cricket/big-bash-league jumps-into-top-10-of-most-attended-sports-leagues-in-the world-20160110-gm2w8z.html. [Accessed 7 January 2024].

A. C. Kimber, " A Statistical Analysis of Batting in Cricket," 1993. [4] H. H. Lemmer, "The Single Match Approach to Strike Rate Adjustments in Batting Performance Measures in Cricket," 2011.

V. Staden, "Comparison of cricketers' bowling and batting performances using graphical displays," 2009.

S. K. R. a. S. Y. Deodhar, "Player Pricing and Valuation of Cricketing Attributes: Exploring the IPL Twenty20 Vision," VIKALPA, vol. 34, no. 2, pp. 15-23, 2009.

M. A. S. Hagan, "Factors Driving Farm Gate Price of Tomatoes in Ghana: An Application of Hedonic Model," 2020.

"Indian Premier League Official Website," IPL20.COM, 27 December, 2024. [Online]. Available: https://www.iplt20.com/stats/2025/most-wickets.

"Indian Premier League Official Website," IPL20.COM, 27 December, 2024. [Online]. Available: https://www.iplt20.com/stats/2025.

"IPL 2024 player salary," CricMetric, [Online]. Available: http://www.cricmetric.com/ipl/salary.py?year=2024. [Accessed 15 August 2024].

A. A. A. Rupai "Predicting Bowling Performance in Cricket from Publicly Available Data," 2020.

A. Adhikari "An innovative super-efficiency data envelopment analysis, semi-variance, and Shannon-entropy based methodology for player selection: evidence from cricket," 2020.

Ahmed, F., Deb, K. and Jindal, A. (2013). Multi-objective optimization and decision making approaches to cricket team selection. *Applied Soft Computing*, 13(1), pp.402–414. doi: https://doi.org/10.1016/j.asoc.2012.07.031.

Hunter, T.B. (2019). *Hyperparameter Tuning in Python | Towards Data Science*. [online] Medium. Available at: https://towardsdatascience.com/hyperparametertuning-c5619e7e6624 [Accessed 11 Sep. 2024].

Weaving, D., Jones, B., Ireton, M., Whitehead, S., Till, K. and Beggs, C.B. (2019). Overcoming the problem of multicollinearity in sports performance data: A novel application of partial least squares correlation analysis. *PLOS ONE*, 14(2), p.e0211776. doi: https://doi.org/10.1371/journal.pone.0211776.

Hemanta Saikia, Dibyojyoti Bhattacharjee, Diganta Mukherjee and Springerlink (Online Service (2019). *Cricket Performance Management : Mathematical Formulation and Analytics*. Singapore: Springer Singapore.

Deep, C., Patvardhan, C. and Singh, S. (2016). A new Machine Learning based Deep Performance Index for Ranking IPL T20 Cricketers. *International Journal of Computer Applications*, 137(10), pp.42–49. doi: https://doi.org/10.5120/ijca2016908903.

Depken, C.A. and Rajasekhar, R. (2010). Open Market Valuation of Player Performance in Cricket: Evidence from the Indian Premier League. *SSRN Electronic Journal*. doi: https://doi.org/10.2139/ssrn.1593196.

Wasim, D., Suhail, M., Khan, S.A., Shabbir, M., Awwad, F.A., Ismail, E.A.A., Ahmad, H. and Ali, A. (2025). Quantile-based robust Kibria–Lukman estimator for linear regression model to combat multicollinearity and outliers: Real life applications using T20 cricket sports and anthropometric data. *Kuwait Journal of Science*, 52(1), p.100336. doi: https://doi.org/10.1016/j.kjs.2024.100336.

Herridge, R., Turner, A. and Bishop, C. (2020). Monitoring Changes in Power, Speed, Agility, and Endurance in Elite Cricketers During the Off-Season Period. *Journal of Strength and Conditioning Research*, 34(8), pp.2285–2293. doi: https://doi.org/10.1519/jsc.0000000000002077.

Breiman, L. and Schapire, R. (2001). Random Forests. 45, pp.5–32.

Kalechofsky, H. (n.d.). *A Simple Framework for Building Predictive Models A Little Data Science Business Guide A Simple Framework for Building Predictive Models | 2*.

Karnik, A. (2009). Valuing Cricketers Using Hedonic Price Models. *Journal of Sports Economics*, 11(4), pp.456–469. doi: https://doi.org/10.1177/1527002509350442.

Klemperer, P. (2004). Auctions: Theory and Practice. *SSRN Electronic Journal*. doi: https://doi.org/10.2139/ssrn.491563.

Malhotra, G. (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(3), pp.149–170. doi: https://doi.org/10.3233/jsa-200580.

Perez, L. (2017). *Principal Component Analysis to Address Multicollinearity*.

Rastogi, S.K. and Deodhar, S.Y. (2009). *Player Pricing and Valuation of Cricketing Attributes: Exploring the IPL Twenty20 Vision*. *Vikalpa: the Journal for Decision Makers*, 34(2), pp.15–24. doi: https://doi.org/10.1177/0256090920090202.

Zhang, Y. and Yang, Y. (2015). Cross-validation for Selecting a Model Selection Procedure. *Journal of Econometrics*, 187(1), pp.95–112. doi: https://doi.org/10.1016/j.jeconom.2015.02.006.

Deep, C., Patvardhan, C. and Vasantha, C. (2016). Data Analytics Based Deep Mayo Predictor for IPL-9. *International Journal of Computer Applications*, 152(6), pp.6–11. doi: https://doi.org/10.5120/ijca2016911875.

Deep Prakash, C. and Verma, S. (2022). A new in-form and role-based Deep Player Performance Index for player evaluation in T20 Cricket. *Decision Analytics Journal*, 2, p.100025. doi: https://doi.org/10.1016/j.dajour.2022.100025.

Hodson, T.O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development*, 15(14), pp.5481–5487. doi: https://doi.org/10.5194/gmd-15-5481-2022.

Biau, G. and Scornet, E. (2016). A random forest guided tour. *TEST*, 25(2), pp.197–227. doi: https://doi.org/10.1007/s11749-016-0481-7

J. Rani P., A. Kulkarni, A. V. Kamath, A. Menon, P. Dhatwalia and D. Rishabh, "Prediction of Player Price in IPL Auction Using Machine Learning Regression Algorithms," *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, Bangalore, India, 2020, pp. 1-6, doi: https://doi.org/10.1109/CONECCT50063.2020.9198668.

Chittibabu, V. and Sundararaman, M. (2023). Base price determination for IPL mega auctions: A player performance-based approach. *Journal of Sports Analytics*, 9(1), pp.77–97. doi: https://doi.org/10.3233/jsa-220633.

Chauhan, N. (n.d.). *ISSN: 2454-132X Impact factor: 4.295 A study of the application of operations research in the valuation of players in IPL*. *International Journal of Advance Research*.

Coluccia, D., Fontana, S. and Solimene, S. (2017). An application of the option-pricing model to the valuation of football player in the Serie A League. *International Journal of Sport Management and Marketing*, 1(1), p.1. doi: https://doi.org/10.1504/ijsmm.2017.10012018.

Davis, J., Perera, H. and Swartz, T.B. (2015). Player evaluation in Twenty20 cricket. *Journal of Sports Analytics*, 1(1), pp.19–31. doi: https://doi.org/10.3233/jsa-150002.

Dey, P.Kr., Banerjee, A., Ghosh, D.N. and Mondal, A.C. (2014). AHP-Neural Network Based Player Price Estimation in IPL. *International Journal of Hybrid Information Technology*, 7(3), pp.15–24. doi: https://doi.org/10.14257/ijhit.2014.7.3.03.

Easton, T. and Newell, S. (2018). Are daily fantasy sports gambling? *Journal of Sports Analytics*, 5(1), pp.35–43. doi: https://doi.org/10.3233/jsa-180240.

Jayanth, S.B., Anthony, A., Abhilasha, G., Shaik, N. and Srinivasa, G. (2018). A team recommendation system and outcome prediction for the game of cricket. *Journal of Sports Analytics*, 4(4), pp.263–273. doi: https://doi.org/10.3233/jsa-170196.

Kapadia, K., Abdel-Jaber, H., Thabtah, F. and Hadi, W. (2020). Sport analytics for cricket game results using machine learning: An experimental study. *Applied Computing and Informatics*, 18(3/4), pp.256–266. doi: https://doi.org/10.1016/j.aci.2019.11.006.

CHAPTER 8:

APPENDIX

8. APPENDIX

APPENDIX 1: Codes and Outputs after using models:

```
import pandas as pd
from sklearn.preprocessing import OneHotEncoder
from sklearn.linear_model import LinearRegression
from sklearn.impute import SimpleImputer
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
import numpy as np
# Step 2: Feature Selection
# Select features and target variable
features = [
    'Country', 'C/U/A', 'Runs_Scored', 'Batting_Average', 'Batting_Strike_Rate', 'Centuries', 'Half_Centuries', 'Catches_Taken', 'Wickets_Taken', 'Bowl
]
target = 'Price Rs (Lakhs)'

X = Xtrain_filtered[features]
y = ytrain_filtered[target]

# Step 3: Train-Test Split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Handle infinite and large values
X_train.replace([np.inf, -np.inf], np.nan, inplace=True)
X_test.replace([np.inf, -np.inf], np.nan, inplace=True)
# Replace NaN values with 0
X_train.fillna(0, inplace=True)
X_test.fillna(0, inplace=True)

# Initialize the Linear Regression model
model = LinearRegression()
# Train the model
model.fit(X_train, y_train)
# Predict on the test set
y_pred_lr = model.predict(X_test)
# Evaluate the model
rmse_lr = np.sqrt(mean_squared_error(y_test, y_pred_lr))
mae_lr = mean_absolute_error(y_test, y_pred_lr)
r2_lr = r2_score(y_test, y_pred_lr)
print("Linear Regression Model:")
print(f"RMSE: {rmse_lr:.3f}")
print(f"MAE: {mae_lr:.3f}")
print(f"R²: {r2_lr:.3f}")
```

```
Linear Regression Model:
RMSE: 33.662
MAE: 26.380
R²: 0.807
```

```
from sklearn.tree import DecisionTreeRegressor
# Initialize the Decision Tree Regressor model
dtree_model = DecisionTreeRegressor(random_state=42)
# Train the model
dtree_model.fit(X_train, y_train)
# Predict on the validation set
y_pred_tree = dtree_model.predict(X_test)
#Evaluate the model
rmse_dt = np.sqrt(mean_squared_error(y_test, y_pred_tree))
mae_dt = mean_absolute_error(y_test, y_pred_tree)
r2_dt = r2_score(y_test, y_pred_tree)
print("Decision Tree Regressor Model:")
print(f"RMSE: {rmse_dt:.3f}")
print(f"MAE: {mae_dt:.3f}")
print(f"R²: {r2_dt:.3f}")
print
```

```
Decision Tree Regressor Model:
RMSE: 38.885
MAE: 22.759
R²: 0.742
```

```
# Model Selection
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)

# Step 5: Cross-Validation
cv_scores = cross_val_score(model, X_train, y_train, cv=5, scoring='neg_mean_squared_error')
cv_rmse = (-cv_scores) ** 0.5  # Convert to RMSE

print(f'Cross-Validated RMSE: {cv_rmse.mean()} ± {cv_rmse.std()}')

# Step 6: Model Training
model.fit(X_train, y_train)

# Step 7: Model Evaluation
y_pred = model.predict(X_test)

# Calculate performance metrics
mse_rf = mean_squared_error(y_test, y_pred)
r2_rf = r2_score(y_test, y_pred)

print(f'Mean Squared Error: {mse_rf}')
print(f'R^2 Score: {r2_rf}')
```

```
Cross-Validated RMSE: 32.22354405321113 ± 3.210097562909922
Mean Squared Error: 1133.121061750659
R^2 Score: 0.806892851265529
```

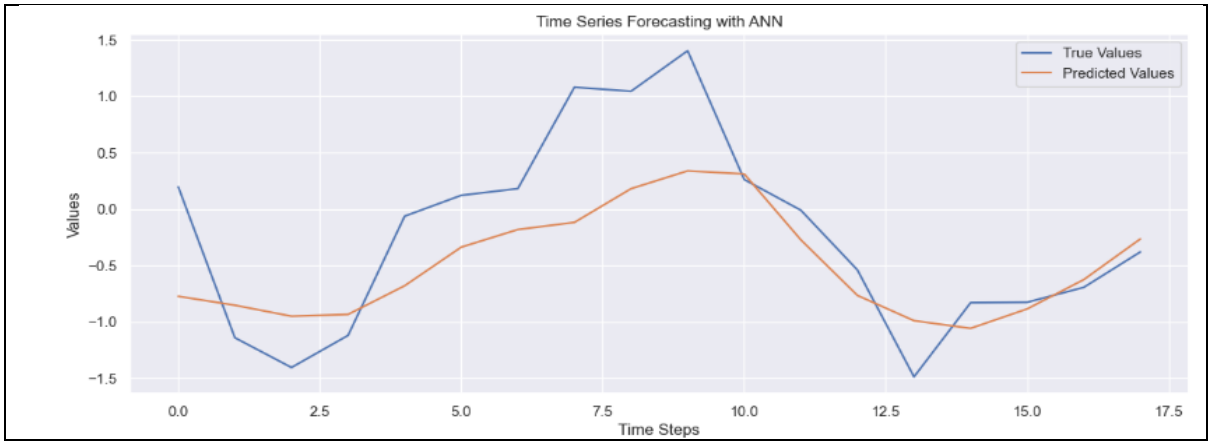
```
# Define the ANN model
def create_model(optimizer='adam', init='uniform'):
    model = Sequential()
    model.add(Input(shape=(X_train.shape[1],)))
    model.add(Dense(64, kernel_initializer=init, activation='relu'))
    model.add(Dense(32, kernel_initializer=init, activation='relu'))
    model.add(Dense(1, kernel_initializer=init))
    model.compile(loss='mean_squared_error', optimizer=optimizer)
    return model
ann_model = create_model()
ann_model.fit(X_train, y_train, epochs=100, batch_size=10, verbose=1)
# Predict on the test set
y_pred_ann = ann_model.predict(X_test)
# Calculate mean squared error and R2 score for ANN
mse_ann = mean_squared_error(y_test, y_pred_ann)
r2_ann = r2_score(y_test, y_pred_ann)
# Print results for ANN
print("ANN Regressor:")
print(f"Mean Squared Error: {mse_ann:.3f}")
print(f"R2 Score: {r2_ann:.3f}")
print()
```

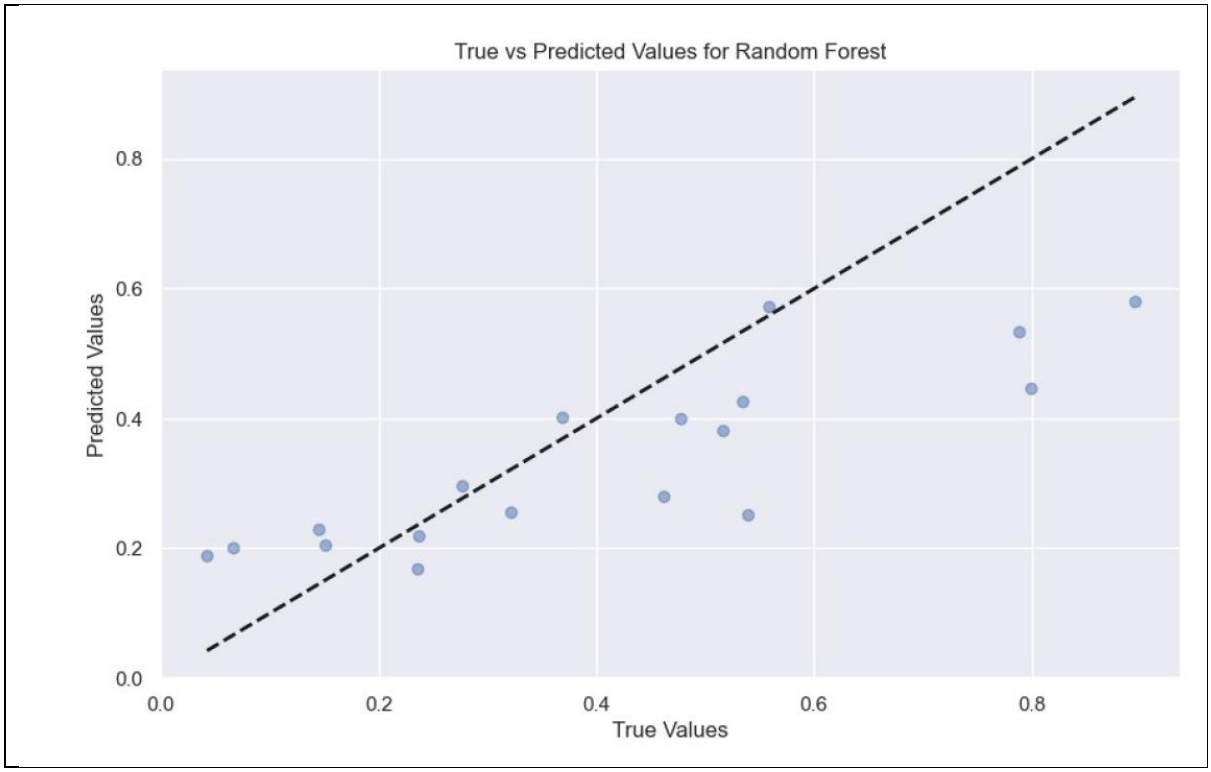
```
12/12 ━━━━━━━━━━━━━━━━━━━━ 0s 8ms/step - loss: 1316.1591
Epoch 95/100
12/12 ━━━━━━━━━━━━━━━━━━━━ 0s 9ms/step - loss: 949.2375
Epoch 96/100
12/12 ━━━━━━━━━━━━━━━━━━━━ 0s 8ms/step - loss: 881.6691
Epoch 97/100
12/12 ━━━━━━━━━━━━━━━━━━━━ 0s 9ms/step - loss: 894.5425
Epoch 98/100
12/12 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 922.3876
Epoch 99/100
12/12 ━━━━━━━━━━━━━━━━━━━━ 0s 8ms/step - loss: 1090.3442
Epoch 100/100
12/12 ━━━━━━━━━━━━━━━━━━━━ 0s 11ms/step - loss: 896.4672
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 191ms/step
ANN Regressor:
Mean Squared Error: 1728.028
R2 Score: 0.706
```

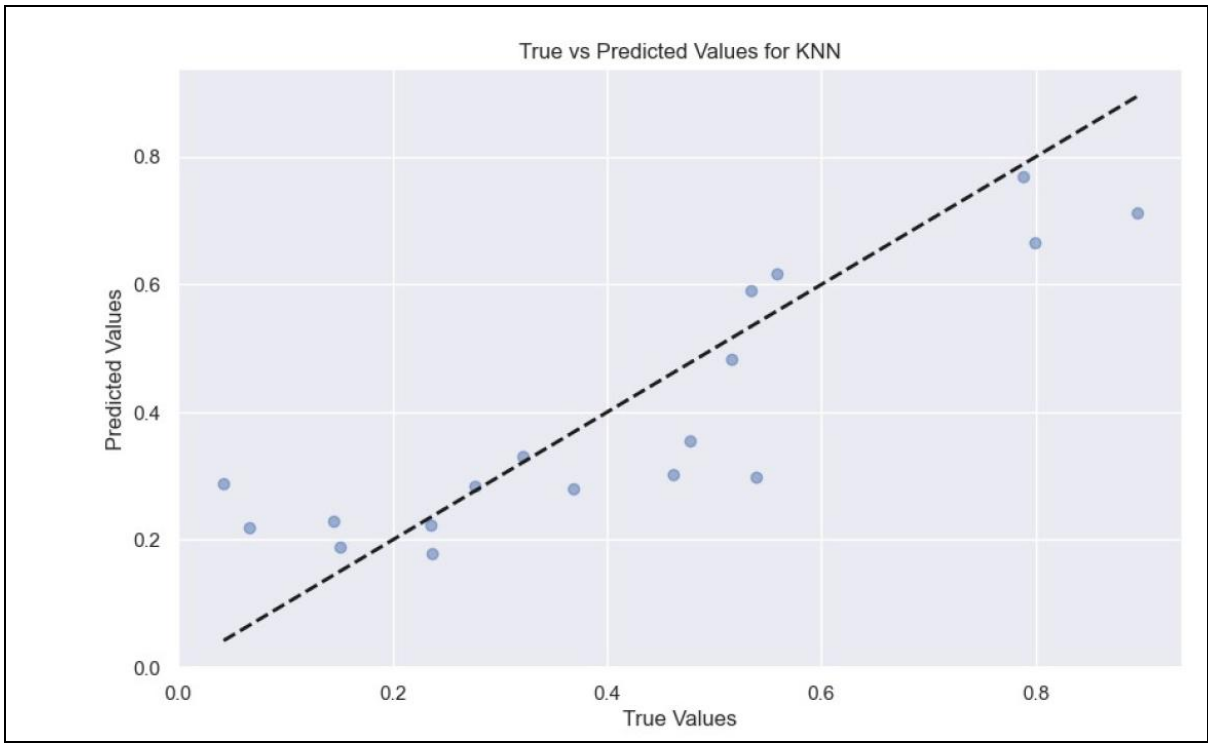
```
svr.fit(X_train, y_train)
# Predict on the test set
y_pred_svr = svr.predict(X_test)
# Evaluate the model
mse_svr = mean_squared_error(y_test, y_pred)
r2_svr = r2_score(y_test, y_pred)
# Print the results
print("Support Vector Regressor:")
print(f"Mean Squared Error: {mse_svr:.3f}")
print(f"R2 Score: {r2_svr:.3f}")

Original shape of X_train: (72, 10)
Original shape of X_test: (18, 10)
Support Vector Regressor:
Mean Squared Error: 0.016
R2 Score: 0.740
```

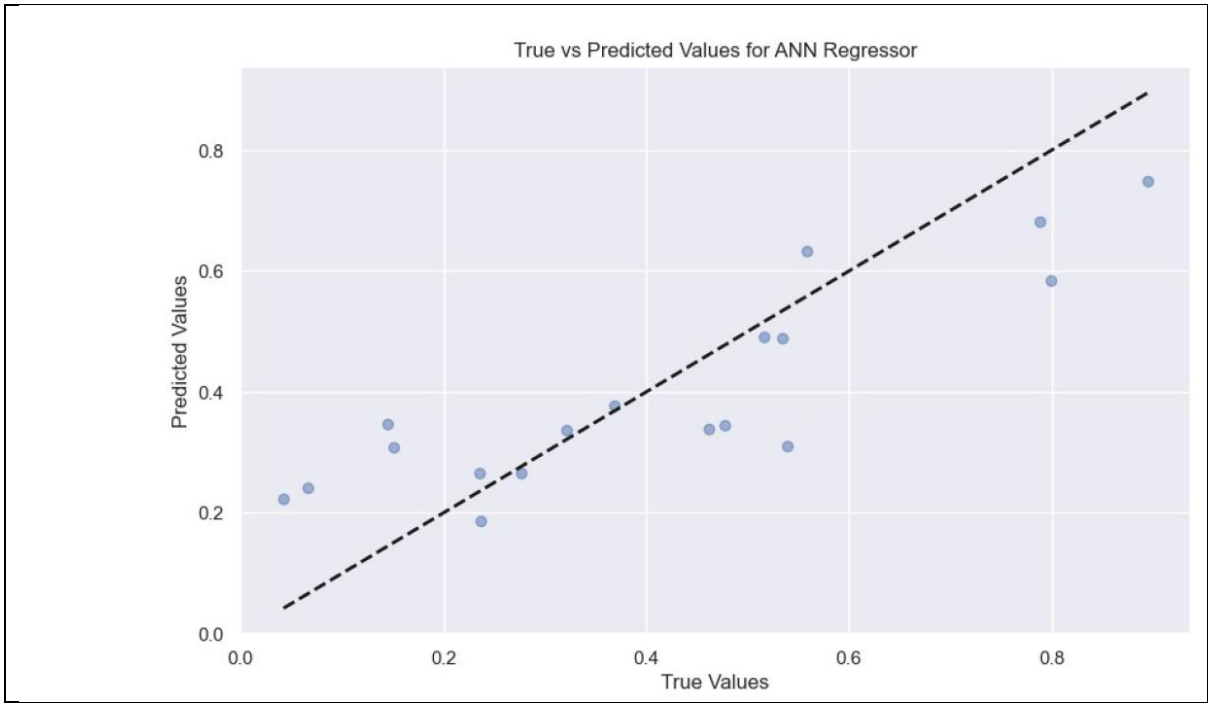
```python
# Create sequences
def create_sequences(data, seq_length):
    X, y = [], []
    for i in range(len(data) - seq_length):
        X.append(data[i:i + seq_length])
        y.append(data[i + seq_length])
    return np.array(X), np.array(y)

seq_length = 10
X, y = create_sequences(data_scaled, seq_length)

# Split the data into training and test sets
split = int(0.8 * len(X))
X_train, X_test = X[:split], X[split:]
y_train, y_test = y[:split], y[split:]

# Build the ANN model
model = Sequential()
model.add(LSTM(50, activation='relu', input_shape=(seq_length, 1)))
model.add(Dense(1))
model.compile(optimizer='adam', loss='mse')

# Train the model
history = model.fit(X_train, y_train, epochs=100, validation_data=(X_test, y_test), verbose=1)

# Make predictions
y_pred = model.predict(X_test)
y_pred_rescaled = scaler.inverse_transform(y_pred)
y_test_rescaled = scaler.inverse_transform(y_test.reshape(-1, 1))

# Plot the results
plt.figure(figsize=(14, 5))
plt.plot(y_test_rescaled, label='True Values')
plt.plot(y_pred_rescaled, label='Predicted Values')
plt.title('Time Series Forecasting with ANN')
plt.xlabel('Time Steps')
plt.ylabel('Values')
plt.legend()
plt.show()
```

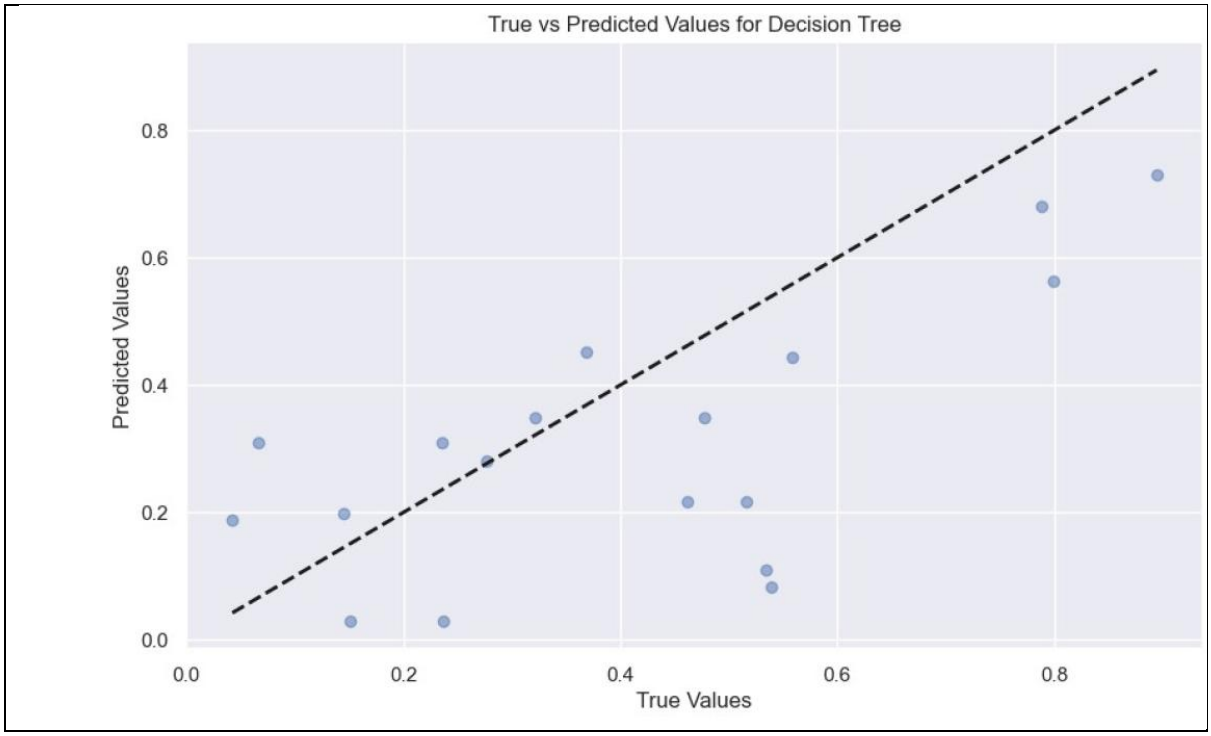
Appendix 2: Output Figures after using machine learning models:




True vs Predicted Values for Random Forest
Predicted Values
True Values
0.0
0.2
0.4
0.6
0.8


True vs Predicted Values for KNN
Predicted Values
True Values
0.0
0.2
0.4
0.6
0.8





C:\Users\bhara\Downloads\Dissertation (1).ipynb

C:\Users\bhara\Downloads\vertopal.com_Dissertation (1).pdf